RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths

Zeyue Xue[*]
The University of Hong Kong
xuezeyue@connect.hku.hk

Guanglu Song[*]
SenseTime Research
songguanglu@sensetime.com

Qiushan Guo
The University of Hong Kong
qsguo@cs.hku.hk

Boxiao Liu
SenseTime Research
liuboxiao@sensetime.com

Zhuofan Zong
SenseTime Research
zongzhuofan@gmail.com

Yu Liu[† ‡]
SenseTime Research
liuyuisanai@gmail.com

Ping Luo[‡]
The University of Hong Kong
Shanghai AI Laboratory
pluo@cs.hku.hk

"When one is painting one does not think."

— *Raffaello Sanzio da Urbino*

Abstract

Text-to-image generation has recently witnessed remarkable achievements. We introduce a text-conditional image diffusion model, termed RAPHAEL, to generate highly artistic images, which accurately portray the text prompts, encompassing multiple nouns, adjectives, and verbs. This is achieved by stacking tens of mixture-of-experts (MoEs) layers, *i.e.,* space-MoE and time-MoE layers, enabling billions of diffusion paths (routes) from the network input to the output. Each path intuitively functions as a "painter" for depicting a particular textual concept onto a specified image region at a diffusion timestep. Comprehensive experiments reveal that RAPHAEL outperforms recent cutting-edge models, such as Stable Diffusion, ERNIE-ViLG 2.0, DeepFloyd, and DALL-E 2, in terms of both image quality and aesthetic appeal. Firstly, RAPHAEL exhibits superior performance in switching images across diverse styles, such as Japanese comics, realism, cyberpunk, and ink illustration. Secondly, a single model with three billion parameters, trained on $1,000$ A100 GPUs for two months, achieves a state-of-the-art zero-shot FID score of 6.61 on the COCO dataset. Furthermore, RAPHAEL significantly surpasses its counterparts in human evaluation on the ViLG-300 benchmark. We believe that RAPHAEL holds the potential to propel the frontiers of image generation research in both academia and industry, paving the way for future breakthroughs in this rapidly evolving field. More details can be found on a webpage: https://raphael-painter.github.io/[§].

[*] Equal contribution. Work done during Zeyue's internship at SenseTime Research.

[†] Project lead.

[‡] Corresponding authors.

[§] More creations can be found in https://miaohua.sensetime.com/zh-CN/picture-selection. Please select the Artist v0.3.5 model to generate. This is our latest version based on RAPHAEL. This information was last updated on Oct. 16th, 2023.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

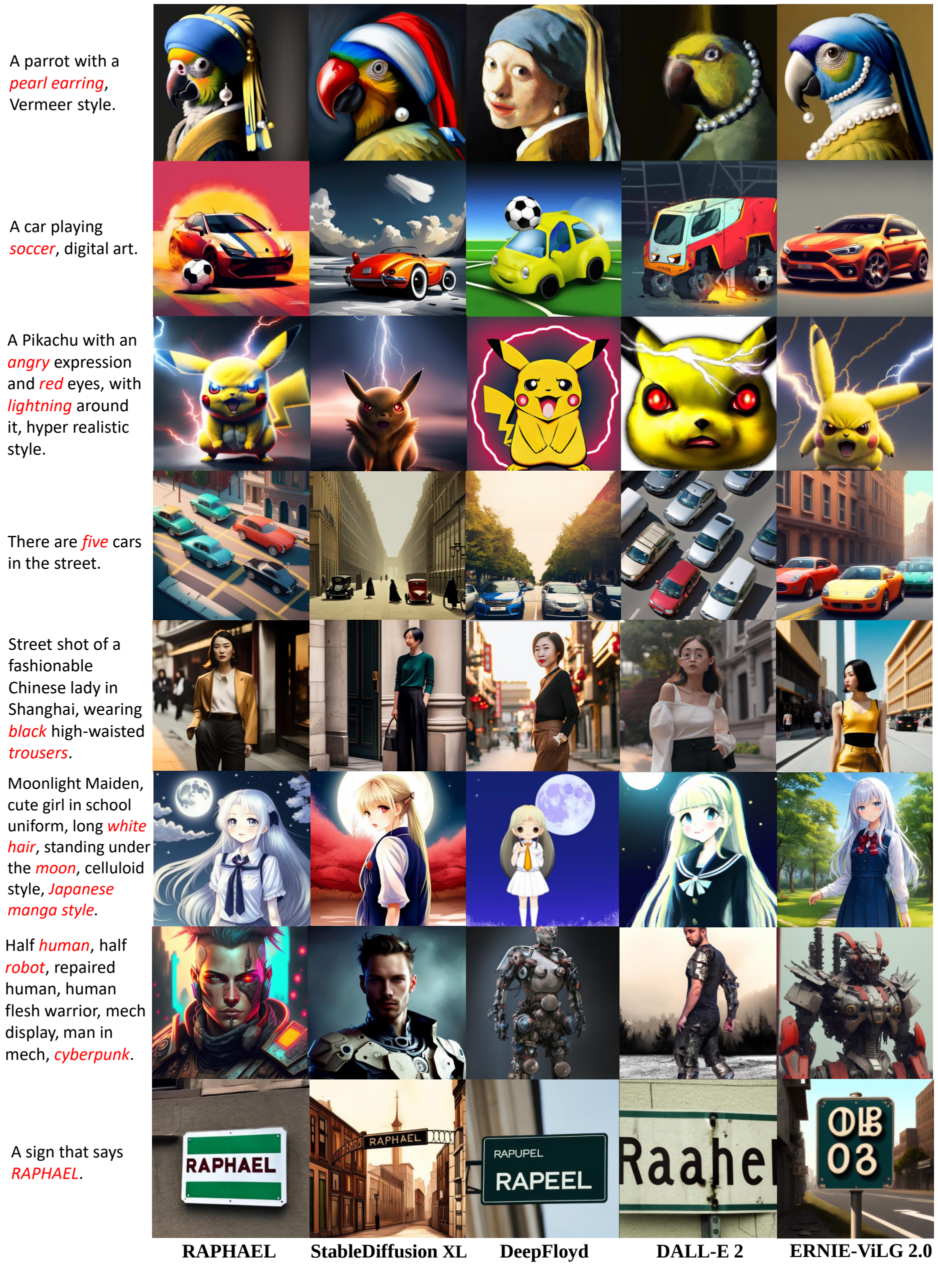

A parrot with a pearl earring, Vermeer style.
A car playing soccer, digital art.
A Pikachu with an angry expression and red eyes, with lightning around it, hyper realistic style.
There are five cars in the street.
Street shot of a fashionable Chinese lady in Shanghai, wearing black high-waisted trousers.
Moonlight Maiden, cute girl in school uniform, long white hair, standing under the moon, celluloid style, Japanese manga style.
Half human, half robot, repaired human, human flesh warrior, mech display, man in mech, cyberpunk.
A sign that says RAPHAEL.
RAPHAEL
StableDiffusion XL
DeepFloyd
DALL-E 2
ERNIE-ViLG 2.0
RAPHAEL
RAPHAEL
RAPUPEL
RAPEEL
Raahel

Figure 1: **Comparisons** of RAPHAEL with recent representative generators, Stable Diffusion XL [2], DeepFloyd, DALL-E 2 [3], and ERNIE-ViLG 2.0 [5]. They are given the same prompts, where the words that the human artists yearn to preserve within the generated images are highlighted in red. These images are not cherry-picked. We see that previous models often fail to preserve the desired concepts. For example, only the RAPHAEL-generated images precisely reflect the prompts such as "pearl earring, Vermeer", "playing soccer", "five cars", "black high-waisted trouser", "white hair, manga, moon", and "sign, RAPHAEL", while other models generate compromised results. **Better zoom in 200%**.

1 Introduction

Recent advancements in text-to-image generators, such as Imagen [1], Stable Diffusion [2], DALL-E 2 [3], eDiff-I [4], and ERNIE-ViLG 2.0 [5], have yielded remarkable success and found wide applications in computer graphics, culture and art, and the generation of medical and biological data.

Despite the substantial progress made in text-to-image diffusion models [1, 2, 3, 4, 5], there remains a pressing need for research to further achieve more precise alignment between text and image. As illustrated in Fig.1, existing models often fail to adequately preserve textual concepts within the generated images. This is primarily due to the reliance on a classic cross-attention mechanism for integrating text descriptions into visual representations, resulting in relatively coarse control of the diffusion process, and leading to compromised results.

To address this issue, we introduce RAPHAEL, a text-to-image generator, which yields images with superior artistry and fidelity compared to prior work, as demonstrated in Fig.2. RAPHAEL, an acronym that stands for “distinct image **r**egions **a**lign with different text **ph**ases in **a**tt**e**ntion **l**earning”, offers an appealing benefit not found in existing approaches.

Specifically, we observe that different text concepts influence distinct image regions during the generation process [6], and the conventional cross-attention layer often struggles to preserve these varying concepts adequately in an image. To mitigate this issue, we employ a diffusion model stacking tens of mixture-of-experts (MoE) layers [7, 8], including both space-MoE and time-MoE layers. Concretely, the space-MoE layers are responsible for depicting different concepts in specific image regions, while the time-MoE layers focus on painting these concepts at different diffusion timesteps.

This configuration leads to billions of diffusion paths from the network input to the output. Naturally, each path can act as a “painter” responsible for rendering a particular concept to an image region at a specific timestep. The result is a more precise alignment between text tokens and image regions, enabling the generated images that accurately represent the associated text prompt. This approach sets RAPHAEL apart from existing models and even sheds light on future studies of the explainability of the generation process. Additionally, we propose an edge-supervised learning module to further enhance the image quality and aesthetic appeal of the generated images.

Extensive experiments demonstrate that RAPHAEL outperforms preceding approaches, such as Stable Diffusion, ERNIE-ViLG 2.0, DeepFloyd, and DALL-E 2. (1) RAPHAEL exhibits superior performance in switching images across diverse styles, such as Japanese comics, realism, cyberpunk, and ink illustration. (2) RAPHAEL establishes a new state-of-the-art with a zero-shot FID-30k score of 6.61 on the COCO dataset. (3) RAPHAEL, a single model with three billion parameters trained on $1,000$ A100 GPUs, significantly surpasses its counterparts in human evaluation on the ViLG-300 benchmark.

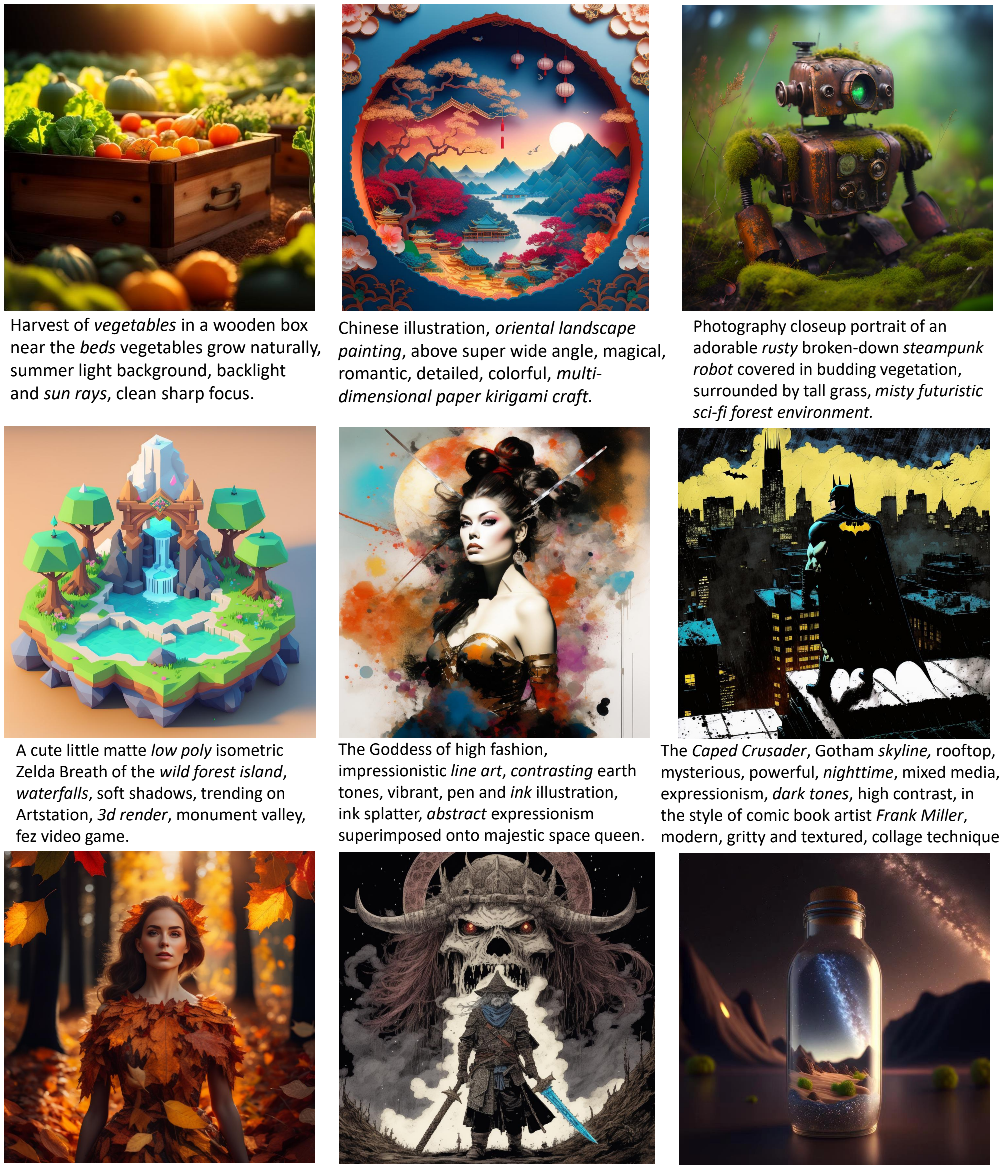
The **contributions** of this work are three-fold: **(i)** We propose a novel text-to-image generator, RAPHAEL, which, through the implementation of several carefully-designed techniques, generates images that more accurately reflect textual prompts than previous works. **(ii)** We thoroughly explore RAPHAEL's potential for switching images in diverse styles, such as Japanese comics, realism, cyberpunk, and ink illustration, and for extension using LoRA [9], ControlNet [10], and SR-GAN [11]. **(iii)** We will release a programming API for RAPHAEL to the public. We believe that RAPHAEL holds the potential to advance the frontiers of image generation in both academia and industry, paving the way for future breakthroughs in this rapidly evolving field.

2 Notation and Preliminary

We present the necessary notations and the Denoising Diffusion Probabilistic Model (DDPM) [12] for text-to-image generation. Given a collection of N images, denoted as $\{\mathbf{x}_i\}_{i=1}^{N}$, the aim is to learn a generative model, $p(\mathbf{x})$, that is capable of accurately representing the underlying distribution.

In forward diffusion, Gaussian noise is progressively introduced into the source images. At an arbitrary timestep t, it is possible to directly sample from the Gaussian distribution following the T-step noise schedule $\{\alpha_t\}_{t=1}^{T}$, without iterative forward sampling. Consequently, the noisy image at timestep t, denoted as $\mathbf{x}_t$, can be expressed as $\mathbf{x}_t = \sqrt{1-\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{\bar{\alpha}_t}\epsilon_t$, where $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. In this expression, $\mathbf{x}_0$ represents the source image, while $\epsilon_t \sim \mathcal{N}(0, I)$ indicates the Gaussian noise at step t. In the reverse process, a denoising neural network, denoted as $D_\theta(\cdot)$, is employed to estimate the additive Gaussian noise. The optimization of this network is achieved by minimizing the loss function, $\mathcal{L}_{\text{denoise}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon\sim\mathcal{N}(0,I)}\left[\|\epsilon - D_\theta\left(\mathbf{x}_t, t\right)\|_2^2\right]$.

By employing the Bayes' theorem, it is feasible to iteratively estimate the image at timestep $t-1$ through sampling from the posterior distribution, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. We have $\mathbf{x}_{t-1} =$



Harvest of *vegetables* in a wooden box near the *beds* vegetables grow naturally, summer light background, backlight and *sun rays*, clean sharp focus.

Chinese illustration, *oriental landscape painting*, above super wide angle, magical, romantic, detailed, colorful, *multi-dimensional paper kirigami craft.*

Photography closeup portrait of an adorable *rusty* broken-down *steampunk robot* covered in budding vegetation, surrounded by tall grass, *misty futuristic sci-fi forest environment.*

A cute little matte *low poly* isometric Zelda Breath of the *wild forest island*, *waterfalls*, soft shadows, trending on Artstation, *3d render*, monument valley, fez video game.

The Goddess of high fashion, impressionistic *line art*, *contrasting* earth tones, vibrant, pen and *ink* illustration, ink splatter, *abstract* expressionism superimposed onto majestic space queen.

The *Caped Crusader*, Gotham *skyline,* rooftop, mysterious, powerful, *nighttime*, mixed media, expressionism, *dark tones*, high contrast, in the style of comic book artist *Frank Miller*, modern, gritty and textured, collage technique.

A beautiful woman dressed in a dress made of *autumn leaves* in the forest, photography, natural lighting, high detail.

A wizard by *Q Hayashida* in the style of *Dorohedoro* for Elden Ring, with biggest most intricate *sword*, on sunlit *battlefield*, breath of the wild, striking illustration.

Milkyway in a *glass bottle*, 4k, unreal engine, octane render.

Figure 2: These examples show that RAPHAEL can generate artistic images with varying text prompts across various styles. The synthesized images have rich details and semantics. The prompts were written by human artists without cherry-picking.

$\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} D_\theta\left(\mathbf{x}_t, t\right)\right) + \sigma_t z$, where σ_t signifies the standard deviation of the newly injected noise into the image at each step, and z represents the Gaussian noise.

In essence, the denoising neural network estimates the score function at varying time steps, thereby progressively recovering the structure of the image distribution. The fundamental insight provided by the DDPM lies in the fact that the perturbation of data points with noise serves to populate regions of low data density, ultimately enhancing the accuracy of estimated scores. This results in stable training and sampling.

U-Net with Text Prompts. The denoising network is commonly implemented using a U-Net [13] architecture, as depicted in Fig.8 in Appendix 7.3. To incorporate textual prompts (denoted by $\mathbf{y}$) into the U-Net, a text encoder neural network, $E_\theta(\mathbf{y})$, is employed to extract the textual representation.

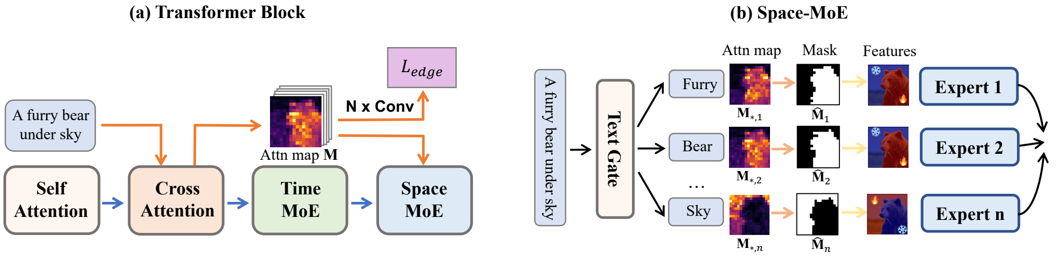


Figure 3: **Framework of RAPHAEL. (a)** Each block contains four primary components including a self-attention layer, a cross-attention layer, a space-MoE layer, and a time-MoE layer. The space-MoE is responsible for depicting different text concepts in specific image regions, while the time-MoE handles different diffusion timesteps. Each block uses edge-supervised cross-attention learning to further improve image quality. **(b)** shows details of space-MoE. For example, given a prompt "a furry bear under sky", each text token and its corresponding image region (given by a binary mask) are directed through distinct space experts, *i.e.,* each expert learns particular visual features at a region. By stacking several space-MoEs, we can easily learn to depict thousands of text concepts.

The extracted text tokens are input into the U-Net through a cross-attention layer. The text tokens possess a size of $n_y \times d_y$, where n_y represents the number of text tokens, and d_y signifies the dimension of a text token (*e.g.,* $d_y = 768$ in [14]).

The cross-attention layer can be formulated as $\mathrm{attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$, where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ correspond to the query, key, and value matrices, respectively. These matrices are computed as $\mathbf{Q} = h\left(\mathbf{x}_t\right)\mathbf{W}_x^{\mathrm{qry}}$, $\mathbf{K} = E_\theta(\mathbf{y})\mathbf{W}_y^{\mathrm{key}}$, and $\mathbf{V} = E_\theta(\mathbf{y})\mathbf{W}_y^{\mathrm{val}}$, where $\mathbf{W}_x^{\mathrm{qry}} \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_y^{\mathrm{key}}, \mathbf{W}_y^{\mathrm{val}} \in \mathbb{R}^{d_y \times d}$ represent the parametric projection matrices for the image and text, respectively. Additionally, d denotes the dimension of an image token, $h(\mathbf{x}_t) \in \mathbb{R}^{n_x \times d}$ indicates the flattened intermediate representation within the U-Net, with n_x being the number of tokens in an image. A cross-attention map between the text and image, $\mathbf{M} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \in \mathbb{R}^{n_x \times n_y}$, is defined, which plays a crucial role in the proposed approach, as described in the following sections.

3 Our Approach

The overall framework of RAPHAEL is illustrated in Fig.3, with the network configuration details provided in the Appendix 7.1. Employing a U-Net architecture, the framework consists of 16 transformer blocks, each containing four components: a self-attention layer, a cross-attention layer, a space-MoE layer, and a time-MoE layer. The space-MoE is responsible for depicting different text concepts in specific image regions at a given scale, while the time-MoE handles different diffusion timesteps.

3.1 Space-MoE and Time-MoE

Space-MoE. Regarding the space-MoE layer, distinct text tokens correspond to various regions within an image, as previously mentioned. For instance, when provided with the prompt "a furry bear under the sky", each text token and its corresponding image region (represented by a binary mask) are fed into separate experts, as illustrated in Fig.3b. The space-MoE layer's output is the mean of all experts, calculated using the following formula: $\frac{1}{n_y}\sum_{i=1}^{n_y} e_{\mathrm{route}(\mathbf{y}_i)}\left(h'(\mathbf{x}_t) \circ \widehat{\mathbf{M}}_i\right)$. In this equation, $\widehat{\mathbf{M}}_i$ is a binary two-dimensional matrix, indicating the image region the i-th text token should correspond to, as shown in Fig.3b. Here, $\circ$ represents hadamard product, and $h'(\mathbf{x}_t)$ is the features from time-MoE. The gating (routing) function $\mathrm{route}(\mathbf{y}_i)$ returns the index of an expert in the space-MoE, with $\{e_1, e_2, \dots, e_k\}$ being a set of k experts.

Text Gate Network. The Text Gate Network is employed to distribute an image region to a specific expert, as shown in Fig.3b. The function $\mathrm{route}(\mathbf{y}_i) = \mathrm{argmax}\left(\mathrm{softmax}\left(\mathcal{G}\left(E_\theta(\mathbf{y}_i)\right) + \epsilon\right)\right)$ is used, where $\mathcal{G} : \mathbb{R}^{d_y} \mapsto \mathbb{R}^k$ is a feed forward network, which uses a text token representation $E_\theta(\mathbf{y}_i)$ as input and assigns a space expert. To prevent mode collapse, random noise ϵ is incorporated. The

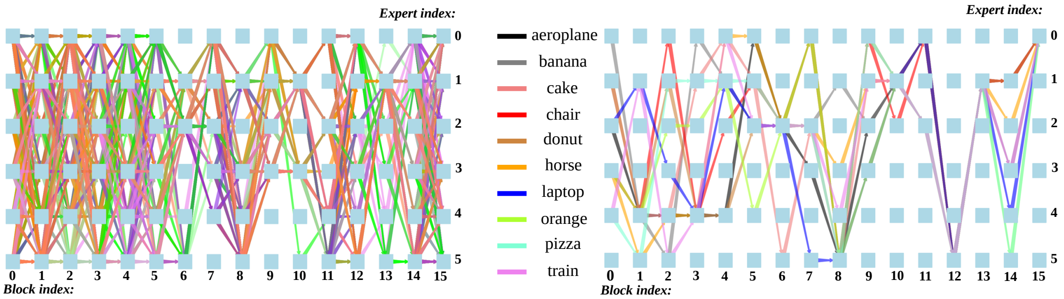


Figure 4: **Left:** We visualize the diffusion paths (routes) from the network input to the output, utilizing 16 space-MoE layers, each containing 6 spatial experts. These paths are closely associated with 100 adjectives, such as "scenic", "peaceful", and "majestic", which represent the most frequently occurring adjectives for describing artworks as suggested by GPT-3.5 [15, 16]. Given that GPT-3.5 has been trained on trillions of tokens, we believe that these adjectives reflect a diverse, real-world distribution. Our findings indicate that different paths distinctively represent various adjectives. **Right:** We depict the diffusion paths for ten categories (*i.e.,* nouns) within the COCO dataset. Our observations reveal that different categories activate distinct paths in a heterogeneous manner. The display colors blend together where the routes overlap.

argmax function ensures that one expert exclusively handles the corresponding image region for each text token, without increasing computational complexity.

From Text to Image Region. Recall that $\mathbf{M}$ is the cross-attention map between text and image, where each element, $\mathbf{M}_{j,i}$, represents a correspondence value between the j-th image token and the i-th text token. In the space-MoE, each entry in the binary mask $\widehat{\mathbf{M}}_i$ equals "1" if $\mathbf{M}_{j,i} \geq \eta_i$, otherwise "0" if $\mathbf{M}_{j,i} < \eta_i$, as illustrated in Fig.3b. A thresholding mechanism is introduced to determine the values in the mask. The threshold value $\eta_i = \alpha \max(\mathbf{M}_{*,i})$ is defined, where $\max(\mathbf{M}_{*,i})$ represents the maximum correspondence between text token i and all image regions. The hyper-parameter α will be evaluated through an ablation study.

Discussions. The insight behind the space-MoE is to effectively model the intricate relationships between text tokens and their corresponding regions in the image, accurately reflecting concepts in the generated images. As illustrated in Fig.4, the employment of 16 space-MoE layers, each containing 6 experts, results in billions of spatial diffusion paths (*i.e.,* 6^{16} possible routes). It is evident that each diffusion path is closely associated with a specific textual concept.

To investigate this further, we generate 100 prevalent adjectives that are the most frequently occurring adjectives for describing artworks as suggested by GPT-3.5 [15, 16]. Given that GPT-3.5 has been trained on trillions of tokens, we posit that these adjectives reflect a diverse, real-world distribution. We input each adjective into the RAPHAEL model to generate 100 distinct images and collect their corresponding diffusion paths. Consequently, we obtain ten thousand paths for the 100 words. By treating these pathways as features (*i.e.,* each path is a vector of 16 entries), we train a straightforward classifier (*e.g.,* XGBoost [17]) to categorize the words. The classifier after 5-fold cross-validation achieves over 93% accuracy for open-world adjectives, demonstrating that different diffusion paths distinctively represent various textual concepts. We observe analogous phenomena within the 80 object categories of the COCO dataset. Further details on verbs and visualization are provided in the Appendix 7.5.

Time-MoE. We can further enhance the image quality by employing a time-mixture-of-experts (time-MoE) approach, which is inspired by previous works such as [4, 5]. Given that the diffusion process iteratively corrupts an image with Gaussian noise over a series of timesteps $t = 1, \ldots, T$, the image generator is trained to denoise the images in reverse order from $t = T$ to $t = 1$. All timesteps aim to denoise a noisy image, progressively transforming random noise into an artistic image. Intuitively, the difficulty of these denoising steps varies depending on the noise ratio presented in the image. For example, when $t = T$, the denoising network's input image $\mathbf{x}_t$ is highly noisy. When $t = 1$, the image $\mathbf{x}_t$ is closer to the original image.

To address this issue, we employ a time-MoE before each space-MoE in each transformer block. In contrast to [4, 5] , which necessitate hand-crafted time expert assignments, we implement an additional gate network to automatically learn to assign different timesteps to various time experts. Further details can be found in the Appendix 7.3.

3.2 Edge-supervised Learning

In order to further enhance the image quality, we propose incorporating an edge-supervised learning strategy to train the transformer block. By implementing an edge detection module, we aim to extract rich boundary information from an image. These intricate boundaries can serve as supervision to guide the model in preserving detailed image features across various styles.

Consider a neural network module, $P_\theta(\mathbf{M})$, with parameters of N convolutional layers (*e.g.,* $N = 5$). This module is designed to predict an edge map given an attention map $\mathbf{M}$ (refer to Fig.7a in the Appendix 7.2). We utilize the edge map of the input image, denoted as $\mathbf{I}_{\text{edge}}$, to supervise the network P_θ. $\mathbf{I}_{\text{edge}}$ can be obtained by the holistically-nested edge detection algorithm [18] (Fig.7b). Intuitively, the network P_θ can be trained by minimizing the loss function, $\mathcal{L}_{\text{edge}} = \text{Focal}(P_\theta(\mathbf{M}), \mathbf{I}_{\text{edge}})$, where $\text{Focal}(\cdot, \cdot)$ denotes the focal loss [19] employed to measure the discrepancy between the predicted and the "ground-truth" edge maps. Moreover, as discussed in [5, 6], the attention map $\mathbf{M}$ is prone to becoming vague when the timestep t is large. Consequently, it is essential to adopt a timestep threshold value to inactivate (pause) edge-supervised learning when t is large. This timestep threshold value (T_c) is a hyper-parameter that will be evaluated through an ablation study.

Overall, the RAPHAEL model is trained by combining two loss functions, $\mathcal{L} = \mathcal{L}_{\text{denoise}} + \mathcal{L}_{\text{edge}}$. As demonstrated in Fig.7d in the Appendix 7.2, edge-supervised learning substantially improves the image quality and aesthetic appeal of the generated images.

4 Experiments

This section presents the experimental setups, the quantitative results compared to recent state-of-the-art models, and the ablation study to demonstrate the effectiveness of RAPHAEL. More artistic images generated by RAPHAEL and comparisons between RAPHAEL and other diffusion models can be found in Appendix 7.6 and 7.7.

Dataset. The training dataset consists of a subset of LAION-5B [20] and some internal datasets, including 730M text-images pairs in total. To collect training data from LAION-5B, we filter the images using the aesthetic scorer same as Stable Diffusion [2] and remove the image-text pairs that have scores smaller than 4.7. We remove the images with watermarks either. Since the text descriptions in LAION-5B are noisy, we clean them by removing useless information such as URLs, HTML tags, and email addresses, inspired by [2, 4, 21].

Multi-scale Training. To improve text-image alignment, instead of cropping images to a fixed scale [2], we resize an image to its nearest size into different buckets, which has 9 different image scales. Additionally, the GPU resources will be automatically allocated to each bucket depending on the number of images it contains, enabling effective use of computational resources*.

Implementations. To reduce training and sampling complexity, we use a Variational Autoencoder (VAE) [22, 23] to compress images using Latent Diffusion Model [2]. We first pre-train an image encoder to transform an image from pixel space to a latent space, and an image decoder to convert it back. Unlike previous works, the cross-attention layers in RAPHAEL are augmented with space-MoE and time-MoE layers. The entire model is implemented in PyTorch [24], and is trained by AdamW [25] optimizer with a learning rate of $1e-4$, a weight decay of 0, a batch size of $2,000$, on $1,000$ NVIDIA A100s for two months. More details on the hyper-parameter settings can be found in the Appendix 7.1.

4.1 Comparisons

Results on COCO. Following previous works [1, 2, 4], we evaluate RAPHAEL on the COCO 256×256 dataset using zero-shot Frechet Inception Distance (FID), which measures the quality and diversity of images. Similar to [1, 2, 4, 5, 32], $30,000$ images are randomly selected from the validation set for evaluation. Table 1 shows that RAPHAEL achieves a new state-of-the-art

*The dimensions of each bucket are as follows: [448, 832], [512, 768], [512, 704], [640, 640], [576, 640], [640, 576], [704, 512], [768, 512], and [832, 448]. For instance, when images are resized, those with an aspect ratio of 1.0 will be assigned to the bucket of size [640, 640]. GPUs will be allocated to each bucket, based on the images it contains. All GPUs will have the same batch size and will select images from its associated bucket.

Table 1: **Comparisons** of RAPHAEL with the recent representative text-to-image generation models on the MS-COCO 256×256 using zero-shot FID-30k. We see that RAPHAEL outperforms all previous works in image quality, even a commercial product released recently.

Approach	Venue/Date	Model Type	FID-30K	Zero-shot FID-30K
DF-GAN [26]	CVPR'22	GAN	21.42	-
DM-GAN + CL [27]	CVPR'19	GAN	20.79	-
LAFITE [28]	CVPR'22	GAN	8.12	-
Make-A-Scene [29]	ECCV'22	Autoregressive	7.55	-
LDM [2]	CVPR'22	Diffusion	-	12.63
GLIDE [30]	ICML'22	Diffusion	-	12.24
DALL-E 2 [3]	arXiv, April 2022	Diffusion	-	10.39
GigaGAN [31]	CVPR'23	GAN	-	9.09
Stable Diffusion [2]	CVPR'22	Diffusion	-	8.32
Muse-3B [32]	arXiv, Jan. 2023	Non-Autoregressive	-	7.88
Imagen [1]	NeurIPS'22	Diffusion	-	7.27
eDiff-I [4]	arXiv, Nov. 2022	Diffusion Experts	-	6.95
ERNIE-ViLG 2.0 [5]	CVPR'23	Diffusion Experts	-	6.75
DeepFloyd	Product, May 2023	Diffusion	-	6.66
RAPHAEL	-	Diffusion Experts	-	**6.61**

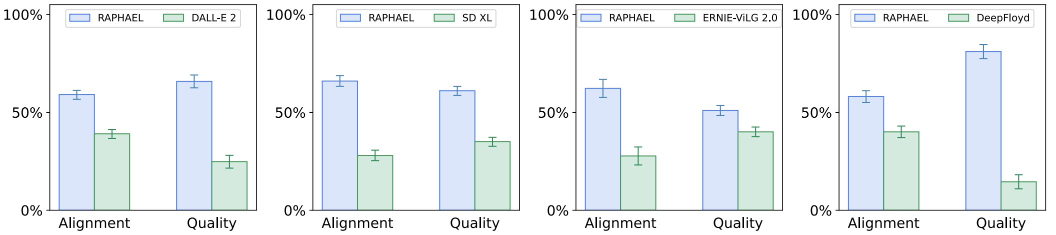


Figure 5: **Comparisons** of RAPHAEL with DALL-E 2, Stable Diffusion XL (SD XL), ERNIE-ViLG 2.0, and DeepFloyd in a user study using the ViLG-300 benchmark. We report the user's preference rates with 95% confidence intervals. We see that RAPHAEL can generate images with higher quality and better conform to the prompts.

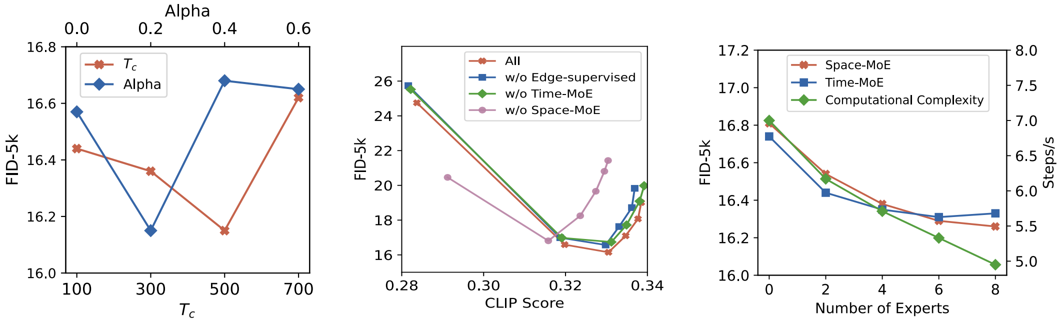
performance of text-to-image generation, with 6.61 zero-shot FID-30k on MS-COCO, surpassing prominent image generators such as Stable Diffusion, Imagen, ERNIE-ViLG 2.0, and DALL-E 2.

Human Evaluations. We employ the ViLG-300 benchmark [5], a bilingual prompt set, which enables to systematically evaluate text-to-image models given various text prompts in Chinese and English. ViLG-300 allows us to convincingly compare RAPHAEL with recent-advanced models including DALL-E 2, Stable Diffusion, ERNIE-ViLG 2.0, and DeepFloyd, in terms of both image quality and text-image alignment. For example, human artists are presented with two sets of images generated by RAPHAEL and a competitor, respectively. They are asked to compare these images from two aspects respectively, including image-text alignment, and image quality and aesthetics. Throughout the entire process, human artists are unaware of which model the image is generated from. Fig.5 shows that RAPHAEL surpasses all other models in both image-text alignment and image quality in the user study, indicating that RAPHAEL can generate high-artistry images that conform to the text.

Extensions to LoRA, ControlNet, and SR-GAN. RAPHAEL can be further extended by incorporating LoRA, ControlNet, and SR-GAN. In Appendix 7.8, we present a comparison between RAPHAEL and Stable Diffusion utilizing LoRA. RAPHAEL demonstrates superior robustness against overfitting compared to Stable Diffusion. We also demonstrate RAPHAEL with a canny-based ControlNet. Furthermore, by employing a tailormade SR-GAN model, we enhance the image resolution to 4096×6144.

4.2 Ablation Study

Evaluate every module in RAPHAEL. We conduct a comprehensive assessment of each module within the RAPHAEL model, utilizing the CLIP [14] score to measure image-text alignment. Given the significance of classifier-free guidance weight in controlling image quality and text alignment, we present ablation results as trade-off curves between CLIP and FID scores across a range of



(a) Impact of α and T_c.

(b) FID-CLIP score curves.

(c) Impact of experts.

Figure 6: **Ablation Study**. (a) examines the selection of α and T_c. (b) presents the trade-off between FID and CLIP scores for the complete RAPHAEL model and its variants without space-MoE, time-MoE, and edge-supervised learning. (c) visualizes the correlation between FID-5k and runtime complexity (measured in terms of the number of DDIM [34] steps for an image per second) as a function of the number of experts employed. Notably, the computational complexity is predominantly influenced by the number of spatial experts.

guidance weights [33], specifically 1.5, 3.0, 4.5, 6.0, 7.5, and 9.0. Fig.6b compares these curves for the complete RAPHAEL model and its variants without space-MoE, edge-supervised learning, and time-MoE, respectively. Our findings indicate that all modules contribute effectively. For example, space-MoE substantially enhances the CLIP score and the optimal guidance weight for the sampler shifts from 3.0 to 4.5. Moreover, at the same guidance weight, space-MoE considerably reduces the FID, resulting in a significant improvement in image quality.

Choice of α and T_c. As depicted in Fig.6a, we observe that $\alpha = 0.2$ delivers the best performance, implying a balance between preserving adequate features and avoiding the use of the entire latent features. An appropriate threshold value for T_c terminates edge-supervised learning when the diffusion timestep is large. Our experiments reveal that a suitable choice for T_c is 500, ensuring the effective learning of texture information.

Performance and Runtime Analysis on Number of Experts. We offer an examination of the number of experts, ranging from 0 to 8, in Fig.6c. For each setting, we employ 100 million training samples. Our results demonstrate that increasing the number of experts improves FID (lower values are preferable). However, adding spatial experts introduces additional computations, with the computational complexity bounded by the total number of experts. Once all available experts have been deployed, the computational complexity ceases to grow. In the right-hand side of Fig.6c, we provide a runtime analysis for 40 input tokens, ensuring the utilization of all space experts. For instance, when the number of experts is 6, the inference speed decreases by 24% but yields superior fidelity. This remains faster than previous diffusion models such as Imagen [1] and eDiff-I [4].

5 Related Work

We review related works from two perspectives, mixture-of-experts and text-to-image generation. More related works can be found in Appendix 7.4. Firstly, the Mixture-of-Experts (MoE) method [7, 8] partitions model parameters into distinct subsets, each termed an "expert". The MoE paradigm finds applicability beyond language processing tasks, extending to visual models [35] and Mixture-of-Modality-Experts within multi-modal transformers [36]. Additionally, efforts are being made to accelerate the training or inference processes for MoE [37, 38]. Secondly, text-to-image generation is to synthesize images from natural language descriptions. Early approaches relied on generative adversarial networks (GANs) [39, 40, 41, 42] to generate images. More recently, with the transformative success of transformers in generative tasks, models such as DALL-E [43], Cogview [44], and Make-A-Scene [29] have treated text-to-image generation as a sequence-to-sequence problem, utilizing auto-regressive transformers as generators and employing text/image tokens as input/output sequences. Recently, another research direction has focused on diffusion models by integrating textual conditioning within denoising steps, like Stable Diffusion [2], DALL-E 2 [3], eDiff-I [4], ERNIE-ViLG 2.0 [5], and Imagen [1].

6 Conclusion

This paper introduces RAPHAEL, a novel text-conditional image diffusion model capable of generating highly-artistic images using a large-scale mixture of diffusion paths. We carefully design space-MoE and time-MoE within an edge-supervised learning framework, enabling RAPHAEL to accurately portray text prompts, enhance the alignment between textual concepts and image regions, and produce images with superior aesthetic appeal. Comprehensive experiments demonstrate that RAPHAEL surpasses previous approaches, such as Stable Diffusion, ERNIE-ViLG 2.0, DeepFloyd, and DALL-E 2, in both FID-30k and the human evaluation benchmark ViLG-300. Additionally, RAPHAEL can be extended using LoRA, ControlNet, and SR-GAN. We believe that RAPHAEL has the potential to advance image generation research in both academia and industry.

Limitation and Potential Negative Societal Impact. We acknowledge some limitations in our paper that require attention. One limitation is the direct binarization of the attention map, which may result in the loss of some information. An adaptive module should be proposed to address this issue effectively. Additionally, the performance may be affected by failure cases of the edge detector, leading to potential degradation. We plan to explore solutions for these limitations in our future work. The potential negative social impact is to use the RAPHAEL API to create images containing misleading or false information. This issue potentially presents in all powerful text-to-image generators. We will solve this issue (*e.g.,* by prompt filtering) before releasing the API to the public.

Acknowledgments and Disclosure of Funding

This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622.

References

[1] 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.

[2] 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.

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

[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.

[5] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *arXiv preprint arXiv:2210.15257*, 2022.

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[7] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[8] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

[9] 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. *arXiv preprint arXiv:2106.09685*, 2021.

[10] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[11] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[14] 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.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[16] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[17] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[18] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403, 2015.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[20] 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. *arXiv preprint arXiv:2210.08402*, 2022.

[21] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*, 2023.

[22] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[23] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[26] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022.

[27] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.

[28] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021.

[29] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022.

[30] 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.

[31] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023.

[32] 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.

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

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[35] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

[36] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023.

[37] Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv:2103.13262*, 2021.

[38] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[39] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[41] 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.

[42] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[43] 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.

[44] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[45] Zeyue Xue, Jianming Liang, Guanglu Song, Zhuofan Zong, Liang Chen, Yu Liu, and Ping Luo. Large-batch optimization for dense visual predictions. *arXiv preprint arXiv:2210.11078*, 2022.

[46] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

[47] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.

[48] Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint arXiv:2208.03306*, 2022.

[49] 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. *arXiv preprint arXiv:2208.12242*, 2022.

[50] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[51] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

[52] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023.

[53] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023.

[54] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023.

[55] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.

[56] Lingting Zhu, Zeyue Xue, Zhenchao Jin, Xian Liu, Jingzhen He, Ziwei Liu, and Lequan Yu. Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–601. Springer, 2023.

[57] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

[58] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[59] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023.

[60] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[61] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023.

[62] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 2023.