

# 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<sup>†‡</sup>**

SenseTime Research  
liuyuisanai@gmail.com

**Ping Luo<sup>†</sup>**

The University of Hong Kong  
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.

<sup>†</sup>Project lead.

<sup>‡</sup>Corresponding authors.

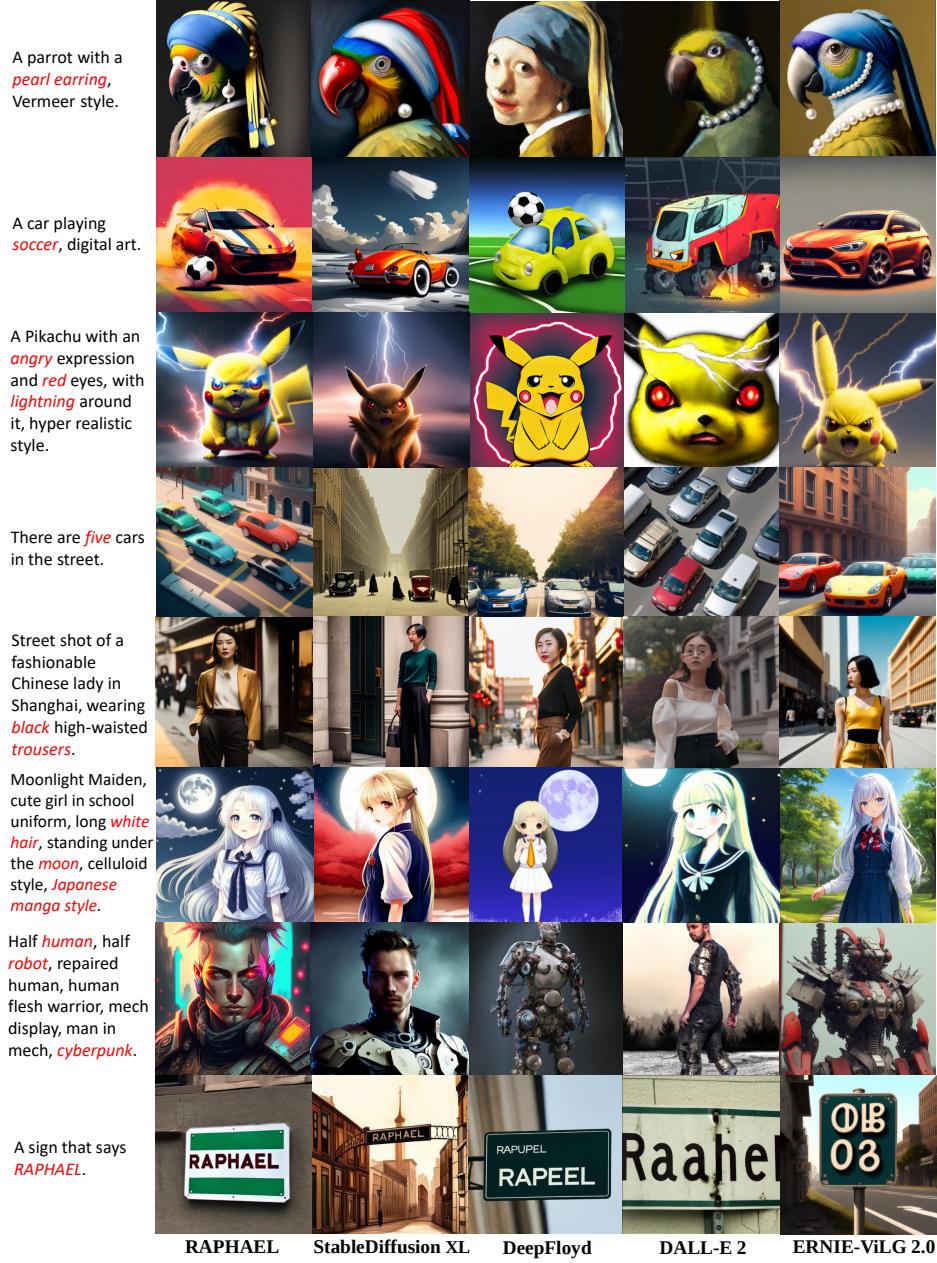


Figure 1: **Comparisons** of RAPHAEL with recent representative generators, Stable Diffusion XL [2], DeepFloyd, DALL-E 2 [3], and ERNIE-ViLG 2.0 [5]. They are given the same prompts, where the words that the human artists yearn to preserve within the generated images are highlighted in red. These images 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 “**d**istinct **i**mage **r**egions **a**lign with different text **p**hases in **a**ttention **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. (4) RAPHAEL is capable of generating images with resolutions up to  $4096 \times 6144$  with rich image contents and details, when combined with a tailor-made SR-GAN model [9].

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 [10], ControlNet [11], and SR-GAN [9]. **(iii)** We have released the demo of the latest version of RAPHAEL to the public\*, which has been fine-tuned on more high aesthetics datasets. 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  $\mathbf{x}_t = \sqrt{1 - \bar{\alpha}_t} \mathbf{x}_0 + \sqrt{\bar{\alpha}_t} \epsilon_t$ , 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,  $\mathcal{L}_{\text{denoise}} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon \sim \mathcal{N}(0, I)} \left[ \|\epsilon - D_\theta(\mathbf{x}_t, t)\|_2^2 \right]$ .

---

\*<https://miaohua.sensetime.com/zh-CN>



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.



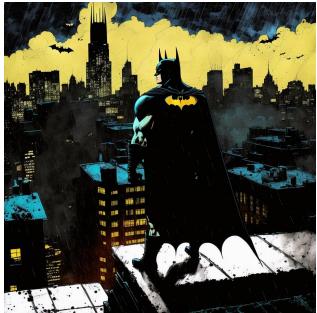
Photography closeup portrait of an adorable *rusty broken-down steampunk robot* covered in budding vegetation, surrounded by tall grass, *misty futuristic sci-fi forest environment*.



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.

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  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} D_\theta(\mathbf{x}_t, t) \right) + \sigma_t z$ , where  $\sigma_t$  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.

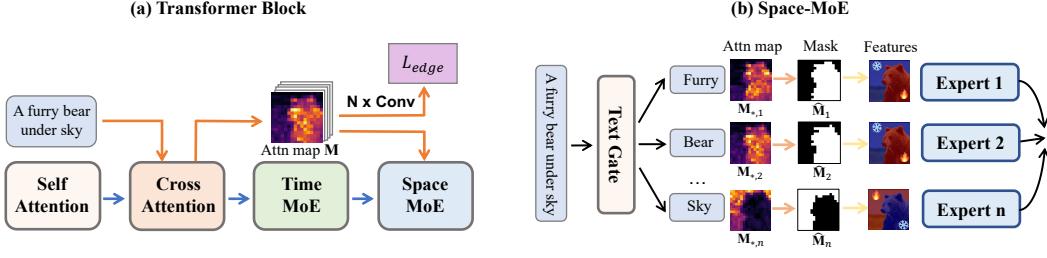


Figure 3: **Framework of RAPHAEL.** (a) Each block contains four primary components including a self-attention layer, a cross-attention layer, a space-MoE layer, and a time-MoE layer. The space-MoE is responsible for depicting different text concepts in specific image regions, while the time-MoE handles 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.

**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  $y$ ) into the U-Net, a text encoder neural network,  $E_\theta(y)$ , is employed to extract the textual representation. 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  $\text{attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$ , where  $\mathbf{Q}$ ,  $\mathbf{K}$ , and  $\mathbf{V}$  correspond to the query, key, and value matrices, respectively. These matrices are computed as  $\mathbf{Q} = h(\mathbf{x}_t)\mathbf{W}_x^{\text{qry}}$ ,  $\mathbf{K} = E_\theta(\mathbf{y})\mathbf{W}_y^{\text{key}}$ , and  $\mathbf{V} = E_\theta(\mathbf{y})\mathbf{W}_y^{\text{val}}$ , where  $\mathbf{W}_x^{\text{qry}} \in \mathbb{R}^{d \times d}$  and  $\mathbf{W}_y^{\text{key}}, \mathbf{W}_y^{\text{val}} \in \mathbb{R}^{d_y \times d}$  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,  $\mathbf{M} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \in \mathbb{R}^{n_x \times n_y}$ , 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:  $\frac{1}{n_y} \sum_{i=1}^{n_y} e_{\text{route}}(\mathbf{y}_i) (h'(\mathbf{x}_t) \circ \widehat{\mathbf{M}}_i)$ . 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  $\text{route}(\mathbf{y}_i)$  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  $\text{route}(\mathbf{y}_i) = \text{argmax}(\text{softmax}(\mathcal{G}(E_\theta(\mathbf{y}_i)) + \epsilon))$  is used,

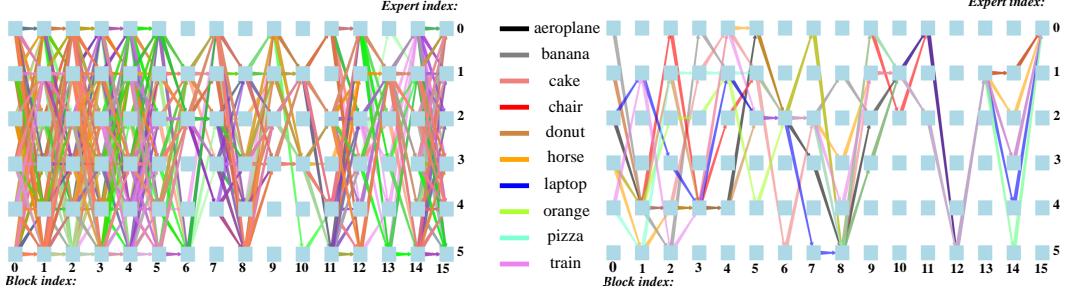


Figure 4: **Left:** We visualize the diffusion paths (routes) from the network input to the output, utilizing 16 space-MoE layers, each containing 6 space experts. These paths are closely associated with 100 adjectives, such as “scenic”, “peaceful”, and “majestic”, which represent the most frequently 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.

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  $\epsilon$  is incorporated. The 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  $\alpha$  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 with prompt templates given by GPT-3.5 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  $t = 1, \dots, T$ , 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  $P_\theta$ .  $\mathbf{I}_{\text{edge}}$  can be obtained by the holistically-nested edge detection algorithm [18] (Fig.7b). Intuitively, the network  $P_\theta$  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 LAION-5B [20] and some internal datasets. 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 in a bucket, which has 9 different image scales<sup>†</sup>. 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, 31], 30,000 images are randomly selected from the validation set for evaluation. Table 1 shows that RAPHAEL achieves a new state-of-the-art

---

<sup>†</sup>The [height, width] for each bucket is [448, 832], [512, 768], [512, 704], [640, 640], [576, 640], [640, 576], [704, 512], [768, 512], and [832, 448].

Table 1: **Comparisons** of RAPHAEL with the recent representative text-to-image generation models on the MS-COCO  $256 \times 256$  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
Stable Diffusion [2]	CVPR'22	Diffusion	-	8.32
Muse-3B [31]	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	-	<b>6.61</b>

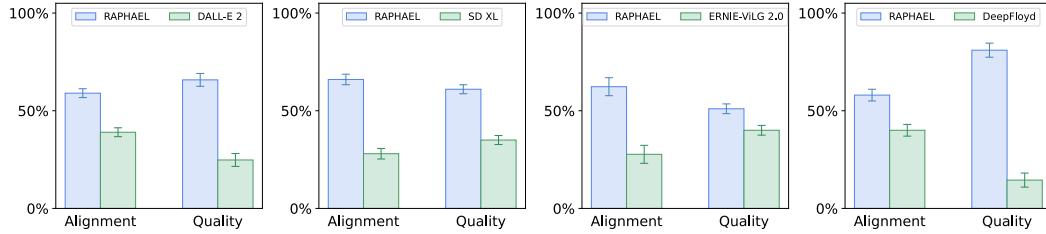


Figure 5: **Comparisons** of RAPHAEL with DALL-E 2, Stable Diffusion XL (SD XL), ERNIE-ViLG 2.0, and DeepFloyd in a user study using the ViLG-300 benchmark. We report the user’s preference rates with 95% confidence intervals. We see that RAPHAEL can generate images with higher quality and better conform 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 tailor-made SR-GAN model, we enhance the image resolution to  $4096 \times 6144$ .

## 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 guidance weights [32], specifically 1.5, 3.0, 4.5, 6.0, 7.5, and 9.0. Fig.6b compares these curves for

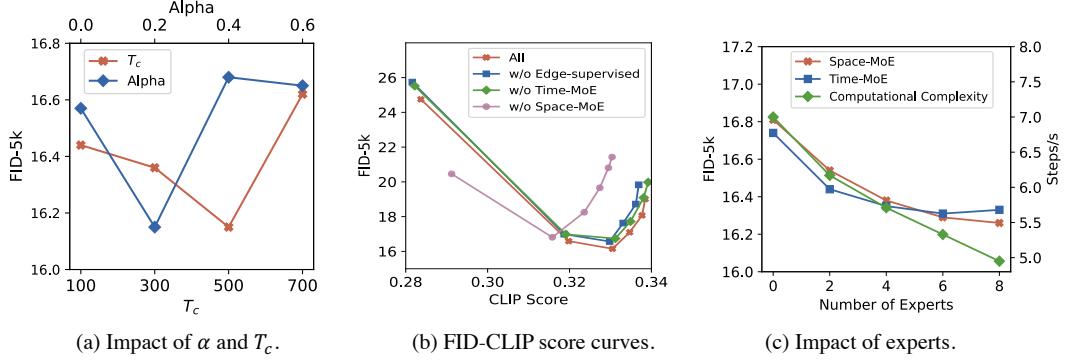


Figure 6: **Ablation Study.** (a) examines the selection of  $\alpha$  and  $T_c$ . (b) presents the trade-off between FID and CLIP scores for the complete RAPHAEL model and its variants without space-MoE, time-MoE, and edge-supervised learning. (c) visualizes the correlation between FID-5k and runtime complexity (measured in terms of the number of DDIM [33] 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 space experts.

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  $\alpha$  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 space 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 [34] and Mixture-of-Modality-Experts within multi-modal transformers [35]. Additionally, efforts are being made to accelerate the training or inference processes for MoE [36, 37]. Secondly, text-to-image generation is to synthesize images from natural language descriptions. Early approaches relied on generative adversarial networks (GANs) [38, 39, 40, 41] to generate images. More recently, with the transformative success of transformers in generative tasks, models such as DALL-E [42], Cogview [43], 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.** 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.

## 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] 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.
- [10] 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.
- [11] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.
- [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] 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.
- [32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] David Eigen, Marc’Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.
- [46] 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.
- [47] 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.
- [48] 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.

- [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.

## 7 Appendix

### 7.1 Hyper-parameters and Values

We give the hyper-parameters and values in Table 2.

Table 2: Hyper-parameters and values in RAPHAEL.

Configs/Hyper-parameters	Values
$T$	1000
$n_y$	77
$d_y$	1024
$T_c$	500
$\alpha$	0.2
Betas of AdamW [25]	(0.9, 0.999)
Weight decay	0.0
Learning rate	1e-4
Number of space experts	6
Number of time experts	4
Warmup steps	20000
Batch size	2000
Number of GPUs	1000
Number of transformer blocks	16
Use checkpoint	True
$\alpha$ in Focal Loss [19]	0.5
$\gamma$ in Focal Loss [19]	2
Text encoder	OpenCLIP-g/14 [14]
Enable multi-scale training	True
Activations in experts and gate network	GELU
Architectures of experts and gate network	FFN

### 7.2 Details of Edge-supervised Learning

We provide some demonstrations of edge-supervised learning in Fig.7.

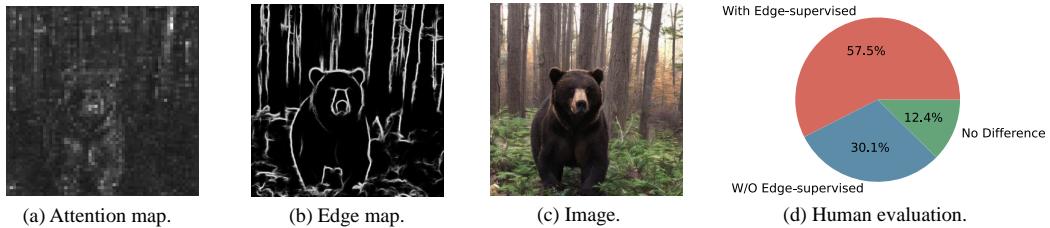


Figure 7: From left to right, we display the attention map corresponding to the pooled token in CLIP, the ground truth edges identified by the edge detection algorithm, and the associated image. In the fourth figure, we present the human evaluation results for models with and without edge-supervised learning on the ViLG-300 benchmark. Evaluators are instructed to compare these images considering image aesthetics and we report the preference rates, and our findings indicate that edge-supervised learning significantly enhances the aesthetic quality of the images.

### 7.3 Details on Time-MoE

The overall architecture of RAPHAEL can be found in Fig.8. The Time-MoE is composed of a Time Gate Network to distribute the features to a specific expert according to the timestep, which can be formulated as  $h'(\mathbf{x}_t) = te_{t\_router(t_i)}(h_c(\mathbf{x}_t))$ . In this equation,  $h_c(\mathbf{x}_t)$  is the features from

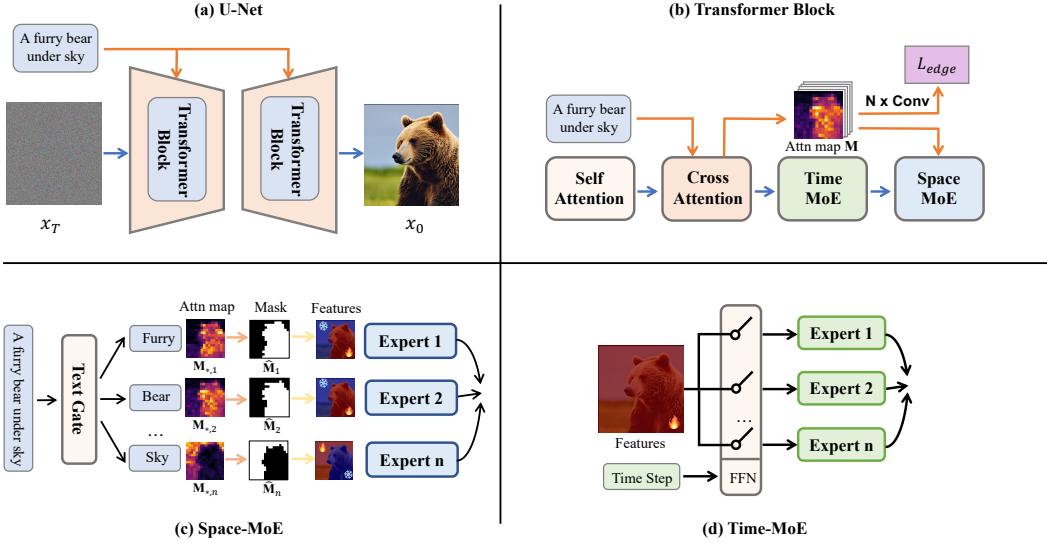


Figure 8: (a) The architecture of U-Net, which consists of many transformer blocks. (b) Each block contains four primary components including a self-attention layer, a cross-attention layer, a space-MoE layer, and a time-MoE layer. The space-MoE is responsible for depicting different text concepts in specific image regions, while the time-MoE handles different diffusion timesteps. Each block uses edge-supervised cross-attention learning to further improve image quality. (c) shows details of space-MoE. For example, given a prompt “a fury 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. (d) For time-MoE, an initial timestep is provided, followed by the selection of an expert responsible for handling the visual features.



Figure 9: The routes of time-MoE in the first transformer block, where the first expert focuses on noisy images, while other experts handle images with low noise levels.

cross-attention module. The gating function  $t\_router$  returns the index of an expert in the Time-MoE, with  $\{te_1, te_2, \dots, te_{n_t}\}$  being a set of  $n_t$  experts. Concretely, the Time Gate Network is implemented by a function,  $t\_router(t_i) = \text{argmax}(\text{softmax}(\mathcal{G}'(E'_\theta(t_i)) + \epsilon))$  at timestep  $t_i$ . To prevent mode collapse, random noise  $\epsilon$  is incorporated. Similar to the Text Gate,  $\mathcal{G}' : \mathbb{R}^{d_t} \mapsto \mathbb{R}^{n_t}$ , is a feed forward network, where  $d_t$  is the dimension of the time embedding  $E'_\theta(t_i)$ .

**Analysis.** In our exploration, we uncover some statistical regularities within the routes of time experts across all transformer blocks, establishing a clear correlation with the timestep dimension. Notably, we observe a distinct division of labor among these experts, specializing in timesteps characterized by varying levels of noise. For instance, as illustrated in Fig.9, in the first transformer block, the first expert predominately focuses on processing noisy images (representing the initial 59% of DDIM sampler steps), while the remaining experts handle images with relatively lower noise levels (representing the final 41% of DDIM sampler steps). This systematic allocation of expertise based on noise characteristics underscores the model’s ability to adapt its computational resources efficiently and effectively.

## 7.4 Related Work

Foundation models have achieved remarkable success in several fields [2, 44, 34, 15], especially for text-to-image generation. We review related works from two perspectives, mixture-of-experts, and text-to-image generation.

**Mixture-of-Experts.** The Mixture-of-Experts (MoE) method [7, 8] in neural networks partitions specific model parameters into distinct subsets, each termed an "expert." During forward propagation, a dynamic routing mechanism assigns these experts to diverse inputs, with each input exclusively interacting with its selected experts. MoE models implement a learned gating function that selectively activates a subset of experts, enabling the input to engage either all experts [45] or a sparse mixture thereof [8, 46], as evidenced in recent expansive language models. While a multitude of models employs experts strictly within the linear layers, other research regards an entire language model as an expert [47]. The MoE paradigm finds applicability beyond language processing tasks, extending to visual models [34] and Mixture-of-Modality-Experts within multi-modal transformers [35]. Additionally, efforts are being made to accelerate the training or inference processes within the MoE paradigm [36, 37].

**Text-to-Image Generation.** Text-to-image generation, the task of synthesizing images from natural language descriptions, has experienced significant progress in recent years. Early approaches relied on generative adversarial networks (GANs) [38, 39, 40, 41, 48] to generate images. More recently, with the transformative success of transformers in generative tasks, models such as DALL-E [42], Cogview [43], 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, framing the task as an iterative denoising process. By integrating textual conditioning within denoising steps, models like Stable Diffusion [2], DALL-E 2 [3], eDiff-I [4], ERNIE-ViLG 2.0 [5], and Imagen [1] have consistently set new benchmarks in text-to-image generation. Specifically, Stable Diffusion and ERNIE-ViLG 2.0 map images into a latent space, following the Latent Diffusion Model paradigm to enhance training and sampling efficiency, while DALL-E 2, eDiff-I, and Imagen operate in pixel space. Furthermore, diffusion models also show great potential in image editing [49, 6, 50, 51], personalized generation [52, 53, 54, 55], and 3D/video/gesture generation [56, 57, 58, 59, 60, 61]. ControlNet [11, 62] is a noteworthy model in the text-to-image generation landscape. It builds upon the concept of controllable image synthesis, wherein generated images can be manipulated based on user-defined constraints or attributes.

## 7.5 More Details on Routers of Space-MoE

We continue to delve into the diffusion paths of both COCO categories and verbs, uncovering intriguing insights. Utilizing the powerful GPT-3.5 [15, 16], we randomly generate 50 verbs to enrich our investigation. Moreover, employing the prompt template randomly generated by GPT-3.5, we generate 100 samples for each COCO category and verb [15, 16]. Similar to Section 3.1, by adopting XGBoost as our classifier, we find that the accuracy rate reaches 94.3% and 97.5%, respectively. We give the routes of COCO and 50 verbs in Fig.10. We also show more visualization results of the attention maps in Fig.11. We provide the adjectives and verbs in the following section.

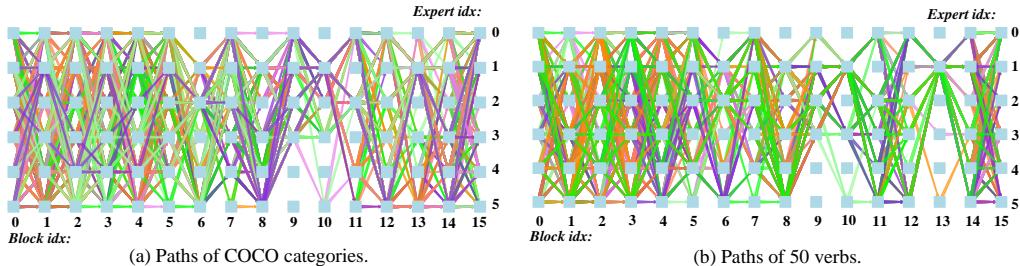


Figure 10: We visualize the diffusion paths (routes) from the network input to the output, utilizing 16 space-MoE layers, each containing 6 space experts. These paths are closely associated with COCO categories and 50 verbs.

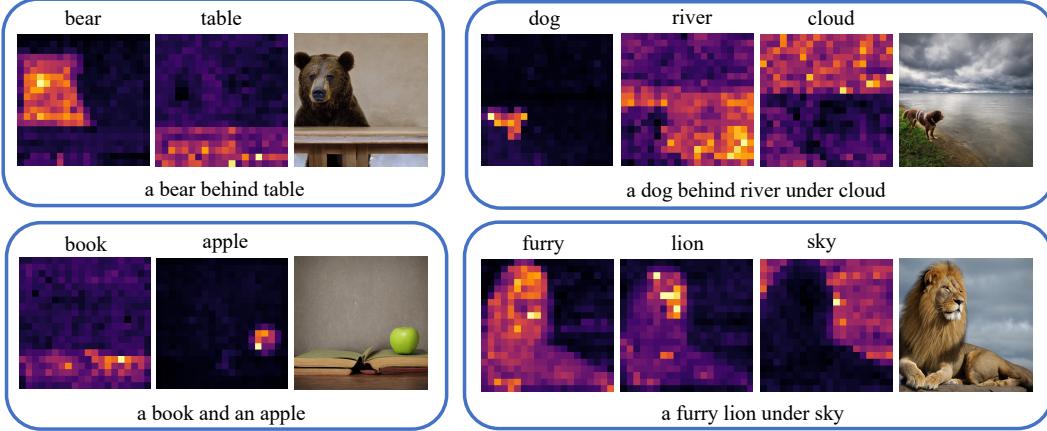


Figure 11: We give the prompts, their associated generated images, and attention maps.

### 7.5.1 Adjectives and Verbs

**Adjectives.** Aesthetic, alluring, artistic, astonishing, attractive, baroque, beautiful, blissful, captivating, chic, classic, coastal, colorful, common, dark, decorative, delicate, dramatic, dreamlike, dreamy, dynamic, eclectic, elegant, emotive, enchanting, energetic, entralling, essential, ethereal, evocative, extraordinary, fascinating, flexible, fragile, futuristic, glamorous, glossy, gorgeous, gothic, grand, harmonious, idyllic, impressive, industrial, innovative, inspiring, intricate, intriguing, joyful, lively, luxurious, magnificent, meditative, mesmerizing, minimal, minimalist, modern, moroccan, mysterious, nostalgic, ordinary, patterned, peaceful, picturesque, plain, playful, practical, quirky, rare, renaissance, retro, rigid, romantic, rough, rustic, satisfying, scandinavian, scenic, serene, serious, shiny, simple, sleek, smooth, sophisticated, static, striking, stunning, sturdy, stylish, textured, traditional, tranquil, unique, unusual, useful, vibrant, victorian, vivid, whimsical.

**Verbs.** Balance, blend, blossom, bond, carve, celebrate, cheer, climb, collaborate, conduct, conquer, cook, craft, create, dance, dream, embrace, experiment, explore, gaze, harmonize, hike, hug, ignite, illuminate, jump, laugh, leap, listen, meander, observe, paint, play, ponder, read, rejoice, relax, ride, run, savor, sculpt, sing, smile, soar, surf, swim, swing, taste, wander, whisper.

## 7.6 More Comparisons between RAPHAEL and Prestigious Diffusion Models

In this section, we provide more comparisons between RAPHAEL and Midjourney, Stable Diffusion XL, DALL-E 2, DeepFloyd, ERNIE-ViLG 2.0 in Fig.12 and 13.



1. A cute little matte low poly isometric **cherry blossom forest island, waterfalls**, lighting, soft shadows, trending on Artstation, 3d render, monument valley, fez video game.
2. A shanty version of Tokyo, new rustic style, **bold colors with all colors palette**, video game, genshin, tribe, fantasy, overwatch.
3. Cartoon characters, mini characters, figures, illustrations, flower fairy, green dress, **brown hair, curly long hair, elf-like wings, many flowers and leaves**, natural scenery, **golden eyes**, detailed light and shadow , a high degree of detail.
4. Cartoon characters, mini characters, hand-made, illustrations, **robot kids**, color expressions, boy, **short brown hair, curly hair, blue eyes**, technological age, **cyberpunk**, big eyes, cute, mini, detailed light and shadow, high detail.

Figure 12: The prompts for each column are given in the figure. We give the comparisons between DALL-E 2 Midjourney v5.1, Stable Diffusion XL, ERNIE ViLG 2.0, DeepFloyd, and RAPHAEL. They are given the same prompts, where the words that the human artists yearn to preserve within the generated images are highlighted in red. Only the RAPHAEL-generated images precisely reflect the prompts, while other models generate compromised results. For images with cartoon styles, we switch Midjourney v5.1 to Nijijourney v5.



1. Landscape, *lake, buildings, bridge*, Chinese ink style.
2. Photo of *an athlete cat* explaining it's latest scandal at a *press conference* to journalists.
3. Impasto, illustration character, illustration poster, techno, cyberpunk, *half human, half robot, repaired human*, human warrior, mech display, woman in mech, purple eyes, *half face destroyed*, trend on Artstation, high detail, detailed light and shadow, 4k.
4. Atmosphere, flat painting illustration, illustration, *unique tall buildings in the city*, irregular design, iron age, *black and orange*, detailed light and shadow, cyberpunk, technological age, high detail, *CG feeling*.

Figure 13: The prompts for each column are given in the figure. We give the comparisons between DALL-E 2, Midjourney v5.1, Stable Diffusion XL, ERNIE ViLG 2.0, DeepFloyd, and RAPHAEL. They are given the same prompts, where the words that the human artists yearn to preserve within the generated images are highlighted in red. Only the RAPHAEL-generated images precisely reflect the prompts, while other models generate compromised results.

## 7.7 More Images Generated by RAPHAEL

We give more cases in Fig.14, 15, 16, 17, and 18.



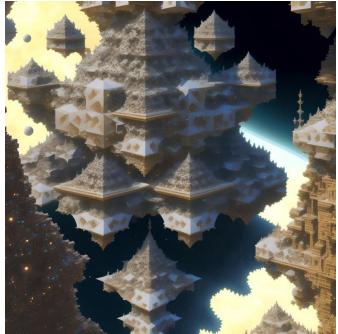
Coloring page for kids, Giant Monster Truck Lifted With Big Wheels, cartoon style, thick lines, low detail.



Watercolor purple Hydrangea clipart  
Hydrangea floral bouquet in a jar, white background, no text.



Sushi shop, isometric, high quality, old tokyo vibes, night, Japanese, 3d.



Spacefaring civilization silicon-based lifeforms megastructures built into fractal mineralized shapes, Menger sponge Sierpinski gasket koch snowflake Mandelbulb, key visual by craig mullins and yoji Ishikawa in the style of fractal core spacewave.



Utopian biophilic eco mansion design in a magical enchanted dream lush pine forest Trend on Artstation, Altered Carbon, waterfalls, streams, pools of golden glowing bioluminescent water, space ship take off, amazing night, photography, cinematic, aerial side ground view, Zoha Hadid.



A crowd of people coloured and light in a theater, christopher balaskas, atey ghalian, mirror rooms, dynamic figure studies, misty, light cyan.



Realistic artistic astrobiological painting, a wide view of the sea floor under the waves of an alien ocean, where life has crawled out from the deep to cling to the sundappled seabed and begin the eternal dance of evolution into new and innovative forms.



A man, a woman and a child hand in hand, walking under blooming flower trees, miniature photo, overlook, paper kirigami craft, high detail.



A red dwarf star showing a large blast of light, alien worlds, red, hyper-realistic water, deconstructive, angura kei, camera tossing, ansel adams, guy aroch, aerial view .

Figure 14: These examples show that RAPHAEL can generate artistic images with varying text prompts across various styles.



Celluloid style, Japanese manga style, cute girl in school uniform, long white hair, under the moon, night, high detail.



Photography, a woman wearing white blazers in a street, chic, copy space, fashion, low light, models, outfit, pose.



Photography, cute child, wearing a down jacket, cat ears, fleshy face, blue high details, detailed light and shadow.



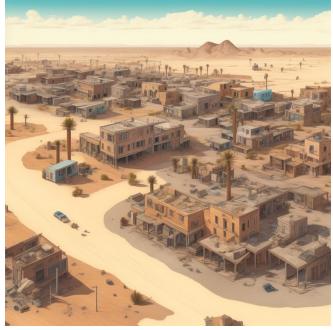
Bright scene, aerial view, ancient city, fantasy, gorgeous light, mirror reflection, high detail, wide angle lens.



Spaceship, a swordsman, magic, fantasy, fog, surreal, bright and brilliant light, gorgeous, glory, epic, high detail.



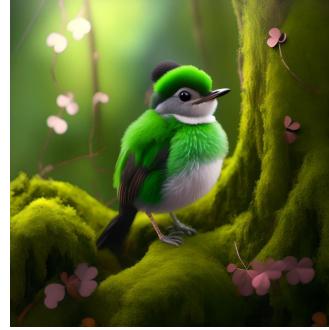
Minimalist sci-fi illustration, futuristic metropolis, tall skyscrapers, floating gardens, crowd, multi-level, detail.



A picture on the browser of an abandoned desert town, collage - oriented, aerial view, ultra detail.



Mayan empire, use Mayan architecture, modern, technological cyberpunk style, photo quality.



A very cute little Shamrock bird in a mossy forest.

Figure 15: These examples show that RAPHAEL can generate artistic images with varying text prompts across various styles.

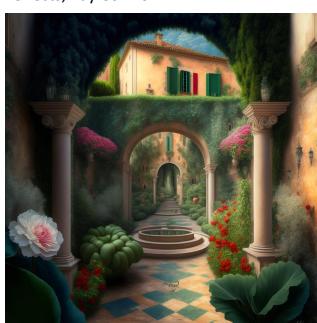
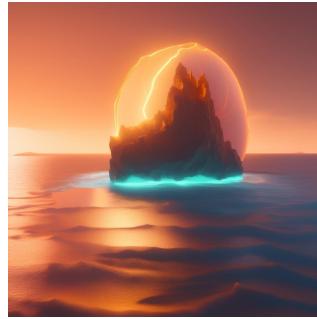


Figure 16: These examples show that RAPHAEL can generate artistic images with varying text prompts across various styles.



Celluloid style, Japanese manga style, flat-painted illustration, due teenager, big bag, short black hair fluttering in the wind, on the road by the sea, high detail.



Impasto, Japanese and Korean manga characters, CG characters, CG avatars, magical girls, black magic hats, with flowers on the hats, elf ears, translucent, long gray hair, Slightly curly hair, trends on Artstation, high detail.



Flat painting, Japanese and Korean manga characters, male protagonist, elf boy, white and green hair, long hair, long ears, sea, bubbles, black suit and green tie, evil smile, green eyes, anime avatar, detailed light and shadow, high detail.



American comic character, flat painting, illustration character, avatar, flat painting illustration, short brown hair, curly hair, green clothes, ID photo pose, jane's style, trends on artstation , detailed light and shadow, high detail.



Flat drawing illustration, flat drawing, illustration, girl in pajamas sitting on the bed, a puppy, sunset, the light of the setting sun enters the house through the window, detailed light and shadow, high detail, 4k.



Cartoon character, mini character, figure, illustration, dango big family, bunny ears, different colors, character design, a set of avatars, cute, mini detail light and shadow, high detail.



Furry character, character design, impasto, cute fox girl, long brown hair, curly long hair, fluttering in the wind, messy, big eyes, smile, petals fluttering in the wind, high detail, CG characters.



Impasto, avatar, illustration, girl with red hair, slightly curly hair, European and American, freckles, jane's style, trends on artstation, crazy colors, light and shadow contrast, high detail.



Impasto, Japanese and Korean manga characters, illustration characters, avatars, CG characters, CG avatars, magical girls, purple magic hats, short silver-white shoulder-length hair, curly hair, two-dimensional, detailed light and shadow, high detail.

Figure 17: These examples show that RAPHAEL can generate artistic images with varying text prompts across various styles.



Character standing drawing, character design, impasto, samurai in black armor with long black hair, holding a long towel in his hand, blindfolded by a long cloth, CG game , game characters, high detail, color contrast.



Impasto, avatar, illustration, color explosion, elf girl on the clouds, reflection, elf ears, colorful clouds, Jane's style, trends on Artstation, light and shadow contrast, high detail.



Impasto, illustration character, illustration poster, tech style, cyberpunk, woman with black long hair, hair up, black leather jacket, mechanical arm, tattoo on face, holding a purple lightsaber, war, trends on Artstation, high detail.



Spike Spiegel wearing Akira Shōtarō Kanesawa's red jacket and outfit, Katsuhiro Otomo, cinematic, extremely detailed and complex, impressive, super resolution, megapixel.



A Greek god, Erebus wearing ancient greek clothing, galaxy with a solar system as background, cinematic, soft studio lighting, backlighting, dark background.



Pixel art, videogame city wallpaper, deskmat. io, low poly, large, wide angle , from the top.



A tall mythical spire mountain with a base of clouds below during golden hour.



Atmosphere, illustration, mecha sports car, mecha era, running on the bustling street, blue, purple, cool.



A cute fluffy sentient alien from planet Axor, in the Andromeda galaxy, the alien have large innocent eyes and is digitigrade, high detail.

Figure 18: These examples show that RAPHAEL can generate artistic images with varying text prompts across various styles.

## 7.8 Extension to LoRA, ControlNet, and SR-GAN

We give the results of LoRA in Fig.19 and 20, ControlNet in Fig.21, and SR-GAN in Fig.22 and 23. The detailed settings are given in captions.



Figure 19: **Results with LoRA.** We use 28 images to finetune RAPHAEL and Stable Diffusion. The prompts are "A spider-man figurine, in the night/on the moon/on the beach/none", only RAPHAEL preserves the concepts in prompts while Stable Diffusion yields compromised results.

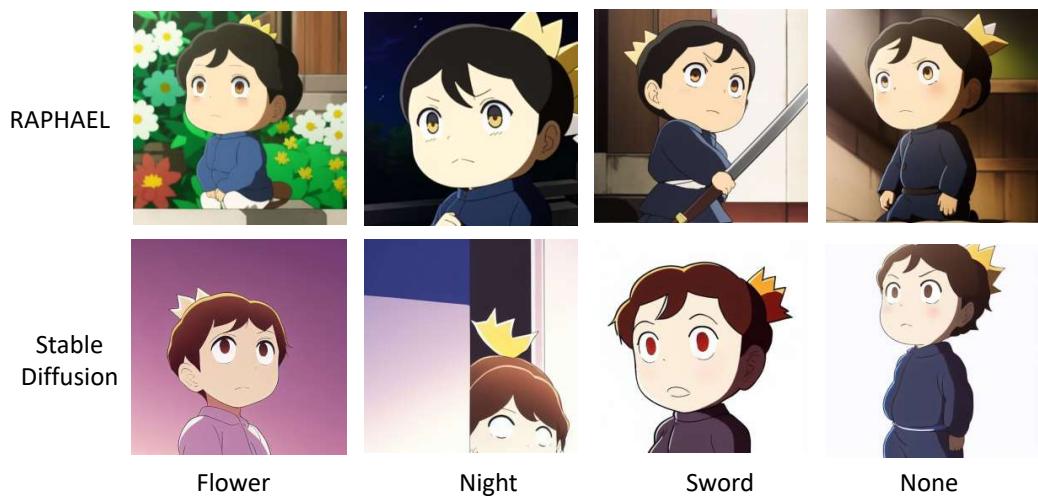
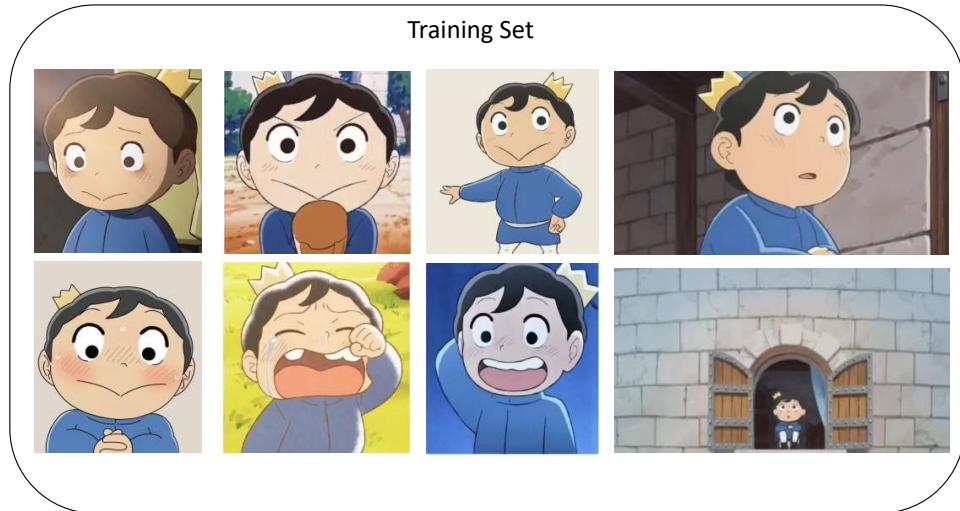


Figure 20: **Results with LoRA.** We use 32 images to finetune RAPHAEL and Stable Diffusion. The prompts are "A boy, flower/night/sword/none", only RAPHAEL preserves the concepts in prompts while Stable Diffusion yields compromised results.

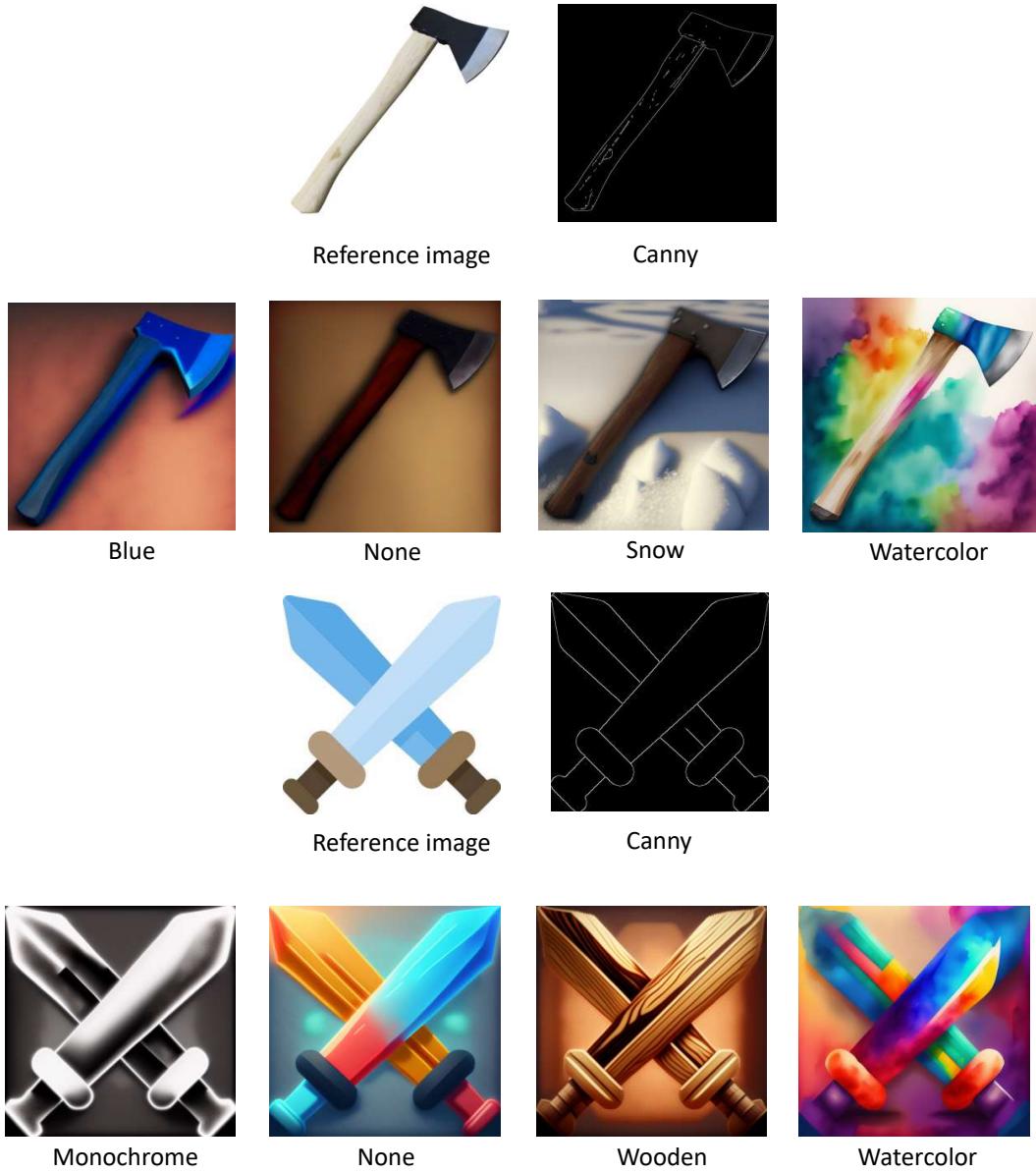


Figure 21: **Results with ControlNet.** We use the reference image to generate canny edges and adopt it as the extra constraint for RAPHAEL. The prompts for each group are "An ox, blue/none/snow/watercolor" and "Icon for game, fighting skill, monochrome/none/wooden/watercolor".



Pirate ship trapped in a cosmic maelstrom nebula, rendered in cosmic beach whirlpool engine, volumetric lighting, spectacular, ambient lights, light pollution, cinematic atmosphere, art nouveau style, illustration art artwork by SenseiJaye, intricate detail.

Figure 22: Result of 4096×6144 image. SR-GAN enhances the resolution of the image generated by RAPHAEL.



A surreal parallel world where mankind avoid extinction by preserving nature, epic trees, water streams, various flowers, intricate details, rich colors, rich vegetation, cinematic, symmetrical, beautiful lighting, V-Ray render, sun rays, magical lights, photography.

Figure 23: Result of  $4096 \times 6144$  image. SR-GAN enhances the resolution of the image generated by RAPHAEL.