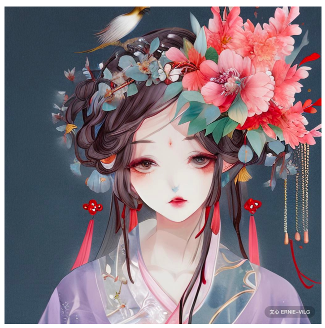
ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts

Zhida Feng[1,2,*], Zhenyu Zhang[1,*], Xintong Yu[1,*], Yewei Fang[1], Lanxin Li[1], Xuyi Chen[1], Yuxiang Lu[1], Jiaxiang Liu[1], Weichong Yin[1], Shikun Feng[1], Yu Sun[1], Li Chen[2], Hao Tian[1], Hua Wu[1], Haifeng Wang[1]
[1]Baidu Inc.
[2]School of Computer Science and Technology, Wuhan University of Science and Technology
{fengzhida, zhangzhenyu07, yuxintong}@baidu.com
{liujiaxiang, yinweichong, fengshikun01, sunyu02}@baidu.com

一艘青花瓷质感的战舰

A warship with blue and white porcelain texture



精致面容的古风少女，头戴百鸟鲜花头冠

An ancient style girl with delicate face, wearing a crown of birds and flowers

一只仙鹤站在平静的湖面上，后面有一轮云雾缭绕的明月

A crane is standing on the calm lake, with a bright moon surrounded by clouds in the background

穿西装的孙悟空的特写镜头

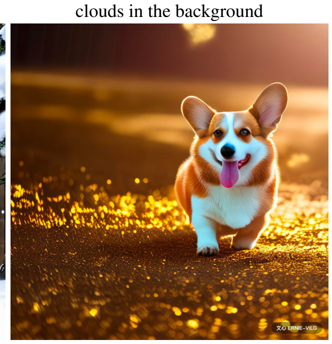
The close up of Sun Wukong in a suit

狮子王身穿紫色皇家大衣，身披红色皇家斗篷，正在发表史诗级的演说

The Lion King, wearing a purple royal coat and a red royal cloak, is delivering an epic speech

雪地里的熊猫宝宝戴着红帽子、黄手套，穿着绿毛衣和蓝裤子

A baby panda in the snow, wearing red hat, yellow gloves, green sweater and blue pants



一只欢快的柯基行走在夕阳里，光从右边打来铺成一地金黄

A cheerful Koji is walking in the sunset, the golden light comes from the right side and spreads all over the ground

泰迪熊穿着戏服，站在太和殿前唱京剧

A teddy bear, wearing a costume, is standing in front of the Hall of Supreme Harmony and singing Beijing opera

Figure 1. Selected 1024×1024 samples with various text inputs, which shows that ERNIE-ViLG 2.0 has powerful capabilities of fine-grained semantic control and high-resolution image synthesis, as well as to produce high-quality creative images of different styles.

Abstract

Recent progress in diffusion models has revolutionized the popular technology of text-to-image generation. While existing approaches could produce photorealistic high-resolution images with text conditions, there are still several open problems to be solved, which limits the further improvement of image fidelity and text relevancy. In this paper, we propose ERNIE-ViLG 2.0, a large-scale Chinese text-to-image

* denotes equal contribution. Work done during Feng's internship at Baidu.

diffusion model, to progressively upgrade the quality of generated images by: (1) incorporating fine-grained textual and visual knowledge of key elements in the scene, and (2) utilizing different denoising experts at different denoising stages. With the proposed mechanisms, ERNIE-ViLG 2.0[1] *not only achieves a new state-of-the-art on MS-COCO with zero-shot FID-30k score of 6.75, but also significantly outperforms recent models in terms of image fidelity and image-text alignment, with side-by-side human evaluation on the bilingual prompt set ViLG-300.*

1. Introduction

Recent years have witnessed incredible progress in text-to-image generation. With large-scale training data and model parameters, kinds of text-to-image generation models are now able to vividly depict the visual scene described by a text prompt, and enable anyone to create exquisite images without sophisticated drawing skills. Among all types of image generation approaches, diffusion models [9] are attracting increasing attention due to their ability to produce highly photorealistic images conditioned on text prompts. Given a text prompt, the models transform a Gaussian noise into an image that conforms to the prompt through iterative denoising steps. In the past years, text-to-image diffusion models such as LDM [25], GLIDE [18], DALL-E 2 [22], and Imagen [26] have achieved impressive performance in both text relevancy and image fidelity. Despite these advances, the exploration of diffusion models by existing methods is still at the initial stage. When we go deep into the principle and implementation of text-to-image diffusion models, there are still many opportunities to improve the quality of generated images further.

First, during the learning process of each denoising step, all text tokens interact with image regions and all the image regions contribute equally to the final loss function. However, a visual scene of text and image contains many elements (i.e., textual words and visual objects), and different elements usually hold different importance for the expression of the scene semantics [42]. The indiscriminate learning process may cause the model to miss some key elements and interactions in the scene, thus facing the risk of text-image misalignment, such as the attribute confusion problem, especially for text prompts containing multiple objects with specific attributes [22]. Second, when opening the horizon from individual step to the whole denoising process, we can found that the requirements of different denoising stages are also not identical. In the early stages, the input images are highly noised, and the model is required to outline the semantic layout and skeleton out of almost pure noise. By contrast, in the later steps close to the image output, denoising mainly means improving the details based on an almost completed image [25]. In practice, existing models usually use one U-Net for all steps, which means that the same set of parameters has to learn different denoising capabilities.

[1] https://wenxin.baidu.com/ernie-vilg

In this paper, we propose ERNIE-ViLG 2.0, an improved text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts, to incorporate extra knowledge about the visual scene and decouple the denoising capabilities in different steps. Specifically, we employ a text parser and an object detector to extract key elements of the scene in the input text-image pair, and then guide the model to pay more attention to their alignment in the learning process, so as to hope the model could handle the relationships among various objects and attributes. Moreover, we divide the denoising steps into several stages and employ specific denoising "experts" for each stage. With the mixture of multiple experts, the model can involve more parameters and learn the data distribution of each denoising stage better, without increasing the inference time, as only one expert is activated in each denoising step.

With the extra knowledge from the visual scene and the mixture-of-denoising-experts mechanism, we train ERNIE-ViLG 2.0 and scale up the model size to 24B parameters. Experiments on MS-COCO show that our model exceeds previous text-to-image models by setting a new state-of-the-art of 6.75 zeros-shot FID-30k score, and detailed ablation studies confirm the contributions of each proposed strategy. Apart from automatic metrics, we also collect 300 bilingual text prompts that could assess the quality of generated images from different aspects and enable a fair comparison between English and Chinese text-to-image models. The human evaluation results again indicate that ERNIE-ViLG 2.0 outperforms other recent methods, including DALL-E 2 [22] and Stable Diffusion [25], by a significant margin both in terms of image-text alignment and image fidelity.

To sum up, the main contributions of this work are: (1) We incorporate textual and visual knowledge into the text-to-image diffusion model, which effectively improves the ability of fine-grained semantic control and alleviates the problem of object-attribute mismatching in generated images. (2) We propose the mixture-of-denoising-experts mechanism to refine the denoising process, which can adapt to the characteristics of different denoising steps and scale up the model to 24B parameters, making it the largest text-to-image model at present. (3) ERNIE-ViLG 2.0 achieves the state-of-the-art zero-shot FID-30k score of 6.75 on MS-COCO, surpasses DALL-E 2 and Stable Diffusion in human evaluation on the Chinese-English bilingual prompt set ViLG-300.

2. Method

During the training process, the text-to-image diffusion model receives paired inputs (x, y) consisting of an image x with its text description y, and the ultimate goal is to gener-

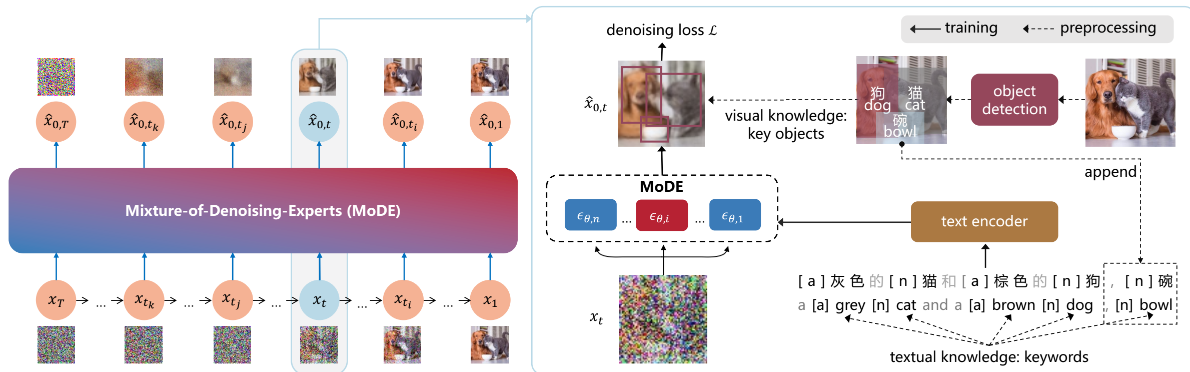


Figure 2. The architecture of ERNIE-ViLG 2.0, which incorporates fine-grained textual and visual knowledge of key elements in the scene and utilizes different denoising experts at different denoising stages.

ate x based on y. To achieve this, a text encoder $f_\theta(\cdot)$ first encodes y as $f_\theta(y)$, then a denoising network $\epsilon_\theta(\cdot)$ conditioned on $f_\theta(y)$ learns to generate x from a Gaussian noise. To help the model generate high-quality images that accurately match the text description (i.e., text prompt), ERNIE-ViLG 2.0 enhances text encoder $f_\theta(\cdot)$ and denoising network $\epsilon_\theta(\cdot)$ with textual and visual knowledge of key elements in the scene. Furthermore, ERNIE-ViLG 2.0 employs mixture-of-denoising-experts to refine the image generation process, where different experts are responsible for different generation steps in the denoising process. The overall architecture of ERNIE-ViLG 2.0 is shown in Figure 2 and the details are described in the following subsections.

2.1. Preliminary

Denoising diffusion probabilistic models (DDPM) are a class of score-based generative models that have recently shown delightful talents in the field of text-to-image generation [9]. The diffusion process of DDPM aims to iteratively add diagonal Gaussian noise to the initial data sample x and turn it into an isotropic Gaussian distribution after T steps:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \quad t \in \{1, \dots, T\} \tag{1}$$

where the sequence $\{x_t\}$ starts with $x_0 = x$ and ends with $x_T \sim \mathcal{N}(0, I)$, the added noise at each step is $\epsilon_t \sim \mathcal{N}(0, I)$, and $\{\alpha_t\}_{1\dots T}$ is a pre-defined schedule [30, 32]. The denoising process is the reverse of diffusion, which converts the Gaussian noise $x_T \sim \mathcal{N}(0, I)$ back into the data distribution x_0 through iterative denoising steps $t = T, \dots, 1$. During training, for a given image x, the model calculates x_t by sampling a Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Given x_t, the target of the denoising network $\epsilon_\theta(\cdot)$ is to restore x_0 by predicting the noise ϵ. It is learned via the loss function

$$\mathcal{L} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,I), t}\left[||\epsilon - \epsilon_\theta(x_t, t)||_2^2 \right]. \tag{3}$$

With the predicted $\epsilon_\theta(x_t, t)$, we can have the prediction of x_0 at step t by converting Equation (2):

$$\hat{x}_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)). \tag{4}$$

In Figure 2, we visualize the sampled x_t and the predicted $\hat{x}_{0,t}$ for several timesteps during training. In the inference process of DDPM, x_0 is unknown, so the model iteratively generates x_{t-1} based on x_t and $\hat{x}_{0,t}$:

$$\begin{aligned} x_{t-1} =& \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\sqrt{\alpha_t} x_t + \frac{1-\alpha_t}{1-\bar{\alpha}_t}\sqrt{\bar{\alpha}_{t-1}}\hat{x}_{0,t} \\ &+ \sqrt{\frac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t}}\epsilon_t', \quad t \in \{T, \dots, 1\}, \end{aligned} \tag{5}$$

where $\epsilon_t' \sim \mathcal{N}(0, I)$ is a sampled Gaussian noise.

The denoising network $\epsilon_\theta(\cdot)$ is typically implemented by U-Net [9]. To allow $\epsilon_\theta(\cdot)$ to condition on text prompts, a text encoder $f_\theta(\cdot)$ first extracts the text representation $f_\theta(y) \in \mathbb{R}^{n_y \times d_y}$, which is then fed into $\epsilon_\theta(\cdot)$ via a cross-modal attention layer [18]. Formally, the U-Net representation $\varphi_i(x_t) \in \mathbb{R}^{n_x \times d}$ is concatenated with the text representation $f_\theta(y)$ after projection, and then goes through an attention layer to achieve cross-modal interaction,

$$\begin{aligned} Q &= \varphi_i(x_t) W_Q^{(i)}, \\ K &= [\varphi_i(x_t) W_{K_x}^{(i)}; f_\theta(y) W_{K_y}^{(i)}], \\ V &= [\varphi_i(x_t) W_{V_x}^{(i)}; f_\theta(y) W_{V_y}^{(i)}], \end{aligned} \tag{6}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{7}$$

where i is the index for U-Net layers, $[;]$ is the concatenation operator, $W_Q^{(i)}, W_{K_x}^{(i)}, W_{V_x}^{(i)} \in \mathbb{R}^{d\times d}$ and $W_{K_y}^{(i)}, W_{V_y}^{(i)} \in \mathbb{R}^{d_y \times d}$ are learnable projection layers, n_x and n_y are the length of encoded image and text, respectively.

During inference, given a text prompt y, the denoising U-Net $\epsilon_\theta(\cdot)$ predicts the image sample x conditioned on the text y with classfier-free guidance [10] and Denoising Diffusion Implicit Models (DDIM) sampling [31].

2.2. Knowledge-Enhanced Diffusion Model

The text-to-image model receives a text prompt that describes the scene of an image, then depicts it with crucial objects and corresponding attribute details. In other words, both text and image are intended to express a visual scene, in which key elements have different expressions, such as keywords in text or salient regions in image. However, naive diffusion model does not distinguish the importance of elements and indiscriminately iterates the denoising process. ERNIE-ViLG 2.0 incorporates extra text and visual knowledge into the learning stage, hoping to enhance the fine-grained semantic perception of diffusion model.

Textual Knowledge. An ideal text-to-image model is expected to focus on all the critical semantics mentioned in the text prompt. To distinguish function words and words describing key semantics, we adopt an off-the-shelf part-of-speech toolkit to extract lexical knowledge of the input prompt, and then improve the learning process by (1) inserting special tokens into the input sequence and (2) increasing the weight of tokens with specific part-of-speech tags in the attention layer. Specifically, we selected 50% of samples and inserted special tokens at the beginning of each word, in which each part-of-speech tag corresponds to a special token. For the selected samples, we also strengthen the attention weight of keywords based on the lexical analysis results. In this way, Equation (7) is modified to,

$$\text{Attention}(Q,K,V)' = \text{softmax}\left(\frac{W_a \cdot (QK^\top)}{\sqrt{d}}\right)V, \quad (8)$$

where $W_a \in \mathbb{R}^{n_x \times (n_x+n_y)}$ is an auxiliary weight matrix that used to scale the vanilla attention, and

$$W_a^{ij} = \begin{cases} 1+w_a & tok_i \in \{x\}, tok_j \in \{x, y_{key}\} \\ 1 & \text{otherwise.} \end{cases} \quad (9)$$

Here w_a^{ij} is the scaling factor of the attention weight between token tok_i and tok_j, w_a is a hyper-parameter, x refers to all the image tokens, and y_{key} denotes the keywords in text[2]. Figure 2 gives an example, where special tokens "`[a]`" and "`[n]`" are inserted for adjectives and nouns, respectively.

Visual Knowledge. Similar to notional words in the text prompt, there are also salient regions in an image, such as people, trees, buildings, and objects mentioned in the input. To extract such visual knowledge, we apply an object detector [1] to 50% of training samples, and then select eye-catching objects from the results with heuristic strategies. Since the loss function of the diffusion model directly acts on the image space, we can assign higher weights to corresponding regions by modifying Equation (3), thus promoting the model to focus on the generation of these objects:

$$\mathcal{L}' = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,I),t}\left[W_l \cdot ||\epsilon - \epsilon_\theta(z_t,t)||_2^2\right], \quad (10)$$

$$W_l^{ij} = \begin{cases} 1+w_l & los_{ij} \in \{x_{key}\} \\ 1 & \text{otherwise.} \end{cases} \quad (11)$$

Here $W_l \in \mathbb{R}^{n_h \times n_w}$ is the weight matrix, n_h and n_w are the height and weight of image space, w_l is a hyper-parameter, los_{ij} is the loss item in i-th row and j-th column of image space, x_{key} is the regions that corresponding to key objects. As Figure 2 illustrates, the regions of "`dog`" and "`cat`" are assigned with larger weights in the calculation of $\mathcal{L}'$.

Now a new problem arises: as a kind of fine-grained knowledge, the selected objects may not appear in the text prompt, thus perplexing the model in learning the alignment between words and objects. An intuitive idea is first to obtain the object and attribute category of each region, then combine corresponding class labels with the original text prompt to achieve fine-grained description, thus ensuring the input contains both coarse and fine granularity information. For instance, as shown in Figure 2, the detected object "`bowl`" is not included in the caption, so we append it to the original description. Besides, we also employ an image captioning model [38] to generate text for images, and randomly replace the original prompt with generated captions, because the generated captions of many images are more concise and reveal more accurate semantics than original prompts.

Most notably, the above strategies are only limited to the training stage. By randomly selecting a part of samples to equip these additional enhancement strategies, the model is supposed to sense the hints of knowledge from various perspectives, and generate higher quality images for the given text in the inference stage, even without special tokens, attention strengthening, or text refinement.

2.3. Mixture-of-Denoising-Experts

Recall that the diffusion process is to iteratively corrupt the image with Gaussian noises by a series of diffusion steps $t = 1, \dots, T$, and DDPM [9] are trained to revert the diffusion process by denoising steps $t = T, \dots, 1$. During the denoising process, all steps aim to denoise a noised input, and they together convert a completely random noise into a meaningful image gradually. Although sharing the same goal, the difficulty of these denoising steps varies according to the noise ratio of input. Figure 2 illustrates such difference by presenting some examples of x_t and the denoising prediction $\hat{x}_{0,t}$ during training. For timesteps t near T, such as $t = T, t_k, t_j$ in the figure, the input of the denoising network x_t is highly noised, and the network of these steps mainly tackles a generation task, i.e., generating an image from a noise. On the contrary, for timesteps t near 1, such as $t = t_i, 1$, the input x_t is close to the original image, and the network of these steps needs to refine the image details.

DDPM makes no specific assumption on the implementation of denoising network, that is, the denoising process does not require the same denoising network for all steps in theory. However, most of the previous text-to-image diffusion ap-

[2] The keywords is defined as notional words in modern Chinese (i.e., nouns, verbs, adjectives, numerals, quantifiers, and pronouns).

proaches [18,22,25,26] follow the vanilla implementation to adopt a denoising network for the whole denoising process. Considering that tasks of different timesteps are different, we conjecture that using the same set of parameters to learn different tasks might lead to suboptimal performance.

In view of this, we further propose Mixture-of-Denoising-Experts (MoDE), where the primary motivation is to employ multiple specialized expert networks to fit different tasks at different timesteps. Since the inputs of adjacent timesteps are similar and so are the denoising tasks, we divide all the timesteps uniformly into n blocks $(S_1, \cdots, S_i, \cdots, S_n)$, in which each block consists of consecutive timesteps and is assigned to one denoising expert. In other words, the timesteps in the same block are denoised by the same group of network parameters. In practice, we share the same text encoder for all denoising experts, and utilize different U-Net experts for different timestep blocks:

$$\epsilon_\theta(x_t, t) = \{\epsilon_{\theta,i}(x_t, t)\}, \quad t \in S_i, \tag{12}$$

where $\epsilon_{\theta,i}(x_t, t)$ is the i-th expert network. Herein, MoDE improves the model performance by adopting expert networks to specially deal with different denoising stages.

Intuitively, when using more experts, each block contains fewer timesteps, so each expert could better focus on learning the characteristics of specific denoising steps assigned to it. Meanwhile, as only one expert network is activated at each step, increasing the number of experts does not affect the computation overhead during inference. Therefore, ERNIE-ViLG 2.0 can flexibly scale up the parameters of diffusion model, allowing the experts to fit the data distribution better without increasing inference time.

3. Experiments

In this section, we first introduce the implementation details of ERNIE-ViLG 2.0. Then we present the comparison of models with automatic metrics and human evaluation. Last, we further analyze the results with quantitative ablation studies and qualitative showcases.

3.1. Implementation Details

To reduce learning complexity, we use diffusion models to generate the representations of images in latent space of an image auto-encoder following Latent Diffusion Models [25]. We first pre-train an image encoder to transform an image $x \in \mathbb{R}^{n_h \times n_w \times 3}$ from pixel space into latent space $\hat{x} \in \mathbb{R}^{n_h^l \times n_w^l \times 4}$ and an image decoder to convert it back. Here n_h/n_h^l and n_w/n_w^l denote the image's original/latent height and width, and we collectively refer to pixel space and hidden space as image space in this paper. Then we fix the auto-encoder and train the diffusion model to generate $\hat{x}$ from text prompt y. During inference, we adopt the pre-trained image decoder to turn $\hat{x}$ into pixel-level image output.

Table 1. Comparison of ERNIE-ViLG 2.0 and representative text-to-image generation models on MS-COCO 256×256 with zero-shot FID-30k. We use classifier-free guidance scale 2.1 for our diffusion model and achieve the best performance.

Model	Zero-Shot FID-30k ↓
DALL-E [23]	27.50
CogView [2]	27.10
LAFITE [46]	26.94
LDM [25]	12.61
ERNIE-ViLG [45]	14.70
GLIDE [18]	12.24
Make-A-Scene [6]	11.84
DALL-E 2 [22]	10.39
CogView2 [3]	24.00
Imagen [26]	7.27
Parti [43]	7.23
ERNIE-ViLG 2.0	**6.75**

ERNIE-ViLG 2.0 contains a transformer-based text encoder with 1.3B parameters and 10 denoising U-Net experts with 2.2B parameters each, which totally add up to about 24B parameters. For hyper-parameters to incorporate knowledge, the attention weight scale w_a is set to 0.01 and the loss weight scale w_l is set to 0.1 (both chosen from [0.01, 0.1, 0.5, 1]). For the MoDE strategy, all timesteps are divided into 10 blocks. The model is optimized by AdamW [15], with a fixed learning rate 0.9×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.01. We train ERNIE-ViLG 2.0 on 320 Tesla A100 GPUs for 18 days.

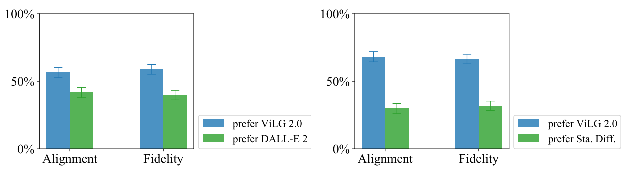
The training data consists of 170M image-text pairs, including publicly available English datasets like LAION [28] and a series of internal Chinese datasets. The image auto-encoder is trained on the same set. For images with English captions, we translate them with Baidu Translate API[3] to get the Chinese version.

3.2. Results

Automatic Evaluation on MS-COCO. Following previous work [22, 25, 26], we evaluate ERNIE-ViLG 2.0 on MS-COCO 256×256 with zero-shot FID-30k, where 30,000 images from the validation set are randomly selected and the English captions are automatically translated to Chinese.

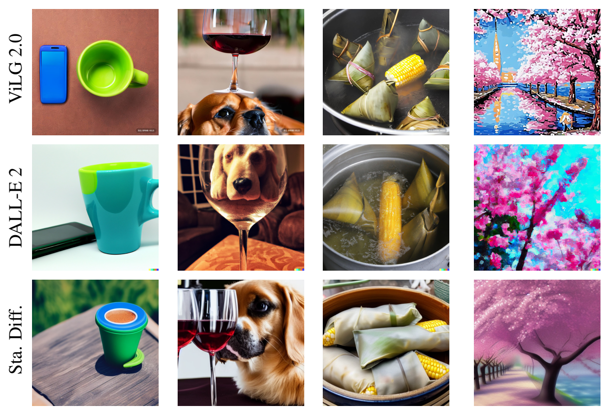
Table 1 shows that ERNIE-ViLG 2.0 achieves new state-of-the-art performance of text-to-image generation, with 6.75 zero-shot FID-30k on MS-COCO. Inspired by DALL-E [23] and Parti [43], we rerank the batch-sampled images (with only 4 images per text prompt, comparing with 512 images used in DALL-E and 16 images used in Parti) based on the image-text alignment score, calculated by a pre-trained CLIP model [21], in which the text is the initial English

[3] https://fanyi.baidu.com



(a) v.s. DALL-E 2. (b) v.s. Stable Diffusion.

Figure 3. Comparison of ERNIE-ViLG 2.0 and DALL-E 2/Stable Diffusion on ViLG-300 with human evaluation. We report the user preference rates with 95% confidence intervals.



(1) A green cup and a blue cell phone (2) A wine glass on top of a dog
(3) Zongzi and corn boiled in the pot (4) Cherry blossom, digital oil painting

Figure 4. Qualitative Comparison of ERNIE-ViLG 2.0 and DALL-E 2/Stable Diffusion on ViLG-300.

caption in MS-COCO and the image is generated from the auto-translated Chinese caption. Besides, even without the reranking strategy, we find that ERNIE-ViLG 2.0 can also beat the latest diffusion-based models like DALL-E 2 [22] and Imagen [26], with the zero-shot FID-30k of 7.23.

Human Evaluation on ViLG-300. ERNIE-ViLG 2.0 takes Chinese prompts as input and generates high-resolution images, unlike recent English-oriented text-to-image models. Herein, we introduce ViLG-300[4], a bilingual prompt set that supports the systematic evaluation and comparison of Chinese and English text-to-image models. ViLG-300 contains 300 prompts from 16 categories, composed of DrawBench [26] (in English) and the prompt set used in ERNIE-ViLG [45] (in Chinese).

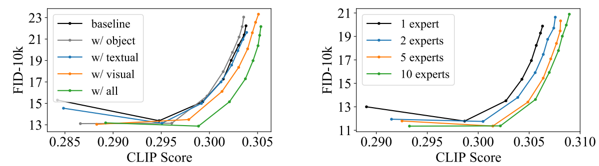
With ViLG-300, we can make convincing comparisons between ERNIE-ViLG 2.0 and DALL-E 2[5], Stable Diffusion[6,7]. For evaluation, five raters are presented with two sets

[4]https://github.com/PaddlePaddle/ERNIE/tree/ernie-kit-open-v1.0/Research/ERNIE-ViLG2/data/ViLG-300.csv

[5]https://openai.com/dall-e-2/

[6]https://beta.dreamstudio.ai/dream

[7]We use DALL-E 2 and Stable Diffusion interfaces to generate images on October 25, 2022, before the CVPR 2023 submission deadline.



(a) Knowledge enhancement. (b) Mixture-of-denoising-experts.

Figure 5. Performance with various strategies in ERNIE-ViLG 2.0. Here we draw pareto curves with guidance scale [2,3,4,5,6,7,8,9].

(1) A white car and a red sheep (1) A single clock is sitting on a table
(2) A panda making latte art (2) An umbrella on top of a spoon
(3) A small red ball in a big green block (3) Wolf in a suit
(4) A burning fish (4) A rabbit with a fox's head

Figure 6. Samples from ViLG-300 with different knowledge enhancement strategies (left) and different number of experts (right).

of images generated by ERNIE-ViLG 2.0 and the compared model. Next, they are asked to compare these images from two dimensions of image-text alignment and image fidelity, and then select the model they prefer, or respond that there is no measurable difference between two models. Throughout the process, raters are unaware of which model the image is generated from, and we do not apply any filtering strategy to the rating results. Figure 3 shows that human raters prefer ERNIE-ViLG 2.0 over all other models in both image-text alignment (56.5%±3.8% and 68.2%±3.8% when compared to DALL-E 2 and Stable Diffusion, respectively) and image fidelity (58.8%±3.6% to DALL-E 2, 66.5%±3.5% to Stable Diffusion, respectively), which again proves that ERNIE-ViLG 2.0 can generate high-quality images that conform to the text, with the help of knowledge enhancement and mixture-of-denoising-experts strategies. Beyond text relevancy and image fidelity, we also observe that ERNIE-ViLG 2.0 can generate images with better sharpness and textures than baseline models.

3.3. Analysis

To examine the effectiveness of our design philosophy, we conduct two groups of ablation studies. Similar to the main experiment, we also provide both automatic metrics and intuitive showcases to demonstrate the advantages of each strategy in ERNIE-ViLG 2.0 here.

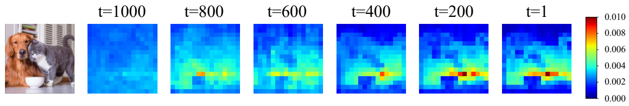


Figure 7. The visualization of cross-attention maps in different denoising timesteps, where each value in the image space is the average of attention from this image token to all text tokens.

Knowledge Enhancement Strategies. In this part, we focus on the impact of various knowledge enhancement strategies by training a series of lightweight models, with 500M text encoders, 870M U-Nets, and 500M training samples. The pareto curves in Figure 5a demonstrate that incorporating knowledge in the learning process brings significant performance gains in image fidelity and image-text alignment. Specifically, (1) the benefits of textual knowledge are mainly reflected in precise fine-grained semantic control (`w/ textual`), (2) only utilizing object knowledge may not be able to steadily promote the performance (`w/ object`), while taking synthetic descriptions into consideration is an effective solution to make full use of visual knowledge (`w/ visual`). Figure 6 provides more visual comparisons to intuitively demonstrate the changes brought by each strategy. When handling complex prompts, `baseline` model faces problems such as the absence of key objects or incorrect assignment of attributes. At this point, textual knowledge helps the model accurately understand the attributes of each object, but the generated images sometimes fall into a new problem of distortion. Complementarily, visual knowledge promotes the generation of high-fidelity images, but it cannot well understand specific entities in text. Eventually, the combination of two kinds of knowledge harmoniously promotes the model from single- and multi-modal views, which ensures high fidelity and boost the image-text alignment in fine-grained visual scene.

Mixture-of-Denoising-Experts Strategies. Based on the above lightweight settings, we further train the `baseline` model with 500M samples, and then train 200M samples for each denoising expert. Figure 5b shows that with the increasing number of denoising experts, the overall performance is gradually improved, proving that scaling the size of U-Net is also an effective solution to achieve better image quality. More showcases are provided in Figure 6. When the number of experts increases from 1 to 10, the model can not only better handle the coupling between different elements but also generate images with more natural textures. For instance, the numbers on clocks become clearer, the proportion of wolf and suit becomes more harmonious, and the model can generate more photorealistic pictures instead of cartoon drawings. We also tried to analyze the impact of the amount of denoising experts and training samples, and found that using more expert networks has better performance than using a network to train more samples.

Figure 7 further visualizes the cross-attention maps from image features to text representations in denoising experts during 1,000-step denoising process, where these steps shown are denoised by different experts. As shown in the illustration, attentions of different denoising timesteps vary. Specifically, the attention maps of timesteps t near 1,000 are almost evenly distributed over the whole image, which is because the input of these steps is close to Gaussian noise and the image layout is unclear, so all the image tokens have to attend to the text prompt to generate image skeleton. When the timesteps are close to 1, attention maps concentrate more on foreground objects. For these timesteps, the input to denoising network is close to the final image and the layout is clear, and only a few parts of the image need to focus on the text to fill in the details of object. These observations again illustrate the difference among denoising timesteps and demonstrate the need to disentangle different timesteps with multiple experts.

4. Related Work

Text-to-Image Generation. Text-to-image generation is the task of synthesizing images according to natural language descriptions. Early works adopted generative adversarial networks [7] to produce images based on text [36, 41, 44, 47]. Inspired by the success of transformers in various generation tasks [37], models such as ERNIE-ViLG [45], DALL-E [23], Cogview [2], Make-A-Scene [6], and Parti [43] have also explored text-to-image generation as a sequence-to-sequence problem, with auto-regressive transformers as generators and text/image tokens as input/output sequences. Recently, another line of works have applied diffusion models [30], shaping it as an iterative denoising task [9, 27, 31]. By adding text condition in the denoising steps, practices such as LDM [25], DALL-E 2 [22], and Imagen [26] constantly set new records in text-to-image generation. Based on diffusion models as the backbone, ERNIE-ViLG 2.0 proposes incorporating knowledge of scene and mixture-of-denoising-experts mechanism into the denoising process.

Knowledge-Enhanced Pre-trained Models. While transformers benefit from pre-training on large-scale data, many attempts have been adding knowledge to guide them to focus on key elements during learning. For language-based tasks, knowledge-enhanced models used knowledge masking strategy [11, 34] or knowledge-aware pre-training tasks [33, 35] to understand the language data distribution. As for vision-language multi-modal discrimination models, OSCAR [14], ERNIE-ViL [42] and ERNIE-Layout [19] leveraged object tags, scene graphs, and document layouts as extra knowledge to help the models better align language and vision modalities. Among multi-modal generation models, Make-A-Scene [6] emphasized the importance of object and face regions by integrating domain-specific perceptual knowledge.

While current text-to-image diffusion models suffer from attribute misalignment problems [22], they have not employed any specific knowledge of objects. Herein, ERNIE-ViLG 2.0 utilizes the knowledge of key elements in images and text to enhance diffusion models, leading to better fine-grained image-text alignment in generated pictures.

Mixture-of-Expert. Mixture-of-Experts (MoE) in neural networks means dividing specific parts of the parameters into subsets, each of which is called an expert [4, 29]. During the forward pass, a router assigns experts to different input, and each input only interacts with the experts assigned to. The router is a critical part of MoE. In language tasks, the most common strategy is a matching algorithm that assigns each text token to several experts in the linear feed-forward layer [5, 8, 12, 24]. While most practices formulate multiple experts in only the linear layers, some works also use an entire language model as an expert [13]. Beyond the natural language processing tasks, the idea of MoE have also been applied to vision models [20] and Mixture-of-Modality-Expert in multi-modal transformers [17, 39, 40]. In ERNIE-ViLG 2.0, the MoDE mechanism takes multiple denoising U-Nets as experts. It uses the denoising step index as the fixed router to determine which expert to use.

5. Risks, Limitations, and Future Work

Model Usage and Data Bias. Text-to-image generation models trained by large-scale image-text data have all faced similar risks regarding to inappropriate usage of generation and data bias [22, 25, 26]. Considering that text-to-image models help people realize their imagination with less effort, the malicious use of models may result in unexpected deceptive or harmful outcomes. Moreover, since the models are trained on datasets consisting of images and their alt-text crawled from websites, the generated images may exhibit social and cultural bias in the datasets and websites.

Character Rendering. Figure 8 shows two successful character rendering cases (a, b) and one failure case (c). Character rendering is a challenging task for ERNIE-ViLG 2.0 for two reasons. First, the training data contains both Chinese text-image pairs and English text-image pairs translated into Chinese. When a text prompt mentions characters, the characters in the image could be in Chinese or English, and it is hard for the model to learn corresponding characters in both languages simultaneously. In the cases of successful character rendering that we observed, the characters could be words that are common in Chinese and do not have an exact match in English, such as "福" ("blessing, happiness, good luck" in English) in Figure 8a, or numbers which are the same in English and Chinese images, such as "20" in Figure 8b. The second reason that makes character rendering difficult is probably that Chinese characters are complex combinations of strokes without basic components like English letters. In Figure 8c, the model does learn that it should write some

(a) 国风福字挂饰
A hanging ornament with "福" in Chinese fashion

(b) 生日蛋糕上有蜡烛"20"
A birthday cake with candles of "20" on it

(c) 国画葡萄
Chinese painting of grapes

Figure 8. Examples of character rendering. The model successfully renders simple characters specified in the prompt, while for more difficult cases, the model only learns the position for now.

Chinese characters in the top right corner, but it only paints meaningless strokes there.

Variation of Mixture-of-Denoising-Experts. Section 3.3 shows that using more denoising experts leads to better model performance. It indicates that using parallel U-Net experts is an effective way to augment the denoising network. Due to the computation limitation, we only try using up to 10 experts in this work, while we believe that exploring more denoising experts and multiple text encoders as experts is a meaningful future direction. Herein, we can further scale up the model and allow it to learn data distribution better with similar inference time.

6. Conclusions

We present ERNIE-ViLG 2.0, the first Chinese large-scale text-to-image generation model based on diffusion models. To improve the fine-grained control of scene semantics, we incorporate visual and textual knowledge of the scene into diffusion models. To disentangle the model parameters for different denoising timesteps, we introduce MoDE and scale up the model parameters to 24B with a relatively short inference time. Experiments show that ERNIE-ViLG 2.0 achieves state-of-the-art on MS-COCO and each proposed mechanism contributes to the final results. To allow fair comparisons between Chinese and English text-to-image models, we collect a bilingual prompt set ViLG-300, and human evaluation indicates that ERNIE-ViLG 2.0 is preferred over strong baselines in both text relevancy and image fidelity. Further analysis suggests that different knowledge sources improve the generation in different aspects, and using more experts results in better image quality. In the future, we intend to enrich external image-text alignment knowledge and expand the usage of multiple experts to advance the generation. See also Appendix for more details on training and evaluation.

Acknowledgement This work was partly supported by National Natural Science Foundation of China (62271359). We thank Yehan Zheng for help on dataset collection; Weixin Liu and Bin Shan for discussions on paper.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society, 2018. 4, 12

[2] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 19822–19835, 2021. 5, 7

[3] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *CoRR*, abs/2204.14217, 2022. 5

[4] William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *CoRR*, abs/2209.01667, 2022. 8

[5] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961, 2021. 8

[6] 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. *CoRR*, abs/2203.13131, 2022. 5, 7, 13

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. 7

[8] Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah Smith, and Luke Zettlemoyer. Demix layers: Disentangling domains for modular language modeling. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5557–5576. Association for Computational Linguistics, 2022. 8

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3, 4, 7

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[11] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77, 2020. 7

[12] 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. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 8

[13] 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. *CoRR*, abs/2208.03306, 2022. 8

[14] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. 7

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5

[16] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of DALL-E 2. *CoRR*, abs/2204.13807, 2022. 13

[17] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *CoRR*, abs/2206.02770, 2022. 8

[18] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 2022. 2, 3, 5, 13

[19] Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *CoRR*, abs/2210.06155, 2022. 7

[20] Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, Cédric Renggli, André Susano Pinto, Sylvain Gelly, Daniel Keysers, and Neil Houlsby. Scalable transfer learning with

expert models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 8

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 5

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 2, 5, 6, 7, 8, 12, 13

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021. 5, 7, 13

[24] Stephen Roller, Sainbayar Sukhbaatar, Arthur Szlam, and Jason Weston. Hash layers for large sparse models. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17555–17566, 2021. 8

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. 2, 5, 7, 8, 13

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *CoRR*, abs/2205.11487, 2022. 2, 5, 6, 7, 8, 12, 13

[27] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 7

[28] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 5

[29] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 8

[30] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015. 3, 7

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3, 7

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[33] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137, 2021. 7

[34] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223, 2019. 7

[35] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8968–8975. AAAI Press, 2020. 7

[36] Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16494–16504. IEEE, 2022. 7

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 7

[38] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and

Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 2022. 4

[39] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *CoRR*, abs/2208.10442, 2022. 8

[40] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *CoRR*, abs/2111.02358, 2021. 8

[41] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1316–1324. Computer Vision Foundation / IEEE Computer Society, 2018. 7

[42] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3208–3216. AAAI Press, 2021. 2, 7

[43] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *CoRR*, abs/2206.10789, 2022. 5, 7, 12, 13

[44] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 833–842. Computer Vision Foundation / IEEE, 2021. 7

[45] Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *CoRR*, abs/2112.15283, 2021. 5, 6, 7, 12, 13

[46] 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. *CoRR*, abs/2111.13792, 2021. 5

[47] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5802–5810. Computer Vision Foundation / IEEE, 2019. 7