Algorithm 1 Training Process

Require: Paired (image, text) inputs, POS toolkit, object detector.
1: Find keywords in text and salient regions in image. ▷ Pre-processing
2: For training step s in $(0, 350000]$: ▷ First stage
3: Sample a denoising timestep t from $[0, 1000)$,
4: Optimize both the text encoder and the denoising network.
5: Init 10 denoising networks with current parameters.
6: For training step s in $(350000, 440000]$: ▷ Second stage
7: Sample a denoising timestep t from $[0, 1000)$,
8: Optimize the $\lfloor \frac{t}{100} \rfloor$-th denoising network.

Table 2. Hyperparameters and Configuration of ERNIE-ViLG 2.0.

Text Encoder (Transformer)	
Vocab size	21128
Text encoder context	77
Text encoder width	2,048
Text encoder depth	24
Text encoder heads	32
Denoising Network (U-Net)	
Noise schedule	linear
Diffusion steps	1,000
Sampling steps	50
Model channels	512
Head channels	64
Channel multiplier	[1,2,3,4]
Attention resolutions	[2,4,8]
ResNet number	3
Dropout	0

A. Detailed Training Process

Algorithm 1 shows the pseudo code for training ERNIE-ViLG 2.0. With (image, text) pairs as input, we first find the keywords in texts with an open-source POS toolkit "jieba" and salient regions in images with an object detector [1]. These additional information are then used in the knowledge-enhanced training. The training process consists of two stages. In the first stage, we train a U-Net with 2.2B parameters and a text encoder with 1.3B parameters for $350,000$ steps. In the second stage, the text encoder is shared, and we train 10 denoising experts for $90,000$ steps that inherit U-Net parameters from the first stage. The unaccomplished hyper-parameters we use for ERNIE-ViLG 2.0 is provided in Table 2.

B. Detailed Automatic Evaluation

Table 3 presents a detailed comparison on the automatic evaluation scores, including model sizes and reranking strategies of models. At the end of the first training stage, the FID-30K metric of our 3.5B model with only one denoising expert is 8.07 (w/o reranking), which is better than DALL-E 2 [22] (10.39) with a similar model size, and worse than our final 24B model with 10 experts (7.23 w/o reranking). After the complete training process, our 24B model (6.75 w/ 4 reranking images) outperforms Parti [43] (7.23 w/ 16 reranking images) with a similar number of parameters and a smaller number of reranking images. These comparisons indicate that both extra knowledge and model scaling contribute to the final performance of our model.

C. Detailed Human Evaluation

In this section, we supplement the part about human evaluation omitted in main content, including the construction process of ViLG-300, the performance comparison on various categories, and the example qualitative comparison of different models.

C.1. The Construction of ViLG-300

To construct ViLG-300, we first remove the language-related prompts in DrawBench [26] (`text rendering`, `rare words`, `misspelled` prompts) and `MS-COCO` prompts in ERNIE-ViLG [45], leaving 162 and 398 prompts, respectively, then randomly sampled 150 prompts from these two parts, manually translated and proofread these prompts to achieve the final parallel Chinese and English set. Specifically, we remove language-related text prompts in DrawBench since these are not comparable inputs for models in different languages. For text rendering in Chinese, we also have discussed in detail in Section 5. We also remove the MS-COCO category in ERNIE-ViLG, because MS-COCO has been used in the automatic evaluation, and the prompts are relatively simple for current text-to-image models, especially when evaluating the models' ability to understand complex scene. Note that there are two similar categories (i.e., `Conflicting` and `Counterfactual`) in DrawBench and ERNIE-ViLG that we do not align and merge. The reason is that the `Conflicting` category focuses on the impossible combination of things, while `Counterfactual` contains many prompts with negative descriptions, both of which are now difficult problems.

C.2. Detailed Results on ViLG-300

Figure 9 shows the detailed performance comparison between ERNIE-ViLG 2.0 and DALL-E 2/Stable Diffusion on ViLG-300, and example qualitative comparisons are shown in Figure 13 and 14. The most important conclusion is that ERNIE-ViLG 2.0 is quite skilled in dealing with text prompts with colors and complex scenes, and also has impressive performance in many categories, such as `Geography`, `Scene`, and `Cartoon`. Intuitively, we attribute the excellent performance to the knowledge injection that endows the model with the ability to perceive and understand various named entities and detailed descriptions, as well as the increase in the number of parameters brought by the mixture-of-denoising-experts strategies also makes the model even more powerful.

Table 3. Detailed comparison of ERNIE-ViLG 2.0 and representative text-to-image generation models on MS-COCO 256×256 with zero-shot FID-30k.

Model	#params	FID w/o reranking	FID/#reranking images
DALL-E [23]	12B	34.6	27.5/512
LDM [25]	1.45B	12.61	-
GLIDE [18]	6B	12.24	-
Make-A-Scene [6]	4B	11.84	-
DALL-E 2 [22]	4.5B	10.39	-
Imagen [26]	6.6B	7.27	-
Parti [43]	20B	-	7.23/16
ERNIE-ViLG 2.0 w/ 1 denoising expert	3.5B	8.07	7.62/4
ERNIE-ViLG 2.0 w/ 10 denoising experts	24B	**7.23**	**6.75**/4

Table 4. Detailed categories and statistics of ViLG-300.

Source	Category	Description	Number
DrawBench [26]	Color	objects with specified colors	22
	Counting	objects with specified numbers	18
	Positional	objects with specified spatial positioning	16
	Conflicting	objects with conflicting interactions	10
	Description	complex and long prompts describing an objects	20
	DALL-E case	prompts from DALL-E [23]	19
	Marcus	prompts from Marcus et al. [16]	9
	Reddit	prompts from DALL-E 2 Reddit	36
ERNIE-ViLG [45]	Simple	single-object with specified attributes	18
	Complex	multi-objects with specified attributes and relationships	23
	Counterfactual	objects with impossible interactions or negative words	23
	Geography	specific geographic entities	24
	View	objects with specified view angles	16
	Scene	objects with specified time and scenes	14
	Style	objects with specified styles	16
	Cartoon	anthropomorphic animals or cartoon characters	16

At the same time, we also propose that further understanding of the number of objects and the relationship between them can be the focus of future text-to-image models.

D. Detailed Ablation Study

Here we attach more analysis to the ablation study and more showcases in Figure 15,16.

D.1. Knowledge Enhancement Ablation

Figure 10 provides the convergence curves of various models, it is obvious that the knowledge enhancement strategies significantly accelerate the convergence process of diffusion models. Notably, at the very beginning of training, the knowledge-enhanced model reaches or even exceeds the performance that the baseline model with two times of training samples (i.e., 100M v.s. 200M, 200M v.s. 400M).

To quantitatively measure the improvement brought by each knowledge source, we calculate the CLIP score between ViLG-300 prompts and generated images[8]. Table 5 presents the top five categories with maximum performance gain of each strategy against baseline. It can be found that different strategies result in improvements in different categories, indicating that they help model absorb knowledge and improve text-image alignment in complementary aspects. In addition, we also notice that CLIP could not well capture the relationships between multiple objects (e.g., counterfactual), so we leave the accurate automatic evaluation method for fine-grained semantic control as a valuable future work.

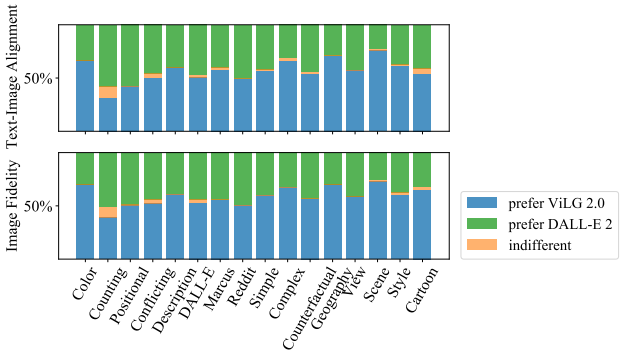
D.2. Mixture-of-Denoising-Experts Ablation

To explore the impact of training samples or model size (i.e., the number of denoising experts) on performance, we train the setting of `1 expert` with 400M/1B/2B samples following Section 3.3, which aligns with the number of sam-

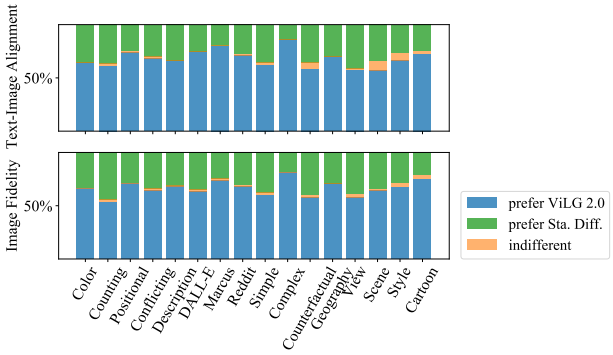
[8]We feed ERNIE-ViLG 2.0 with Chinese prompts and calculate the CLIP score between generated images and corresponding English prompts.

Table 5. Top five ViLG-300 categories with the maximum CLIP Score improvement for each knowledge enhancement strategy.

No.	w/ textual		w/ visual		w/ all	
	Prompt category	ΔCLIP Score	Prompt category	ΔCLIP Score	Prompt category	ΔCLIP Score
1	Counterfactual	0.0051	Counterfactual	0.0080	Complex	0.0074
2	Color	0.0041	Counting	0.0047	Counterfactual	0.0073
3	Marcus	0.0038	Color	0.0035	Cartoon	0.0069
4	Style	0.0022	Cartoon	0.0032	Color	0.0066
5	Positional	0.0018	Complex	0.0016	Style	0.0061



(a) ERNIE-ViLG 2.0 v.s. DALL-E 2



(b) ERNIE-ViLG 2.0 v.s. Stable Diffusion

Figure 9. Detailed comparison of ERNIE-ViLG 2.0 and DALL-E 2/Stable Diffusion on ViLG-300 with human evaluation. We do not apply any filtering strategy and report the initial results here.

ples trained by `2/5/10 experts`. Figure 11 shows that the performance of `1 expert` with 400M and `2 experts` with 200M each is basically equal, while the performance of `1 expert` lags behind that of `2 experts` as training goes on. This shows that decoupling the denoising capability of different stages is an effective strategy, and reasonably scaling the size of U-Net is able to further boost the performance of text-to-image model.

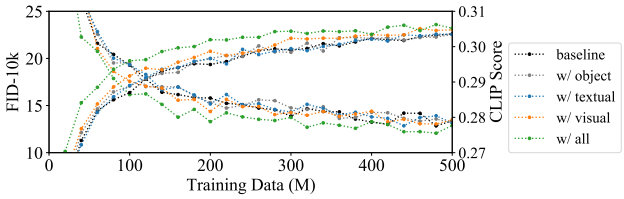


Figure 10. The convergence curves of various models with knowledge enhancement strategies. We choose guidance scale 3 and 8 to draw the curves of FID-10k and CLIP Score, respectively.

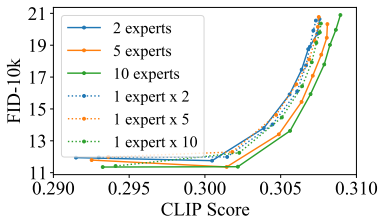


Figure 11. The performance comparison of different amount of denoising experts and training samples, and models see the same number of training samples at comparing points.

D.3. Comparison of Image Quality

Figure 12 compares the image details of ERNIE-ViLG 2.0 and baseline models by zooming in small regions of generated images. Technically, both ERNIE-ViLG 2.0 and Stable Diffusion generate image latent representation with diffusion models conditioned on text. While Stable Diffusion only produces 512×512 sized images, ERNIE-ViLG 2.0 could directly output images with 1024×1024 resolution. Therefore, the magnified parts of ERNIE-ViLG 2.0 are clearer than those of Stable Diffusion. As for DALL-E 2, it employs cascaded generation by first producing 64×64 images with text and then scaling it up to 1024×1024 resolution with two super-resolution models. Although it generates images of the same resolution as ERNIE-ViLG 2.0, the output of DALL-E 2 sometimes contains unnatural textures, such as fluffy trees and rain drops in the magnified regions. Contrary to DALL-E 2, the textures of our model's outcome are more natural and photorealistic.

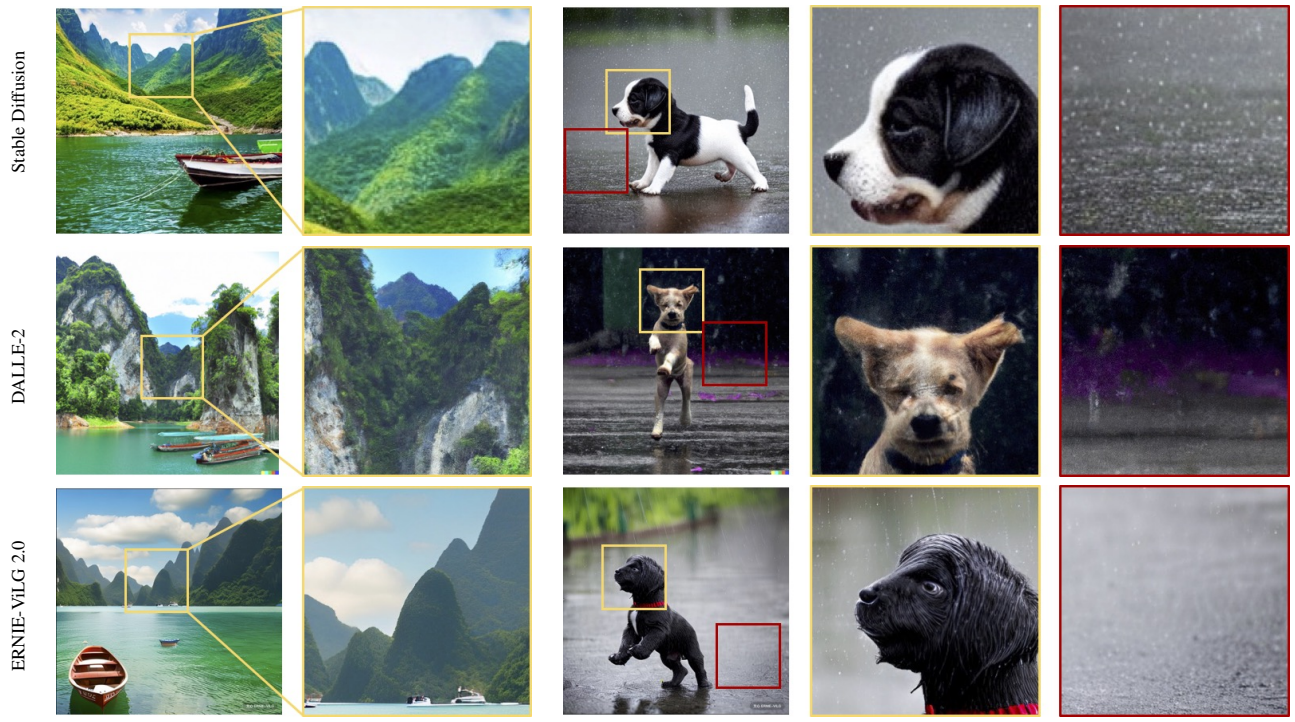


Figure 12. Comparison of image quality by magnifying parts of generated images. ERNIE-ViLG 2.0 enables the generation of sharper 1024×1024 images with more natural details.

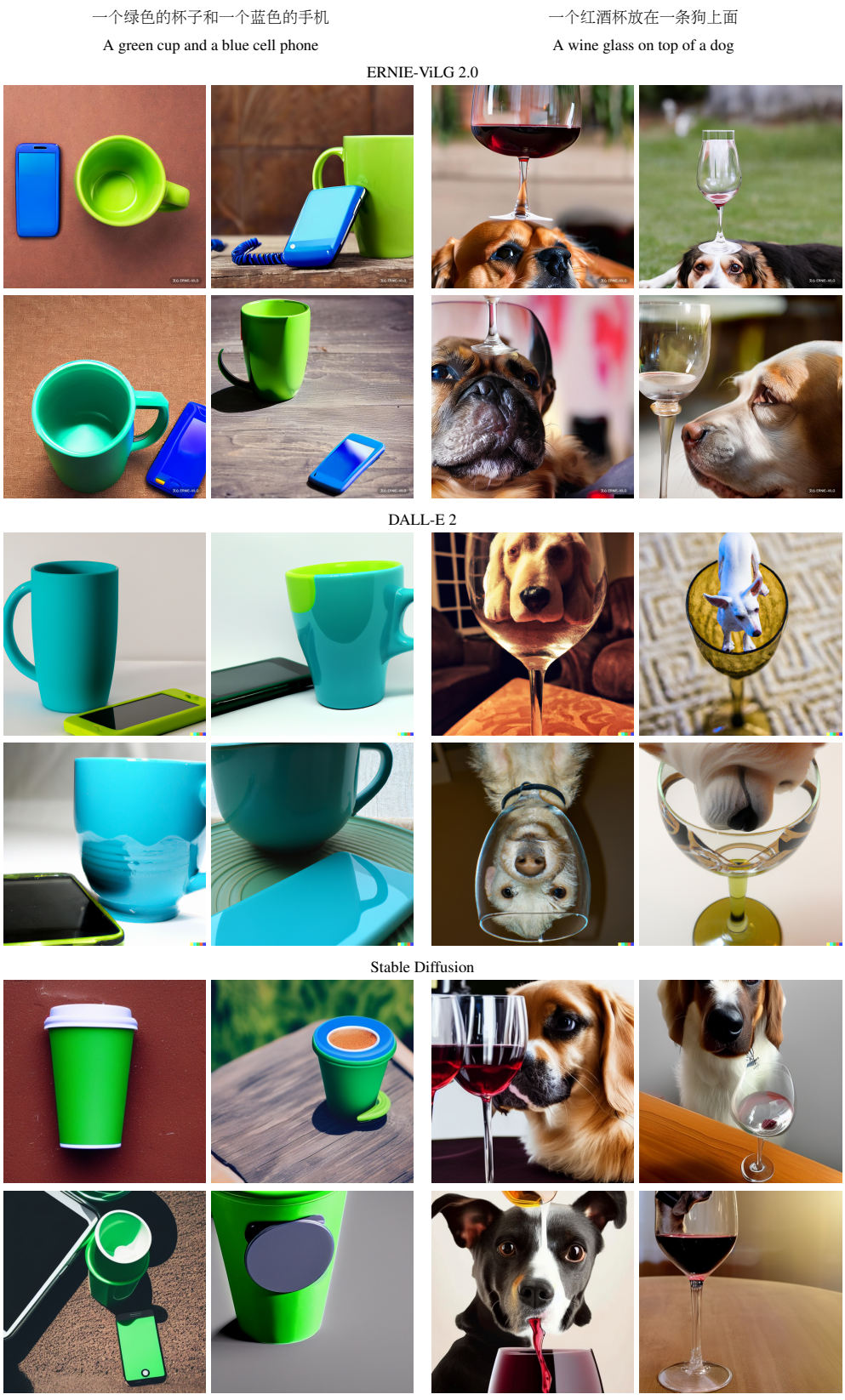


Figure 13. Example qualitative comparisons between ERNIE-ViLG 2.0 and DALL-E 2/Stable Diffusion on DrawBench prompts from ViLG-300.

Figure 14. Example qualitative comparisons between ERNIE-ViLG 2.0 and DALL-E 2/Stable Diffusion on ERNIE-ViLG prompts from ViLG-300.

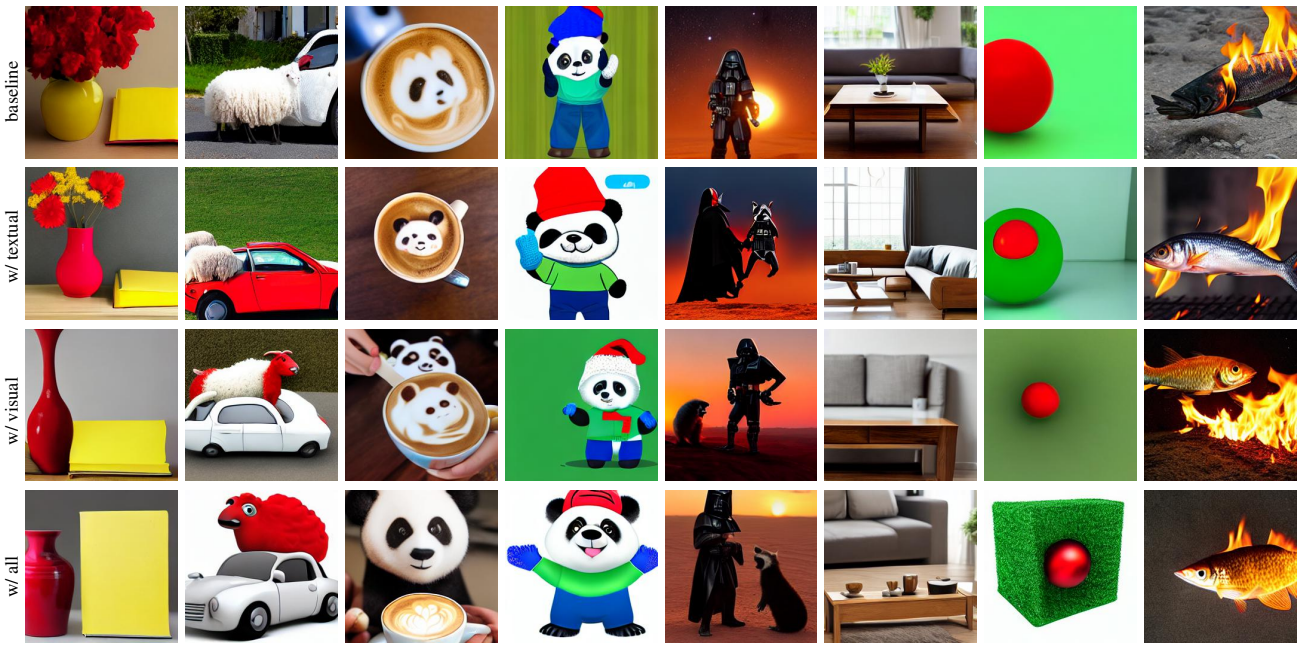


Figure 15. Samples from ViLG-300 with different knowledge enhancement strategies. It can be found that the impacts of textual and visual knowledge do not seem to overlap, and the combination of them is an effective solution to facilitate accurate semantic control and high image fidelity.

Figure 16. Samples from ViLG-300 with different number of denoising experts. When increasing the experts, the most noticeable evolution is that the texture of generated image becomes more natural and photorealistic. Limited by the layout, we simplify the prompt in the second column, and the input received by model actually is "三个立方体堆叠。一个红色立方体在顶部，放在一个红色立方体上。这个红色立方体在中间，放在一个绿色立方体上。这个绿色立方体在底部。(A stack of 3 cubes. A red cube is on the top, sitting on a red cube. The red cube is in the middle, sitting on a green cube. The green cube is on the bottom.)".