Bridging Different Language Models and Generative Vision Models for Text-to-Image Generation

Shihao Zhao[1], Shaozhe Hao[1], Bojia Zi[2], Huaizhe Xu[3], and Kwan-Yee K. Wong[1⋆]

[1] The University of Hong Kong
{shzhao,szhao,kykwong}@cs.hku.hk
[2] The Chinese University of Hong Kong
bjzi@se.cuhk.edu.hk
[3] The Hong Kong University of Science and Technology
hxubr@connect.ust.hk

Abstract. Text-to-image generation has made significant advancements with the introduction of text-to-image diffusion models. These models typically consist of a language model that interprets user prompts and a vision model that generates corresponding images. As language and vision models continue to progress in their respective domains, there is a great potential in exploring the replacement of components in text-to-image diffusion models with more advanced counterparts. A broader research objective would therefore be to investigate the integration of any two unrelated language and generative vision models for text-to-image generation. In this paper, we explore this objective and propose LaVi-Bridge, a pipeline that enables the integration of diverse pre-trained language models and generative vision models for text-to-image generation. By leveraging LoRA and adapters, LaVi-Bridge offers a flexible and plug-and-play approach without requiring modifications to the original weights of the language and vision models. Our pipeline is compatible with various language models and generative vision models, accommodating different structures. Within this framework, we demonstrate that incorporating superior modules, such as more advanced language models or generative vision models, results in notable improvements in capabilities like text alignment or image quality. Extensive evaluations have been conducted to verify the effectiveness of LaVi-Bridge. Code is available at https://github.com/ShihaoZhaoZSH/LaVi-Bridge.

Keywords: Diffusion model · Text-to-image generation

1 Introduction

In recent years, there have been remarkable advancements in the field of text-to-image generation, specifically through the use of diffusion models [9, 18, 43, 45].

⋆ Corresponding authors

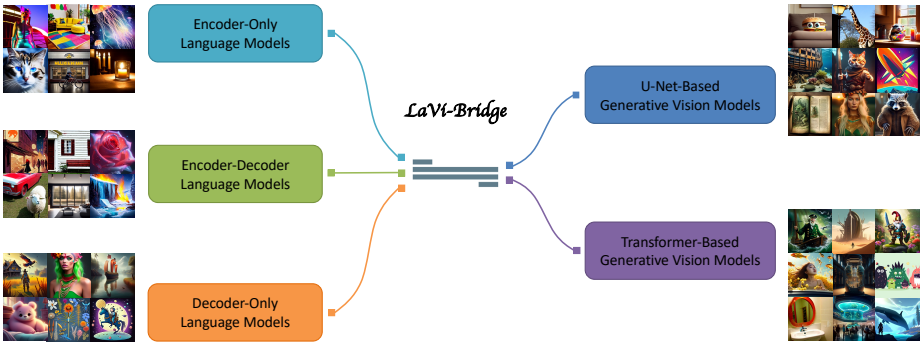

Fig. 1: Overview of LaVi-Bridge. LaVi-Bridge is capable of integrating various language models and generative vision models. On the left side, we keep the vision model fixed and experiment with different language models in our pipeline. On the right side, we keep the language model fixed and try out different vision models. We display the visualization results alongside each combination.

These models have made significant contributions and have gained considerable attention for their exceptional performance. By leveraging large-scale training datasets alongside large deep models, text-to-image diffusion models are capable of producing high-quality images that faithfully align with the textual descriptions provided by users. This has rendered them highly applicable in real-world scenarios such as content creation and architectural design.

Text-to-image diffusion models [2, 6, 29, 35, 38, 41, 50] typically consist of two key components, namely a language model and a generative vision model. The language model is responsible for comprehending the input prompts, whereas the vision model is tasked with generating images that align with the extracted context. Existing text-to-image diffusion models employ various language models and generative vision models and have gained widespread usage. For instance, Stable Diffusion (SD) [38] is a highly popular text-to-image diffusion model that employs the CLIP text encoder [31] as its language model and a U-Net [39] as its generative vision model. Another example is PixArt [6], a recently proposed text-to-image diffusion model that adopts the T5 [34] as its language model and a Vision Transformer (ViT) [11] as its generative vision model. These models are trained on a vast amount of text-image pairs, enabling seamless collaboration between their language modules and vision modules.

The advancements in deep language models and deep vision models have witnessed rapid progress in recent years, with both fields experiencing continuous developments and the introduction of more powerful models. However, this rapid development poses a challenge for the research in text-to-image generation when it comes to integrating more advanced language or vision models into existing text-to-image diffusion models. The problem of how to integrate any two unrelated language and vision models is unexplored, and the impact of newly developed models on text-to-image generation capabilities also remains

uncertain. The current situation highlights the presence of a gap between the language or vision modules within text-to-image diffusion models and the state-of-the-art models in their respective domains. Therefore, it has become crucial to address this gap and explore ways to incorporate more advanced language or vision models into existing text-to-image diffusion models. Furthermore, the broader challenge of integrating any pre-trained language model with any generative vision model deserves a thorough investigation.

In this paper, our objective is to delve into the aforementioned problem. We propose LaVi-Bridge, a flexible framework that facilitates the integration of diverse well-trained language models and generative vision models to achieve text-to-image generation. Our framework enables the integration of two unrelated language and vision models that have not been previously trained together, as shown in Fig. 1. Importantly, LaVi-Bridge does not require modifying the original weights of the language and vision models. Instead, it injects LoRA [20] into the language and vision models separately and utilizes an adapter to bridge these two modules. Moreover, LaVi-Bridge only necessitates a relatively small dataset to integrate different language models and generative vision models for text-to-image generation.

We summarize the advantages and features of LaVi-Bridge as follows:

1. LaVi-Bridge is designed for text-to-image diffusion models and serves as a bridge, capable of connecting various pre-trained language models and generative vision models. Our framework can accommodate different model structures, including encoder-only, encoder-decoder, and decoder-only language models, as well as U-Net-based and Transformer-based generative vision models.
2. LaVi-Bridge utilizes LoRA and adapters, eliminating the need to modify the original weights of the models. It is more flexible and requires relatively small computing resources compared to training the entire diffusion model.
3. We evaluated various text-image alignment and image quality metrics on short prompts, long prompts, and compositional prompts. We also conducted extensive visualization. We then drew several conclusions. For instance, integrating superior models leads to improved performance in the corresponding modality, such as enhanced semantic understanding with advanced language models or improved image quality with more powerful generative vision models. Additionally, the diffusion model utilizing Llama-2 demonstrates exceptional semantic understanding, while the diffusion model utilizing the transformer in PixArt yields images with enhanced aesthetics.

2 Related Work

2.1 Language Models and Generative Vision Models

The mainstream Large Language Models (LLMs) [8, 32–34] are built based on the transformer structure [48], with three main types of architectures, namely

encoder-only, encoder-decoder, and decoder-only. All these three belong to Sequence to Sequence (Seq2Seq) [46]. The encoder-only architecture is exemplified by BERT [8]. CLIP text encoder [31] is based on BERT and further trained to align with the image domain. Models of this type excel at understanding the content of the input and generating outputs tailored to specific tasks. On the other hand, the encoder-decoder framework is adept at handling tasks that involve complex mappings between input and output sequences. Examples include T5 [34] and BART [24]. Recently, due to the tremendous success of ChatGPT, attention has been drawn to models that consist solely of a decoder, like GPT-3 [4] and Llama-2 [47]. The decoder-only architecture demonstrates exceptional performance in semantic understanding. For text-to-image generation, all three types of LLMs can provide effective semantic information to serve as conditions for image generation in diffusion models. In this paper, we explore and compare all these three types of language models.

A generative vision model refers to a vision model with the ability to generate images or visual contents. There are two common types of structures, namely U-Net-based [39] and Transformer-based [11]. Generative Adversarial Networks (GANs) [14, 37, 51] employ a framework consisting of a discriminator and a generator, with the generator's structure based on U-Net. On the other hand, motivated by the success of GPT models, recent works have attempted to use the Transformer architecture for image generation in an autoregressive manner, with notable examples being DALLE [36] and CogView [10]. Another popular class of generative models is diffusion models [7, 18, 43, 45], which are based on the diffusion process and gradually denoise to produce natural images. Early diffusion models often employed U-Net as their generative vision model, such as Stable Diffusion, which scaled up the Latent Diffusion Model (LDM) [38] with larger data scales. Some recent works have started to replace the U-Net in diffusion models with Vision Transformer and have made significant progress, such as DiT [30], U-ViT [1] and PixArt [6]. In this paper, we focus on diffusion models and explore both U-Net-based and Transformer-based vision models.

2.2 Text-to-Image Diffusion Models

Text-to-image diffusion models [2,9,29,38,41,44] are capable of generating images based on user prompts. These models consist of two main components, namely a language model and a vision model. The language module is responsible for understanding the text input provided by the user, extracting contextual information, and injecting it into the vision module to generate the desired image. Text-to-image diffusion models have paved the way for various exciting research areas, including image editing [3,22,27], controllable image generation [28,52,53], personalized object generation [12, 16, 40], as well as other interesting applications [5,13,15]. In the extensive exploration of diffusion models, researchers have utilized different language models and vision models. For instance, Stable Diffusion [38] employs CLIP text encoder [31] as its language model and U-Net as its vision model. Imagen [41] utilizes T5 [34] as its language model, which claims to enhance both sample fidelity and image-text alignment. ParaDiffusion [49]

focuses on paragraph-to-image generation and leverages the powerful semantic understanding capability of Llama-2 [47] to comprehend lengthy sentences. PixArt [6], on the other hand, utilizes a ViT [11] as its vision model and achieves high image fidelity while being trained at a lower cost.

After training on a large dataset of text-image pairs [42], the language and vision models in the text-to-image diffusion model become closely intertwined. This tight coupling ensures a strong alignment between the provided text description and the generated image, but at the same time also limits the flexibility of the diffusion model. For instance, if a more advanced language or vision model becomes available, it may have the potential to enhance the text-to-image task. However, decoupling the language and vision modules in existing text-to-image diffusion models and replacing a module with a new one is nontrivial. Therefore, this paper explores the dilemma faced by text-to-image generation and proposes a framework that enables efficient integration of various language models and generative vision models.

3 Method

3.1 Preliminary

A diffusion model is based on the diffusion process for image generation. This process consists of two stages, namely the forward process and the reverse process. During the forward process, Gaussian noise is progressively added to a natural image until the image becomes completely noisy. After that, during the reverse process, the noise is gradually eliminated over a series of time steps, resulting in a natural image. In the reverse process, a trainable vision model is used to predict and remove the noise. By employing this denoising model, we are able to obtain a natural image from Gaussian noise by denoising. Within a text-to-image diffusion model, there are two components at each time step, namely a language model f and a vision model g. The language model converts user input text y into embeddings, which capture the semantic meaning of the text. On the other hand, the vision model, which is the denoising model aforementioned, encodes image features z, extracting relevant visual information from the input images. The interaction between text embeddings and image features is achieved through cross-attention layers, which can be formulated as

$$c = f(y), \tag{1}$$

$$Q = W_q(z), K = W_k(c), V = W_v(c), \tag{2}$$

$$CrossAttention(Q, K, V) = softmax(Q \cdot K^T) \cdot V, \tag{3}$$

where W_q, W_k and W_v are projection matrices.

3.2 Language and Vision Alignment

LaVi-Bridge enables the integration of any two pre-trained language and generative vision models, even though these models are not related and have been

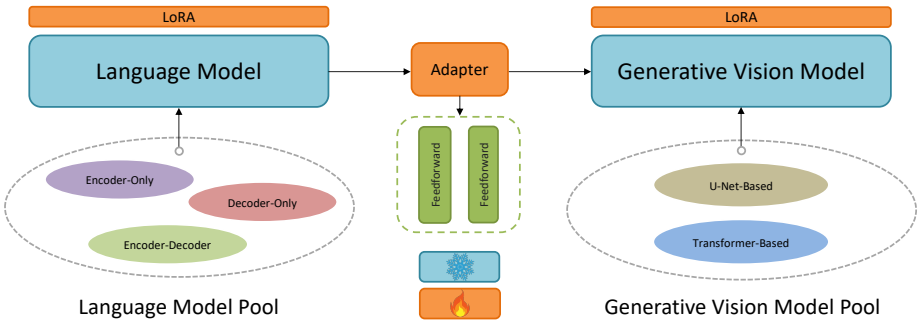

Fig. 2: Pipeline of LaVi-Bridge. We select one model each from the language and vision model pools. We then freeze the pre-trained language and vision models and incorporate LoRA into both models. The connection between the language and vision models is established through an adapter. The only weights we need to train are the ones introduced by LoRA and the adapter.

trained separately. Here, we denote the language model as f and the vision model as g, as mentioned previously. If we directly interact the textual information and image information using Eq. (1), considering that f and g are trained independently, the parameters in the cross-attention layers of g cannot comprehend the text embedding output by f, resulting in meaningless model outputs.

To establish a connection between them, LaVi-Bridge keeps the pre-trained language and vision models fixed and utilizes LoRA to introduce trainable parameters $\Delta\theta$ into both the language model and the vision model. In this context, we denote the language and vision models with LoRA as $f^{\theta_1+\Delta\theta_1}$ and $g^{\theta_2+\Delta\theta_2}$, where θ_1 and θ_2 are the original parameters of f and g, respectively. Furthermore, we introduce an adapter as a bridge between the language model and vision model to facilitate better alignment. The adapter consists of stacked feedforward layers, denoted as h. Consequently, the cross-attention layer can be expressed as

$$c = f^{\theta_1+\Delta\theta_1}(y), \tag{4}$$

$$Q = W_q^{\theta_2+\Delta\theta_2}(z), K = {W_k}^{\theta_2+\Delta\theta_2}(h(c)), V = W_v^{\theta_2+\Delta\theta_2}(h(c)), \tag{5}$$

$$CrossAttention(Q, K, V) = softmax(Q \cdot K^T) \cdot V. \tag{6}$$

Now, we only need to train $\Delta\theta_1$, $\Delta\theta_2$, and h on a relatively small amount of text-image pairs. After training, the language and generative vision models can effectively collaborate to generate meaningful images. We present the framework of LaVi-Bridge in Fig. 2. LaVi-Bridge is very straightforward, with both LoRA and the adapter being its crucial and indispensable components.

3.3 Design Details

LaVi-Bridge is designed to accommodate a wide range of language model structures, including encoder-only, encoder-decoder, decoder-only, as well as generative vision model structures such as U-Net and ViT. In the language model, we inject LoRA into all linear layers of the attention layers. Likewise, in a transformer-based vision model, LoRA is injected into all linear layers of the attention layers. In a U-Net-based vision model, LoRA is injected into all linear layers and convolutional layers of the ResBlocks, attention, and cross-attention layers. To address the dimension disparity between the output embedding of the language model and the dimensions handled by the cross-attention of the vision model, we employ two feedforward layers for the adapter. The input dimension of the adapter matches the output text embedding dimension of the language model, while the output dimension aligns with the dimensions received by the cross-attention of the vision model.

For training, we first select the language and generative vision models that we choose to integrate. We keep their original weights fixed and train LoRA and the adapter on text-image pairs following the design mentioned above. The trained LoRA and adapter have fewer parameters compared to the original model weights, which makes LaVi-Bridge highly flexible. For evaluation, we used various metrics to assess text alignment and image quality across short prompts, long prompts, and compositional prompts.

4 Experiments

4.1 Experimental Settings

In this section, we explored the performance of different language models and generative vision models under LaVi-Bridge. We also tested the impact of LoRA and adapters. We trained on a dataset consisting of a total of 1 million text-image pairs, including around 600k text-image pairs from the COCO2017 [25] train set and 400k text-image pairs from an internal dataset with high-quality images and captions. For each setting, we set the LoRA rank to 32, image resolution to 512×512 and the batch size to 256. We used the AdamW optimizer [26] with a learning rate of 1×10^{-4} and trained for a total of $50k$ steps. During inference, we employed the DDIM sampler [43] for sampling with the number of time steps set to 50 and the classifier free guidance scale [19] set to 7.5.

As mentioned above, we conducted our quantitative evaluation on short prompts, long prompts, and compositional prompts. Specifically,

1. For short prompts, we evaluated using the COCO2014 [25] validation set. We randomly sampled 30k images and tested image quality and text alignment within this subset. We used FID [17] and aesthetic score [23] as evaluation metrics for image quality and CLIP score for text alignment.
2. For long prompts, we employed the same 30k-subset of COCO2014 and utilized Llama-2 to generate expanded captions ranging from 20 to 70 words to

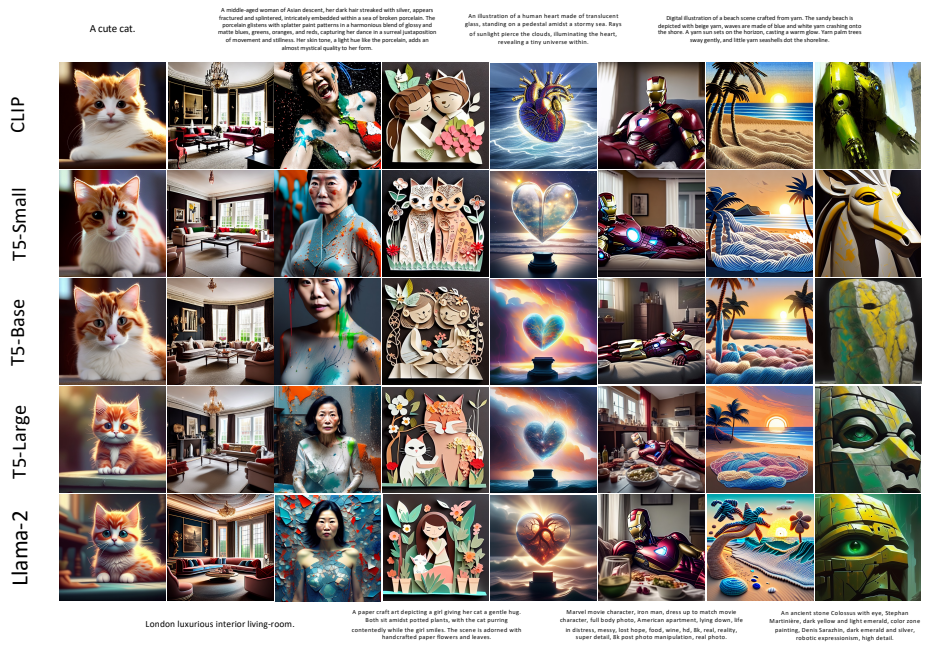

Fig. 3: Visualization results of LaVi-Bridge with different language models. The first row to the fifth row present the results with CLIP text encoder, T5-Small, T5-Base, T5-Large, and Llama-2, respectively. The prompts are displayed at the top or bottom of each column.

construct a dataset of 30k long prompts. Since the caption expansion process does not refer to the content of the reference image, we solely used aesthetic score to evaluate image quality and CLIP score for text alignment.

3. For compositional prompts, we utilized the benchmark proposed by Compbench [21]. Compositional prompts were mainly used to test the model's understanding of textual attributes, such as generating correct object properties like color and shape, as well as accurate relationships between objects, such as spatial positioning.

We conducted a user study on different combinations of language and vision models. For each combination, we evaluated two metrics, namely image quality and text alignment. Users were asked to rank the generated images based on these evaluation criteria. The image ranked last received a score of 1, the second-to-last received a score of 2, and so on. We then calculated the percentage of scores for each model. We selected 20 prompts and included 30 users participated in the testing. In addition to quantitative evaluation and user study, we provided ample visualization results in each section to offer a more intuitive understanding of the performance of each model.

Table 1: Quantitative evaluation of LaVi-Bridge with different language models. "Short", "Long" and "Comp" denote short prompts, long prompts, and compositional prompts respectively. The best results are in **bold**.

	CLIP	T5-Small	T5-Base	T5-Large	Llama-2
Short - FID	23.57	22.98	22.62	23.11	**21.80**
Short - Aesthetics	5.609	5.813	**5.888**	5.881	5.883
Short - CLIP Score	0.3102	0.3122	0.3149	0.3156	**0.3172**
Long - Aesthetics	6.003	6.206	6.284	6.305	**6.355**
Long - CLIP Score	0.3120	0.3111	0.3179	0.3193	**0.3231**
Comp - Color	0.3578	0.3368	0.3856	0.3889	**0.4859**
Comp - Shape	0.3752	0.2962	0.3266	0.3552	**0.4285**
Comp - Texture	0.4506	0.3728	0.4132	0.4524	**0.5055**
Comp - Spatial	0.1296	0.1456	0.1569	0.1582	**0.1914**
Comp - Non-Spatial	0.3009	0.2984	0.3054	0.3068	**0.3106**
Comp - Complex	0.2985	0.2728	0.3055	0.3072	**0.3094**

4.2 Evaluation on Different Language Models

This section evaluates the performance of LaVi-Bridge with different language models. We fixed the vision model to the U-Net of Stable Diffusion V1.4 and integrated it with different language models under LaVi-Bridge. We considered CLIP text encoder, based on the encoder-only framework, T5 series (T5-Small, T5-Base, T5-Large), based on the encoder-decoder framework, and Llama-2-7B, based on the decoder-only framework. We present the visualization results in Fig. 3, quantitative evaluation in Tab. 1, and user study in Figure 1 in the supplementary material.

Visualization From Fig. 3, we can observe that with LaVi-Bridge, all these language models can effectively integrate with U-Net of Stable Diffusion V1.4 and generate meaningful results, such as cases of the cat and living room in Fig. 3. This demonstrates the great generalization ability of LaVi-Bridge for various language models. Additionally, we notice that the performance of different model structures varies when the provided prompts contain more complex semantics. We find that the text-to-image diffusion model corresponding to Llama-2 can perfectly describe semantic information. For example, in the third column, Llama-2's generated result effectively integrates a woman into the sea of fragmented porcelain. In the fourth column, it correctly understands and generates both the girl and the cat in a paper craft art. In the seventh column, it even portrays an entire beach scene using yarn. These examples surpass the capabilities of those models with CLIP and T5. Furthermore, we observe that T5-Large and Llama-2 accurately generate food and wine in the case of Iron Man, and in the last column, they successfully generate "an ancient stone with eyes in dark yellow and emerald". Models with CLIP text encoder, T5-Small, and T5-Base are not able to capture these cases accurately.

Fig. 4: Visualization results of LaVi-Bridge under different generative vision models. The first row to the third row present the results with U-Net in Latent Diffusion Model, U-Net in Stable Diffusion V1.4 and transformer in PixArt, respectively. The prompts are displayed at the top or bottom of each column.

Quantitative Evaluation From Tab. 1, we can observe that Llama-2 achieves the best results for all the metrics used to evaluate text alignment ability, under the setting of all the short prompts, long prompts, and compositional prompts. Besides, Llama-2 also performs the best on most of the metrics evaluating image quality. On the other hand, as the model capacity increases, in general circumstances T5-Large usually outperforms T5-Base, and T5-Base outperforms T5-Small in the area of Natural Language Processing. This conclusion also holds true for LaVi-Bridge. For all the metrics used to evaluate text alignment ability in Tab. 1, T5-Large is superior to T5-Base, and T5-Base is superior to T5-Small. This tells us that incorporating a better language model into the text-to-image diffusion model under LaVi-Bridge can lead to improved text alignment. This makes one of the motivations of LaVi-Bridge meaningful, which is that replacing the model in the existing text-to-image diffusion model with a better model can lead to performance improvements.

User Study We follow the settings described in Sec. 4.1, and the results are presented in Section A in the supplementary material.

4.3 Evaluation on Different Vision Models

This section evaluates the performance of LaVi-Bridge with different vision models. We fixed the language model to T5-Large and integrated it with different generative vision models under LaVi-Bridge. We considered the well-trained U-Nets in the Latent Diffusion Model and Stable Diffusion V1.4, as well as the Vision Transformer in PixArt, totally three models. We present the visualization results in Fig. 4, quantitative evaluation in Tab. 2, and user study in Figure 1 in the supplementary material.

Table 2: Quantitative evaluation of LaVi-Bridge under different generative vision models. "Short", "Long" and "Comp" denote short prompts, long prompts, and compositional prompts respectively. The best results are in **bold**.

	U-Net(LDM)	U-Net(SD)	Transformer(PixArt)
Short - FID	25.94	23.11	**23.02**
Short - Aesthetics	5.703	5.881	**6.145**
Short - CLIP Score	0.3126	0.3156	**0.3172**
Long - Aesthetics	6.122	6.305	**6.406**
Long - CLIP Score	0.3189	0.3193	**0.3210**
Comp - Color	**0.4099**	0.3889	0.3689
Comp - Shape	**0.3724**	0.3552	0.3316
Comp - Texture	**0.5046**	0.4524	0.4553
Comp - Spatial	0.1550	0.1582	**0.1725**
Comp - Non-Spatial	0.3004	0.3068	**0.3098**
Comp - Complex	0.3060	**0.3072**	0.3014

Visualization From Fig. 4, we can see that all these three vision models integrate well with T5-Large and generate relatively accurate images based on the given text prompts. From these cases, we can observe that the images generated by the transformer model based on PixArt exhibit richer details compared to the images generated by the other two models based on U-Net. For example, the forest in the first column, the hull of the pirate ship in the third column, and the bushes at the foot of the mountain in the sixth column are very intricate and realistic. Additionally, we can observe from these cases that the images generated by the model with U-Net of Stable Diffusion V1.4 have more detailed features compared to the images generated by the model with U-Net of Latent Diffusion Model. Furthermore, we can also find that, for the PixArt-based model, text alignment is better in some cases. For instance, in the image of the fifth column, only the model that is based on the transformer of PixArt generates the "aged car" mentioned in the prompt. Similarly, in the seventh column, the garden warrior holding a sword and shield is highly consistent with the prompt description.

Quantitative Evaluation From Tab. 2, it can be observed that for all the metrics measuring image quality, LaVi-Bridge with the PixArt vision model achieves the best results. Additionally, PixArt also achieves the best text alignment for both short and long prompts. This reflects the use of PixArt's transformer as a vision model can also improves the model's understanding of semantics to some extent. Additionally, it is noteworthy that the U-Net in Stable Diffusion, an enhanced version of the U-Net utilized in the Latent Diffusion Model, still outperforms Latent Diffusion Model's U-Net under LaVi-Bridge on all the metrics measuring image quality. This aligns with our previous discussion in Sec. 4.2 and further validates the underlying motivation behind our proposed LaVi-Bridge.

User Study We follow the settings described in Sec. 4.1, and the results are presented in Section A in the supplementary material.

4.4 Ablation Study

In this section, we investigate two sets of ablation experiments. The first set aims to explore the impact of training LaVi-Bridge on the original pre-trained text-to-image diffusion model. The second set of experiments is to study the effects of LoRA and adapters in LaVi-Bridge. For both sets of experiments, we present visualization results in Fig. 5 and provide quantitative evaluations in Tab. 3.

Training with LaVi-Bridge We investigate the impact of our LaVi-Bridge training framework on the original pre-trained text-to-image diffusion model. Specifically, we consider Stable Diffusion V1.4 which adopts CLIP text encoder as its language model and U-Net as its vision model. We incorporate LoRA and an adapter and apply LaVi-Bridge to the same language and vision models with identical structures and weights to those in Stable Diffusion V1.4. We then compare the performance of the model under LaVi-Bridge with the original Stable Diffusion V1.4.

The visualization results are shown in the first two rows of Fig. 5. For these two models, there is no significant difference in image quality and text alignment, varying on a case-by-case basis. In some cases, Stable Diffusion performs better, such as in the third column, where Stable Diffusion successfully generates the case of a "Fox bracelet made of buckskin with fox features", while the model trained under LaVi-Bridge only generates the fox and fails to understand the bracelet made of buckskin. Similarly, in the case of Marvel's Hulk playing basketball, Stable Diffusion generates a slam dunk action following the prompt, whereas the model trained under LaVi-Bridge does not. However, in the second column, the model trained under LaVi-Bridge correctly understands the quantity and successfully generates two elephants, while Stable Diffusion only generates one. Moreover, in the last column, the model trained under LaVi-Bridge accurately describes a frog in a spacesuit, while Stable Diffusion fails.

The left two columns of Tab. 3 present the quantitative evaluation results. It can be observed that Stable Diffusion achieves the best image quality and text alignment for both short prompts and long prompts. However, for compositional prompts, the model trained under LaVi-Bridge outperforms Stable Diffusion in four out of six settings.

Based on the visualization results and quantitative evaluations, we can conclude that overall there is no significant improvement or decline in text alignment. Regarding image quality, it should be noted that training with LaVi-Bridge may result in a decrease compared to the original text-to-image diffusion model, if the same models and weights are used. However, it is important to understand that the main purpose of LaVi-Bridge is to establish connections between different language and vision models, enabling the utilization of more advanced models for performance enhancement. It is not intended to be directly applied to the original text-to-image diffusion models using the same models and weights.

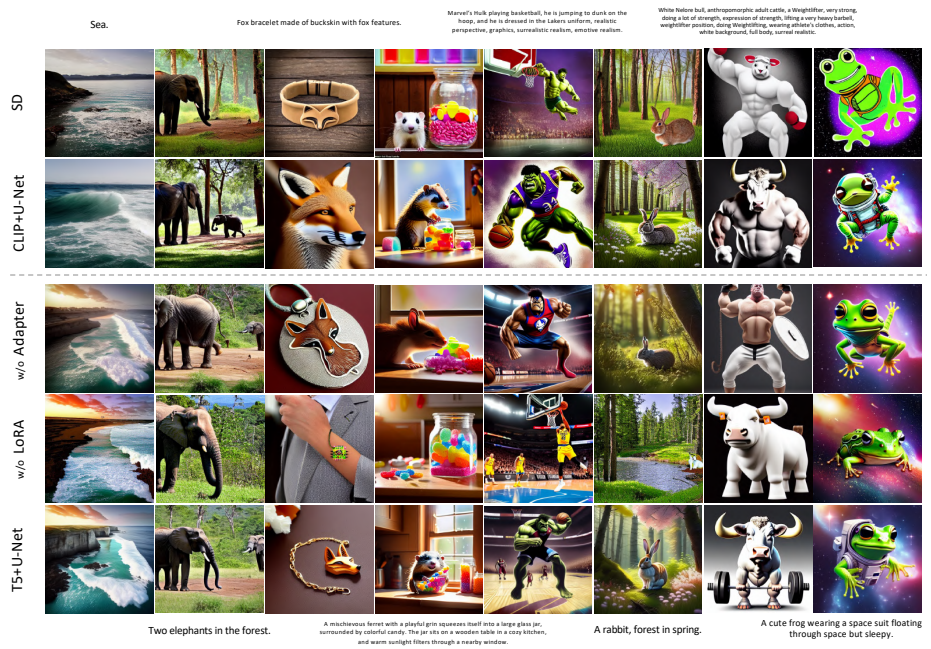

Fig. 5: Visualization results of the ablation study. The top two rows show the impact of LaVi-Bridge on the original pre-trained text-to-image diffusion models. The bottom three rows illustrate the influence of the adapter and LoRA. The prompts are displayed at the top or bottom of each column.

LoRA and Adapter Here, we investigate the role of LoRA and adapters in LaVi-Bridge. We use T5-Large as the language model and Stable Diffusion V1.4's U-Net as the vision model. For the LoRA experiments, we kept the language and vision models fixed without introducing LoRA, and only trained the adapter. For the adapter experiments, considering the mismatch in the dimensions of text embeddings from the language model and the input embeddings acceptable by the vision model, we aligned the dimensions between the language and vision models using a single linear layer instead of stacked feedforward layers which include non-linear activation layers. Under this setting, we trained both LoRA and this linear layer.

The visualization results are shown in the bottom three rows of Fig. 5. We can observe that both image quality and text alignment are significantly affected when LoRA and adapters are not used. For example, in the case of "Bull Fit Athlete" in the seventh column, without LoRA or the adapter, the model cannot understand and integrate these two less related elements, and the image quality is much lower compared to results generated by the original setting. We also found that the results without LoRA are worse than those without the adapter. For instance in the fourth column, there is not even a ferret present in the image in the absence of LoRA.

Table 3: Quantitative evaluation of the ablation study. The left two columns present the impact of LaVi-Bridge on the original pre-trained text-to-image diffusion models. The right three columns demonstrate the influence of the adapter and LoRA. "Short", "Long" and "Comp" denote short prompts, long prompts, and compositional prompts respectively. The best results are in **bold**.

	SD	CLIP+U-Net	w/o Adapter	w/o LoRA	T5+U-Net
Short - FID	**20.32**	23.57	23.81	**22.35**	23.11
Short - Aesthetics	**5.899**	5.609	5.807	5.829	**5.881**
Short - CLIP Score	**0.3132**	0.3102	0.3147	0.3107	**0.3156**
Long - Aesthetics	**6.120**	6.003	6.131	6.273	**6.305**
Long - CLIP Score	**0.3171**	0.3120	0.3106	0.3097	**0.3193**
Comp - Color	0.3570	**0.3578**	0.3550	0.2485	**0.3889**
Comp - Shape	0.3563	**0.3752**	0.3044	0.2944	**0.3552**
Comp - Texture	0.4028	**0.4506**	0.4001	0.3190	**0.4524**
Comp - Spatial	0.1225	**0.1296**	**0.1651**	0.0956	0.1582
Comp - Non-Spatial	**0.3104**	0.3009	0.3065	0.2998	**0.3068**
Comp - Complex	**0.3042**	0.2985	0.2878	0.2687	**0.3072**

The right three columns of Tab. 3 present the quantitative evaluation results. We find that our default setting, which utilizes both LoRA and the adapter, achieves the best performance in most cases. Additionally, overall, the absence of LoRA has a significant impact on text alignment, with many text alignment evaluation metrics being much lower compared to the absence of the adapter.

5 Conclusion

In this paper, we propose LaVi-Bridge, which works on text-to-image diffusion models. LaVi-Bridge is capable of connecting various language models and generative vision models for text-to-image generation. It is highly versatile and can adapt to different structures. LaVi-Bridge is also flexible, as it achieves integration without modifying the original weights of language and vision models. Instead, it utilizes LoRA and an adapter for fine-tuning. Additionally, under LaVi-Bridge, using superior language or vision models can enhance the text comprehension capability or image quality. These advantages enable LaVi-Bridge to help text-to-image diffusion models leverage the latest advancements in the areas of Natural Language Processing and Computer Vision, to enhance text-to-image generation. We believe that this task holds significant research value and requires further exploration. LaVi-Bridge allows designers, artists, and others to flexibly utilize existing language and vision models to achieve their creative goals. It is of utmost importance to avoid misuse and mitigate potential negative social impacts. In practical deployment, it is crucial to standardize its usage, improve model transparency.

References

1. Bao, F., Li, C., Cao, Y., Zhu, J.: All are worth words: a vit backbone for score-based diffusion models. CVPR (2023)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf` (2023)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. CVPR (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS 2020
5. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. NeurIPS (2024)
6. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. ICLR (2024)
7. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. TPAMI (2023)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT (2018)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS (2021)
10. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. NeurIPS (2021)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
13. Ge, S., Park, T., Zhu, J.Y., Huang, J.B.: Expressive text-to-image generation with rich text. ICCV (2023)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS (2014)
15. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. ICLR (2024)
16. Hao, S., Han, K., Zhao, S., Wong, K.Y.K.: Vico: Detail-preserving visual condition for personalized text-to-image generation. arXiv preprint arXiv:2306.00971 (2023)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. NeurIPS Workshop on Deep Generative Models and Downstream Applications (2021)

20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. ICLR (2022)
21. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. arXiv preprint arXiv: 2307.06350 (2023)
22. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. CVPR (2023)
23. LAION-AI: aesthetic-predictor. `https://github.com/LAION-AI/aesthetic-predictor` (2022)
24. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. ECCV (2014)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. ICLR (2019)
27. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. ICLR (2022)
28. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
29. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. ICML (2021)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. ICCV (2023)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. ICML (2021)
32. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. OpenAI blog (2018)
33. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog (2019)
34. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020)
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
36. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. ICML (2021)
37. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. ICML (2016)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. CVPR (2022)
39. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. MICCAI (2015)
40. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. CVPR (2023)

41. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)
42. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. ICLR (2021)
44. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. ICML (2023)
45. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. ICLR (2021)
46. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. NeurIPS (2014)
47. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
49. Wu, W., Li, Z., He, Y., Shou, M.Z., Shen, C., Cheng, L., Li, Y., Gao, T., Zhang, D., Wang, Z.: Paragraph-to-image generation with information-enriched diffusion model. arXiv preprint arXiv:2311.14284 (2023)
50. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. NeurIPS (2024)
51. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. ICCV (2017)
52. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. ICCV (2023)
53. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Unicontrolnet: All-in-one control to text-to-image diffusion models. NeurIPS (2024)

Bridging Different Language Models and Generative Vision Models for Text-to-Image Generation (Supplementary Material)

A User Study

We conducted a user study following the settings described in Section 4.1 in the main paper, and the results are presented in Fig. 1.

The two disk diagrams on the left side of Fig. 1 shows the user's scoring result on different language models. We can find that the model using Llama-2 demonstrates the best performance in terms of both image quality and text alignment, with a particularly pronounced advantage in text alignment. On the other hand, CLIP and T5-Small exhibit noticeably poorer performance on both image quality and text alignment compared to other models.

The two disk diagrams on the right side of Fig. 1 shows the user's scoring result on different generative vision models. We can find that the model using the transformer from PixArt demonstrates the best performance in terms of both image quality and text alignment, with a notably significant advantage in image quality. Additionally, the U-Net in Stable Diffusion outperforms the U-Net in the Latent Diffusion Model overall.

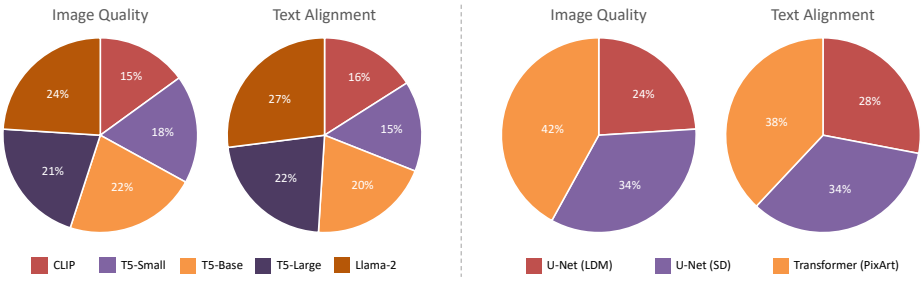

Fig. 1: User study. The two disk diagrams on the left display the user's scoring results on different language models, while the two disk diagrams on the right display the user's scoring results on different generative vision models. The percentage represents the proportion of the score obtained by a model out of the total score of all models.

B Long Prompts

In this section, we provided a comprehensive explanation of the evaluation for long prompts mentioned in the main paper. To conduct this evaluation, we uti-

lized the subset of 30k text-image pairs from COCO2014, which was originally used for evaluating short prompts, and employed the Llama-2 to extend the captions within this subset to a length of 20-70 words.

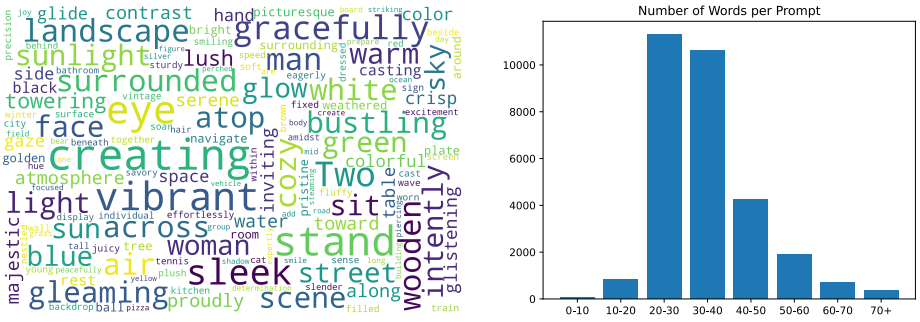

Fig. 2: Statistics regarding the long prompts utilized in evaluations in the main paper. The figure on the left visualizes word frequency. The histogram on the right presents the distribution of sentence length, where horizontal axis represents the range of word counts in the prompts, and the vertical axis represents the number of prompts falling within each sentence length range.

Specifically, we used the Llama-2-7b-chat model and provided the following prompt: *"Please expand the caption to 20-70 words to enrich its semantic meaning: '**placeholder**'. Just give me one answer."* Here, the term "***placeholder***" represents the short prompts that we aimed to expand. In this way, we successfully generated 30k long prompts for evaluation. Here are a few examples of the generated long prompts:

1. Short Prompt: *"a brown, white and yellow bird standing in the grass."* ⟶ Long Prompt: *"A striking, multi-hued bird with warm brown plumage, crisp white patches on its wings, and vibrant yellow feathers stands gracefully in lush green grass, creating a picturesque scene."*
2. Short Prompt: *"two sheep standing in the snow with one looking for food"* ⟶ Long Prompt: *"Two fluffy white sheep stand stately in the pristine snow, their wool glistening under the crisp sunlight. One of them eagerly scans the ground, sniffing out potential nourishment amidst the frozen landscape, while the other stands watchful and still, seemingly lost in thought."*
3. ···

We presented statistical analysis of the generated long prompts in the Fig. 2. The figure on the left visualizes word frequency, while histogram on the right illustrates the distribution of sentence lengths. We observed that the vocabulary diversity in long prompts is quite rich, and the majority of long prompts typically have sentence lengths ranging from 20 to 40 words, and there is also a certain portion that falls within the 40-60 words range.

C Training Steps

As mentioned in the main paper, we trained the LaVi-Bridge for 50k steps. In this section, we present the generated images as training progresses. We utilized the T5-Large as the language model and the U-Net from Stable Diffusion V1.4 as the vision model.

Fig. 3 illustrates the evolution of the model's performance. Initially, during the first 1k steps, the image quality was poor, and the model struggled to comprehend the given prompt. However, as the training progressed to 10k steps, there was a significant improvement in image quality. By the time it reached 20k steps, the model exhibited enhanced semantic understanding. Finally, at 50k steps, the model demonstrated further optimization compared to the model at 20k step, showcasing the best performance.

Fig. 3: The results for different training steps. Prompts are displayed above, while the number of training steps is shown below.

D Training Cost

As mentioned in the main paper, LaVi-Bridge does not require modifying the original weights of the language and vision models. Instead, it introduces and trains LoRA and adapters. This approach significantly reduces the training cost compared to training the entire text-to-image diffusion model. For the training of LaVi-Bridge, we utilized 8 A100 GPUs with a batch size of 256 and completed the training in less than 2 days. Furthermore, we provided a comparison of the number of parameters for different language and vision model combinations in Tab. 1. The leftmost column shows the language and vision models used. It can be observed that training only the LoRA and the adapter leads to a significant reduction in the number of trainable parameters compared to training the original language and vision model. Additionally, both LoRA and the adapter are plug-and-play components, making LaVi-Bridge highly flexible.

E Training Set

As mentioned in our main paper, we conducted training using the COCO2017 [1] train set, which consists of around 600k text-image pairs, along with an addi-

Table 1: Comparison of number of parameters.

	Language Model	Vision Model	**Sum**	Adapter	Language LoRA	Vision LoRA	**Sum**
CLIP U-Net(SD)	123M	860M	**983M**	14M	2M	28M	**44M**
T5-Small U-Net(SD)	35M	860M	**895M**	9M	0.8M	28M	**38M**
T5-Base U-Net(SD)	110M	860M	**970M**	14M	2M	28M	**44M**
T5-Large U-Net(SD)	335M	860M	**1195M**	21M	6M	28M	**55M**
Llama-2 U-Net(SD)	6738M	860M	**7598M**	229M	34M	28M	**291M**
T5-Large U-Net(LDM)	335M	872M	**1207M**	30M	6M	29M	**65M**
T5-Large Transformer (PixArt)	335M	611M	**946M**	113M	6M	17M	**136M**

tional 400k internal data. The COCO2017 dataset primarily comprises highly realistic images with short and straightforward captions. In order to enhance the diversity and quality of the training data, we collected an additional 400k text-image pairs that exhibit a wide range of artistic styles and high-quality visuals, accompanied by accurate and detailed captions. We provide some examples of the COCO2017 dataset and the internal dataset in Fig. 4.

In Fig. 5, we present a comparison of the results obtained from training on COCO2017 alone and training on both COCO2017 and our internal dataset. We used T5-Large [2] as the language model and Stable Diffusion V1.4's U-Net [3] as the vision model for LaVi-Bridge. It can be observed in Fig. 5 that in the first three columns, the model trained solely on COCO2017 performs well in terms of both image quality and text alignment. The model accurately understands quantities, as seen in the case of two elephants, and attributes, such as the black cat and white pillow or the brown bench in front of the white building, and so on. Furthermore, the model is capable of generating images using complex prompts, as demonstrated in columns 4-6. However, when the model is tasked with generating images depicting fanciful scenarios, as shown in the seventh column, or images with a non-realistic style, as shown in the last column, the model trained only on COCO2017 struggles to produce such images.

As mentioned earlier, the text-image pairs in COCO2017, from a certain perspective, lack diversity. This is because the images in COCO2017 have a highly realistic style, resulting in a lack of variety, and their quality is relatively low. Additionally, all the captions in COCO2017 are short and very direct, as illus-

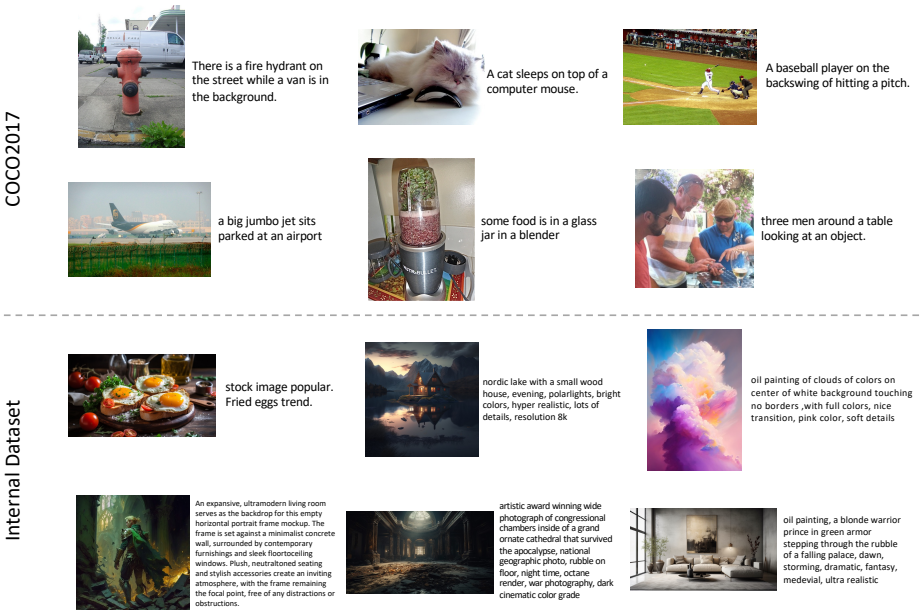

Fig. 4: Examples from COCO2017 train set and the internal dataset.

trated in Fig. 4. Consequently, it is expected that models trained on COCO2017 will face challenges in generating images of whimsical scenarios or non-realistic styles. Fortunately, there are now many high-quality text-image datasets available, such as [4], and we highly recommend incorporating these datasets into training in order to achieve better image generation.

Here, we want to emphasize the contribution of LaVi-Bridge again. LaVi-Bridge is a framework designed to bridge various language and vision models. Even though it was trained only on COCO2017, as can be seen from the first six columns in Fig. 5, LaVi-Bridge is still effective. The reason for the poor performance in the last two columns in the first row of Fig. 5 is due to the limited diversity of the COCO2017 training set as previously mentioned. If LaVi-Bridge is trained on a more general text-image dataset, it will become more versatile, enabling better generation on a wide variety of images.

F LaVi-Bridge vs. Whole Model Training

In this section, we explored the performance gap between LaVi-Bridge and fine-tuned the whole T2I model without using LoRA. We evaluated on one combination: CLIP+U-Net-SD and compared the trained model with the corresponding version of LaVi-Bridge. We showcase the results in Fig. 6 and Tab. 2. It can be observed that fine-tuning the whole model achieves better performance in 4 out of 5 settings for short and long prompts. However, for compositional prompts,

Fig. 5: Comparison of training only on COCO2017 and training on both COCO2017 and the internal dataset.

LaVi-Bridge outperforms in 5 out of 6 settings. And the performance gap between training the whole model and LaVi-Bridge is not substantial in most settings favoring the former.

Fig. 6: Visualization results of different training strategies. The prompts can be found in Figure 5 in the main paper.

G FID-CLIP Curves

To provide more comprehensive measurements, we reported the FID-CLIP curves in Fig. 7 on short prompts as CFG varied.

H More Visualization Results

In this section, we present additional visualization results in Fig. 8. In columns one through five, we kept the vision model fixed as U-Net from Stable Diffusion V1.4 and employed different language models. In columns four, six, and seven, we kept the language model fixed as T5-Large and utilized different vision models.

Table 2: Quantitative evaluation of different training strategies.

	Whole Model	LaVi-Bridge
Short - FID	**22.73**	23.57
Short - Aesthetics	**5.792**	5.609
Short - CLIP Score	0.3089	**0.3102**
Long - Aesthetics	**6.052**	6.003
Long - CLIP Score	**0.3131**	0.3120
Comp - Color	0.3559	**0.3578**
Comp - Shape	0.3612	**0.3752**
Comp - Texture	0.4021	**0.4506**
Comp - Spatial	0.1218	**0.1296**
Comp - Non-Spatial	**0.3070**	0.3009
Comp - Complex	0.2980	**0.2985**

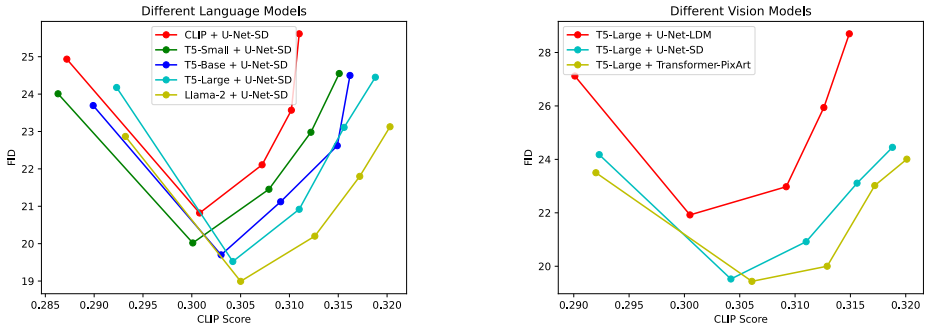

Fig. 7: FID-CLIP curves. CFG: [1.5, 3, 5, 7.5, 9]

Fig. 8: More visualization results. The first column to the seventh column present the results of different combinations using LaVi-Bridge, where the language and vision models used are indicated at the top of each column. The prompts for each row are displayed either on the right or left.

References

1. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. ECCV (2014)
2. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020)
3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. CVPR (2022)
4. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. NeurIPS (2024)