PanGu-Draw: Advancing Resource-Efficient Text-to-Image Synthesis with Time-Decoupled Training and Reusable Coop-Diffusion

Guansong Lu[1], Yuanfan Guo[1], Jianhua Han[1], Minzhe Niu[1], Yihan Zeng[1], Songcen Xu[1], Zeyi Huang[2], Zhao Zhong[2], Wei Zhang[1], and Hang Xu[1⋆]

[1] Huawei Noah's Ark Lab [2] Huawei

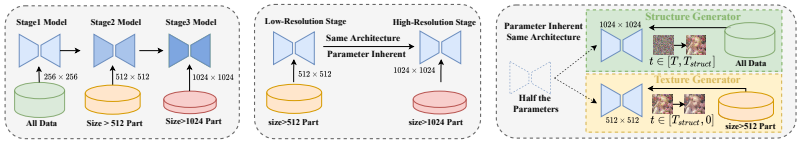
Abstract. Current large-scale diffusion models represent a giant leap forward in conditional image synthesis, capable of interpreting diverse cues like text, human poses, and edges. However, their reliance on substantial computational resources and extensive data collection remains a bottleneck. On the other hand, the integration of existing diffusion models, each specialized for different controls and operating in unique latent spaces, poses a challenge due to incompatible image resolutions and latent space embedding structures, hindering their joint use. Addressing these constraints, we present **"PanGu-Draw"**, a novel latent diffusion model designed for resource-efficient text-to-image synthesis that adeptly accommodates multiple control signals. We first propose a resource-efficient Time-Decoupling Training Strategy, which splits the monolithic text-to-image model into structure and texture generators. Each generator is trained using a regimen that maximizes data utilization and computational efficiency, cutting data preparation by 48% and reducing training resources by 51%. Secondly, we introduce **"Coop-Diffusion"**, an algorithm that enables the cooperative use of various pre-trained diffusion models with different latent spaces and predefined resolutions within a unified denoising process. This allows for multi-control image synthesis at arbitrary resolutions without the necessity for additional data or retraining. Empirical validations of Pangu-Draw show its exceptional prowess in text-to-image and multi-control image generation, suggesting a promising direction for future model training efficiencies and generation versatility. The largest 5B T2I **PanGu-Draw** model is released on the Ascend platform. Project page: https://pangu-draw.github.io

Keywords: Text-to-Image Synthesis · Resource-Efficient · Diffusion Model

1 Introduction

The Denoising Diffusion Probabilistic Models (DDPMs) [11] and their subsequent enhancements [5, 14, 21] have established diffusion models as a leading approach for image generation. These advancements excel in the application of diffusion models to text-to-image synthesis, yielding high-fidelity results with large-scale models and datasets, supported by substantial computational resources [13, 20, 24, 26, 29]. These foundational models, capable of understanding

⋆ Corresponding author: xu.hang@huawei.com



Resource Efficiency \ Training Strategy	Cascaded Training (e.g. DeepFlyod-IF, GLIDE...)	Resolution Boost Training (e.g. Stable Diffusion, AltDiffusion...)	Time-decoupling(Ours)	Ours vs. Resolution Boost Training
Data Efficiency	High All training data	Low Drop low resolution data	High All training data	+ 48% Data Usage
Training Efficiency	Low 3x time training	Mid High cost for High-Resolution Stage	High Half the parameters, Texture Generator train on low res.	- 51% Training Cost
Inference Efficiency	Low 3x times for cascaded models inference	Mid Single large model across all timesteps	High Half the parameters	- ~50% Infer Time Similar Performance

(d) Comparison on Resource Efficiency

Fig. 1: Illustration of three multi-stage training strategies and comparison between them in resource efficiency in data, training and inference aspects. Our time-decoupling training strategy significantly surpasses the representative methods in Cascaded Training [20, 31] and Resolution Boost Training [26, 37] in resource efficiency.

and rendering complex semantics, have paved the way for diverse image generation tasks, accommodating various control signals such as reference images, edges [39], and poses [39].

However, the extensive computational demand and significant data collection required by these models pose a substantial challenge. The ambitious goal of higher fidelity and increased resolution in image synthesis pushes the boundaries of model and dataset sizes, escalating computational costs, and environmental impact. Moreover, the aspiration for versatile control and multi-resolution in image generation introduces additional complexity. Existing diffusion models, each tailored for specific controls and operating within distinct latent spaces, face the challenge of integration due to incompatible image resolutions and latent space embeddings, obstructing their concurrent utilization.This incompatibility not only leads to more resource consumption of retraining but also impedes the joint synthesis of images controlled by multiple factors, thereby limiting the scalability and practical application of such existing generative models. In response to these challenges, our work introduces a novel paradigm named **"PanGu-Draw"** that judiciously conserves training resources while enhancing data efficiency, thereby proposing a resource-efficient pathway forward for diffusion model scalability.

As shown in Figure 1, the training strategies of predecessors like DeepFloyd [31] and GLIDE [20], which employ a cascaded approach, excel in leveraging data across resolutions but suffer from inefficient inference due to their reliance on multiple models. Alternatively, Stable Diffusion [26] and AltDiffusion [37] use a Resolution Boost Training strategy aiming for cost-effectiveness by refining a single model. However, this strategy falls short on data efficiency.

In light of these considerations, our PanGu-Draw framework advances the field by presenting a Time-Decoupling Training Strategy that segments the training of a comprehensive text-to-image model into two distinct generators: one dedicated to structural outlines and another to textural details. This division not only concentrates on training efforts but also enhances data efficacy. The structural generator is adept at crafting the initial outlines of images, offering

flexibility in data quality and enabling training across a spectrum of data calibers; the textural generator, in contrast, is fine-tuned using low-resolution data to infuse these outlines with fine-grained details, ensuring optimal performance even during high-resolution synthesis. This focused approach not only accelerates the training process of our **5B model** but also significantly reduces the reliance on extensive data collection and computational resources, as evidenced by a 48% reduction in data preparation and a 51% reduction in resource consumption.

Furthermore, we introduce a pioneering algorithm named **Coop-Diffusion**, which facilitates the cooperative integration of diverse pre-trained diffusion models. Each model, conditioned on different controls and pre-defined resolutions, contributes to a seamless denoising process. The first algorithmic sub-module addresses inconsistencies in VAE decoders that arise during the denoising process across different latent spaces, ensuring cohesive image quality by effectively reconciling disparate latent space representations. The second sub-module confronts the challenges associated with multi-resolution denoising. Traditional bilinear upsampling for the intermediate noise map, introduced during the denoising process, can undesirably amplify the correlation between pixels. This amplification deviates from the initial Independent and Identically Distributed (IID) assumption, leading to severe artifacts in the final output image. However, our innovative approach circumvents this issue with a single-step sampling method that preserves the integrity of pixel independence, thus preventing the introduction of artifacts. **Coop-Diffusion** obviates the need for additional data or model retraining, addressing the challenges of multi-control and multi-resolution image generation with scalability and efficiency.

PanGu-Draw excels in text-to-image (T2I) generation, outperforming established models like DALL-E 2 and SDXL, as evidenced by its FID of 7.99 in English T2I. It also leads in Chinese T2I across metrics like FID, IS, and CN-CLIP-score. User feedback highlights a strong preference for PanGu-Draw, aligning well with human visual perceptions.

In summary, our contributions are manifold:

• **PanGu-Draw**: A resource-efficient diffusion model with a Time-Decoupling Training Strategy, reducing data and training resources for T2I synthesis.

• **Coop-Diffusion**: A novel approach for integrating multiple diffusion models, enabling efficient multi-control image synthesis at multi-resolutions within a unified denoising process.

• Comprehensive evaluations demonstrate **PanGu-Draw**'s (5B model) can produce high-quality images aligned with text and various controls, advancing the scalability and flexibility of diffusion-based image generation.

2 Related Work

Text-to-Image Generation. The integration of diffusion models into the realm of text-to-image (T2I) generation marks a significant stride in computational creativity [5, 7, 12, 13, 20, 21, 24, 26, 29, 31–33]. Models like GLIDE [20] and DALL-E 2 [24] have significantly advanced in generating diverse and semantically aligned images from textual descriptions. LDM [26] addresses computational challenges by creating images from text-conditioned low-dimensional latent representations.

ControlNet [39] introduces spatial conditioning controls, offering flexibility in image generation under varied conditions like edges and depth. Despite the proliferation of such specialized models, a unified framework that consolidates these disparate capabilities remains absent, limiting the potential for multi-control and complex editing in image synthesis.

Model Efficient Training and Scaling Up Strategies. Efficient training and scaling of models are pivotal for advancing large-scale neural networks. Previous models like DeepFloyd [31] and GLIDE [20] capitalize on cascaded approaches that proficiently utilize data across various resolutions, which results in less efficient inference processes. Contrastingly, models like Stable Diffusion [26] and AltDiffusion [37] adopt Resolution Boost Training strategies that refine a single model for cost-effectiveness, which however do not fully exploit data efficiency. In scaling up strategies, training efficiency is also important. Efficient adaptation and scaling are explored in [3] through distillation, and in [23] by marrying model expansion with domain-specific prompts. Serial scaling and knowledge distillation reduce training times significantly as demonstrated by [8], while [6] proposes progressive network expansion for faster training with minimal loss. We propose a novel Time-Decoupling training strategy to diffusion model scaling that enhances efficiency. eDiff-I [1] similarly proposes to split a diffusion model across time steps to boost generation quality without increasing inference cost. However, they do not consider about data and training efficiencies.

3 Preliminary

Given an image x_0, diffusion models first produce a series of noisy images $x_1, ..., x_T$ by adding Gaussian noise to x_0 according to some noise schedule given by $\bar{\alpha}_t$ as follows:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

Diffusion models then learn a denoising model $\epsilon_\theta(x_t, t)$ to predict the added noise of a noisy image x_t with the following training objective:

$$\mathcal{L} = \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,I), t \sim [1,T]} \left\| \epsilon - \epsilon_\theta \left(x_t, t \right) \right\|^2, \tag{2}$$

where t is uniformly sampled from $\{1, ..., T\}$. Once the denoising model $\epsilon_\theta(x_t, t)$ is learned, starting from a random noise $x_T \sim \mathcal{N}(0, I)$, one can iteratively predict and reduce the noise in x_t to get a real image x_0. During the sampling process, we can predict the clean data x_0 from $\epsilon_\theta \left(x_t, t \right)$ with single-step sampling as:

$$\hat{x}_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t)). \tag{3}$$

Our text-to-image generation model is built on the model architecture proposed in Latent Diffusion Model [26]. In this model, a real image x_0 is first down-sampled 8 times as a lower-dimension latent code z_0 with an image encoder model E, which can be decoded with a latent decoder model D back to a real image x_0. The denoising network $\epsilon_\theta(z_t, t, c)$ is parameterized as a U-Net [28]

model, where embedding of time step t is injected with adaptive normalization layers and embedding of input text c is injected with cross-attention layers.

4 PanGu-Draw

In this section, we first illustrate our resource-efficient 5B text-to-image generation model, trained with a time-decoupling training strategy and further enhanced with a prompt enhancement LLM. Then, we present our *Coop-Diffusion* algorithm for the cooperative integration of diverse pre-trained diffusion models, enabling multi-control and multi-resolution image generation.

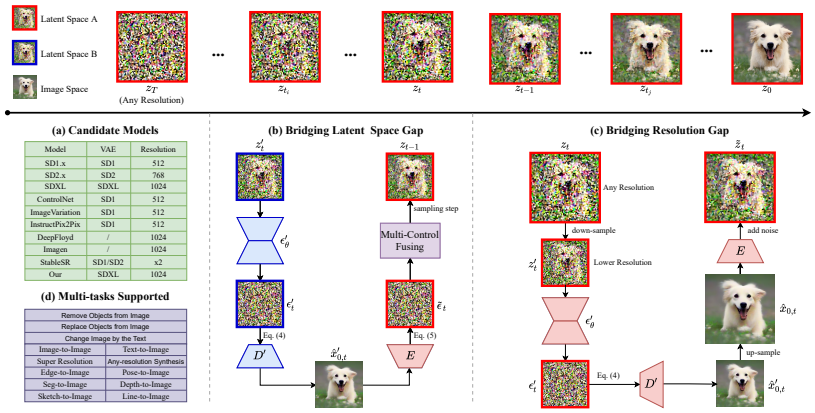
4.1 Time-Decoupling Training Strategy

Enhancing data, training, and inference efficiency is vital for text-to-image models' practical use. Figure 1 shows two existing training strategies: (a) Cascaded Training, using three models to incrementally improve resolution, is data-efficient but triples training and inference time. (b) Resolution Boost Training starts at 512×512 and then 1024×1024 resolution, discarding lower resolution data and offering moderate efficiency with higher training costs and single-model inference across all timesteps. These approaches differ from our time-decoupling strategy, detailed below.

Responding to the need for enhanced efficiencies, we draw inspiration from the denoising trajectory of diffusion processes, where initial denoising stages primarily shape the image's structural foundation, and later stages refine its textural complexity. With this insight, we introduce the Time-Decoupling Training Strategy. This approach divides a comprehensive text-to-image model, denoted as ϵ_θ, into two specialized sub-models operating across different temporal intervals: a structure generator, ϵ_{struct}, and a texture generator, $\epsilon_{texture}$. Each sub-model is half the size of the original, thus enhancing manageability and reducing computational load.

As illustrated in Figure 1(c), the structure generator, ϵ_{struct}, is responsible for early-stage denoising across larger time steps, specifically within the range $T, ..., T_{struct}$, where $0 < T_{struct} < T$. This stage focuses on establishing the foundational outlines of the image. Conversely, the texture generator, $\epsilon_{texture}$, operates during the latter, smaller time steps, denoted by $T_{struct}, ..., 0$, to elaborate on the textural details. Each generator is trained in isolation, which not only alleviates the need for high-memory computation devices but also avoids the complexities associated with model sharding and its accompanying inter-machine communication overhead.

In the inference phase, ϵ_{struct} initially constructs a base structural image, $z_{T_{struct}}$, from an initial random noise vector, z_T. Subsequently, $\epsilon_{texture}$ refines this base to enhance textural details, culminating in the final output, z_0. This sequential processing facilitates a more resource-efficient workflow, significantly reducing the hardware footprint and expediting the generation process without compromising the model's performance or output quality, as demonstrated in our ablated experiment in Sec. 5.3.



Model	VAE	Resolution
SD1.x	SD1	512
SD2.x	SD2	768
SDXL	SDXL	1024
ControlNet	SD1	512
ImageVariation	SD1	512
InstructPix2Pix	SD1	512
DeepFloyd	/	1024
Imagen	/	1024
StableSR	SD1/SD2	x2
Our	SDXL	1024

Remove Objects from Image	
Replace Objects from Image	
Change Image by the Text	
Image-to-Image	Text-to-Image
Super Resolution	Any-resolution Synthesis
Edge-to-Image	Pose-to-Image
Seg-to-Image	Depth-to-Image
Sketch-to-Image	Line-to-Image

Fig. 2: Visualization of our *Coop-Diffusion* algorithm for the cooperative integration of diverse pre-trained diffusion models. (a) Existing pre-trained diffusion models, each tailored for specific controls and operating within distinct latent spaces and image resolutions. (b) This sub-module bridges the gap arising from different latent spaces by transforming ϵ'_t in latent space B to the target latent space A as $\tilde{\epsilon}_t$. (c) This sub-module bridges the gap arising from different resolutions by performing upsampling on the predicted clean data $\hat{x}'_{0,t}$.

Resource-Efficient Specialized Training Regime. We further adopt specialized training designs for the above two models. The structure generator ϵ_{struct}, which derives image structures from text, requires training on an extensive dataset encompassing a wide range of concepts. Traditional methods, like Stable Diffusion, often eliminate low-resolution images, discarding about 48% of training data and thereby inflating dataset costs. Contrarily, we integrate high-resolution images with upscaled lower-resolution ones. This approach, as proven by our ablated experiments in Sec. 5.3, shows no performance drop, as the predicted $z_{T_{struct}}$ still contains substantial noise. In this way, we achieve higher data efficiency and avoid the problem of semantic degeneration.

Since the image structure is determined in $z_{T_{struct}}$ and the texture generator $\epsilon_{texture}$ focuses on refining texture, we propose training $\epsilon_{texture}$ at a lower resolution while still sampling at high resolution. This strategy, as demonstrated in our ablated experiments in Sec. 5.3, results in no performance drop and no structural problems (e.g., repetitive presentation [15]). Consequently, we achieved an overall 51% improvement in training efficiency. Figure 1 summarizes the data, training, and inference efficiency of different training strategies. Besides higher data and training efficiency, our strategy also achieves higher inference efficiency with fewer inference steps compared to the Cascaded Training strategy and a smaller per-step model compared to the Resolution Boost Training strategy.

4.2 Coop-Diffusion: Multi-Diffusion Fusion

As shown in Figure 2(a), there are numerous pre-trained diffusion models, such as various SD, ControlNet, image variation, etc., each tailored for specific con-

trols and image resolutions. It is promising to fuse these pre-trained models for multi-control or multi-resolution image generation without needing to train a new model. However, the different latent spaces and resolutions of these models impede joint synthesis of images controlled by different models, thereby limiting their practical applications. In response to these challenges, we propose the *Coop-Diffusion* algorithm with two key sub-modules, as shown in Figures 2(b) and (c), to bridge the latent space gap and the resolution gap, and to unite the denoising process in the same space.

Bridging the Latent Space Gap. To bridge the latent space gap between two different latent spaces A and B, we propose to unify the model prediction in latent space A by transforming the model prediction ϵ'_t in latent space B to latent space A using the image space as an intermediate. This is done in the following way: first, we predict the clean data $\hat{z}'_{0,t}$ using Equation (3) as:

$$\hat{z}'_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(z'_t - \sqrt{1-\bar{\alpha}_t}\epsilon'_t), \tag{4}$$

which is then decoded into a pixel-level image $\hat{x}'_{0,t}$ using the latent decoder model D'. This image is encoded into latent space A using the image encoder model E, as $\tilde{z}_{0,t} = E(\hat{x}'_{0,t})$, and finally transformed into a model prediction by inverting Equation (3) as:

$$\tilde{\epsilon}_t = \frac{1}{\sqrt{1-\bar{\alpha}_t}}(z_t - \sqrt{\bar{\alpha}_t}\tilde{z}_{0,t}). \tag{5}$$

With the united $\tilde{\epsilon}_t$, we can now perform multi-control fusion between $\tilde{\epsilon}_t$ and ϵ_t (the prediction from model ϵ_θ with z_t in latent space A, omitted in Figure 2 for brevity). In this paper, we adopt the fusion method proposed in Composable-Diffusion [19] as: $\epsilon_{t,fuse} = d \cdot \tilde{\epsilon}_t + (1-d) \cdot \epsilon_t$, where d and $1-d$ are the guidance strengths of each model with $d \in [0, 1]$, to guide the denoising process jointly with these two models for multi-control image generation. Algorithm 1 further illustrates this unification and fusion process.

Bridging Resolution Gap. To integrate the denoising processes of a low-resolution model with a high-resolution model, upsampling and/or downsam-

(a) Upsampling from intermediate z_t. (b) Our upsampling algorithm.

Fig. 3: Results of fusing a low-resolution model and a high-resolution model with different upsampling methods. Upsampling from intermediate z_t results in severe artifacts, while our upsampling algorithm results in high-fidelity image.

Algorithm 1 Coop-Diffusion: Multi-Diffusion Fusing

Sub-Module 1. Bridging Latent Space Gap

Input: random noise $z_T \sim \mathcal{N}(0, I)$, diffusion model ϵ_θ, decoder D, encoder E in latent space A; random noise $z'_T = z_T$, diffusion model ϵ'_θ, decoder D', encoder E' in latent space B; guidance strength d, sampling method S.

for $t = T, \dots, 1$ **do**
  $\epsilon'_t = \epsilon'_\theta(z'_t)$, $\hat{z}'_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(z'_t - \sqrt{1-\bar{\alpha}_t}\epsilon'_t)$
  $\hat{x}'_{0,t} = D'(\hat{z}'_{0,t})$, $\tilde{z}_{0,t} = E(\hat{x}'_{0,t})$
  $\tilde{\epsilon}_t = \frac{1}{\sqrt{1-\bar{\alpha}_t}}(z_t - \sqrt{\bar{\alpha}_t}\tilde{z}_{0,t})$
  $\epsilon_t = \epsilon_\theta(z_t)$, $\epsilon_{t,fuse} = d \cdot \tilde{\epsilon}_t + (1-d) \cdot \epsilon_t$
  $z_{t-1} = S(z_t, t, \epsilon_{t,fuse})$
  $z'_{t-1} = S(z'_t, t, \epsilon'_{t,fuse})$ ▷ $\epsilon'_{t,fuse}$ from $\epsilon_{t,fuse}$ similar to the process from ϵ'_t to $\tilde{\epsilon}_t$, omitted for brevity
end for
return $D(z_0)$

Sub-Module 2. Bridging Resolution Gap

Input: diffusion model ϵ_θ, decoder D, encoder E in high-resolution space; random noise $z'_T \sim \mathcal{N}(0, I)$, diffusion model ϵ'_θ, decoder D', encoder E' in low-resolution space; low-resolution sampling end step T_{low}, sampling method S.

for $t = T, \dots, T_{low} + 1$ **do**
  $\epsilon'_t = \epsilon'_\theta(z'_t)$, $z'_{t-1} = S(z'_t, t, \epsilon'_t)$
end for
$\hat{z}'_{0,T_{low}} = \frac{1}{\sqrt{\bar{\alpha}_{T_{low}}}}(z'_{T_{low}} - \sqrt{1-\bar{\alpha}_{T_{low}}}\epsilon'_{T_{low}})$
$\hat{x}'_{0,T_{low}} = D'(\hat{z}'_{0,T_{low}})$, $\hat{x}_{0,T_{low}} =$Upsample$(\hat{x}'_{0,T_{low}})$
$\hat{z}_{0,T_{low}} = E(\hat{x}_{0,T_{low}})$
$z_{T_{low}} = \sqrt{\bar{\alpha}_{T_{low}}}\hat{z}_{0,T_{low}} + \sqrt{1-\bar{\alpha}_{T_{low}}}\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$
for $t = T_{low}, \dots, 1$ **do**
  $\epsilon_t = \epsilon_\theta(z_t)$, $z_{t-1} = S(z_t, t, \epsilon_t)$
end for
return $D(z_0)$

pling is necessary. Traditional bilinear upsampling, often applied to the intermediate result z_t during the denoising process, can undesirably amplify pixel correlation. This amplification deviates from the initial Independent and Identically Distributed (IID) assumption, leading to severe artifacts in the final images, as shown in Figure 3(a). Conversely, downsampling does not present this issue. To address the IID issue in upsampling, we propose a new upsampling algorithm that preserves the IID assumption, thereby bridging the resolution gap between models with different pre-trained resolutions.

Figure 2(c) visualizes our upsampling algorithm. Specifically, for a low-resolution z'_t, we use the image space as an intermediate space to transform z'_t in low-resolution space into high-resolution space as $\tilde{z}_t$. We first predict the noise ϵ'_t with the denoising model ϵ'_θ and then predict the clean data $\hat{z}'_{0,t}$ as described in Eq. 4. This is decoded into an image $\hat{x}'_{0,t}$ using decoder D'. We then perform upsampling on $\hat{x}'_{0,t}$ to obtain its high-resolution counterpart $\hat{x}_{0,t}$. Finally, $\hat{x}_{0,t}$ is encoded into the latent space with encoder E as $\hat{z}_{0,t}$, and t-step noise is added to get the final result $\tilde{z}_t$ using Eq. 1.

With the unified $\tilde{z}_t$, we can now perform multi-resolution fusion. First, we denoise with a low-resolution model to obtain the intermediate z'_t and its high-

resolution counterpart $\tilde{z}_t$. Then, we perform denoising with a high-resolution model starting from $\tilde{z}_t$, and vice versa. This approach allows us to conduct one-stage super-resolution without undergoing all the low-resolution denoising steps, thereby improving inference efficiency. Algorithm 1 further illustrates this unification and fusion process.

5 Experiments

Implementation Details. We adopt the pretrained Variational Autoencoder (VAE) model from SDXL [22], and we build our structure and texture generator based on the architecture of its U-Net model with the following modifications. To achieve bilingual text-to-image generation (Chinese and English), we pretrain a Chinese text encoder [9, 36] on our Chinese training dataset. We then concatenate the text embeddings from this Chinese text encoder with those from a pretrained English text encoder, serving as the final text embeddings for the denoising models. For multi-resolution image generation, we select a range of image resolutions around 1024×1024 and further condition the denoising model on the sinusoidal positional embeddings corresponding to the index of image resolutions. The T_{struct} parameter is set to 500, as suggested by our ablation study.
Dataset Construction. To encompass the abundant concepts in the world, we collect images in various styles from multiple sources, including Noah-Wukong [9], LAION [27], and others, such as photography, cartoons, portraits, and gaming assets. The collected images are filtered based on CLIP score, aesthetic score, watermark presence, resolution, and aspect ratio. To improve the semantic alignment of PanGu-Draw, we discard parts of the noisy captions that are meaningless or mismatched to the image, sourced from the Internet. Instead, we recaption the collected images by first employing an open-vocabulary detector [35] to locate the primary subjects within the images. These subjects are then processed by LLaVA [18], a high-performance vision-language model, along with prompting templates, to yield detailed image descriptions. These English annotations are subsequently translated into Chinese.
Evaluation Metrics. We evaluate PanGu-Draw's text-to-image generation on COCO [17] with 30k images for English, and COCO-CN [16] with 10k images for Chinese. The Frechet Inception Distance (FID [10]) is utilized to evaluate image quality and diversity. For Chinese, additional metrics include the Inception Score (IS [30]) and CN-CLIP-score [34], assessing image quality and text-image alignment. Besides, a user study is conducted to evaluate image-text alignment, fidelity, and aesthetics using ImageEval-prompt[1] across 339 prompts.

5.1 Text-to-Image Generation

Evaluation on COCO. As shown in Table 1, PanGu-Draw achieves a FID of 7.99, which is superior to compared methods such as DALL-E 2 and SDXL. It also achieves competitive FID with SOTA methods, indicating the effectiveness

[1] https://github.com/FlagOpen/FlagEval/tree/master/imageEval

Table 1: Comparisons of PanGu-Draw with recent representative English text-to-image generation models on COCO dataset in terms of FID.

Method	FID↓	Model Size	Release
DALL-E [25]	27.50	12B	N
LDM [26]	12.63	1.5B	Y
GLIDE [20]	12.24	5B	N
SDXL [22]	11.93	2.5B	Y
PixArt-α [4]	10.65	0.6B	Y
DALL-E 2 [24]	10.39	5.5B	N
Imagen [29]	7.27	3B	N
RAPHAEL [33]	**6.61**	3B	N
PanGu-Draw	7.99	5B	Y

Table 2: Comparisons of PanGu-Draw with Chinese text-to-image generation models on COCO-CN dataset in terms of FID, IS and CN-CLIP-score. The classifier-free guidance scales are set as 9 following AltDiffusion [37].

Model	FID↓	IS↑	CN-CLIP-score↑
AltDiffusion [37]	25.31	29.16	35.12
Taiyi-Bilingual [38]	24.61	34.29	32.26
Taiyi-CN [38]	23.99	34.29	34.22
PanGu-Draw	**21.81**	**37.00**	**36.62**

Table 3: Results of a User study on ImageVal-prompt in terms of image-text alignment, image fidelity, and aesthetics.

Method	Align↑	Fidelity↑	Aesthetics↑	Ave↑
DALL-E 3 [2]	**4.72**	**4.59**	**4.76**	**4.69**
MJ 5.2	4.63	4.54	4.75	4.64
SDXL [22]	4.41	4.37	4.59	4.46
SD [26]	4.17	3.99	4.20	4.12
PanGu-Draw	4.5	4.52	4.72	4.58

of our time-decoupling training strategy and its outstanding data and training efficiencies. Our 5B PanGu model is the best-released model in terms of FID.

Evaluation on COCO-CN. As shown in Table 2, PanGu-Draw outperforms other released Chinese text-to-image models, including Taiyi-CN, Taiyi-Bilingual, and AltDiffusion, across all three metrics. This performance highlights PanGu-Draw's exceptional Chinese text-to-image generation capabilities and the effectiveness of our bilingual text encoder architecture.

User Study. We conducted a user study to compare PanGu-Draw with top-performing methods, including SDXL [22], Midjourney 5.2, and DALL-E 3 [2]. As shown in Table 3, PanGu-Draw achieves better results than SD and SDXL across all three metrics. It also attains approximately 99%/98% of the performance of

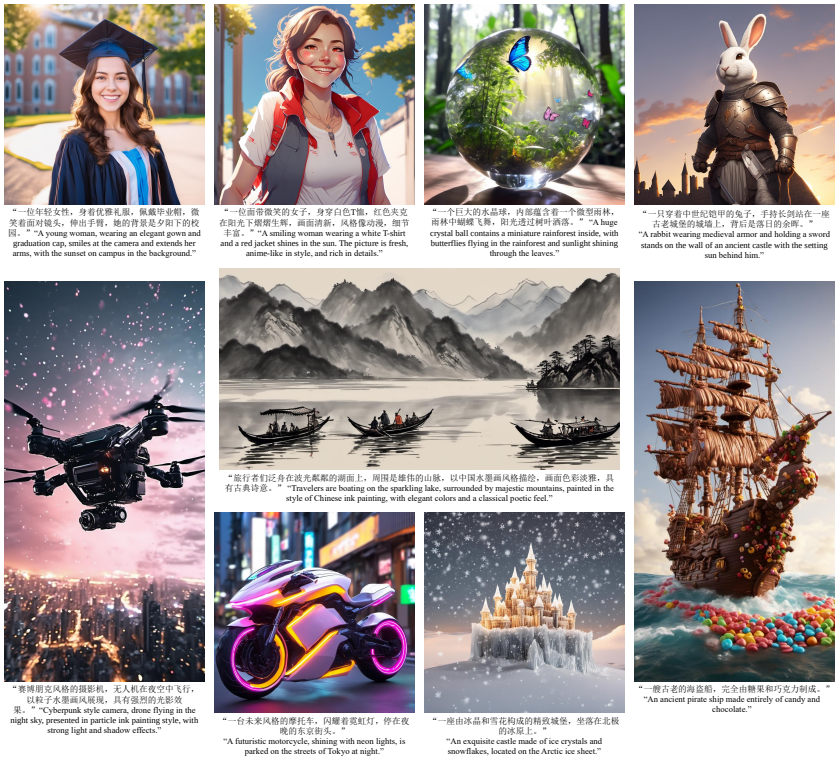


Fig. 4: Images generated with PanGu-Draw, our 5B multi-lingual text-to-image generation model. PanGu-Draw is able to generate multi-resolution high-fidelity images semantically aligned with the input prompts.

Midjourney 5.2 and DALL-E 3, respectively, indicating PanGu-Draw's excellent text-to-image capabilities. Figure 4 shows some high-fidelity multi-resolution images generated by PanGu-Draw. As we can see, the generated images of PanGu-Draw are of high aesthetics and semantically aligned with the input prompts.

5.2 Multi-Diffusion Fusing Results

Multi-Control Image Generation. To demonstrate the effectiveness of the proposed reusable multi-diffusion fusing algorithm, *Coop-Diffusion*, we first present multiple results of multi-control image generation. Figure 5 displays results from fusing an image variation model[2] with PanGu-Draw. The fusing results maintain a style similar to that of the reference image, matching the texture described by the input prompt. Figure 6 shows results from fusing PanGu-Draw with a pose/edge-to-image ControlNet model, which operates in guess mode without input prompts. Here, the fusing results combine the structure of the pose/edge image with the texture described by the input prompt.

[2] https://huggingface.co/lambdalabs/sd-image-variations-diffusers

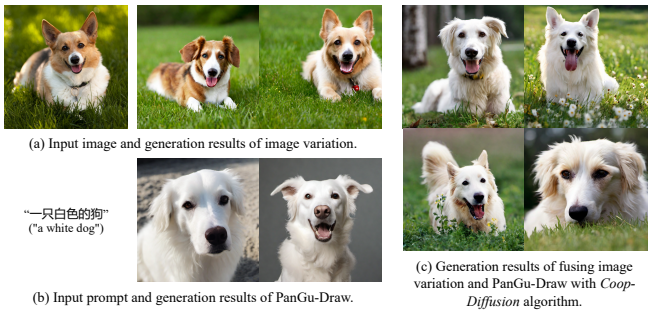


Fig. 5: Generation results of the fusing of an image variation model and PanGu-Draw and with the proposed *Coop-Diffusion* algorithm.

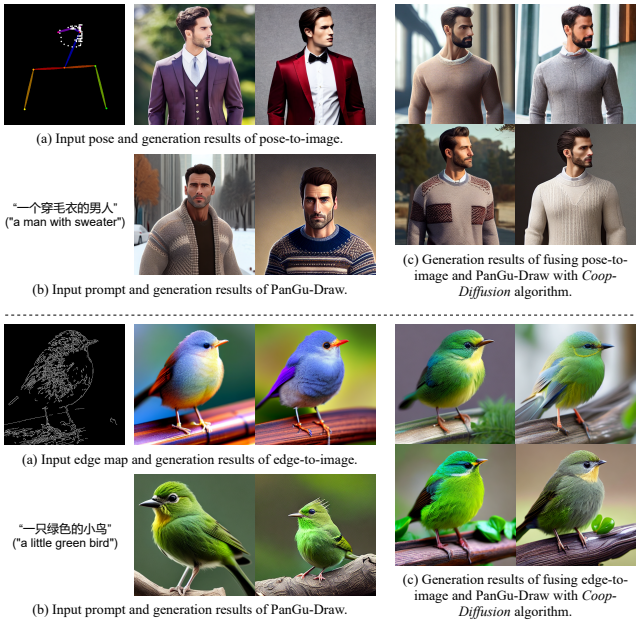


Fig. 6: Generation results guided by fusing signals of text and pose/edge map by our *Coop-Diffusion*.

Multi-Resolution Image Generation. We also present multi-resolution image generation results of fusing PanGu-Draw with low-resolution text-to-image and edge-to-image ControlNet model by first denoising with the low-resolution model to get the intermediate z'_t and the high-resolution counterpart $\tilde{z}_t$, and then perform denoising in high resolution with PanGu-Draw. Figure 7 shows the results from the low-resolution model and our fusing algorithm *Coop-Diffusion*. As we can see, PanGu-Draw adds much details to the low-resolution predictions

Fig. 7: Images generated with a low-resolution (LR) model (1st row: T2I model; 2nd row: edge-to-image ControlNet) and the fusion of the LR model and HR PanGu-Draw with our *Coop-Diffusion*. The resolutions are 512×512 and 1024×1024 respectively. This allows for single-stage super-resolution for better details and higher inference efficiency.

leading to high-fidelity high-resolution results. Besides, compared with the common practice of super-resolution with diffusion model, which carries out all the low-resolution denoising steps, our method achieve higher inference efficiency.

5.3 Ablation Study

In this section, we perform ablation studies to analyze our time-decoupling training strategy. The baseline model has $1B$ parameters while the structure and texture generators both have $0.5B$ parameters. During the training process, the latter two models only train half the steps of the baseline model with T_{struct} set as 500. Both settings of the models are trained from scratch on a subset of the LAION dataset containing images with all sizes. After training, FID, IS and CLIP-score on COCO are reported for comparison.

Time-Decoupling Training Strategy. We compare the performance of models trained with the Resolution Boost strategy and our time-decoupling strategy in Table 4. We found that models trained with our strategy achieves better performance in all three criteria, indicating the effectiveness of our strategy.

Training Designs. The structure and texture generators (ϵ_{struct} and $\epsilon_{\text{texture}}$) are designed to train on different resolutions to improve data and training efficiency. However, this approach may negatively influence the final performance. In Table 5, we compare such a design with a traditional training process, where ϵ_{struct} discards low-resolution images, or $\epsilon_{\text{texture}}$ trains with high resolution. Results on COCO show that ϵ_{struct} benefits from these extra up-scaled data, and $\epsilon_{\text{texture}}$ learns enough texture patterns at a smaller resolution.

Table 4: Comparison of models across Resolution Boost (1B parameters) and Time-Decoupling training strategies (0.5B parameters for structure and texture generators)

Model	FID↓	IS↑	CLIP-score↑
Resolution Boost	106.12	10.46	22.9
Time-Decoupling	**87.66**	**11.07**	**23.4**

Table 5: Performance of structure and texture models training with images of different resolutions.

Structure Data	Texture Resolution	FID↓	IS↑	CLIP-score↑
All data	256	**87.66**	**11.07**	**23.4**
Only high resolution	256	89.52	10.96	23.2
All data	512	90.98	10.59	23.3

Table 6: Comparisons of PanGu-Draw inference performance with different time step splitting point T_{struct} settings.

T_{struct}	FID↓	IS↑	CLIP-score↑
200	105.08	10.59	22.98
300	98.08	10.72	23.12
500	**87.66**	**11.07**	**23.40**
700	89.48	11.02	23.32

Timestep Splitting Point. The timestep splitting point T_{struct} between the structure and texture generators also influences the final performance. To this end, we set T_{struct} to 200, 300, 500, and 700, while keeping the other settings of the structure and texture generators unchanged. As shown in Table 6, as T_{struct} increases from 200 to 700, the performance initially increases and then decreases continuously. $T_{\text{struct}} = 500$ is the optimal value, and we adopt it as the default setting in all other experiments.

6 Conclusion

In this paper, we present "PanGu-Draw", a new latent diffusion model for efficient text-to-image generation that effectively integrates multiple control signals. Our approach includes a Time-Decoupling Training Strategy to separate the text-to-image process into structure and texture generation, enhancing data use and computational efficiency. Additionally, "Coop-Diffusion" is introduced, an algorithm allowing cooperative use of different pre-trained diffusion models in a unified denoising process for multi-control image synthesis at various resolutions without extra data or retraining. PanGu-Draw outperforms models like DALL-E 2 and SDXL in English T2I, achieves superior FID, IS, and CN-CLIP-scores in Chinese T2I, and receives favorable user feedback. This positions PanGu-Draw as a versatile and efficient state-of-the-art method.

References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf (2023)
3. Chen, C., Yin, Y., Shang, L., Jiang, X., Qin, Y., Wang, F., Wang, Z., Chen, X., Liu, Z., Liu, Q.: bert2bert: Towards reusable pretrained language models. arXiv preprint arXiv:2110.07143 (2021)
4. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
6. Ding, N., Tang, Y., Han, K., Xu, C., Wang, Y.: Network expansion for practical training acceleration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20269–20279 (2023)
7. Feng, Z., Zhang, Z., Yu, X., Fang, Y., Li, L., Chen, X., Lu, Y., Liu, J., Yin, W., Feng, S., et al.: Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10135–10145 (2023)
8. Fu, C., Huang, H., Jiang, Z., Ni, Y., Nai, L., Wu, G., Cheng, L., Zhou, Y., Li, S., Li, A., et al.: Triple: Revisiting pretrained model reuse and progressive learning for efficient vision transformer scaling and searching. In: ICCV (2023)
9. Gu, J., Meng, X., Lu, G., Hou, L., Minzhe, N., Liang, X., Yao, L., Huang, R., Zhang, W., Jiang, X., et al.: Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. Advances in Neural Information Processing Systems **35**, 26418–26431 (2022)
10. 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. Advances in neural information processing systems **30** (2017)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
13. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**, 47–1 (2022)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free diffusion model adaptation for variable-sized text-to-image synthesis. arXiv preprint arXiv:2306.08645 (2023)
16. Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., Xu, J.: Coco-cn for cross-lingual image tagging, captioning, and retrieval. IEEE Transactions on Multimedia **21**(9), 2347–2360 (2019)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)

18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
19. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
20. 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. arXiv preprint arXiv:2112.10741 (2021)
21. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171. PMLR (2021)
22. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
23. Qin, Y., Zhang, J., Lin, Y., Liu, Z., Li, P., Sun, M., Zhou, J.: Elle: Efficient lifelong pre-training for emerging data. arXiv preprint arXiv:2203.06311 (2022)
24. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
25. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
29. 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. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
30. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29**, 2234–2242 (2016)
31. Shonenkov, A., Konstantinov, M., Bakshandaeva, D., Schuhmann, C., Ivanova, K., Klokova, N.: Deepfloyd if: A powerful text-to-image model that can smartly integrate text into images (2023), `https://www.deepfloyd.ai/deepfloyd-if`, online; accessed 16-November-2023
32. Xu, X., Wang, Z., Zhang, E., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. arXiv preprint arXiv:2211.08332 (2022)
33. 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. arXiv preprint arXiv:2305.18295 (2023)
34. Yang, A., Pan, J., Lin, J., Men, R., Zhang, Y., Zhou, J., Zhou, C.: Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335 (2022)

35. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23497–23506 (2023)
36. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
37. Ye, F., Liu, G., Wu, X., Wu, L.: Altdiffusion: A multilingual text-to-image diffusion model (2023)
38. Zhang, J., Gan, R., Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R., Wang, H., Wu, X., Zeng, Z., Chen, C.: Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. CoRR **abs/2209.02970** (2022)
39. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)