

# CogView3: Finer and Faster Text-to-Image Generation via Relay Diffusion

Wendi Zheng<sup>1\*†</sup>, Jiayan Teng<sup>1\*‡</sup>, Zhuoyi Yang<sup>1‡</sup>, Weihan Wang<sup>1‡</sup>,  
Jidong Chen<sup>1‡</sup>, Xiaotao Gu<sup>2</sup>, Yuxiao Dong<sup>1†</sup>, Ming Ding<sup>2†</sup>, and Jie Tang<sup>1†</sup>

<sup>1</sup> Tsinghua University

{zhengwd23@mails., tengjy20@mails., yuxiaod@, jietang@mail.}tsinghua.edu.cn

<sup>2</sup> Zhipu AI

mingding.thu@gmail.com

**Abstract.** Recent advancements in text-to-image generative systems have been largely driven by diffusion models. However, single-stage text-to-image diffusion models still face challenges, in terms of computational efficiency and the refinement of image details. To tackle the issue, we propose CogView3, an innovative cascaded framework that enhances the performance of text-to-image diffusion. CogView3 is the first model implementing relay diffusion in the realm of text-to-image generation, executing the task by first creating low-resolution images and subsequently applying relay-based super-resolution. This methodology not only results in competitive text-to-image outputs but also greatly reduces both training and inference costs. Our experimental results demonstrate that CogView3 outperforms SDXL, the current state-of-the-art open-source text-to-image diffusion model, by 77.0% in human evaluations, all while requiring only about 1/2 of the inference time. The distilled variant of CogView3 achieves comparable performance while only utilizing 1/10 of the inference time by SDXL.

**Keywords:** Text-to-Image Generation · Diffusion Models

## 1 Introduction

Diffusion models have emerged as the mainstream framework in today’s text-to-image generation systems [3, 5, 17, 19, 21]. In contrast to the paradigm of auto-regressive models [6, 20, 31] and generative adversarial networks [12], the diffusion models conceptualize the task of image synthesis as a multi-step denoising process that starts from an isotropic Gaussian noise [8]. With the surge in the volume of training data and computation cost of neural networks, the framework of diffusion models has achieved effectiveness in the realm of visual generation, able to follow user instructions and generate images with commendable details.

Current state-of-the-art text-to-image diffusion models mostly operate in a single stage, conducting the diffusion process at high image resolutions, such as

---

\*equal contribution

†corresponding authors

‡work was done when interned in Zhipu AI.



**Fig. 1:** Showcases of CogView3 generation of resolution  $2048 \times 2048$  (**top**) and  $1024 \times 1024$  (**bottom**). All prompts are sampled from Partiprompts [31].

$1024 \times 1024$  [3, 5, 17]. The direct modeling on high resolution images aggravates the inference costs since every denoising step is performed on the high resolution space. To address such an issue, Luo *et al.* [14] and Sauer *et al.* [23] propose to distill diffusion models to significantly reduce the number of sampling steps. However, the generation quality tends to degrade noticeably during diffusion distillation, unless a GAN loss is introduced, which otherwise complicates the distillation and could lead to instability of training.

In this work, we propose CogView3, a novel text-to-image generation system that employs relay diffusion [27]. Relay diffusion is a new cascaded diffusion framework, decomposing the process of generating high-resolution images into multiple stages. It first generates low-resolution images and subsequently performs relaying super-resolution generation. Unlike previous cascaded diffusion frameworks that condition every step of the super-resolution stage on low-resolution generations [9, 19, 21], relaying super-resolution adds Gaussian noise to the low-resolution generations and starts diffusion from these noised images. This enables the super-resolution stage of relay diffusion to rectify unsatisfactory artifacts produced by the previous diffusion stage. In CogView3, we apply relay diffusion in the latent image space rather than at pixel level as the original version, by utilizing a simplified linear blurring schedule and a correspondingly formulated sampler. By the iterative implementation of the super-resolution stage, CogView3 is able to generate images with extremely high resolutions such as  $2048 \times 2048$ .

Given that the cost of lower-resolution inference is quadratically smaller than that of higher-resolution, CogView3 can produce competitive generation results at significantly reduced inference costs by properly allocating sampling steps between the base and super-resolution stages. Our results of human evaluation show that CogView3 outperforms SDXL [17] with a win rate of 77.0%. Moreover,

through progressive distillation of diffusion models, CogView3 is able to produce comparable results while utilizing only 1/10 of the time required for the inference of SDXL. Our contributions can be summarized as follows:

- We propose CogView3, the first text-to-image system in the framework of relay diffusion. CogView3 is able to generate high quality images with extremely high resolutions such as  $2048 \times 2048$ .
- Based on the relaying framework, CogView3 is able to produce competitive results at a significantly reduced time cost. CogView3 achieves a win rate of 77.0% over SDXL with about 1/2 of the time during inference.
- We further explore the progressive distillation of CogView3, which is significantly facilitated by the relaying design. The distilled variant of CogView3 delivers comparable generation results while utilizing only 1/10 of the time required by SDXL.

## 2 Background

### 2.1 Text-to-Image Diffusion Models

Diffusion models, as defined by Ho *et al.* [8], establish a forward diffusion process that gradually adds Gaussian noise to corrupt real data  $\mathbf{x}_0$  as follows:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad t \in \{1, \dots, T\}, \quad (1)$$

where  $\beta_t$  defines a noise schedule in control of diffusion progression. Conversely, the backward process generates images from pure Gaussian noise by step-by-step denoising, adhering a Markov chain.

A neural network is trained at each time step to predict denoised results based on the current noised images. For text-to-image diffusion models, an additional text encoder encodes the text input, which is subsequently fed into the cross attention modules of the main network. The training process is implemented by optimizing the variational lower bound of the backward process, which is written as

$$\mathbb{E}_{\mathbf{x}_0 \sim p_{data}} \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} \|\mathcal{D}(\mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, t, c) - \mathbf{x}_0\|^2, \quad (2)$$

where  $\sigma_t$  denotes the noise scale controlled by the noise schedule.  $c$  denotes input conditions including the text embeddings.

Recent works [3, 17] consistently apply diffusion models to the latent space, resulting in a substantial saving of both training and inference costs. They first use a pretrained autoencoder to compress the image  $\mathbf{x}$  into a latent representation  $\mathbf{z}$  with lower dimension, which is approximately recoverable by its decoder. The diffusion model learns to generate latent representations of images.

### 2.2 Relay Diffusion Models

Cascaded diffusion [9, 21] refers to a multi-stage diffusion generation framework. It first generates low-resolution images using standard diffusion and subsequently

performs super-resolution. The super-resolution stage of the original cascaded diffusion conditions on low-resolution samples  $\mathbf{x}^L$  at every diffusion step, by channel-wise concatenation of  $\mathbf{x}^L$  with noised diffusion states. Such conditioning necessitates augmentation techniques to bridge the gap in low-resolution input between real images and base stage generations.

As a new variant of cascaded diffusion, the super-resolution stage of relay diffusion [27] instead starts diffusion from low-resolution images  $\mathbf{x}^L$  corrupted by Gaussian noise  $\sigma_{T_r}\epsilon$ , where  $T_r$  denotes the starting point of the blurring schedule in the super-resolution stage. The forward process is formulated as:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | F(\mathbf{x}_0, t), \sigma_t^2 \mathbf{I}), \quad t \in \{0, \dots, T\}, \quad (3)$$

where  $F(\cdot)$  is a pre-defined transition along time  $t$  from high-resolution images  $\mathbf{x} = \mathbf{x}_0$  to the upsampled low-resolution images  $\mathbf{x}^L$ . The endpoint of  $F$  is set as  $F(\mathbf{x}_0, T_r) = \mathbf{x}^L$  to ensure a seamless transition. Conversely, the backward process of relaying super-resolution is a combination of denoising and deblurring.

This design allows relay diffusion to circumvent the need for intricate augmentation techniques on lower-resolution conditions  $\mathbf{x}^L$ , as  $\mathbf{x}^L$  is only inputted at the initial sampling step of super-resolution stage and is already corrupted by Gaussian noise  $\sigma_{T_r}\epsilon$ . It also enables the super-resolution stage of relay diffusion to possibly rectify some unsatisfactory artifacts produced by the previous diffusion stage.

### 2.3 Diffusion Distillation

Knowledge distillation [7] is a training process aiming to transfer a larger teacher model to the smaller student model. In the context of diffusion models, distillation has been explored as a means to reduce sampling steps thus saving computation costs of inference, while preventing significant degradation of the generation performance [14, 22, 23, 26].

As one of the prominent paradigms in diffusion distillation, progressive distillation [22] trains the student model to match every two steps of the teacher model with a single step in each training stage. This process is repeated, progressively halving the sampling steps. On the other hand, consistency models [14, 26] propose a fine-tuning approach for existing diffusion models to project every diffusion step to the latest one to ensure step-wise consistency, which also reduces sampling steps of the model. While previous diffusion distillation methods mostly compromise on the quality of generation, adversarial diffusion distillation [23] mitigates this by incorporating an additional GAN loss in the distillation. However, this makes the process of distillation more challenging due to the instability of GAN training.

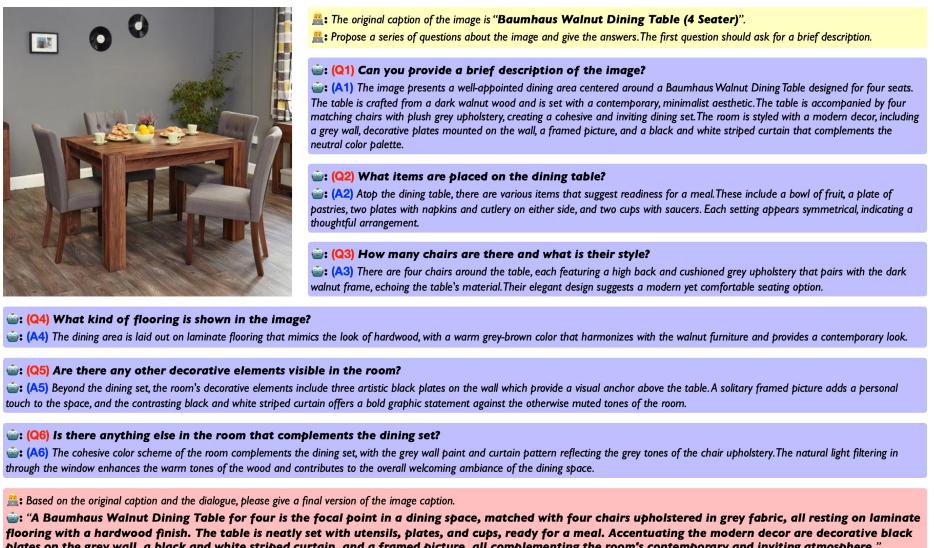
## 3 Method

### 3.1 Text Preprocessing

**Image Recaption** Following DALL-E-3 [3], we develop an automatic pipeline to re-caption images from the training dataset. While DALL-E-3 derives instruction-tuning data of the re-caption model from human labelers, we extract triplets

of `<image, old_cap, new_cap>` by automatically prompting GPT-4V [1], as shown in Figure 2. Generally, we prompt GPT-4V to propose several questions about the content of the uploaded image. The first question is forced to be about a brief description. Finally, we instruct the model to combine the answers together with the original caption to build a new caption.

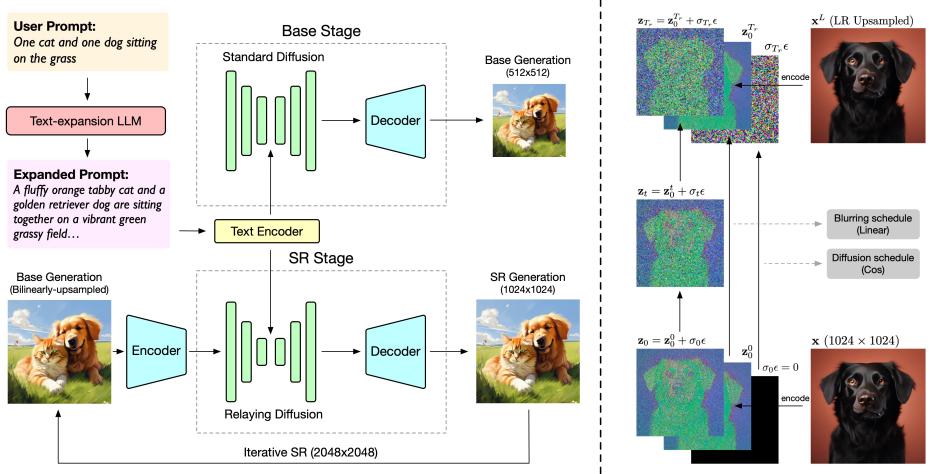
We collect approximately 70,000 recaption triplets with this paradigm and finetune CogVLM-17B [28] by these examples to obtain a recaption model. We finetune the model by a moderate degree, with batch size 256 and 1,500 steps to prevent model from severe overfitting. Eventually the model is utilized to re-caption the whole training dataset. The re-caption results provide comprehensive, graceful and detailed descriptions of images, in contrast to the original short and less relevant captions from the dataset. The prefix statement we use to prompt GPT-4V and the template we use in fine-tuning the recaption model are both provided in Appendix B.



**Fig. 2:** An example of re-caption data collection from GPT-4V.

**Prompt Expansion** On account that CogView3 is trained on datasets with comprehensive re-captions while users of text-to-image generation systems may tend to provide brief prompts lacking descriptive information, this introduces an explicit misalignment between model training and inference [3]. Therefore, we also explore to expand user prompts before sampling with the diffusion models. We prompt language models to expand user prompts into comprehensive descriptions, while encouraging the model generation to preserve the original intention from users. With human evaluation, we find results of the expanded prompts to achieve higher preference. We provide the template and showcases of our prompt expansion in Appendix B.

### 3.2 Model Formulation



**Fig. 3:** (left) The pipeline of CogView3. User prompts are rewritten by a text-expansion language model. The base stage model generates  $512 \times 512$  images, and the second stage subsequently performs relaying super-resolution. (right) Formulation of relaying super-resolution in the latent space.

**Model Framework** The backbone of CogView3 is a 3-billion parameter text-to-image diffusion model with a 3-stage UNet architecture. The model operates in the latent image space, which is  $8\times$  compressed from the pixel space by a variational KL-regularized autoencoder. We employ the pretrained T5-XXL [18] encoder as the text encoder to improve model’s capacity for text understanding and instruction following, which is frozen during the training of the diffusion model. To ensure alignment between training and inference, user prompts are first rewritten by language models as mentioned in the previous section. We set the input token length for the text encoder as 225 to facilitate the implementation of the expanded prompts.

As shown in Figure 3(left), CogView3 is implemented as a 2-stage relay diffusion. The base stage of CogView3 is a diffusion model that generates images at a resolution of  $512 \times 512$ . The second stage model performs  $2\times$  super-resolution, generating  $1024 \times 1024$  images from  $512 \times 512$  inputs. It is noteworthy that the super-resolution stage can be directly transferred to higher resolutions and iteratively applied, enabling the final outputs to reach higher resolutions such as  $2048 \times 2048$ , as cases illustrated from the top line of Figure 1.

**Training Pipeline** We use Laion-2B [24] as our basic source of the training dataset, after removing images with politically-sensitive, pornographic or violent contents to ensure appropriateness and quality of the training data. The filtering

process is executed by a pre-defined list of sub-strings to block a group of source links associated with unwanted images. In correspondence with Betker *et al.* [3], we replace 95% of the original data captions with the newly-produced captions.

Similar to the training approach used in SDXL [17], we train Cogview3 progressively to develop multiple stages of models. This greatly reduced the overall training cost. Owing to such a training setting, the different stages of CogView3 share a same model architecture.

The base stage of CogView3 is trained on the image resolution of  $256 \times 256$  for 600,000 steps with batch size 2048 and continued to be trained on  $512 \times 512$  for 200,000 steps with batch size 2048. We finetune the pretrained  $512 \times 512$  model on a highly aesthetic internal dataset for 10,000 steps with batch size 1024, to achieve the released version of the base stage model. To train the super-resolution stage of CogView3, we train on the basis of the pretrained  $512 \times 512$  model on  $1024 \times 1024$  resolution for 100,000 steps with batch size 1024, followed by a 20,000 steps of finetuning with the loss objective of relaying super-resolution to achieve the final version.

### 3.3 Relaying Super-resolution

**Latent Relay Diffusion** The second stage of CogView3 performs super-resolution by relaying, starting diffusion from the results of base stage generation. While the original relay diffusion handles the task of image generation in the pixel level [27], we implement relay diffusion in the latent space and utilize a simple linear transformation instead of the original patch-wise blurring. The formulation of latent relay diffusion is illustrated by Figure 3(right). Given an image  $\mathbf{x}_0$  and its low-resolution version  $\mathbf{x}^L = \text{Downsample}(\mathbf{x}_0)$ , they are first transformed into latent space by the autoencoder as  $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ ,  $\mathbf{z}^L = \mathcal{E}(\mathbf{x}^L)$ . Then the linear blurring transformation is defined as:

$$\mathbf{z}_0^t = \mathcal{F}(\mathbf{z}_0, t) = \frac{T_r - t}{T_r} \mathbf{z}_0 + \frac{t}{T_r} \mathbf{z}^L, \quad (4)$$

where  $T_r$  denotes the starting point set for relaying super-resolution and  $\mathbf{z}_0^{T_r}$  matches exactly with  $\mathbf{z}^L$ . The forward process of the latent relay diffusion is then written as:

$$q(\mathbf{z}_t | \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t | \mathbf{z}_0^t, \sigma_t^2 \mathbf{I}), \quad t \in \{1, \dots, T_r\}. \quad (5)$$

The training objective is accordingly formulated as:

$$\mathbb{E}_{\mathbf{x}_0 \sim p_{data}} \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \in \{0, \dots, T_r\}} \|\mathcal{D}(\mathbf{z}_0^t + \sigma_t \epsilon, t, c_{text}) - \mathbf{z}_0\|^2, \quad (6)$$

where  $\mathcal{D}$  denotes the UNet denoiser function and  $c_{text}$  denotes the input text condition.

**Sampler Formulation** Next we introduce the sampler designed for the relaying super-resolution. Given samples  $X^L$  generated in the base stage, we bilinearly upsample  $X^L$  into  $\mathbf{x}^L$ . The starting point of relay diffusion is defined as  $\mathbf{z}_{T_r} =$

$\mathbf{z}_0^{T_r} + \sigma_{T_r}\epsilon$ , where  $\epsilon$  denotes a unit isotropic Gaussian noise and  $\mathbf{z}_0^{T_r} = \mathcal{E}(\mathbf{x}^L)$  is the latent representation of the bilinearly-upsampled base-stage generation. Corresponding to the forward process of relaying super-resolution formulated in Equation 5, the backward process is defined in the DDIM [25] paradigm:

$$q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_{t-1}|a_t\mathbf{z}_t + b_t\mathbf{z}_0 + c_t\mathbf{z}_0^t, \delta_t^2\mathbf{I}), \quad (7)$$

where  $a_t = \sqrt{\sigma_{t-1}^2 - \delta_t^2}/\sigma_t$ ,  $b_t = 1/t$ ,  $c_t = (t-1)/t - a_t$ ,  $\mathbf{z}_0^t$  is defined in Equation 4 and  $\delta_t$  represents the random degree of the sampler. In practice, we simply set  $\delta_t$  as 0 to be an ODE sampler. The procedure is shown in Algorithm 1. A detailed proof of the consistency between the sampler and the formulation of latent relay diffusion is shown in Appendix A.

---

**Algorithm 1** latent relay sampler

---

```

Given  $\mathbf{x}^L, \mathbf{z}_0^{T_r} = \mathcal{E}(\mathbf{x}^L)$ 
 $\mathbf{z}_{T_r} = \mathbf{z}_0^{T_r} + \sigma_{T_r}\epsilon$             $\triangleright$  transform into the latent space and add noise for relaying
for  $t \in \{T_r, \dots, 1\}$  do
     $\tilde{\mathbf{z}}_0 = \mathcal{D}(\mathbf{z}_t, t, c_{text})$             $\triangleright$  predict  $\mathbf{z}_0$ 
     $\mathbf{z}_0^{t-1} = \mathbf{z}_0^t + (\tilde{\mathbf{z}}_0 - \mathbf{z}_0^t)/t$             $\triangleright$  linear blurring transition
     $a_t = \sigma_{t-1}/\sigma_t, b_t = 1/t, c_t = (t-1)/t - a_t$             $\triangleright$  coefficient of each item
     $\mathbf{z}_{t-1} = a_t\mathbf{z}_t + b_t\tilde{\mathbf{z}}_0 + c_t\mathbf{z}_0^t$             $\triangleright$  single sampling step
end for
 $\mathbf{x}_0 = \text{Decode}(\mathbf{z}_0)$ 

```

---

### 3.4 Distillation of Relay Diffusion

We combine the method of progressive distillation [15] and the framework of relay diffusion to achieve the distilled version of CogView3. While the base stage of CogView3 performs standard diffusion, the distillation procedure follows the original implementation.

For the super-resolution stage, we merge the blurring schedule into the diffusion distillation training, progressively halving sampling steps by matching two steps from the latent relaying sampler of the teacher model with one step of the student model. The teacher steps are formulated as:

$$\begin{aligned} \mathbf{z}_{t-1} &= a_t\mathbf{z}_t + b_t\tilde{\mathbf{z}}_0(\mathbf{z}_t, t)_{teacher} + c_t\mathbf{z}_0^t, \\ \mathbf{z}_{t-2} &= a_{t-1}\mathbf{z}_{t-1} + b_{t-1}\tilde{\mathbf{z}}_0(\mathbf{z}_{t-1}, t-1)_{teacher} + c_{t-1}\mathbf{z}_0^{t-1}, \end{aligned} \quad (8)$$

where  $(a_k, b_k, c_k)$ ,  $k \in \{0, \dots, T_r\}$  refers to the item coefficients defined in Algorithm 1. One step of the student model is defined as:

$$\tilde{\mathbf{z}}_{t-2} = \frac{\sigma_{t-2}}{\sigma_t}\mathbf{z}_t + \frac{\tilde{\mathbf{z}}_0(\mathbf{z}_t, t)_{student}}{t} + \left(\frac{t-2}{t} - \frac{\sigma_{t-2}}{\sigma_t}\right)\mathbf{z}_0^t. \quad (9)$$

The training objective is defined as the mean square error between  $\tilde{\mathbf{z}}_{t-2}$  and  $\mathbf{z}_{t-2}$ . Following Meng *et al.* [15], we incorporate the property of the classifier-free guidance (CFG) [10] strength  $w$  into the diffusion model in the meantime of

distillation by adding learnable projection embeddings of  $w$  into timestep embeddings. Instead of using an independent stage for the adaptation, we implement the incorporation at the first round of the distillation and directly condition on  $w$  at subsequent rounds.

The inference costs of the low-resolution base stage are quadratically lower than the high-resolution counterparts, while it ought to be called from a complete diffusion schedule. On the other hand, the super-resolution stage starts diffusion at an intermediate point of the diffusion schedule. This greatly eases the task and reduces the potential error that could be made by diffusion distillation. Therefore, we are able to distribute final sampling steps for relaying distillation as 8 steps for the base stage and 2 steps for the super-resolution stage, or even reduce to 4 steps and 1 step respectively, which achieves both greatly-reduced inference costs and mostly-retained generation quality.

## 4 Experiments

### 4.1 Experimental Setting

We implement a comprehensive evaluation process to demonstrate the performance of CogView3. With an overall diffusion schedule of 1000 time steps, we set the starting point of the relaying super-resolution at 500, a decision informed by a brief ablation study detailed in Section 4.4. To generate images for comparison, we sample 50 steps by the base stage of CogView3 and 10 steps by the super-resolution stage, both utilizing a classifier-free guidance [10] of 7.5, unless specified otherwise. The comparison is all conducted at the image resolution of  $1024 \times 1024$ .

**Dataset** We choose a combination of image-text pair datasets and collections of prompts for comparative analysis. Among these, MS-COCO [13] is a widely applied dataset for evaluating the quality of text-to-image generation. We randomly pick a subset of 5000 image-text pairs from MS-COCO, named as COCO-5k. We also incorporate DrawBench [21] and PartiPrompts [31], two well-known sets of prompts for text-to-image evaluation. DrawBench comprises 200 challenging prompts that assess both the quality of generated samples and the alignment between images and text. In contrast, PartiPrompts contains 1632 text prompts and provides a comprehensive evaluation critique.

**Baselines** In our evaluation, we employ state-of-the-art open-source text-to-image models, specifically SDXL [17] and Stable Cascade [16] as our baselines. SDXL is a single-stage latent diffusion model capable of generating images at and near a resolution of  $1024 \times 1024$ . On the other hand, Stable Cascade implements a cascaded pipeline, generating  $16 \times 24 \times 24$  priors at first and subsequently conditioning on the priors to produce images at a resolution of  $1024 \times 1024$ . We sample SDXL for 50 steps and Stable Cascade for 20 and 10 steps respectively for its two stages. In all instances, we adhere to their recommended configurations of the classifier-free guidance.

**Evaluation Metrics** We use Aesthetic Score (Aes) [24] to evaluate the image quality of generated samples. We also adopt Human Preference Score v2 (HPS v2) [29] and ImageReward [30] to evaluate text-image alignment and human preference. Aes is obtained by an aesthetic score predictor trained from LAION datasets, neglecting alignment of prompts and images. HPS v2 and ImageReward are both used to predict human preference for images, including evaluation of text-image alignment, human aesthetic, etc. Besides machine evaluation, we also conduct human evaluation to further assess the performance of models, covering image quality and semantic accuracy.

## 4.2 Results of Machine Evaluation

Table 1 shows results of machine metrics on DrawBench and Partiprompts. While CogView3 has the lowest inference cost, it outperforms SDXL and Stable Cascade in most of the comparisons except for a slight setback to Stable Cascade on the ImageReward of PartiPrompts. Similar results are observed from comparisons on COCO-5k, as shown in Table 2. The distilled version of CogView3 takes an extremely low inference time of 1.47s but still achieves a comparable performance. The results of the distilled variant of CogView3 significantly outperform the previous distillation paradigm of latent consistency model [14] on SDXL, as illustrated in the table.

Model	Steps	Time Cost	DrawBench			PartiPrompts		
			Aes↑	HPS v2↑	ImageReward↑	Aes↑	HPS v2↑	ImageReward↑
SDXL [17]	50	19.67s	5.54	<u>0.288</u>	0.676	5.78	0.287	0.915
StableCascade [16]	20+10	10.83s	5.88	0.285	0.677	5.93	0.285	<b>1.029</b>
<b>CogView3</b>	50+10	<b>10.33s</b>	<b>5.97</b>	<b>0.290</b>	<b>0.847</b>	<b>6.15</b>	<b>0.290</b>	<b>1.025</b>
LCM-SDXL [14]	4	2.06s	5.45	0.279	0.394	5.59	0.280	0.689
<b>CogView3-distill</b>	4+1	<b>1.47s</b>	5.87	<u>0.288</u>	<u>0.731</u>	6.12	0.287	0.968
<b>CogView3-distill</b>	8+2	1.96s	<u>5.90</u>	0.285	0.655	<u>6.13</u>	<u>0.288</u>	0.963

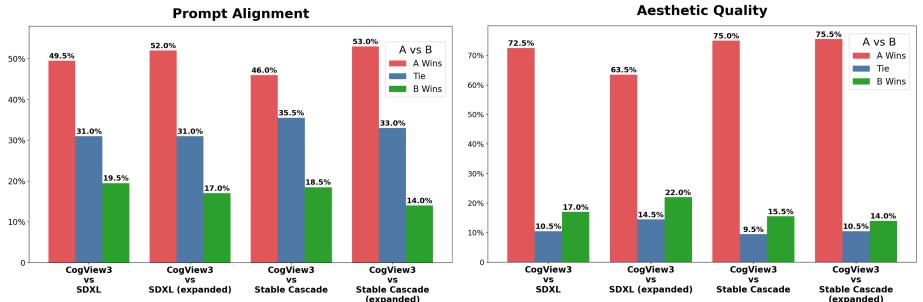
**Table 1:** Results of machine metrics on DrawBench and PartiPrompts. All samples are generated on  $1024 \times 1024$ . The time cost is measured with a batch size of 4.

COCO-5k						
Model	Steps	Time Cost	FID↓	Aes↑	HPS v2↑	ImageReward↑
SDXL [17]	50	19.67s	<b>26.29</b>	5.63	0.291	0.820
StableCascade [16]	20+10	10.83s	36.59	5.89	0.283	0.734
<b>CogView3</b>	50+10	<b>10.33s</b>	31.63	<b>6.01</b>	<b>0.294</b>	<b>0.967</b>
LCM-SDXL [14]	4	2.06s	<u>27.16</u>	5.39	0.281	0.566
<b>CogView3-distill</b>	4+1	<b>1.47s</b>	34.03	5.99	0.292	0.920
<b>CogView3-distill</b>	8+2	1.96s	35.53	<u>6.00</u>	<u>0.293</u>	<u>0.921</u>

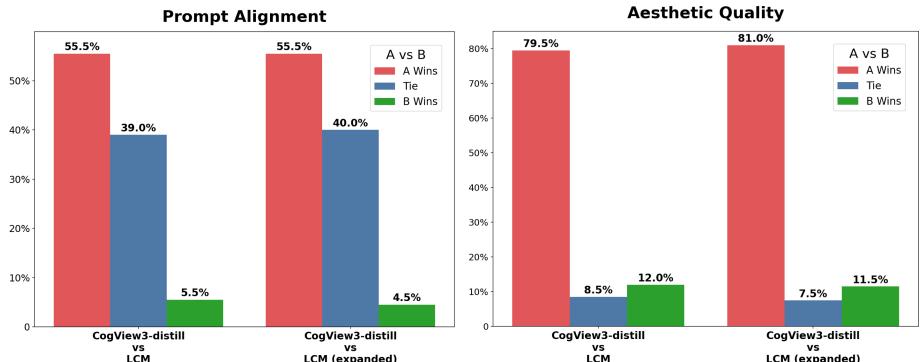
**Table 2:** Results of machine metrics on COCO-5k. All samples are generated on  $1024 \times 1024$ . The time cost is measured with a batch size of 4.

The comparison results demonstrate the performance of CogView3 for generating images of improved quality and fidelity with a remarkably reduced cost. The distillation of CogView3 succeeds in preserving most of the generation quality while reduces the sampling time to an extreme extent. We largely attribute the aforementioned comparison results to the relaying property of CogView3. In the following section, we will further demonstrate the performance of CogView3 with human evaluation.

### 4.3 Results of Human Evaluation



**Fig. 4:** Results of human evaluation on DrawBench generation. (**left**) Comparison results about prompt alignment, (**right**) comparison results about aesthetic quality. “(expanded)” indicates that prompts used for generation is text-expanded.



**Fig. 5:** Results of human evaluation on Drawbench generation for distilled models. (**left**) Comparison results about prompt alignment, (**right**) comparison results about aesthetic quality. “(expanded)” indicates that prompts used for generation is text-expanded. We sample 8+2 steps for CogView3-distill and 4 steps for LCM-SDXL.

We conduct human evaluation for CogView3 by having annotators perform pairwise comparisons. The human annotators are asked to provide results of win, lose or tie based on the prompt alignment and aesthetic quality of the generation. We use DrawBench [21] as the evaluation benchmark. For the generation of CogView3, we first expand the prompts from DrawBench to detailed descriptions

as explained in Section 3.1, using the expanded prompts as the input of models. For a comprehensive evaluation, we compare CogView3 generation with SDXL and Stable Cascade by both the original prompts and the expanded prompts.

As shown in Figure 4, CogView3 significantly outperforms SDXL and Stable Cascade in terms of both prompt alignment and aesthetic quality, achieving average win rates of 77.0% and 78.1% respectively. Similar results are observed on comparison with SDXL and Stable Cascade generation by the expanded prompts, where CogView3 achieves average win rates of 74.8% and 82.1% respectively.

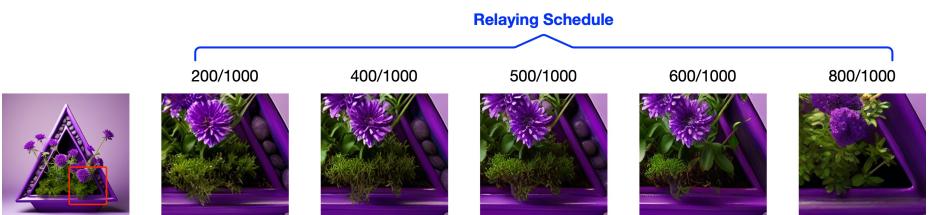
To evaluate the distillation, we compare the distilled CogView3 with SDXL distilled in the framework of latent consistency model [14]. As shown in Figure 5, the performance of the distilled CogView3 significantly surpasses that of LCM-distilled SDXL, which is consistent with the results from Section 4.2.

#### 4.4 Additional Ablations

**Starting Points for Relaying Super-resolution** We ablate the selection of starting point for relaying super-resolution as shown in Table 3, finding that a midway point achieves the best results. The comparison is also illustrated with a qualitative case in Figure 6. An early starting point tends to produce blurring contents, as shown by the flower and grass in case of 200/1000, while in contrast, a late starting point introduces artifacts, as shown by the flower and edge in case of 800/1000, suggesting a midway point to be the best choice. Based on the results of comparison, we choose 500 as our finalized starting point.

Starting Point	200/1000	400/1000	500/1000	600/1000	800/1000
HPS v2 ↑	0.288	0.289	<b>0.290</b>	0.289	0.286
ImageReward ↑	0.829	0.835	<b>0.847</b>	0.836	0.812

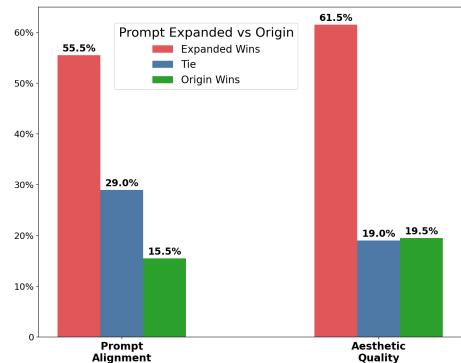
**Table 3:** Ablation of starting points on DrawBench.



**Fig. 6:** Comparison of results from super-resolution stages with different relaying starting points. Sampling steps are all set  $\sim 10$  by controlling the number of steps from the complete diffusion schedule.

**Alignment Improvement with Text Expansion** While prompt expansion hardly brings an improvement for the generation of SDXL and Stable Cascade, we highlight its significance for the performance of CogView3. Figure 7 shows the results of comparison with and without prompt expansion, explicitly demonstrating that prompt expansion significantly enhances the ability of prompt instruction following for CogView3.

Figure 8 shows qualitative comparison between before and after the prompt expansion. The expanded prompts provide more comprehensive and in-distribution descriptions for model generation, largely improving the accuracy of instruction following for CogView3. Similar improvement is not observed on the generation of SDXL. The probable reason may be that SDXL is trained on original captions and only has an input window of 77 tokens, which leads to frequent truncation of the expanded prompts. This corroborates the statement in Section 3.1 that prompt expansion helps bridge the gap between model inference and training with re-captioned data.



**Fig. 7:** Human evaluation results of CogView3 before and after prompt expansion on DrawBench.

CogView3 provides more comprehensive and in-distribution descriptions for model generation, largely improving the accuracy of instruction following for CogView3. Similar improvement is not observed on the generation of SDXL. The probable reason may be that SDXL is trained on original captions and only has an input window of 77 tokens, which leads to frequent truncation of the expanded prompts. This corroborates the statement in Section 3.1 that prompt expansion helps bridge the gap between model inference and training with re-captioned data.



**Fig. 8:** Comparison of the effect of prompt expansion for CogView3 and SDXL.

**Methods of Iterative Super-Resolution** Although straightforward implementation of the super-resolution stage model on higher image resolutions achieves desired outputs, this introduces excessive requirements of the CUDA memory, which is unbearable on the resolution of  $4096 \times 4096$ . Tiled diffusion [2] [11] is a series of inference methods for diffusion models tackling such an issue. It separates an inference step of large images into overlapped smaller blocks and mix them together to obtain the overall prediction of the step. As shown in Figure 9, comparable results can be achieved with tiled inference. This enables CogView3 to generate images with higher resolution by a limited CUDA memory usage. It is also possible to generate  $4096 \times 4096$  images with tiled methods, which we leave for future work.



**Fig. 9:** Comparison of direct higher super-resolution and tiled diffusion on  $2048 \times 2048$ . We choose Mixture of Diffusers [11] in view of its superior quality of integration. Original prompts are utilized for the inference of all blocks.

## 5 Conclusion

In this work, we propose CogView3, the first text-to-image generation system in the framework of relay diffusion. CogView3 achieves preferred generation quality with greatly reduced inference costs, largely attributed to the relaying pipeline. By iteratively implementing the super-resolution stage of CogView3, we are able to achieve high quality images of extremely high resolution as  $2048 \times 2048$ .

Meanwhile, with the incorporation of data re-captioning and prompt expansion into the model pipeline, CogView3 achieves better performance in prompt understanding and instruction following compared to current state-of-the-art open-source text-to-image diffusion models.

We also explore the distillation of CogView3 and demonstrate its simplicity and capability attributed to the framework of relay diffusion. Utilizing the progressive distillation paradigm, the distilled variant of CogView3 reduces the inference time drastically while still preserves a comparable performance.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
3. 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> **2**(3), 8 (2023)
4. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
5. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
6. 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. Advances in Neural Information Processing Systems **34**, 19822–19835 (2021)
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022)
10. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
11. Jiménez, Á.B.: Mixture of diffusers for scene composition and high resolution image generation. arXiv preprint arXiv:2302.02412 (2023)
12. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
13. 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)
14. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
15. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14297–14306 (2023)
16. Pernias, P., Rampas, D., Richter, M.L., Pal, C.J., Aubreville, M.: Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models (2023)
17. 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)
18. 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. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)

19. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
20. 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)
21. 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)
22. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
23. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023)
24. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
25. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
26. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
27. Teng, J., Zheng, W., Ding, M., Hong, W., Wangni, J., Yang, Z., Tang, J.: Relay diffusion: Unifying diffusion process across resolutions for image synthesis. arXiv preprint arXiv:2309.03350 (2023)
28. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)
29. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
30. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imageward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
31. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)

## A Sampler Derivation

In this section, we aim to demonstrate that our designed latent relay sampler matches with the forward process of latent relay diffusion. That is, we need to prove that if the joint distribution holds,

$$q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_{t-1}|a_t \mathbf{z}_t + b_t \mathbf{z}_0 + c_t \mathbf{z}_0^t, \delta_t^2 \mathbf{I}), \quad (10)$$

where  $a_t = \sqrt{\sigma_{t-1}^2 - \delta_t^2}/\sigma_t$ ,  $b_t = 1/t$ ,  $c_t = (t-1)/t - a_t$ , then the marginal distribution holds,

$$\begin{aligned} q(\mathbf{z}_t|\mathbf{z}_0) &= \mathcal{N}(\mathbf{z}_t|\mathbf{z}_0^t, \sigma_t^2 \mathbf{I}), \quad t \in \{1, \dots, T_r\}, \\ \mathbf{z}_0^t &= \mathcal{F}(\mathbf{z}_0, t) = \frac{T_r - t}{T_r} \mathbf{z}_0 + \frac{t}{T_r} \mathbf{z}^L. \end{aligned} \quad (11)$$

*proof.*

Given that  $q(\mathbf{z}_{T_r}|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}^L, \sigma_{T_r}^2 \mathbf{I})$ , we employ mathematical induction to prove it. Assuming that for any  $t \leq T_r$ ,  $q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_0^t, \sigma_t^2 \mathbf{I})$ . Next we only need to prove that  $q(\mathbf{z}_{t-1}|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_0^{t-1}, \sigma_{t-1}^2 \mathbf{I})$  holds, then it holds for all  $t$  from  $T_r$  to 1 according to the induction hypothesis.

First, based on

$$q(\mathbf{z}_{t-1}|\mathbf{z}_0) = \int q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0) q(\mathbf{z}_t|\mathbf{z}_0) d\mathbf{z}_t, \quad (12)$$

we have that

$$q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_{t-1}|a_t \mathbf{z}_t + b_t \mathbf{z}_0 + c_t \mathbf{z}_0^t, \delta_t^2 \mathbf{I}) \quad (13)$$

and

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_0^t, \sigma_t^2 \mathbf{I}). \quad (14)$$

Next, from Bishop and Nasrabadi [4], we know that  $q(\mathbf{z}_{t-1}|\mathbf{z}_0)$  is also Gaussian, denoted as  $\mathcal{N}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ . So, from Equation 12, it can be derived that

$$\begin{aligned} \boldsymbol{\mu}_{t-1} &= a_t \mathbf{z}_0^t + b_t \mathbf{z}_0 + c_t \mathbf{z}_0^t \\ &= \frac{\sqrt{\sigma_{t-1}^2 - \delta_t^2}}{\sigma_t} \mathbf{z}_0^t + \frac{\mathbf{z}_0}{t} + \left( \frac{t-1}{t} - \frac{\sqrt{\sigma_{t-1}^2 - \delta_t^2}}{\sigma_t} \right) \mathbf{z}_0^t \\ &= \frac{\mathbf{z}_0^t}{t} + \frac{t-1}{t} \mathbf{z}_0^t \\ &= \mathbf{z}_0^{t-1} \quad (\text{based on Equation 4}) \end{aligned} \quad (15)$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_{t-1} &= a_t^2 \sigma_t^2 + \delta_t^2 \\ &= \left( \frac{\sigma_{t-1}^2 - \delta_t^2}{\sigma_t^2} \right) \sigma_t^2 + \delta_t^2 \\ &= \sigma_{t-1}^2 \end{aligned} \quad (16)$$

In summary,  $q(\mathbf{z}_{t-1}|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_0^{t-1}, \sigma_{t-1}^2 \mathbf{I})$ . The inductive proof is complete.

## B Supplements of Text Expansion

We use the following passage as our template prompting GPT-4V to generate the growth truth of the recaption model:

```

**Objective**: **Give a highly descriptive image caption. **. As an expert, delve deep into the image with a discerning eye, leveraging rich creativity, meticulous thought. Generate a list of multi-round question-answer pairs about the image as an aid and final organise a highly descriptive caption. Image has a simple description.

**Instructions**:
- **Simple description**: Within following double braces is the description: {{<CAPTION>}}.
- Please note that the information in the description should be used cautiously. While it may provide valuable context such as artistic style, useful descriptive text and more, it may also contain unrelated, or even incorrect, information. Exercise discernment when interpreting the caption.
- Proper nouns such as character's name, painting's name, artistic style should be incorporated into the caption.
- URL, promoting info, garbled code, unrelated info, or info that relates but is not beneficial to our descriptive intention should not be incorporated into the caption.
- If the description is misleading or not true or not related to describing the image like promoting info, url, don't incorporate that in the caption.

- **Question Criteria**:
  - **Content Relevance**: Ensure questions are closely tied to the image's content.
  - **Diverse Topics**: Ensure a wide range of question types
  - **Keen Observation**: Emphasize questions that focus on intricate details, like recognizing objects, pinpointing positions, identifying colors, counting quantities, feeling moods, analyzing description and more.
  - **Interactive Guidance**: Generate actionable or practical queries based on the image's content.
  - **Textual Analysis**: Frame questions around the interpretation or significance of textual elements in the image.

- **Note**:
  - The first question should ask for a brief or detailed description of the image.
  - Count quantities only when relevant.
  - Questions should focus on descriptive details, not background knowledge or causal events.
  - Avoid using an uncertain tone in your answers. For example, avoid words like "probably, maybe, may, could, likely".
  - You don't have to specify all possible details, you should specify those that can be specified naturally here. For instance, you don't need to count 127 stars in the sky.
  - But as long as it's natural to do so, you should try to specify as many details as possible.
  - Describe non-English textual information in its original language without translating it.

- **Answering Style**:
Answers should be comprehensive, conversational, and use complete sentences. Provide context where necessary and maintain a certain tone.

Incorporate the questions and answers into a descriptive paragraph. Begin directly without introductory phrases like "The image showcases" "The photo captures" "The image shows" and more. For example, say "A woman is on a beach", instead of "A woman is depicted in the image".

**Output Format**:
```json
{
  "queries": [
    {
      "question": "[question text here]",
      "answer": "[answer text here]"
    },
    {
      "question": "[question text here]",
      "answer": "[answer text here]"
    }
  ],
  "result": "[highly descriptive image caption here]"
}
```
Please strictly follow the JSON format, akin to a Python dictionary with keys: "queries" and "result". Exclude specific question types from the question text.

```

In the prompt we fill <CAPTION> with the original caption, the prompt is used along with the input of images. On finetuning the recaption model, we use a template as:

```
<IMAGE> Original caption: <OLD_CAPTION>. Can you provide a more comprehensive description of the image? <NEW_CAPTION>.
```

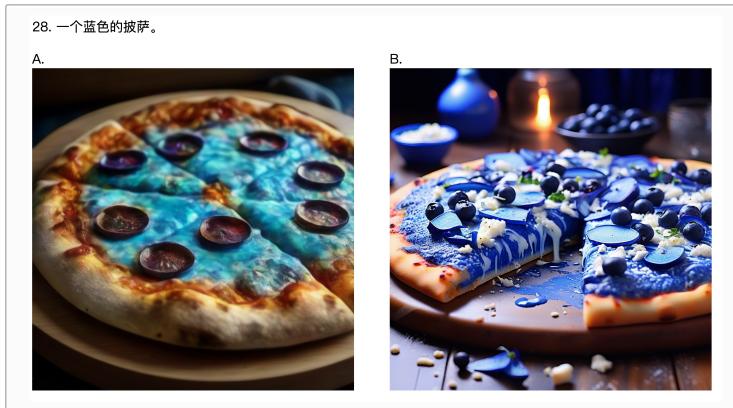
Figure 10 shows additional examples of the finalized recaption model.

## C Details of Human Evaluation

Figure 11 shows a case of the interface for the human evaluation. We shuffle the order of the comparison pairs by A/B in advance and provide human annotators with equal pairs from all the comparative groups. The annotators are asked to scroll down the interface and record their preference for each pair of comparison.

| Image | Original Caption   | New Caption   |
|-------|--|---|
|       | Tablou Canvas - Gold Lion - Tablomag   | A canvas painting titled 'Gold Lion' by Tablomag is prominently displayed in a modern living room, featuring a detailed and expressive depiction of a lion's face. The artwork stands out against the room's neutral palette, with the lion's mane and eyes highlighted in gold, creating a striking contrast. This piece serves as an elegant focal point and adds a touch of wild sophistication to the space.  |
|       | Flower offering to a child :: Victor Gabriel Gilbert - Street and market genre scenes ðòî              | In an open-air flower market, depicted by Victor Gabriel Gilbert, a woman in traditional attire offers a small white hat to a child dressed in a pink dress, who holds it gently. Another woman in a dark blue apron observes with a smile. The lush array of potted flowers, including reds, whites, and purples, enhances the warmth of the scene, while the soft lighting suggests a time of day when the sun casts gentle shadows. In the background, other individuals contribute to the vibrant atmosphere of this everyday moment. |
|       | Budapest's Matthias Church Hosts Exhibition On Habsburg Charles IV, Hungary's Last King post's picture | Matthias Church in Budapest stands majestically under a clear blue sky, its Gothic spires reaching towards the heavens. The intricate patterned roof tiles in red, orange, and white, along with the stone facade adorned with pointed arches and ornate windows, reflect the historical significance tied to Hungary's last king, Charles IV. A modern vehicle parked beside the entrance and a well-manicured garden with vibrant flowers add a touch of contemporary life to this ancient edifice.                                     |
|       | Brunch Mount Engadine Canmore2   | A wooden ledge in Mount Engadine, Canmore presents a sumptuous brunch with a glass of orange juice and a strawberry garnish, alongside a plate of pancakes dusted with powdered sugar and accompanied by fresh berries. The tranquil backdrop features a meandering river, dense forests, and towering mountains under an overcast sky, evoking a sense of peaceful solitude and natural beauty.  |

**Fig. 10:** Examples of the recaption model results.



**Fig. 11:** Interface showcases of the human evaluation. The original prompts is translated to Chinese, the mother language of our human annotators, for evaluation.

## D Additional Qualitative Comparisons

### D.1 Qualitative model Comparisons



**Fig. 12:** Qualitative comparisons of CogView3 with SDXL, Stable Cascade and DALL-E 3. All prompts are sampled from Partiprompts.

## D.2 Qualitative comparisons Between Distilled Models



**Fig. 13:** Qualitative comparisons of CogView3-distill with LCM-SDXL, recent model of diffusion distillation capable of generating  $1024 \times 1024$  samples. The first column shows samples from the original version of CogView3.