

# Faster Diffusion: Rethinking the Role of the Encoder for Diffusion Model Inference

Senmao Li<sup>1\*</sup>, Taihang Hu<sup>1\*</sup>, Joost van de Weijer<sup>2</sup>, Fahad Shahbaz Khan<sup>3,4</sup>, Tao Liu<sup>1</sup>

Linxuan Li<sup>1</sup>, Shiqi Yang<sup>5</sup>, Yaxing Wang<sup>1†</sup>, Ming-Ming Cheng<sup>1</sup>, Jian Yang<sup>1</sup>

<sup>1</sup>VCIP, CS, Nankai University, <sup>2</sup>Computer Vision Center, Universitat Autònoma de Barcelona

<sup>3</sup>Mohamed bin Zayed University of AI, <sup>4</sup>Linkoping University, <sup>5</sup>Independent Researcher, Tokyo

{senmaonk, hutaihang00, ltolcy0, linxuanli520, shiqi.yang147.jp}@gmail.com  
joost@cvc.uab.es, fahad.khan@liu.se, {yaxing, cmm, csjyang}@nankai.edu.cn

<https://sen-mao.github.io/FasterDiffusion>

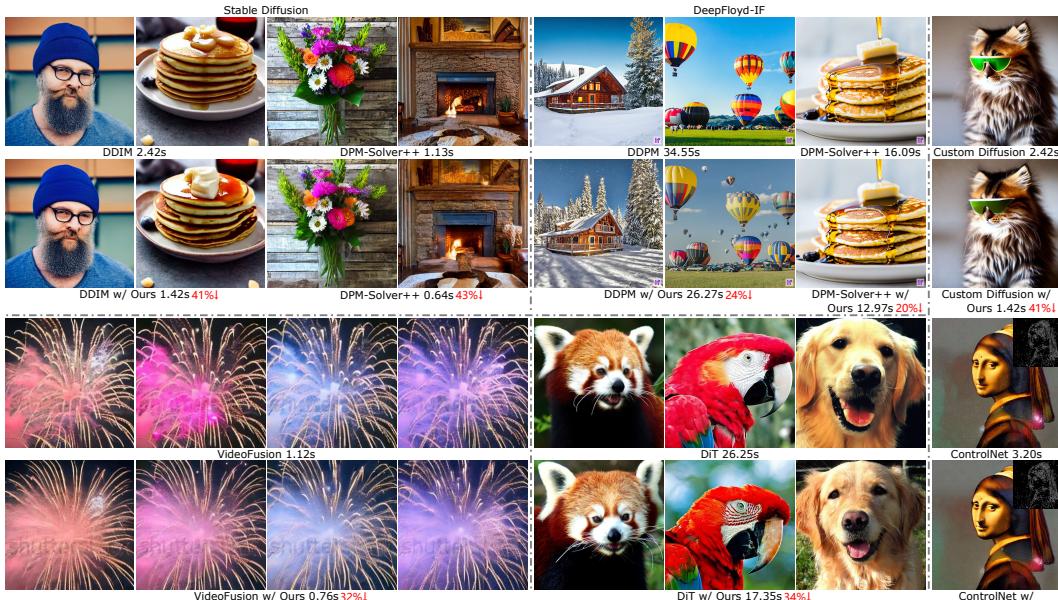


Figure 1: Results of our method for a diverse set of generation tasks. We significantly increase the image generation speed (**second/image**).

## Abstract

One of the main drawback of diffusion models is the slow inference time for image generation. Among the most successful approaches to addressing this problem are distillation methods. However, these methods require considerable computational resources. In this paper, we take another approach to diffusion model acceleration. We conduct a comprehensive study of the UNet encoder and empirically analyze the encoder features. This provides insights regarding their changes during the inference process. In particular, we find that encoder features change minimally, whereas the decoder features exhibit substantial variations across different time-steps. This insight motivates us to omit encoder computation at certain adjacent time-steps and reuse encoder features of previous time-steps as input to the decoder in multiple time-steps. Importantly, this allows us to perform decoder computation in parallel, further accelerating the denoising process. Additionally, we introduce a prior noise injection method to improve the texture details in the generated image. Besides the standard text-to-image task, we also validate our approach on

\*Equal contribution. Author ordering determined by coin flip over a Google Hangout.

†The corresponding author.

other tasks: text-to-video, personalized generation and reference-guided generation. Without utilizing any knowledge distillation technique, our approach accelerates both the Stable Diffusion (SD) and DeepFloyd-IF model sampling by 41% and 24% respectively, and DiT model sampling by 34%, while maintaining high-quality generation performance.

## 1 Introduction

One of the popular paradigms in image generation, Diffusion Models (DMs) [1, 2, 3] have recently achieved significant breakthroughs in various domains, including text-to-video generation [4, 5, 6], personalized image generation [7, 8, 9] and reference-guided image generation [10, 11, 12]. While diffusion models produce images of exceptional visual quality, their primary drawback lies in the prolonged inference time. The original diffusion model had an inference time several orders of magnitude slower than, for instance, GANs. One of the challenges hindering the acceleration of diffusion models is their inherent sequential denoising process, which limits the possibilities of effective parallelization.

To improve inference time speed of diffusion models, several methods have been developed, that can roughly be divided in two sets of approaches. Firstly, involving step reduction, the aim is to reduce the number of sampling steps within diffusion model inference, such as DDIM [13] and DPM-Solver [14], which have significantly reduced the number of sampling steps. Secondly, in contrast, knowledge distillation progressively distills a slow (many-step) teacher model into a faster (few-step) student model [15, 16]. Some recent works [17, 18, 19] excel at generating high-fidelity images in a few-step sampling scenario but face challenges in maintaining quality and diversity in one-step sampling. The main drawback of the distillation methods is that they require retraining to perform the distillation into faster diffusion models.

Orthogonal to these methods, we take a closer look at the sequential nature of the denoising process. We focus on the characteristics of the encoder in pretrained diffusion models (e.g., the SD and the DiT [21]<sup>2</sup>) Interestingly, based on our analysis presented in Sec 3.2, we discover that encoder features change minimally (Fig. 3a) and have a high degree of similarity (Fig. 2 (top)), whereas the decoder features exhibit substantial variations across different time-steps (Fig. 3a and Fig. 2 (bottom)). This insight is relevant because it allows us to circumvent the computation of the encoder during multiple time-steps. As a consequence, the decoder computations which are based on the same encoder input can be performed in parallel. Instead, we reuse the computed encoder features computed at one time-step (since these change minimally) as input to adjacent decoders during the following time-steps. More recently, both DeepCache[22] and CacheMe[23] leverage feature similarity to achieve acceleration. However, they rely on sequential denoising, as well as CacheMe requires fine-tuning. Unlike these methods, our approach supports parallel processing, which leads to faster inference (Tab. 2).

We show that the proposed propagation scheme accelerates the SD sampling by 24%, DeepFloyd-IF sampling by 18%, and DiT sampling by 27%. Furthermore, since the same encoder features (from previous time-steps) can be used as the input to the decoder of multiple later time-steps, this makes it possible to conduct multiple time-steps decoding concurrently. This parallel procession accelerates SD sampling by 41%, DeepFloyd-IF sampling by 24%, and DiT sampling by 34%. Furthermore, to alleviate the deterioration of the generated quality, we introduce a prior noise injection strategy to preserve the texture details in the generated images. With these contributions, our proposed method achieves improved sampling efficiency while maintaining high image generation quality.

<sup>1</sup>See the visualisation of the hierarchical features for all blocks in Appendix A

<sup>2</sup>We define the first several transformer blocks of DiT as the encoder, and the remaining ones of the transform blocks as the decoder.

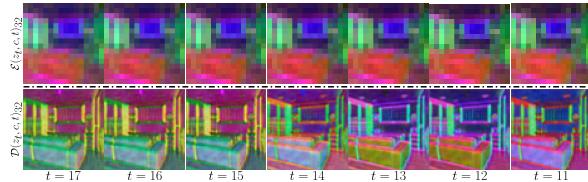


Figure 2: Visualising the hierarchical features <sup>1</sup>. We applied PCA to the hierarchical features following PnP [20] and used the top three leading components as an RGB image for visualization. The encoder features change minimally and have similarities at many time-steps (top), while the decoder features exhibit substantial variations across different time-steps (bottom).

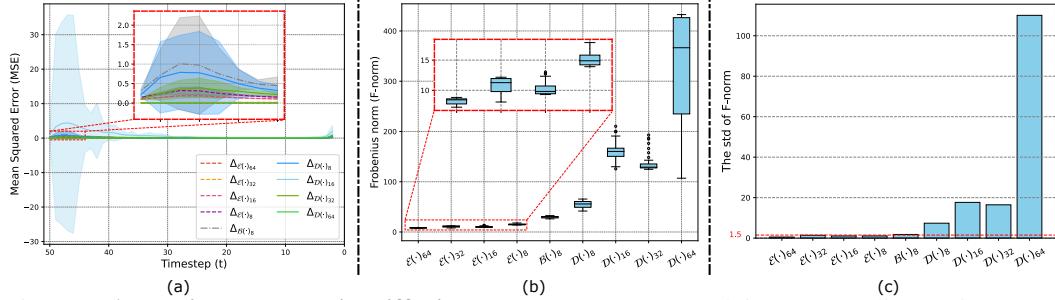


Figure 3: **Analyzing the UNet in Diffusion Model.** (a) Feature evolving across adjacent time-steps is measured by MSE. (b) We extract the hierarchical features output of different layers of the UNet at each time-step, average them along the channel dimension to obtain two-dimensional hierarchical features, and then calculate their *Frobenius norm*. (c) The hierarchical features of the UNet encoder show a lower standard deviation, while those of the decoder exhibit a higher standard deviation.

Importantly, our method can be combined with several approaches existing to speed up DMs. The main advantage of our method with respect to distillation-based approaches, is that our method can be applied at inference time, and does not require retraining a new faster distillation model; a process that is computationally very demanding and infeasible for actors with limited computational budget. Finally, we evaluate the effectiveness of our approach across a wide range of conditional diffusion-based tasks, including text-to-video generation (e.g., Text2Video-zero [4] and VideoFusion [5]), personalized image generation (e.g., Dreambooth [7]) and reference-guided image generation (e.g., ControlNet [10]).

To summarize, we make the following contributions:

- We conduct a thorough empirical study of the features of the UNet in the diffusion model showing that encoder features vary minimally (whereas decoder feature vary significantly).
- We propose a parallel strategy for diffusion model sampling at adjacent time-steps that significantly accelerates the denoising process. Importantly, our method does not require any training or fine-tuning technique.
- Furthermore, we also present a prior noise injection method to improve the image quality (mainly improving the quality of high-frequency textures).
- Our method can be combined with existing methods (like DDIM, and DPM-solver) to further accelerate diffusion model inference time.

## 2 Related Work

**Denoising diffusion model.** Recently, Text-to-image diffusion models [1, 24, 25, 26] have made significant advancements. Notably, Stable Diffusion and DeepFloyd-IF stand out as two of the most successful diffusion models available within the current open-source community. These models, building upon the UNet architecture, are versatile and can be applied to a wide range of tasks, including image editing [27, 28], super-resolution [29, 30], segmentation [31, 32], and object detection [33, 34]. Given the strong scalability of transformer networks, DiT [21] investigates the transformer backbone for diffusion models.

**Diffusion model acceleration.** Diffusion models use iterative denoising with UNet for image generation, which is time-consuming. There are plenty of works trying to address this issue. One strategy involves employing efficient diffusion model solvers, such as DDIM [13] and DPM-Solver [14], which have demonstrated notable reductions in sampling steps. Additionally, ToMe [35] exploits token redundancy to minimize the computations necessary for attention operations [36]. Conversely, knowledge distillation methods, exemplified by techniques like progressive simplification by student models [15, 16], aim to streamline existing models. Some recent studies combine model compression with distillation to achieve faster sampling [37, 38]. Orthogonal to these approaches, we introduce a novel method for enhancing sampling efficiency in DMs inference. We show that our method can be combined with several existing speed-up methods for further acceleration.

DeepCache[22] and CacheMe[23] are two recent works that leverage feature similarity to achieve acceleration. DeepCache [22] adopts a strategy of rudely reusing features cached from the previous step, requiring iterative denoising. Furthermore, CacheMe [23] requires additional fine-tuning for

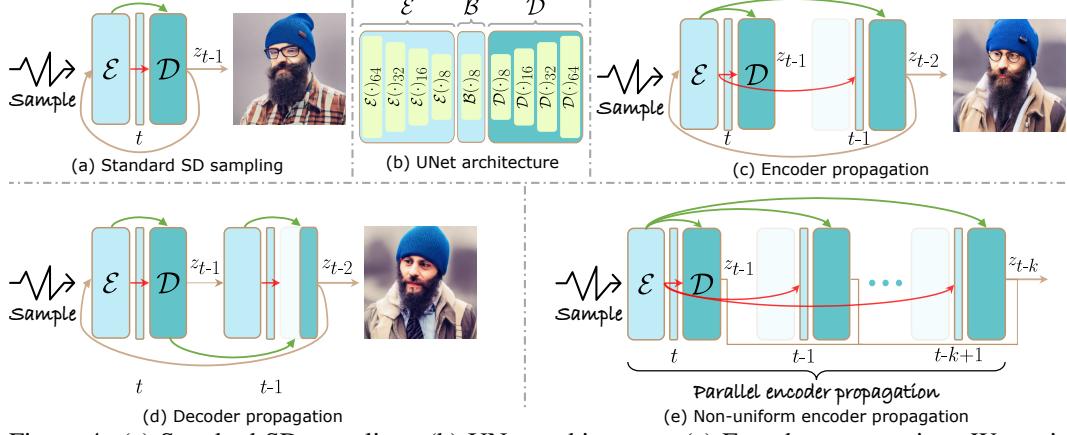


Figure 4: (a) Standard SD sampling. (b) UNet architecture. (c) Encoder propagation. We omit the encoder at certain adjacent time-steps and reuse in parallel the encoder features in the previous time-steps for the decoder. Applying encoder propagation for uniform strategy every two iterations. Note, at time-step  $t-1$ , predicting noise does not require  $z_{t-1}$  (i.e., Eq. 1:  $z_{t-2} = \sqrt{\frac{\alpha_{t-2}}{\alpha_{t-1}}} z_{t-1} + \sqrt{\alpha_{t-2}} \left( \sqrt{\frac{1}{\alpha_{t-2}} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1} \right) \cdot \epsilon_\theta(z_{t-2}, t-1, c)$ ). (d) Decoder propagation. The generated images often fail to cover some specific objects in the text prompt. For example, given one prompt case “A man with a beard wearing glasses and a beanie”, this method fails to generate the **glasses** subject. See Appendix F for quantitative evaluation. (e) Applying encoder propagation for non-uniform strategy. By benefiting from our propagation scheme, we are able to perform the decoder in parallel at certain adjacent time-steps.

better performance. In contrast, our approach enables parallel processing, leading to considerably faster inference.

### 3 Method

We first briefly revisit the architecture of the Stable Diffusion (SD) (Sec. 3.1), and then conduct a comprehensive analysis for the hierarchical features of the UNet (Sec. 3.2). Our analysis shows that it is possible to parallelize the diffusion model denoising process partially. Thus, we introduce a novel method to accelerate the diffusion sampling while still largely maintaining the generation quality and fidelity (Sec. 3.3).

#### 3.1 Latent Diffusion Model

In the diffusion inference stage, the denoising network  $\epsilon_\theta$  takes as input a text embedding  $c$ , a latent code  $z_t$  and a time embedding, predicts noise, resulting in a latent  $z_{t-1}$  using the DDIM scheduler [13]:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \sqrt{\alpha_{t-1}} \left( \sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \epsilon_\theta(z_t, t, c), \quad (1)$$

where  $\alpha_t$  is a predefined scalar function at time-step  $t$  ( $t = T, \dots, 1$ ). The typical denoising network uses a UNet-based architecture. It consists of an encoder  $E$ , a bottleneck  $B$ , and a decoder  $D$ , respectively (Fig. 4b). The hierarchical features extracted from the encoder  $E$  are injected into the decoder  $D$  by a skip connection (Fig. 4a). For convenience of description, we divide the UNet network into specific blocks:  $E = \{E(\cdot)_s\}$ ,  $B = \{B(\cdot)_8\}$ , and  $D = \{D(\cdot)_s\}$ , where  $s \in \{8, 16, 32, 64\}$  (see Fig. 4b). Both  $E(\cdot)_s$ <sup>3</sup> and  $D(\cdot)_s$  represent the block layers with input resolution  $s$  in both encoder and decoder, respectively.

Diffusion Transformer (DiT) [21] is a novel architecture for diffusion models. It replaces the UNet backbone with a transformer, which consists of 28 blocks. Based on our observations, we define the first 18 blocks as the encoder, and the remaining 10 blocks as the decoder (see Appendix A.3).

#### 3.2 Analyzing the UNet in Diffusion Model

In this section, we take the UNet-based diffusion model as example to analyze the properties of the pretrained diffusion model. We delve into the UNet which consists of the encoder  $E$ , the bottleneck  $B$ ,

<sup>3</sup>Once we replace the  $\cdot$  with specific inputs in  $E(\cdot)_s$ , we define that it represents the feature of  $E(\cdot)_s$

and the decoder  $\mathcal{D}$ , for a deeper understanding of the different parts of the UNet. Note the following observed properties also exist in DiT (see Appendix A.3).

**Feature evolution across time-steps.** We experimentally observe that the encoder features exhibit a subtle variation at adjacent time-steps, whereas the decoder features exhibit substantial variations across different time-steps (see Fig. 3a and Fig. 2). Specifically, given a pretrained diffusion model, we iteratively produce a latent code  $z_t$  (see Eq. 1), and the corresponding hierarchical features:  $\{\mathcal{E}(z_t, c, t)_s\}$ ,  $\{\mathcal{B}(z_t, c, t)_8\}$ , and  $\{\mathcal{D}(z_t, c, t)_s\}$  ( $s \in \{8, 16, 32, 64\}$ )<sup>4</sup>, as shown in Fig. 4b. Here, we analyze how the hierarchical features change at adjacent time-steps. To achieve this goal, we quantify the variation of the hierarchical features as follows:

$$\Delta_{\mathcal{E}(\cdot)_s} = \frac{1}{d \times s^2} \|\mathcal{E}(z_t, c, t)_s - \mathcal{E}(z_{t-1}, c, t-1)_s\|_2^2, \quad (2)$$

where  $d$  represents the number of channels in  $\mathcal{E}(z_t, c, t)_s$ . Similarly, we also compute  $\Delta_{\mathcal{B}(\cdot)_8}$  and  $\Delta_{\mathcal{D}(\cdot)_s}$ .

As illustrated in Fig. 3a, for both the encoder  $\mathcal{E}$  and the decoder  $\mathcal{D}$ , the curves exhibit a similar trend: in the wake of an initial increase, the variation reaches a plateau and then decreases, followed by a continuing growth towards the end. However, the extent of change in  $\Delta_{\mathcal{E}(\cdot)_s}$  and  $\Delta_{\mathcal{D}(\cdot)_s}$  is quantitatively markedly different. For example, the maximum value and variance of the  $\Delta_{\mathcal{E}(\cdot)_s}$  are less than 0.4 and 0.05, respectively (Fig. 3a (zoom-in area)), while the corresponding values of the  $\Delta_{\mathcal{D}(\cdot)_s}$  are about 5 and 30, respectively (Fig. 3a). Furthermore, we find that  $\Delta_{\mathcal{D}(\cdot)_64}$ , the change of the last layer of the decoder, is close to zero. This is due to the output of the denoising network being similar at adjacent time-steps [39]. In conclusion, the overall feature change  $\Delta_{\mathcal{E}(\cdot)_s}$  is smaller than  $\Delta_{\mathcal{D}(\cdot)_s}$  throughout the inference phase.

**Feature evolution across layers.** We experimentally observe that the feature characteristics are significantly different between the encoder and the decoder across all time-steps. For the encoder  $\mathcal{E}$  the intensity of the change is slight, whereas it is very drastic for the decoder  $\mathcal{D}$ . Specifically we calculate the *Frobenius norm* for hierarchical features  $\mathcal{E}(z_t, c, t)_s$  across all time-steps, dubbed as  $\mathcal{F}_{\mathcal{E}(\cdot)_s} = \{\mathcal{F}_{\mathcal{E}(z_T, c, T)_s}, \dots, \mathcal{F}_{\mathcal{E}(z_1, c, 1)_s}\}$ . Similarly, we compute  $\mathcal{F}_{\mathcal{B}(\cdot)_8}$  and  $\mathcal{F}_{\mathcal{D}(\cdot)_s}$ , respectively.

Fig. 3b shows the feature evolution across layers with a boxplot<sup>5</sup>. Specifically, for  $\{\mathcal{F}_{\mathcal{E}(\cdot)_s}\}$  and  $\{\mathcal{F}_{\mathcal{B}(\cdot)_8}\}$ , the box is relatively compact, with a narrow range between their first-quartile and third-quartile values. For example, the maximum box height ( $\mathcal{F}_{\mathcal{E}(\cdot)_32}$ ) of these features is less than 5 (see Fig. 3b (zoom-in area)). This indicates that the features from both the encoder  $\mathcal{E}$  and the bottleneck  $\mathcal{B}$  slightly change. In contrast, the box heights corresponding to  $\{\mathcal{D}(\cdot)_s\}$  are relatively large. For example, for the  $\mathcal{D}(\cdot)_64$  the box height is over 150 between the first quartile and third quartile values (see Fig. 3b). Furthermore, we also provide a standard deviation (Fig. 3c), which exhibits similar phenomena to Fig. 3b. These results show that the encoder features have relatively small discrepancy and a high degree of similarity across all layers. However, the decoder features evolve drastically.

**Could we omit the Encoder at certain time-steps?** As indicated by the previous experimental analysis, we observe that, during the denoising process, the decoder features change drastically, whereas the encoder  $\mathcal{E}$  features change minimally, and have a high degree of similarities at certain adjacent time-steps. Therefore, as shown in Fig. 4c, We propose to omit the encoder at certain time-steps and use the same encoder features for several decoder steps. This allows us to compute these multiple decoder steps in parallel.

Specifically, we delete the encoder at time-step  $t-1$  ( $t-1 < T$ ), and the corresponding decoder (including the skip connections) takes as input the hierarchical outputs of the encoder  $\mathcal{E}$  from the previous time-step  $t$ , instead of the ones from the current time-step  $t-1$  like the standard SD sampling (for more detail, see Sec. 3.3).

When omitting the encoder at a certain time-step, we are able to generate similar images (Fig. 4c) like standard SD sampling (Fig. 4a, Tab. 1 (the first and second rows) and additional results in Appendix F). Alternatively, if we use a similar strategy for the decoder (i.e., *decoder propagation*), we find the

<sup>4</sup>The feature resolution is half of the previous one in the encoder and two times in the decoder. Note that the feature resolutions of  $\mathcal{E}(\cdot)_8$ ,  $\mathcal{B}(\cdot)_8$  and  $\mathcal{D}(\cdot)_64$  do not change in the SD model.

<sup>5</sup>Each boxplot contains the minimum (0th percentile), the maximum (100th percentile), the median (50th percentile), the first quartile (25th percentile) and the third quartile (75th percentile) values of the feature Frobenius norm (e.g.,  $\{\mathcal{F}_{\mathcal{E}(z_T, c, T)_s}, \dots, \mathcal{F}_{\mathcal{E}(z_1, c, 1)_s}\}$ ).

generated images often fail to cover some specific objects in the text prompt (Fig. 4d). For example, when provided with prompt “A man with a beard wearing glasses and a beanie”, the SD model fails to synthesize “glasses” when applying decoder propagation. This is due to the fact that the semantics are mainly contained in the features from the decoder rather than the encoder [40].

The *encoder propagation*, which uses encoder outputs from previous time-step as the input to the current decoder, could speed up the diffusion model sampling at inference time. In the following Sec. 3.3, we give further elaborate on encoder propagation.

### 3.3 Encoder propagation

Diffusion sampling, combining iterative denoising with transformers, is time-consuming. Therefore we propose a novel and practical diffusion sampling acceleration method. During the diffusion sampling process  $t = \{T, \dots, 1\}$ , we refer to the time-steps where encoder propagation is deployed, as *non-key* time-steps denoted as  $t^{non-key} = \{t_0^{non-key}, \dots, t_{N-1}^{non-key}\}$ . The remaining time-steps are dubbed as  $t^{key} = \{t_0^{key}, t_1^{key}, \dots, t_{T-1-N}^{key}\}$ . In other words, we do not use the encoder at time-steps  $t^{non-key}$ , and leverage the hierarchical features of the encoder from the time-step  $t^{key}$ . Note we utilize the encoder  $\mathcal{E}$  at the initial time-step ( $t_0^{key} = T$ ). Thus, the diffusion inference time-steps could be reformulated as  $\{t^{key}, t^{non-key}\}$ , where  $t^{key} \cup t^{non-key} = \{T, \dots, 1\}$  and  $t^{key} \cap t^{non-key} = \emptyset$ . In the following, we introduce both *uniform encoder propagation* and *non-uniform encoder propagation* strategies.

As shown in Fig. 3a, The encoder feature change is larger in the initial inference phase compared to the later phases throughout the inference process. Therefore, we select more *key* time-steps in the initial inference phase, and less *key* time-steps in later phases. We experimentally define the *key* time-steps as  $t^{key} = \{50, 49, 48, 47, 45, 40, 35, 25, 15\}$  for SD model with DDIM, and  $t^{key} = \{100, 99, 98, \dots, 92, 91, 90, 85, 80, \dots, 25, 20, 15, 14, 13, \dots, 2, 1\}$ <sup>6</sup>,  $\{50, 49, \dots, 2, 1\}$  and  $\{75, 73, 70, 66, 61, 55, 48, 40, 31, 21, 10\}$  for three stage of DeepFloyd-IF (see detail *key* time-steps selection in Appendix F.2). The remaining time-steps are categorized as *non-key* time-steps. We define this strategy as non-uniform encoder propagation (see Fig. 4e). As shown in Fig. 4c, we also explore the *non-key* time-step selection with fix stride (e.g, 2), dubbed as *uniform encoder propagation*.

Note that our method does not reduce the number of sampling steps. During encoder propagation, the decoder computes for all time-steps, necessitating time embedding inputs for each time-step to maintain temporal coherence (see detail in Appendix D).

Tab. 5 reports the results of the ablation study, considering various combinations of *key* and *non-key* time-steps. These results indicate that the set of non-uniform *key* time-steps performs better in generating images.

#### Parallel non-uniform encoder propagation.

When applying the non-uniform encoder propagation strategy, at time-step  $t \in t^{non-key}$  the decoder inputs do not rely on the encoder outputs at time-step  $t$  (see Fig. 4e). Instead, it relies on the encoder output at the previous nearest *key* time-step. This allows us to perform *parallel non-uniform encoder propagation* at these adjacent time-steps in  $t^{non-key}$ . We perform decoding in parallel from  $t$  to  $t - k + 1$  time-steps. This technique further improves the inference efficiency since the decoder forward in multiple time-steps could be conducted concurrently. We indicate this as *parallel-batch non-key* time-steps. As shown in Fig. 5 (right), this further reduces evaluation time by 41% for the SD model.



Figure 5: Comparing with SD (left), encoder propagation reduces the sampling time by 24% (middle). Furthermore, parallel encoder propagation achieves a 41% reduction in sampling time (right).

**Prior noise injection.** Although the encoder propagation could improve the efficiency in the inference phase, we observe that it leads to a slight loss of texture information in the generated results (see Fig. 6 (left, middle)). Inspired by related works [41, 42], we propose a prior noise

<sup>6</sup>The ellipsis in  $t^{key}$  denotes each time-step between the time-steps on either side of the ellipsis. For example, 80...25 means that every time-step between 80 and 25 is included.

injection strategy. It combines the initial latent code  $z_T$  into the generative process at subsequent time-step (i.e.,  $z_t$ ), following  $z_t = z_t + \alpha \cdot z_T$ , if  $t < \tau$ , where  $\alpha = 0.003$  is the scale parameter to control the impact of  $z_T$ . And we start to use this injection mechanism from  $\tau = 25$  step. This strategic incorporation successfully improves the texture information. Importantly, it demands almost negligible extra computational resources. We observe that the loss of texture information occurs in all frequencies of the frequency domain (see Fig. 6 (right, red, and blue curves)). This approach ensures a close resemblance of generated results in the frequency domain to both SD and  $z_T$  injection (see Fig. 6 (right, red and green curves)), with the generated images maintaining the desired fidelity (see Fig. 6 (left, bottom)).

## 4 Experiments

In our experiments, we assess the speed-up of our method compared to others for inference acceleration. We also explore combining our method with these approaches. We do not directly compare our method with distillation methods, which offer superior results but involve computationally expensive retraining.

**Datasets and evaluation metrics.** We randomly select 10K prompts from the MS-COCO2017 validation dataset [43] and feed them into the text-to-image diffusion model to obtain 10K generated images. For the transformer architecture diffusion model, we randomly generate 50K images from 1000 ImageNet [44] class labels. For other tasks, we use the same settings as baselines (e.g., Text2Video-zero [4], VideoFusion [5], Dreambooth [7] and ControlNet [10]). We use the Fréchet Inception Distance (FID) [45] metric to assess the visual quality of the generated images, and the Clipscore [46] to measure the consistency between image content and text prompt. Furthermore, we report the average values for both the computational workload (GFLOPs/image) and sampling time (s/image) to represent the resource demands for a single image. See more detailed implementation information on Appendix A.

### 4.1 Text-to-image Generation

We first evaluate the proposed encoder propagation method for the standard text-to-image generation task on both the latent space (i.e., SD) and pixel space (i.e., DeepFloyd-IF) diffusion models. As shown in Tab. 1, we significantly accelerate the diffusion sampling with negligible performance degradation. Specifically, our proposed method decreases the computational burden (GFLOPs) by a large margin (27%) and greatly reduces sampling time to 41% when compared to standard DDIM sampling in SD. Similarly, in DeepFloyd-IF, the reduction in both computational burden and time reaches 15% and 24%, respectively. Furthermore, our method can be combined with the latest sampling techniques like DPM-Solver [14], DPM-Solver++ [47], and ToMe [35]. Our method enhances sampling efficiency while preserving good model performance, with negligible variations of both FID and Clipscore values (Tab. 1 (the third to eighth rows)). Our method achieves good performance across different sampling steps (Fig. 7 and see Appendix D for quantitative results.). Importantly, these results show that our method is orthogonal and compatible

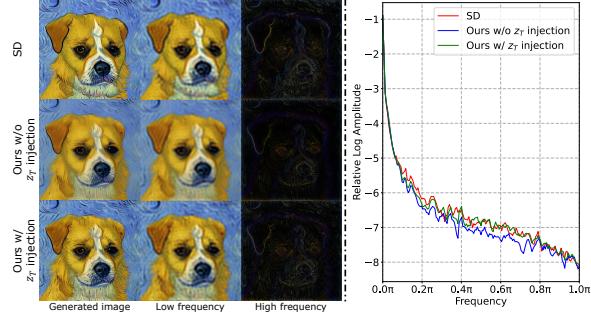


Figure 6: (left) We keep the image content through  $z_T$  injection, slightly compensating for the texture information loss caused by encoder propagation. (right) The amplitudes of the generated image through  $z_T$  injection closely resemble those from SD.

Table 1: Quantitative evaluation<sup>7</sup> for both SD and DeepFloyd-IF diffusion models.

DM	Sampling Method	T	FID $\downarrow$	Clip-score $\uparrow$	s/image $\downarrow$	
					GFLOPs/ image $\downarrow$	Unet of DM
Stable Diffusion	DDIM	50	21.75 0.773	37050	2.23	2.42
	DDIM w/ Ours	50	21.08 0.783	<b>27350</b> <sub>27%<math>\downarrow</math></sub>	<b>1.21</b> <sub>45%<math>\downarrow</math></sub>	<b>1.42</b> <sub>41%<math>\downarrow</math></sub>
DPM-Solver	DPM-Solver	20	21.36 0.780	14821	0.90	1.14
	DPM-Solver w/ Ours	20	21.25 0.779	<b>11743</b> <sub>21%<math>\downarrow</math></sub>	<b>0.46</b> <sub>48%<math>\downarrow</math></sub>	<b>0.64</b> <sub>43%<math>\downarrow</math></sub>
DPM-Solver++	DPM-Solver++	20	20.51 0.782	14821	0.90	1.13
	DPM-Solver++ w/ Ours	20	20.76 0.781	<b>11743</b> <sub>21%<math>\downarrow</math></sub>	<b>0.46</b> <sub>48%<math>\downarrow</math></sub>	<b>0.64</b> <sub>43%<math>\downarrow</math></sub>
DDIM + ToMe	DDIM + ToMe	50	22.32 0.782	35123	2.07	2.26
	DDIM + ToMe w/ Ours	50	20.73 0.781	<b>26053</b> <sub>26%<math>\downarrow</math></sub>	<b>1.15</b> <sub>44%<math>\downarrow</math></sub>	<b>1.33</b> <sub>41%<math>\downarrow</math></sub>
DeepFloyd-IF	DDPM	225	23.89 0.783	734825	33.91	34.55
	DDPM w/ Ours	225	23.73 0.782	<b>626523</b> <sub>15%<math>\downarrow</math></sub>	<b>25.61</b> <sub>25%<math>\downarrow</math></sub>	<b>26.27</b> <sub>24%<math>\downarrow</math></sub>
	DPM-Solver++100	100	20.79 0.784	370525	15.19	16.09
	DPM-Solver++100 w/ Ours	100	20.85 0.785	<b>313381</b> <sub>15%<math>\downarrow</math></sub>	<b>12.02</b> <sub>21%<math>\downarrow</math></sub>	<b>12.97</b> <sub>20%<math>\downarrow</math></sub>

<sup>7</sup>We use the official implementation of Clipscore [46] to obtain around 0.75 but around 0.3. See Appendix C.

Table 2: Comparison with DeepCache and CacheMe. CacheMe is not open-source.

Sampling Method	T	Parallel	FID ↓	Clipscore ↑	s/image
DDIM	50	✗	21.75	0.773	2.42
DDIM w/ DeepCache	50	✗	21.53	0.770	1.05 <small>56%↓</small>
DDIM w/ CacheMe	50	✗	—	—	1.30 <small>44%↓</small>
DDIM w/ Ours	50	✓	21.62	0.775	0.56 <small>77%↓</small>

Table 3: Quantitative evaluation for DiT.

Sampling Method	T	Image Res.	FID ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑	s/image
DiT	250	256	2.27	4.60	278.24	0.83	0.57	5.13
DiT w/ Ours	250	256	<b>2.31</b>	<b>4.55</b>	<b>276.05</b>	<b>0.82</b>	<b>0.57</b>	<b>3.62</b> <small>29%↓</small>
DiT	250	512	3.04	5.02	240.82	0.84	0.54	26.25
DiT w/ Ours	250	512	<b>3.25</b>	<b>5.05</b>	<b>245.13</b>	<b>0.83</b>	<b>0.51</b>	<b>17.35</b> <small>34%↓</small>

Table 4: Quantitative evaluation on text-to-video, personalized generation and reference-guided generation tasks. † and ‡ indicate “edges” and “scribble” conditions, respectively.

Method	T	FID ↓	Clip-score ↑	Gflops/image ↓	s/image ↓	Unet of SD	SD
Text2Video-zero	50	-	0.732	39670	12.59/8	13.65/8	
Text2Video-zero w/ Ours	50	-	0.731	<b>30690</b> <small>22%↓</small>	<b>9.46/8</b> <small>25%↓</small>	<b>10.54/8</b> <small>23%↓</small>	
VideoFusion	50	-	0.700	224700	16.71/16	17.93/16	
VideoFusion w/ Ours	50	-	0.700	<b>148680</b> <small>33%↓</small>	<b>11.1/16</b> <small>34%↓</small>	<b>12.2/16</b> <small>32%↓</small>	
ControlNet (†)	50	13.78	0.769	49500	3.09	3.20	
ControlNet (†) w/ Ours	50	14.65	0.767	<b>31400</b> <small>37%↓</small>	<b>1.43</b> <small>54%↓</small>	<b>1.52</b> <small>51%↓</small>	
ControlNet (‡)	50	16.17	0.775	56850	3.85	3.95	
ControlNet (‡) w/ Ours	50	16.42	0.775	<b>35990</b> <small>37%↓</small>	<b>1.83</b> <small>53%↓</small>	<b>1.93</b> <small>51%↓</small>	
Dreambooth	50	-	0.640	37050	2.23	2.42	
Dreambooth w/ Ours	50	-	0.660	<b>27350</b> <small>27%↓</small>	<b>1.21</b> <small>45%↓</small>	<b>1.42</b> <small>41%↓</small>	
CustomDiffusion	50	-	0.640	37050	2.21	2.42	
CustomDiffusion w/ Ours	50	-	0.650	<b>27350</b> <small>27%↓</small>	<b>1.21</b> <small>45%↓</small>	<b>1.42</b> <small>41%↓</small>	

Table 5: Quantitative evaluation in various propagation strategies on MS-COCO 2017 10K subset. FTC=FID×Time/Clipscore.

Propagation strategy	FID ↓		Clipscore ↑	GFLOPs /image ↓		s/image ↓		FTC ↓
	Unet of SD	SD	Unet of SD	SD	SD	SD	SD	
Uniform	21.75	0.773	37050	2.23	2.42	68.1		
I	$t^{key} = \{50, 48, 46, 44, 42, 40, 38, 36, 34, 32, 30, 28, 26, 24, 22, 20, 18, 16, 14, 12, 10, 8, 6, 4, 2\}$	21.55	0.775	31011 <small>16%↓</small>	<b>1.62</b> <small>27%↓</small>	<b>1.81</b> <small>25%↓</small>	50.3	
II	$t^{key} = \{50, 44, 38, 32, 26, 20, 14, 8, 2\}$	21.54	0.773	27350 <small>27%↓</small>	<b>1.26</b> <small>43%↓</small>	<b>1.46</b> <small>40%↓</small>	40.7	
III	$t^{key} = \{50, 38, 26, 14, 2\}$	24.61	0.766	26370 <small>29%↓</small>	<b>1.12</b> <small>50%↓</small>	<b>1.36</b> <small>44%↓</small>	43.7	
Non-uniform	$t^{key} = \{50, 40, 39, 38, 30, 25, 20, 15, 5\}$	22.94	0.776	27350 <small>27%↓</small>	<b>1.26</b> <small>43%↓</small>	<b>1.42</b> <small>41%↓</small>	41.9	
I	$t^{key} = \{50, 30, 25, 20, 15, 14, 5, 4, 3\}$	35.25	0.742	27350 <small>27%↓</small>	<b>1.25</b> <small>43%↓</small>	<b>1.42</b> <small>41%↓</small>	67.4	
II	$t^{key} = \{50, 41, 37, 35, 22, 21, 18, 14, 5\}$	22.14	0.778	27350 <small>27%↓</small>	<b>1.22</b> <small>45%↓</small>	<b>1.42</b> <small>41%↓</small>	40.4	
III	$t^{key} = \{50, 49, 48, 47, 45, 40, 35, 25, 15\}$	21.08	0.783	27350 <small>27%↓</small>	<b>1.21</b> <small>45%↓</small>	<b>1.42</b> <small>41%↓</small>	38.2	
IV (Ours)								

with these acceleration techniques. As shown in Fig. 1, we visualize the generated images with different sampling techniques. Our method still generates high-quality results (see Appendix F for additional results).

Our method allows to use multi-GPU to generate one image. With multi-GPU parallel, our proposed method further accelerates the SD sampling by 77%, whereas DeepCache [22] and CacheMe [23] achieve speedups of 56% and 44%, respectively (See Tab. 2). These results indicate that we achieve superior acceleration compared to DeepCache [22] and CacheMe [23].

## 4.2 Diffusion Transformer

We also evaluate our approach on DiT. As reported in Tab. 3, we achieve accelerations of about 29% and 34% for DiT sampling with image resolution of 256 and 512, respectively, while preserving high-quality results (see Figs. 1 and 18).

## 4.3 Other tasks with text-guided diffusion model

Besides the standard text-to-image task, we also validate our proposed approach on other tasks: *text-to-video generation*, *personalized generation*, and *reference-guided image generation*.

**Text-to-video.** To evaluate our method, we combine it with both Text2Video-zero [4] and VideoFusion [5]. As reported in Tab. 4 (the second and fourth rows), when combined with our method, the two methods have a reduction of approximately 22% to 33% in both computational burden and generation time. These results indicate that we are able to enhance the efficiency of generative processes in the text-to-video task while preserving video fidelity at the same time (Fig. 1 (left, bottom)). As an example, when generating a video using the prompt “Fireworks bloom in the night sky”, the VideoFusion model takes 17.92 seconds with 16 frames for the task (1.12s/frame), when combined with our method it only takes 12.27s (0.76s/frame) to generate a high-quality video (Fig. 1 (left, bottom)).

**Personalized image generation.** Dreambooth [7] and Custom Diffusion [8] are two approaches for customizing tasks by fine-tuning text-to-image diffusion models. As reported in Tab. 4 (the ninth to twelfth rows), our method, combining with the two customization approaches, accelerates image



Figure 7: Generated images at different time-steps.

generation and reduces computational demands. Visually, it maintains the ability to generate images with specific contextual relationships based on reference images. (Fig. 1 (right))

**Reference-guided image generation.** ControlNet [10] incorporates a trainable encoder, successfully generates a text-guided image, and preserves similar content with conditional information. Our approach can be applied concurrently to two encoders of ControlNet. In this paper, we validate the proposed method with two conditional controls: *edge* and *scribble*. Tab. 4 (the fifth to eighth row) reports quantitative results. We observe that it leads to a significant decrease in both generation time and computational burden. Furthermore, Fig. 1 (middle, bottom) qualitatively shows that our method successfully preserves the given structure information and achieves similar results as ControlNet.

**User study.** We conducted a user study, as depicted in Fig. 8, and asked subjects to select results. We apply pairwise comparisons (forced choice) with 18 users (35 pairs of images or videos/user). The results demonstrate that our method performs equally well as the baseline methods.

#### 4.4 Ablation study

We ablate the results with different selections of both uniform and non-uniform encoder propagation. Tab. 5 reports that the performance of the non-uniform setting outperforms the uniform one in terms of both FID and Clipscore (see Tab. 5 (the third and eighth rows)). Furthermore, we explore different configurations within the non-uniform strategy. The strategy, using the set of *key* time-steps we established, yields better results in the generation process (Tab. 5 (the eighth row)). We further present qualitative results stemming from the above choices. As shown in Fig. 9, given the same number of *key* time-steps, the appearance of nine-step non-uniform strategy II, III and IIII settings do not align with the prompt “Fireflies dot the night sky”. Although the generated image in the two-step setting exhibits a pleasing visual quality, its sampling efficiency is lower than our chosen setting (see Tab. 5 (the second and eighth rows)).

**Effectiveness of prior noise injection.** We evaluate the effectiveness of injecting initial  $z_T$ .

As reported in Tab. 6, the differences in FID and Clipscores without  $z_T$  (the third column), when compared to DDIM and Ours (the second and fourth columns), are approximately 0.01%, which can be considered negligible. While this is not the case for the visual expression of the generated image, it is observed that the output contains complete semantic information with smoothing texture (refer to Fig. 6 (left, the second row)). Injecting the  $z_T$  aids in maintaining fidelity in the generated results during encoding propagation (see Fig. 6 (left, the third row) and Fig. 6 (right, red and green curves)).

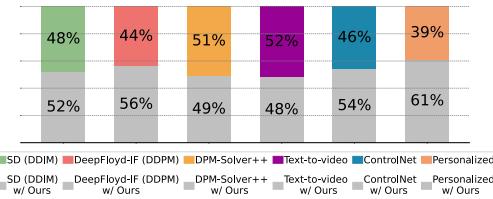


Figure 8: User study results.

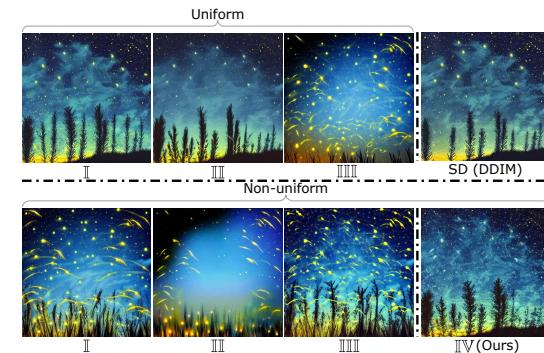


Figure 9: Generating image with uniform and non-uniform encoder propagation. The result of uniform strategy III yields smooth and loses textual compared with SD. Both uniform strategy III and non-uniform strategy II, III and IIII generate images with unnatural saturation levels.

Table 6: Quantitative evaluation for prior noise injection.

Sampling Method	SD (DDIM)	SD (DDIM) + Ours w/o $z_T$ injection	SD (DDIM) + Ours w/ $z_T$ injection
<b>FID ↓</b>	21.75	21.71	21.08
<b>Clipscore ↑</b>	0.773	0.779	0.783

## 5 Conclusion

In this work, We explore the characteristics of the encoder and decoder in UNet of the text-to-image diffusion model and find that encoder feature variation is minimal for many time-steps, while the decoder plays a significant role across all time-steps. Building upon this finding, we propose encoder propagation for efficient diffusion sampling, reducing time on both the UNet-based and the transform-based diffusion models on a diverse set of generation tasks. We conduct extensive experiments and validate that our approach can achieve improved sampling efficiency while maintaining image quality. **Limitations:** Although our approach achieves efficient diffusion sampling, it faces challenges in generating quality when using a limited number of sampling steps (e.g., 5). In addition, even though our proposed parallelization can also be applied to network distillation approaches [18, 17, 19], we have not explored this direction in this paper and leave it to future research.

## Acknowledgements

This work was supported by NSFC (NO. 62225604) and Youth Foundation (62202243). We acknowledge project PID2022-143257NB-I00, financed by the Spanish Government MCIN/AEI/10.13039/501100011033 and FEDER. We acknowledge "Science and Technology Yongjiang 2035" key technology breakthrough plan project (2024Z120). Computation is supported by the Supercomputing Center of Nankai University (NKSC).

We would like to thank Kai Wang, a postdoctoral researcher at the Computer Vision Center, Universitat Autònoma de Barcelona, for his helpful discussions and comments during the rebuttal period.

## References

- [1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [3] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [4] Levon Khachatryan, Andranik Mousisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.
- [5] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023.
- [6] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.
- [7] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.
- [8] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.
- [9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[11] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[12] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[14] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[15] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[16] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[17] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[18] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

[19] Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. *arXiv preprint arXiv:2312.05239*, 2023.

[20] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[21] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

[22] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. *arXiv preprint arXiv:2312.00858*, 2023.

[23] Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, et al. Cache me if you can: Accelerating diffusion models through block caching. *arXiv preprint arXiv:2312.03209*, 2023.

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[25] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[26] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[27] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[28] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023.

[29] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023.

[30] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12413–12422, 2022.

[31] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[32] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022.

[33] Walter HL Pinaya, Mark S Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H Mah, Andrew D MacKinnon, James T Teo, Rolf Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 705–714. Springer, 2022.

[34] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023.

[35] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4598–4602, 2023.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[37] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. On architectural compression of text-to-image diffusion models. *arXiv preprint arXiv:2305.15798*, 2023.

[38] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *arXiv preprint arXiv:2306.00980*, 2023.

[39] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023.

[40] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene? *arXiv preprint arXiv:2310.06836*, 2023.

[41] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022.

[42] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[45] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[46] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[47] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

[48] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023.

- [49] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022.
- [50] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- [51] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023.
- [52] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023.
- [53] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin C.K. Chan, and Ziwei Liu. ReVersion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023.
- [54] Enshu Liu, Xuefei Ning, Zinan Lin, Huazhong Yang, and Yu Wang. Oms-dpm: Optimizing the model schedule for diffusion probabilistic models. In *International Conference on Machine Learning*, pages 21915–21936. PMLR, 2023.
- [55] Lijiang Li, Huixia Li, Xiawu Zheng, Jie Wu, Xuefeng Xiao, Rui Wang, Min Zheng, Xin Pan, Fei Chao, and Rongrong Ji. Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7105–7114, 2023.
- [56] Shuai Yang, Yukang Chen, Luozhou Wang, Shu Liu, and Yingcong Chen. Denoising diffusion step-aware models. *arXiv preprint arXiv:2310.03337*, 2023.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Related Work</b>	<b>3</b>
<b>3</b>	<b>Method</b>	<b>4</b>
3.1	Latent Diffusion Model	4
3.2	Analyzing the UNet in Diffusion Model	4
3.3	Encoder propagation	6
<b>4</b>	<b>Experiments</b>	<b>7</b>
4.1	Text-to-image Generation	7
4.2	Diffusion Transformer	8
4.3	Other tasks with text-guided diffusion model	8
4.4	Ablation study	9
<b>5</b>	<b>Conclusion</b>	<b>10</b>
<b>Appendix</b>		<b>15</b>
<b>A</b>	<b>Implementation Details</b>	<b>15</b>
A.1	Configure	15
A.2	Details about the layers in the UNet	15
A.3	Details about the blocks in the DiT	15
A.4	Time and memory consumption ratios	17
A.5	GFLOPs	19
A.6	Baseline Implementations	19
<b>B</b>	<b>Parameter Count and FLOPs of SD</b>	<b>20</b>
<b>C</b>	<b>The methodology for computing Clipscore</b>	<b>20</b>
<b>D</b>	<b>Different from step-reduction methods</b>	<b>20</b>
<b>E</b>	<b>Difference between “prior noise injection” and “churn”</b>	<b>22</b>
<b>F</b>	<b>Ablation Experiments and Additional Results</b>	<b>22</b>
F.1	Comparion with DeepCache	22
F.2	The definition of <i>key</i> time-steps in various tasks	23
F.3	The effectiveness of encoder propagation	23
F.4	User study details	23
F.5	Additional metrics	24
F.6	Additional results	25

F.7 Additional tasks . . . . .	25
<b>G Automatic selection for key time steps</b>	<b>25</b>
<b>H Impact Statements</b>	<b>25</b>

## Appendix

In the appendix, we provide a detailed description of the experimental implementation (Appendix A). Subsequently, an analysis of the parameter quantities for the encoder and decoder is conducted (Appendix B).

Following that, we explain the methodology used to compute Clipscore (Appendix C), as well as highlight the distinction between our approach and methods that reduce sampling steps (Appendix D). Then, we present additional experiments and results, including more detailed ablation studies and comparative experiments (Appendix F). In the end, we provide a statement of the potential broader impact of our work (Appendix H).

## A Implementation Details

As shown in Code. 1, in the standard SD sampling code, adding 3 lines of code with green comments can achieve encoder propagation.

Code 1: Encoder propagation for SD (DDIM)

```
from diffusers import StableDiffusionPipeline
import torch
from utils import register_parallel_pipeline, register_faster_forward # 1. import package

model_id = "runwayml/stable-diffusion-v1-5"
pipe = StableDiffusionPipeline.from_pretrained(model_id, torch_dtype=torch.float16)
pipe = pipe.to("cuda")

register_parallel_pipeline(pipe) # 2. enable parallel
register_faster_forward(pipe.unet) # 3. encoder propagation

prompt = "a photo of an astronaut riding a horse on mars"
image = pipe(prompt).images[0]
image.save("astronaut_rides_horse.png")
```

### A.1 Configure

We use the Stable Diffusion v1.5 pre-trained model <sup>8</sup> and DeepFloyd-IF <sup>9</sup>. All of our inference experiments are conducted using an A40 GPU (48GB of VRAM).

We randomly select 100 captions from the MS-COCO 2017 validation dataset [43] as prompts for generating images. The analytical results presented in Sec. 3.2 are based on the statistical outcomes derived from these 100 generated images.

### A.2 Details about the layers in the UNet

The UNet in Stable Diffusion (SD) consists of an encoder  $\mathcal{E}$ , a bottleneck  $\mathcal{B}$ , and a decoder  $\mathcal{D}$ , respectively. We divide the UNet into specific blocks:  $\mathcal{E} = \{\mathcal{E}(\cdot)_s\}$ ,  $\mathcal{B} = \{\mathcal{B}(\cdot)_8\}$ , and  $\mathcal{D} = \{\mathcal{D}(\cdot)_s\}$ , where  $s \in \{8, 16, 32, 64\}$ .  $\mathcal{E}(\cdot)_s$  and  $\mathcal{D}(\cdot)_s$  represent the block layers with input resolution  $s$  in the encoder and decoder, respectively. Tab. 7 presents detailed information about the block architecture. Fig. 10 illustrates the hierarchical features for these blocks.

### A.3 Details about the blocks in the DiT

<sup>8</sup><https://huggingface.co/runwayml/stable-diffusion-v1-5>

<sup>9</sup><https://github.com/deep-floyd/IF>

Table 7: Detailed information about the layers of the encoder  $\mathcal{E}$ , bottleneck  $\mathcal{B}$  and decoder  $\mathcal{D}$  in the UNet of SD.

UNet	Layer number	Type of layer	Layer name	Input resolution of layer	Output resolution of layer
$\mathcal{E}(\cdot)$	0	resnets	down_blocks.0.resnets.0	(320, <b>64, 64</b> )	(320, 64, 64)
	1	attention	down_blocks.0.attentions.0	(320, 64, 64)	(320, 64, 64)
	2	resnet	down_blocks.0.resnets.1	(320, 64, 64)	(320, 64, 64)
	3	attention	down_blocks.0.attentions.1	(320, 64, 64)	(320, 64, 64)
	4	downsamplers	down_blocks.0.downsamplers.0	(320, 64, 64)	(320, 32, 32)
	5	resnet	down_blocks.1.resnets.0	(320, <b>32, 32</b> )	(640, 32, 32)
	6	attention	down_blocks.1.attentions.0	(640, 32, 32)	(640, 32, 32)
	7	resnet	down_blocks.1.resnets.1	(640, 32, 32)	(640, 32, 32)
	8	attention	down_blocks.1.attentions.1	(640, 32, 32)	(640, 32, 32)
	9	downsamplers	down_blocks.1.downsamplers.0	(640, 32, 32)	(640, 16, 16)
	10	resnet	down_blocks.2.resnets.0	(640, <b>16, 16</b> )	(1280, 16, 16)
	11	attention	down_blocks.2.attentions.0	(1280, 16, 16)	(1280, 16, 16)
	12	resnet	down_blocks.2.resnets.1	(1280, 16, 16)	(1280, 16, 16)
	13	attention	down_blocks.2.attentions.1	(1280, 16, 16)	(1280, 16, 16)
	14	downsamplers	down_blocks.2.downsamplers.0	(1280, 16, 16)	(1280, 8, 8)
$\mathcal{B}(\cdot)$	15	resnet	mid_blocks.resnets.0	(1280, <b>8, 8</b> )	(1280, 8, 8)
	16	resnet	mid_blocks.resnets.1	(1280, 8, 8)	(1280, 8, 8)
$\mathcal{D}(\cdot)$	17	resnet	mid_blocks.attentions.0	(1280, <b>8, 8</b> )	(1280, 8, 8)
	18	attention	mid_blocks.attentions.0	(1280, 8, 8)	(1280, 8, 8)
	19	resnet	mid_blocks.attentions.1	(1280, 8, 8)	(1280, 8, 8)
	20	resnet	up_blocks.0.resnets.0	(1280+1280, <b>8, 8</b> )	(1280, 8, 8)
	21	resnet	up_blocks.0.resnets.1	(1280+1280, 8, 8)	(1280, 8, 8)
	22	resnet	up_blocks.0.resnets.2	(1280+1280, 8, 8)	(1280, 8, 8)
	23	upsamplers	up_blocks.0.upsamplers.0	(1280, 8, 8)	(1280, 16, 16)
	24	resnet	up_blocks.1.resnets.0	(1280+1280, <b>16, 16</b> )	(1280, 16, 16)
	25	attention	up_blocks.1.attentions.0	(1280, 16, 16)	(1280, 16, 16)
	26	resnet	up_blocks.1.resnets.1	(1280+1280, 16, 16)	(1280, 16, 16)
	27	attention	up_blocks.1.attentions.1	(1280, 16, 16)	(1280, 16, 16)
	28	resnet	up_blocks.1.resnets.2	(1280+640, 16, 16)	(1280, 16, 16)
	29	attention	up_blocks.1.attentions.2	(1280, 16, 16)	(1280, 16, 16)
	30	upsamplers	up_blocks.1.upsamplers.0	(1280, 16, 16)	(1280, 32, 32)
$\mathcal{D}(\cdot)$	31	resnet	up_blocks.2.resnets.0	(1280+640, <b>32, 32</b> )	(640, 32, 32)
	32	attention	up_blocks.2.attentions.0	(640, 32, 32)	(640, 32, 32)
	33	resnet	up_blocks.2.resnets.1	(640+640, 32, 32)	(640, 32, 32)
	34	attention	up_blocks.2.attentions.1	(640, 32, 32)	(640, 32, 32)
	35	resnet	up_blocks.2.resnets.2	(640+320, 32, 32)	(640, 32, 32)
	36	attention	up_blocks.2.attentions.2	(640, 32, 32)	(640, 32, 32)
	37	upsamplers	up_blocks.2.upsamplers.0	(640, 32, 32)	(640, 64, 64)
	38	resnet	up_blocks.3.resnets.0	(640+320, <b>64, 64</b> )	(320, 64, 64)
	39	attention	up_blocks.3.attentions.0	(320, 64, 64)	(320, 64, 64)
	40	resnet	up_blocks.3.resnets.1	(320+320, 64, 64)	(320, 64, 64)
	41	attention	up_blocks.3.attentions.1	(320, 64, 64)	(320, 64, 64)
	42	resnet	up_blocks.3.resnets.2	(320+320, 64, 64)	(320, 64, 64)
	43	attention	up_blocks.3.attentions.2	(320, 64, 64)	(320, 64, 64)

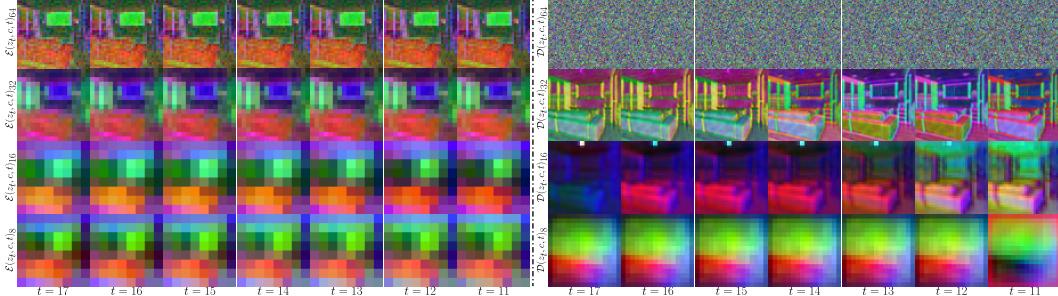


Figure 10: Visualising the hierarchical features. We applied PCA to the hierarchical features following PnP [20] and used the top three leading components as an RGB image for visualization. The encoder features changes minimally and has similarity at many time-steps (left), while the decoder features exhibit substantial variations across different time-steps (right).

Fig. 12 visualizes the hierarchical features for DiT [21] blocks, which includes 28 transformer blocks. Through visualization and statistical analysis (see Fig. 12 and Fig. 11), we observe that the features in the first several transformer blocks change minimally (i.e., the first 18 blocks), similar to the Encoder in SD, while the features in the remaining transformer blocks exhibit substantial variations (i.e., the remaining 10 blocks), akin to the Decoder in SD. For ease of presentation, we refer to the first several transformer blocks of DiT as the Encoder and the remaining transformer blocks as the Decoder. In Tab. 3, we demonstrate the accelerating performance of our method while applied to DiT-based generation models.

#### A.4 Time and memory consumption ratios

We report the run time and GPU memory consumption ratios for text-to-image task. As shown in Tab. 8, we significantly accelerate the diffusion sampling with encoder propagation while maintaining a comparable memory demand to the baselines (Tab. 8 (the last two columns)). Specifically, our proposed method reduces the spending time (s/image) by 24% and requires a little additional memory compared to standard DDIM sampling in SD (DDIM vs. Ours: 2.62GB vs. 2.64GB). The increased GPU memory requirement is for caching the features of the encoder from the previous time-step. Though applying parallel encoder propagation results in an increase in memory requirements by 51%, it leads to a more remarkable acceleration of 41% (DDIM vs. Ours: 2.62GB vs. 3.95GB). In conclusion, applying encoder propagation reduces the sampling time, accompanied by a negligible increase in memory requirements. Parallel encoder propagation on text-to-image tasks yields a sampling speed improvement of 20% to 43%, requiring an additional acceptable amount of memory.

Besides the standard text-to-image task, we also validate our proposed approach on other tasks: *text-to-video generation* (i.e., Text2Video-zero [4] and VideoFusion [5]), *personalized generation* (i.e., Dreambooth [7] and Custom Diffusion [8]) and *reference-guided image generation* (i.e., ControlNet [10]). We present the time and memory consumption ratios for these tasks in Tab. 9.

As reported in Tab. 9 (top), when combined with our method, there is a reduction in sampling time by 23% and 32% for Text2Video-zero [4] and VideoFusion [5], respectively, while the memory requirements increased slightly by 3% (0.2GB) and 0.9% (0.11GB).

The time spent by reference-guided image generation (i.e., ControlNet [10]) is reduced by more than 20% with a negligible increase in memory (1%). When integrated with our parallel encoder propagation, the sampling time in this task can be reduced by more than half (51%) (Tab. 9 (middle)). Dreambooth [7] and Custom Diffusion [8] are two approaches for customizing tasks by fine-tuning text-to-image diffusion models. As reported in Tab. 9 (bottom), our method, working in conjunction with the two customization approaches, accelerates the image generation with an acceptable increase in memory.

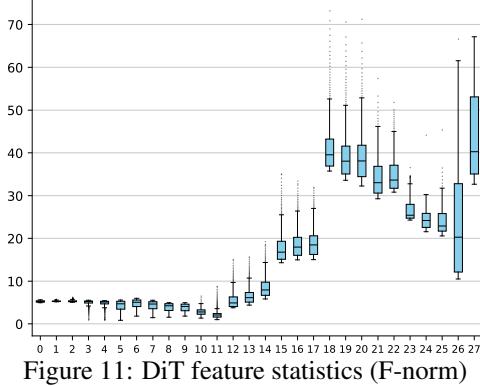


Figure 11: DiT feature statistics (F-norm)

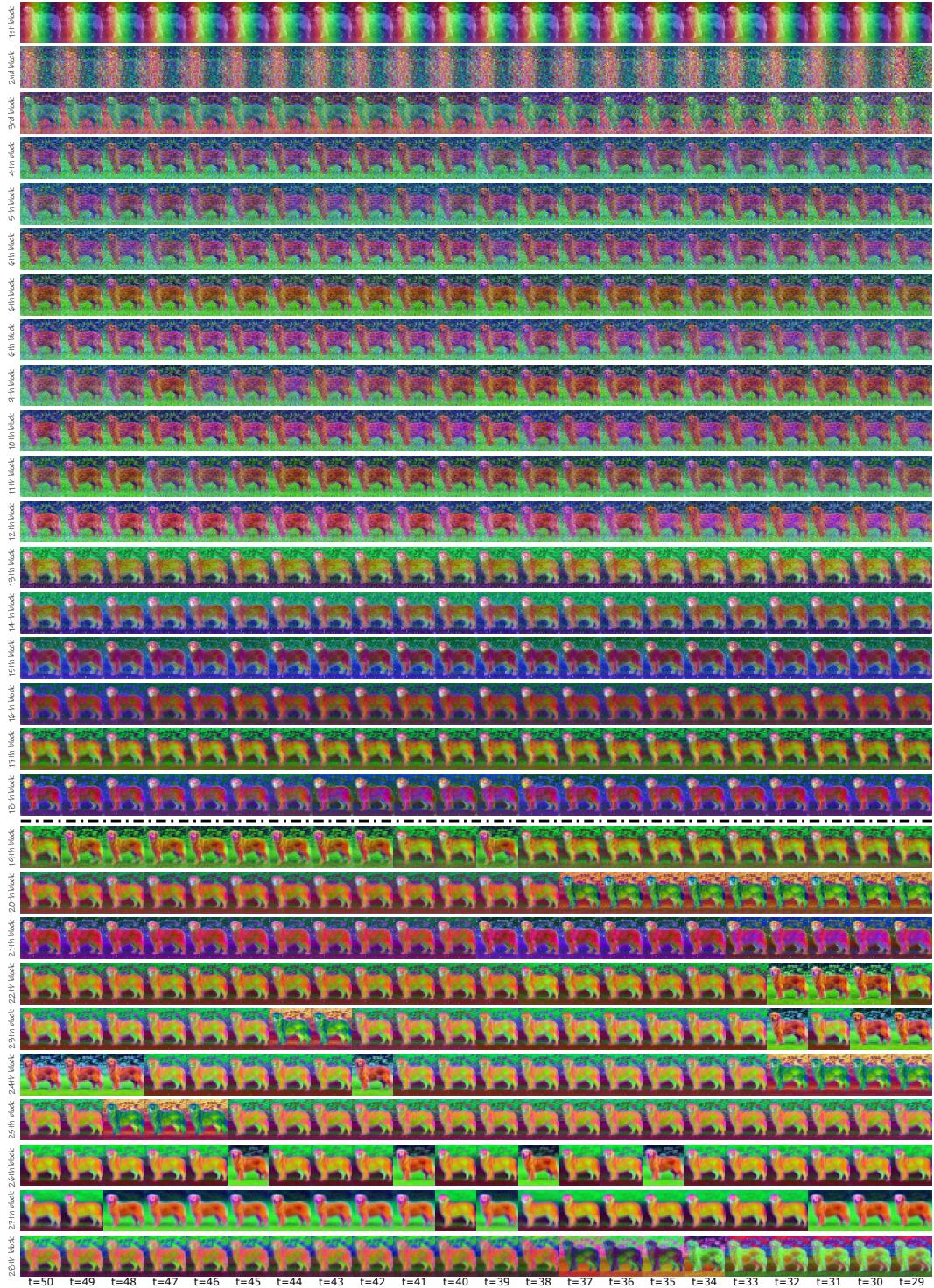


Figure 12: Visualising the hierarchical features. We apply PCA to the hierarchical features following PnP [20], and use the top three leading components as an RGB image for visualization. The output features of the first 18 blocks change slowly, and show similarity at many time steps (top), while the ones in the remaining 10 blocks exhibit substantial variations across different time steps (bottom).

Note that our method, either utilizing encoder propagation or parallel encoder propagation, improves sampling speed without compromising image quality (Sec. 4.1 and Sec. 4.3). We

Table 8: Time and GPU memory consumption ratios in both SD model and DeepFloyd-IF diffusion model.  $\dagger$ : Encoder propagation,  $\ddagger$ : Parallel encoder propagation.

DM	Sampling Method	T	s/image	memory (GB)
Stable Diffusion	DDIM [13]		2.42	2.62
	DDIM [13] w/ Ours	$\dagger$ 50 $\ddagger$	<b>1.82</b> <sub>24%<math>\downarrow</math></sub> <b>1.42</b> <sub>41%<math>\downarrow</math></sub>	2.64 3.95
	DPM-Solver [14]		1.14	2.62
	DPM-Solver [14] w/ Ours	$\dagger$ 20 $\ddagger$	<b>0.92</b> <sub>19%<math>\downarrow</math></sub> <b>0.64</b> <sub>43%<math>\downarrow</math></sub>	2.64 2.69
	DPM-Solver++ [47]		1.13	2.61
	DPM-Solver++ [47] w/ Ours	$\dagger$ 20 $\ddagger$	<b>0.91</b> <sub>19%<math>\downarrow</math></sub> <b>0.64</b> <sub>43%<math>\downarrow</math></sub>	2.65 2.68
	DDIM + ToMe [35]		2.26	2.62
	DDIM + ToMe [35] w/ Ours	$\dagger$ 50 $\ddagger$	<b>1.72</b> <sub>24%<math>\downarrow</math></sub> <b>1.33</b> <sub>41%<math>\downarrow</math></sub>	2.64 3.95
	DDPM [2]		34.55	40.5
	DDPM [2] w/ Ours	$\dagger$ 225 $\ddagger$	<b>29.45</b> <sub>15%<math>\downarrow</math></sub> <b>26.27</b> <sub>24%<math>\downarrow</math></sub>	41.1 41.1
DeepFloyd-IF	DPM-Solver++ [47]		16.09	40.5
	DPM-Solver++ [47] w/ Ours	$\dagger$ 100 $\ddagger$	<b>14.13</b> <sub>12%<math>\downarrow</math></sub> <b>12.97</b> <sub>20%<math>\downarrow</math></sub>	40.8 40.8

Table 9: Time and GPU memory consumption ratios in text-to-video, personalized generation and reference-guided generated tasks.  $\dagger$ : Encoder propagation,  $\ddagger$ : Parallel encoder propagation.

Sampling Method	T	s/image	memory (GB)
Text2Video-zero [4]	$\dagger$ 50 $\ddagger$	13.65	6.59
Text2Video-zero [4] w/ Ours		<b>10.54</b> <sub>23%<math>\downarrow</math></sub> -	6.79 -
VideoFusion [5]	$\dagger$ 50 $\ddagger$	17.93	11.87
VideoFusion [5] w/ Ours		<b>12.27</b> <sub>32%<math>\downarrow</math></sub> -	11.98 -
ControlNet [10] (edges)		3.20	3.81
ControlNet [10] (edges) w/ Ours	$\dagger$ 50 $\ddagger$	<b>2.01</b> <sub>37%<math>\downarrow</math></sub> <b>1.52</b> <sub>51%<math>\downarrow</math></sub>	3.85 5.09
ControlNet [10] (scribble)		3.95	3.53
ControlNet [10] (scribble) w/ Ours	$\dagger$ 50 $\ddagger$	<b>3.18</b> <sub>20%<math>\downarrow</math></sub> <b>1.93</b> <sub>51%<math>\downarrow</math></sub>	3.57 4.45
Dreambooth [7]		2.42	2.61
Dreambooth [7] w/ Ours	$\dagger$ 50 $\ddagger$	<b>1.81</b> <sub>24%<math>\downarrow</math></sub> <b>1.42</b> <sub>41%<math>\downarrow</math></sub>	2.65 3.93
Custom Diffusion [8]		2.42	2.61
Custom Diffusion [8] w/ Ours	$\dagger$ 50 $\ddagger$	<b>1.82</b> <sub>24%<math>\downarrow</math></sub> <b>1.42</b> <sub>41%<math>\downarrow</math></sub>	2.64 3.94

conducted GPU memory consumption ratios using the official method provided by PyTorch, `torch.cuda.max_memory_allocated`<sup>10</sup>, which records the peak allocated memory since the start of the program.

## A.5 GFLOPs

We use the fvcore<sup>11</sup> library to calculate the GFLOPs required for a single forward pass of the diffusion model. Multiplying this by the number of sampling steps gives us the total computational burden for sampling one image. Additionally, using fvcore, we can determine the computational load required for each layer of the diffusion model. Based on the key time-steps set in our experiments, we subtract the computation we save from the original total computational load, which then represents the GFLOPs required by our method. Similarly, fvcore also supports parameter count statistics.

## A.6 Baseline Implementations

For the comparisons of text-to-image generation in Sec. 4, we use the official implementation of DPM-Solver [14], DPM-Solver++ [47]<sup>12</sup>, and ToMe [35]<sup>13</sup>. For the other tasks with text-guided diffusion model, we use the official implementation of Text2Video-zero [4]<sup>14</sup>, VideoFusion [5]<sup>15</sup>, ControlNet [10]<sup>16</sup>, Dreambooth [7]<sup>17</sup>, and Custom Diffusion [8]<sup>18</sup>. We maintain the original implementations of these baselines and directly integrate code into their existing implementations to implement our method.

<sup>10</sup>[https://pytorch.org/docs/stable/generated/torch.cuda.max\\_memory\\_allocated.html](https://pytorch.org/docs/stable/generated/torch.cuda.max_memory_allocated.html)

<sup>11</sup><https://github.com/facebookresearch/fvcore>

<sup>12</sup><https://github.com/LuChengTHU/dpm-solver>

<sup>13</sup><https://github.com/dbolya/tomesd>

<sup>14</sup><https://github.com/Picsart-AI-Research/Text2Video-Zero>

<sup>15</sup>[https://huggingface.co/docs/diffusers/api/pipelines/text\\_to\\_video](https://huggingface.co/docs/diffusers/api/pipelines/text_to_video)

<sup>16</sup><https://github.com/lillyasviel/ControlNet>

<sup>17</sup><https://github.com/google/dreambooth>

<sup>18</sup><https://github.com/adobe-research/custom-diffusion>

Table 10: Model complexity comparison regarding the encoder  $\mathcal{E}$ , the bottleneck  $\mathcal{B}$  and the decoder  $\mathcal{D}$  in terms of parameter count and FLOPs.

	Parameter (billion)	FLOPs (million)
$\mathcal{E} + \mathcal{B}$	0.25 + 0.097	224.2 + 6.04
$\mathcal{D}$	0.52 <sub>1.47x</sub>	504.4 <sub>2.2x</sub>

## B Parameter Count and FLOPs of SD

We take into account model complexity in terms of parameter count and FLOPs (see Tab. 10). It's noteworthy that the decoder  $\mathcal{D}$  exhibits a significantly greater parameter count, totaling 0.51 billion. This figure is approximately 1.47 times the number of parameter combinations for the encoder  $\mathcal{E}$  (250 million) and the bottleneck  $\mathcal{B}$  (97 million). This substantial parameter discrepancy suggests that the decoder  $\mathcal{D}$  carries a more substantial load in terms of model complexity.

Furthermore, when we consider the computational load, during a single forward inference pass of the SD model, the decoder  $\mathcal{D}$  incurs a considerable 504.4 million FLOPs. This value is notably higher, approximately 2.2 times, than the cumulative computational load of the encoder  $\mathcal{E}$ , which amounts to 224.2 million FLOPs, and the bottleneck  $\mathcal{B}$ , which requires 6.04 million FLOPs. This observation strongly implies that the decoder  $\mathcal{D}$  plays a relatively more important role in processing and transforming the data within the UNet architecture, emphasizing its critical part in the overall functionality of the model.

## C The methodology for computing Clipscore

Clipscore [46] is a metric for computing the consistency between text and image. The formula for computing Clip-score between image embedding  $v$  and text embedding  $c$ , as presented in the original paper, is as follows:

$$\text{CLIP-S} = w * \max(\cos(c, v), 0),$$

where a re-scaling operation is performed using  $w$ , set to 2.5 in the official implementation of Clipscore [46]. The rationale behind the re-scaling operation, as provided by the official paper (excerpted from [46], Section 3, Footnote 6), is as follows:

While the cosine similarity, in theory, can range from  $[-1, 1]$  (1) we never observed a negative cosine similarity; and (2) we generally observe values ranging from roughly zero to roughly .4. The particular value of  $w$  we advocate for,  $w = 2.5$ , attempts to stretch the range of the score distribution to  $[0, 1]$ .

Therefore, adhering to the aforementioned setting, we employed the official implementation of Clipscore for evaluation, yielding Clipscore data distributed around 0.75. Some works [8, 48] employ re-scaling operations during evaluation, yielding Clipscore values around 0.75, while others [10, 4] do not, resulting in Clipscore values around 0.3. In essence, both approaches are equivalent, differing only in whether re-scaling is applied at the end. We evaluate Clipscore using  $w = 2.5$  for re-scaling, hence yielding Clipscore data around 0.75, unless otherwise stated.

## D Different from step-reduction methods

Several efficient diffusion model solvers, such as DDIM [13], DPM-Solver [14] and DPM-Solver++ [47], have significantly reduced sampling steps. Our method does not reduce the number of sampling steps. In the encoder propagation, the decoder needs to compute for all time-steps, and we require time embedding inputs for all time-steps to maintain temporal coherence.

Fig. 13 illustrates the qualitative comparison results of SD (DDIM) showcasing a reduction in steps (i.e., 9 and 25 time-steps). We present the results with 9 steps because the number of *key* time-steps for the SD model with DDIM in our approach is 9. The generation quality notably declines with step reduction, attributed to alterations in image structure and a diminished attention to detail generation,



Figure 13: When lowering the time-steps in inference, the image quality noticeably deteriorates, while ours maintains a similar image quality to the original.



Figure 14: Fewer sampling steps with noise injection.

Table 11: Quantitative comparison for DDIM with fewer steps.

Sampling method	T	<b>FID</b> ↓	<b>Clipscore</b> ↑	s/image ↓
DDIM	50	21.75	0.773	2.42
DDIM	25	22.16	0.761	1.54
DDIM	25	21.89	0.761	1.54
w/ noise injection	9	27.58	0.735	0.96
DDIM	9	27.63	0.736	0.96
w/ noise injection	50	21.08	0.783	1.42

Table 12: Quantitative comparison for DPM-Solver/DPM-solver++ with fewer steps.

Sampling method	T	<b>FID</b> ↓	<b>Clipscore</b> ↑
DPM-Solver	20	21.36	0.780
DPM-Solver	10	22.41	0.768
DPM-Solver w/ ours	20	21.25	0.779
DPM-Solver++	20	20.51	0.782
DPM-Solver++	10	21.25	0.771
DPM-Solver++ w/ ours	20	20.76	0.781

as exemplified by the **hands** in Fig. 13 (the second column). We increase the number of sampling steps (i.e., 25 steps), and find that the sampling results do not perform as well as DDIM with 50 steps and its variation with FasterDiffusion (Fig. 13 and Tab. 11). As shown in Tab. 11, the FID and Clipscore of SD (DDIM) with 25 time-steps on the MS-COCO 2017 10K subset are 22.16 and 0.761, respectively, significantly worse than the results obtained with 50 time-steps and ours (FID↓: 21.75, 21.08; Clipscore↑: 0.773, 0.783). This demonstrates that our method is not simply reducing the number of sampling steps. With fewer steps (T=10), both DPM-Solver and DPM-Solver++ also

exhibit worse FID and Clipscore (see Tab. 12). These results indicate that our method achieves better performance compared to simply reducing the sampling steps.

As shown on Tab. 11, we conduct DDIM scheduler inference time with various steps. Although FasterDiffusion (the last row) is slightly longer than DDIM scheduler with 9 steps, our sampling results are much closer the 50-step DDIM generation quality, whereas the 9-step DDIM results are much inferior (see Tab. 11, Tab. 12, and Fig. 13 for examples).

For directly applying noise injection to the generation phase, we show image generation examples in Fig. 14 in the rebuttal file. Fewer sampling steps equipped with prior noise injection result in almost no improvement in image quality. Tab. 11 further supports this conclusion. While in our case, FasterDiffusion working without the noise injection will lead to smoother textures, as shown in Fig.6. The noise injection technique helps in preserving fidelity in the generated results.

## E Difference between “prior noise injection” and “churn”

In EDM [49] (Karras et al.), they increase the noise level during ODE sampling to improve deterministic sampling, which is referred to as stochastic sampling (i.e., “churn”). Compared to deterministic sampling, it injects noise into the image at each step, which helps to enhance the quality of the generated images. Song et al. [50] first observed that perturbing data with random Gaussian noise makes the data distribution more amenable to score-based generative modeling. Increasing the noise level in EDM is based on Song’s paper.

Their purpose for injecting noise is to perturb the data, whereas our purpose for injecting noise during sampling is to preserve high-frequency details in the image during the denoising process, preventing the diffusion model from removing high-frequency information as noise (see Figure.6 in the main paper). In addition, our method FasterDiffusion differs from them by only inserting noise in the later stage of the diffusion steps, while they insert noise to all time-steps.

## F Ablation Experiments and Additional Results

### F.1 Comparison with DeepCache

DeepCache is developed based on the observation over the temporal consistency between high-level features. In this paper, we have a more thorough analytical study over the SD model features as shown in Fig.3, where we find out the encoder features change less, whereas the decoder features exhibit substantial variations across different time steps. This insight motivates us to omit encoder computation at certain adjacent time-steps and reuse encoder features of previous time-steps as input to the decoder in multiple time-steps. We further determine the key time steps in the T2I inference stage, which helps our method to skip the time steps in a more scientific way.

It is also important to note that DeepCache is not parallelizable on multiple GPUs since DeepCache needs to use all or part of the encoder and decoder at every time step. FasterDiffusion, on the other hand, only uses the encoder at the *key* time steps, which enables parallel processing at these time steps and faster inference time cost.

It is also worth to notice that FasterDiffusion can be further combined with a wide range of diffusion model based tasks, including Text2Video-zero, VideoFusion, Dreambooth, and ControlNet. In this case, DeepCache often shows much slower speed while applied to these tasks. As an example shown in Table. 13, when combined with ControlNet DeepCache is slower by 24% compared to the FasterDiffusion.

More specifically, since the ControlNet model requires an additional encoder, our method FasterDiffusion is able to execute this extra encoder in a parallel manner and reuse it, and that makes the additional time negligible. On the other hand, DeepCache is reusing the decoder feature, which leads the ControlNet to wait for the additional encoder to complete computation at each time-step, resulting in almost no time saved from skipping the encoder in the standard SD models. The T2I generation results are shown in Figure. 15, FasterDiffusion successfully preserves the given structure information and achieves similar results as the original ControlNet model.

Table 13: When combined with ControlNet (Edge) 50-step DDIM, our inference time shows a significant advantage compared to DeepCache.

	Clipscore↑	FID↓	s/image↓
ContrlNet	0.769	13.78	3.20
ContrlNet w/ DeepCache	0.765	14.18	1.89 (1.69x)
ContrlNet w/ Ours	0.767	14.65	<b>1.52 (2.10x)</b>



Figure 15: Combine our method with ControlNet.

## F.2 The definition of *key* time-steps in various tasks

We utilize 50, 20 and 20 time-steps for DDIM scheduler [13], DPM-solver [14] and DPM-solver++ [47], respectively. We follow the official implementation of DeepFloyd-IF, setting the time-steps for the three sampling stages to 100, 50, and 75, respectively.

**Text-to-image generation.** We experimentally define the *key* time-steps as  $t^{key} = \{50, 49, 48, 47, 45, 40, 35, 25, 15\}$  for SD model with DDIM [13], and  $t^{key} = \{100, 99, 98, \dots, 92, 91, 90, 85, 80, \dots, 25, 20, 15, 14, 13, \dots, 2, 1\}$ ,  $\{50, 49, \dots, 2, 1\}$  and  $\{75, 73, 70, 66, 61, 55, 48, 40, 31, 21, 10\}$  for three stages of DeepFloyd-IF with DDPM [2]. For SD with both DPM-Solver [14] and DPM-Solver++ [47], we experimentally set the *key* time-steps to  $t^{key} = \{20, 19, 18, 17, 15, 10, 5\}$ .

**Other tasks with text-guided diffusion model.** In addition to standard text-to-image tasks, we further validate our approach on other tasks with text-guided diffusion model (Sec. 4.2). These tasks are all based on the SD implementation of DDIM [13]. Through experiments, we set the *key* time-steps to  $t^{key} = \{50, 49, 48, 47, 45, 40, 35, 25, 15\}$  for Text2Video-zero [4], VideoFusion [5], and ControlNet [10]. For personalized tasks (i.e., Dreambooth [7] and Custom Diffusion [8]), we set the *key* time steps to  $t^{key} = \{50, 49, 48, 47, 45, 40, 35, 25, 15, 10\}$ .

## F.3 The effectiveness of encoder propagation

In Sec. 3.2, we have demonstrated that encoder propagation (Fig. 16c) can preserve semantic consistency with standard SD sampling (Fig. 16a). However, images generated through decoder propagation often fail to match certain specific objects mentioned in the text prompt (Fig. 16d and Tab. 14 (the third row)).

We extended this strategy to include encoder and decoder propagation (Fig. 16e) as well as decoder and encoder dropping (Fig. 16f). Similarly, encoder and decoder propagation often fail to cover specific objects mentioned in the text prompt, leading to a degradation in the quality of the generated results (Fig. 16e and Tab. 14 (the fourth row)). On the other hand, decoder and encoder dropping are unable to completely denoise, resulting in the generation of images with noise (Fig. 16f and Tab. 14 (the fifth row)). Note that decoder and encoder dropping (Fig. 16f) is different from directly reducing the number of time-steps. Time embedding is required as input for each time-step, even when using only the encoder or decoder at each time-step.

## F.4 User study details

The study participants were volunteers from our college. The questionnaire consisted of 35 questions, each presenting two images: one from the baseline methods (including ControlNet, SD, DeepFloyd-

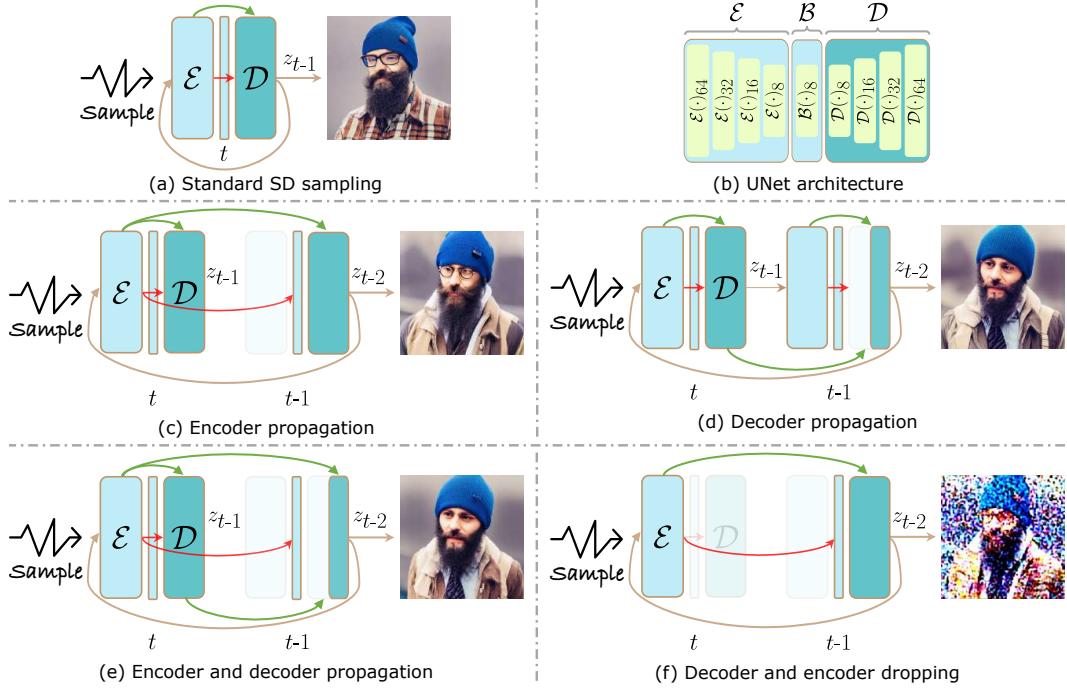


Figure 16: (a) Standard SD sampling. (b) UNet architecture. (c) Encoder propagation. (d) Decoder propagation. (e) Encoder and decoder propagation. (f) Decoder and encoder dropping.

Table 14: Quantitative evaluation for additional strategy on MS-COCO 2017 10K subset. Other propagation strategies can lead to the loss of some semantics under prompts and degradation of image quality (the third to fifth rows).

Sampling method	T	FID $\downarrow$	Clipscore $\uparrow$
DDIM	50	21.75	0.773
DDIM w/ Encoder propagation (Ours)	50	<b>21.08</b>	<b>0.783</b>
DDIM w/ Decoder propagation	50	22.97	0.758
DDIM w/ Encoder and decoder propagation	50	23.69	0.742
DDIM w/ Decoder and encoder dropping	50	199.48	0.679

IF, DPM-Solver++, Dreambooth, Custom Diffusion, Text-to-Video, etc.) and the other from our method FasterDiffusion (one example shown in Fig. 17).

Users were required to select the image where the target was more accurately portrayed, or choose “both equally good”. A total of 18 users participated the questionnaire, resulting in totally 630 samples (35 questions  $\times$  1 option  $\times$  18 users). As the final results shown in Fig.8, the chosen percentages for SD, DeepFloyd-IF, DPM-Solver++, Text-to-Video, ControlNet, and Personalize were 48%, 44%, 51%, 52%, 46%, and 39%, respectively. These results show that our method performs on par with the baseline methods and demonstrate that FasterDiffusion retains the T2I generation quality while reducing the inference time.

## F.5 Additional metrics

We showcase our experimental results with the ImageReward [51] and PickScore [52] evaluation metrics over the MS-COCO2017 10K dataset, as shown in the Tab. 15. Our method FasterDiffusion enhances sampling efficiency while preserving the original model performance.

\* **12** Please select a high-quality image that matches the specified prompt.

\* **30** Please select a high-quality video that matches the specified prompt.

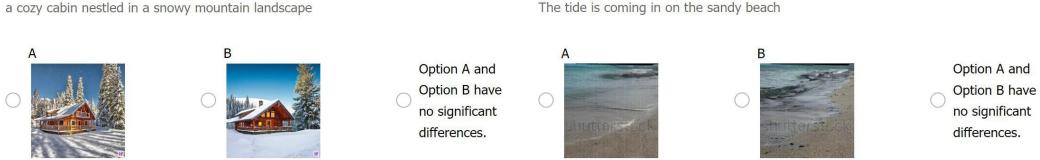


Figure 17: User study examples.

	ImageReward↑	PickScore↑	s/image
SD (DDIM)	0.149	52.10%	2.42
SD (DDIM) w/ Ours	0.162	47.90%	1.42

Table 15: Metrics in ImageReward and PickScore.

## F.6 Additional results

Fig. 18 shows additional results generated by DiT. In Figs. 19, 20 and 21, we show additional text-to-image generation results. Further results regarding other tasks with text-guided diffusion model are illustrated in Figs. 22, 23 and 24.

## F.7 Additional tasks

**ReVersion** [53] is a relation inversion method that relies on the Stable Diffusion (SD). Our method retains the capacity to generate images with specific relations based on exemplar images, as illustrated in Fig. 25 (top).

**P2P** [27] is an image editing method guided solely by text. When combined with our approach, it enhances sampling efficiency while preserving the editing effect (Fig. 25 (bottom)).

## G Automatic selection for key time steps

To identify key time steps, we conducted empirical analysis of feature changes at adjacent time steps in multistep Stable Diffusion models (50-step model as an example). This analysis is based on statistics from the distribution of features across 100 random prompts, which does not impose a high time cost. We found that the encoder features change minimally in later time steps, whereas the encoder features in earlier time steps exhibit substantial variations compared to later ones. Based on analysis as shown in Section 3.2, we determine the *key* time-step as  $t^{key} = \{50, 49, 48, 47, 45, 40, 35, 25, 15\}$  for the 50-step Stable Diffusion model. As a general case, we also applied this *key* time-step configuration to the ControlNet [10], Text2Video-zero [4], VideoFusion [5], Dreambooth [7] and Custom Diffusion [8] models based on the 50-step SD models. That will not impose any additional searching time for the key time steps for these downstream application scenarios.

Existing methods such as OMS-DPM [54], AutoDiffusion [55], and DDSM [56] utilize reinforcement learning or machine learning algorithms to search for optimal schedules, time steps, and model sizes. Each search iteration involves frequent image generation and FID calculations. Executing these search algorithms is time-consuming, often exceeding the training time, as noted in DDSM. For instance, the NSGA-II search algorithm used in DDSM incurs a search cost approximately 1.1 to 1.2 times that of training a standard diffusion model. We plan to incorporate the NSGA-II algorithm for automatically searching  $t^{key}$  time steps as an enhancement to FasterDiffusion in our future research.

## H Impact Statements

Diffusion models generate realistic fake images, which assist individuals in their creative endeavors. However, for those lacking the ability to discern whether the images are generated by a model, it may affect their judgment.



Figure 18: Additional results of DiT (top) and this method in conjunction with our proposed approach (bottom).

Equipped with our technique, using DMs on the edge devices becomes feasible. Also, importantly, our proposed acceleration of DMs does not require access to enormous compute, which is required to the methods based on knowledge distillation [18, 17, 19]. This makes our technique also useful for the many smaller actors in the generative AI field.

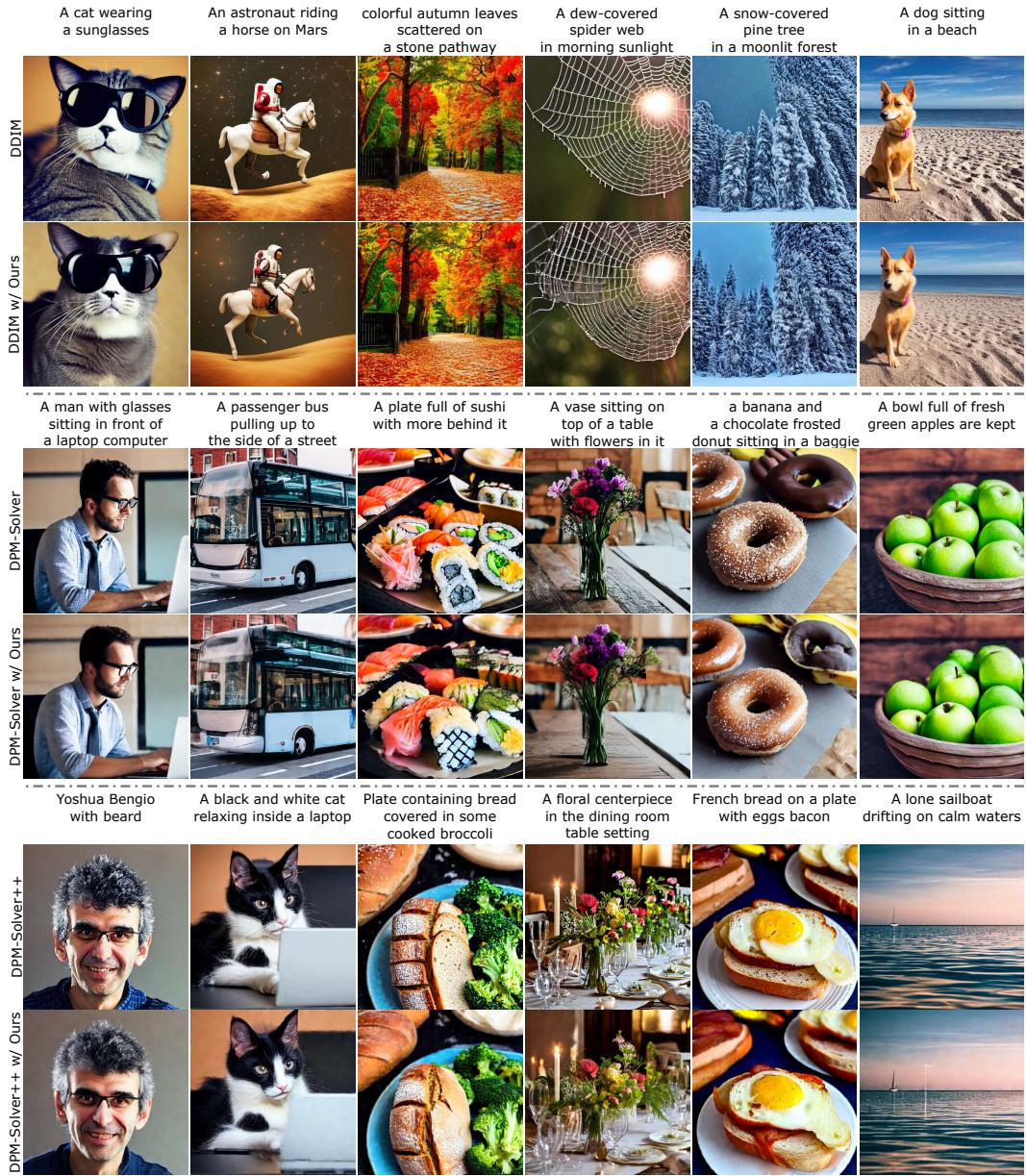


Figure 19: Additional results of text-to-image generation combining SD with DDIM [13], DPM-Solver [14], DPM-Solver++ [47], and these methods in conjunction with our proposed approach.

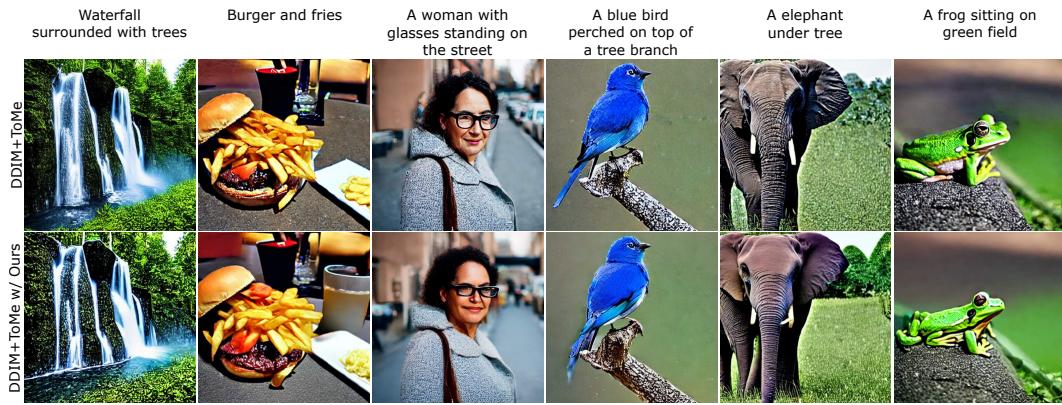


Figure 20: Additional results of text-to-image generation combining SD with DDIM+ToMe [35], and this method in conjunction with our proposed approach.

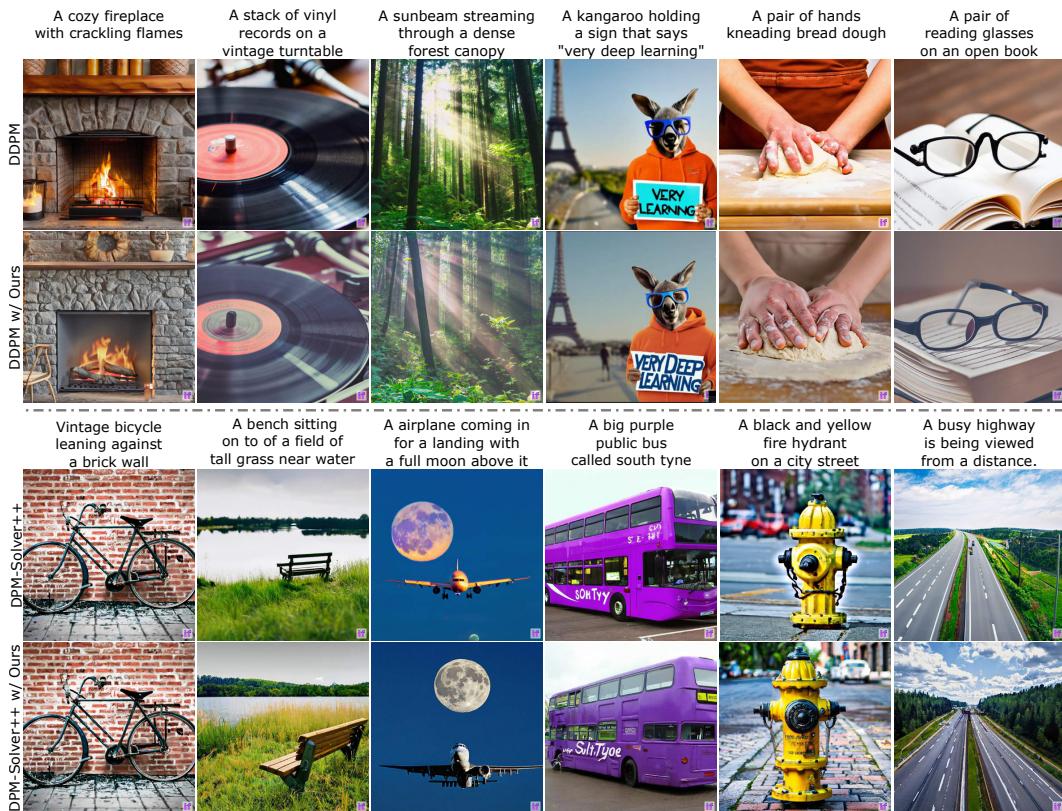


Figure 21: Additional results of text-to-image generation combining DeepFloyd-IF with DDPM [2] and DPM-Solver++ [47], and these methods in conjunction with our proposed approach.

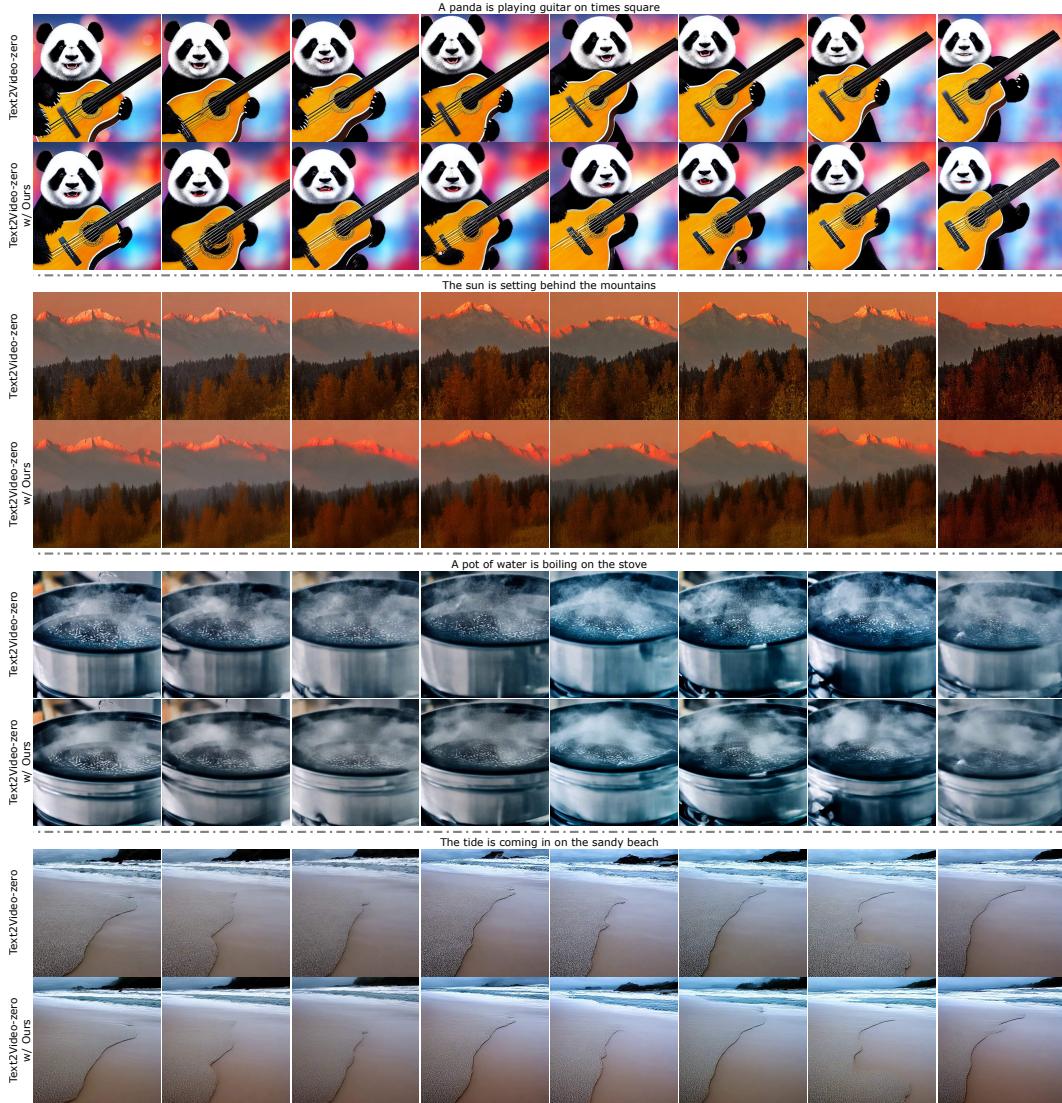


Figure 22: Additional results of Text2Video-zero [4] both independently and when combined with our proposed method.



Figure 23: Additional results of VideoFusion [5] both independently and when combined with our proposed method.

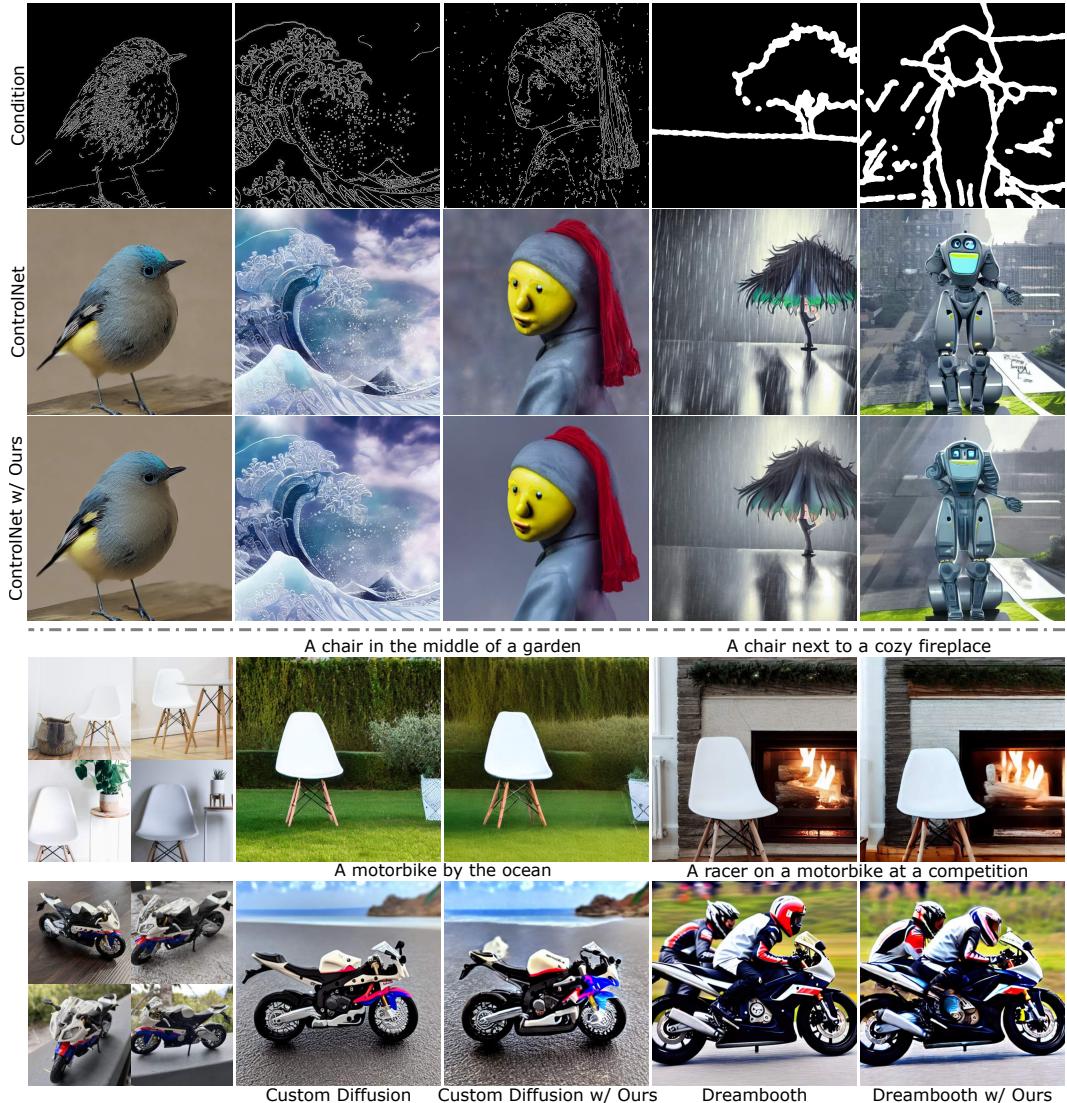


Figure 24: Additional results of ControlNet [10] (top) and personalized tasks [7, 8] (bottom) obtained both independently and in conjunction with our proposed method..

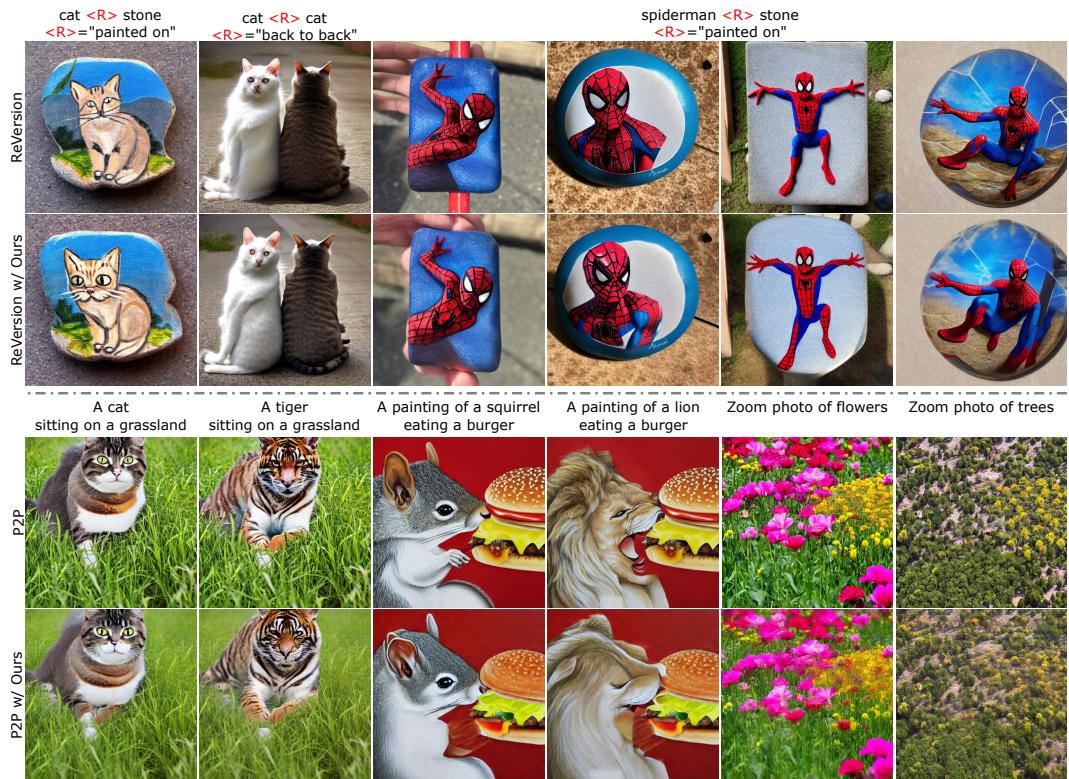


Figure 25: Edited results of ReVersion [53] (top) and P2P [27] (bottom) obtained both independently and in conjunction with our proposed method.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: See Abstract and Introduction (Section 1).

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: See Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: See Section 4 and Appendix A. We will release our code to the public.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: We will provide our code in supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: See Section 4 for the experimental setting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: See Figure 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: See Tables 1, 4, 5, 8 and 9.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: The research conducted in the paper conforms in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: See Appendix H.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite papers that we used in Section 4. The models, datasets, and codes used in the paper are introduced in detail in Appendix A.1 and Appendix A.6.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We provide our source code as an anonymized zip file in the supplemental material.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[Yes\]](#)

Justification: See User Study in Section [4.3](#).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.