Noise Calibration: Plug-and-play Content-Preserving Video Enhancement using Pre-trained Video Diffusion Models

Qinyu Yang[1], Haoxin Chen[2], Yong Zhang[2,*], Menghan Xia[2], Xiaodong Cun[2], Zhixun Su[1,*], and Ying Shan[2]

[1] Dalian University of Technology
[2] Tencent AI Lab
https://github.com/yangqy1110/NC-SDEdit/

Abstract. In order to improve the quality of synthesized videos, currently, one predominant method involves retraining an expert diffusion model and then implementing a noising-denoising process for refinement. Despite the significant training costs, maintaining consistency of content between the original and enhanced videos remains a major challenge. To tackle this challenge, we propose a novel formulation that considers both visual quality and consistency of content. Consistency of content is ensured by a proposed loss function that maintains the structure of the input, while visual quality is improved by utilizing the denoising process of pretrained diffusion models. To address the formulated optimization problem, we have developed a plug-and-play noise optimization strategy, referred to as **Noise Calibration**. By refining the initial random noise through a few iterations, the content of original video can be largely preserved, and the enhancement effect demonstrates a notable improvement. Extensive experiments have demonstrated the effectiveness of the proposed method.

Keywords: Diffusion Models · Video Enhancement · Plug-and-play

1 Introduction

Recently, diffusion models have emerged as a distinct type of generative models, in contrast to traditional Generative Adversarial Networks (GANs) [17] and Variational Autoencoders (VAEs) [29]. These models [22, 54] have demonstrated superior performance across a wide range of applications. In particular, visual synthesis has significantly benefited from the development of diffusion models. A popular subset of methods [2, 11, 19, 40, 42, 54, 74, 76, 79, 90, 96] leverages pretrained Text-to-Image (T2I) models and incorporates additional temporal blocks to extend video generation capabilities. However, obtaining results with both accurate semantics and high visual quality through a single inference often proves

* Corresponding author

Fig. 1: Examples demonstrating video enhancement based on SDEdit

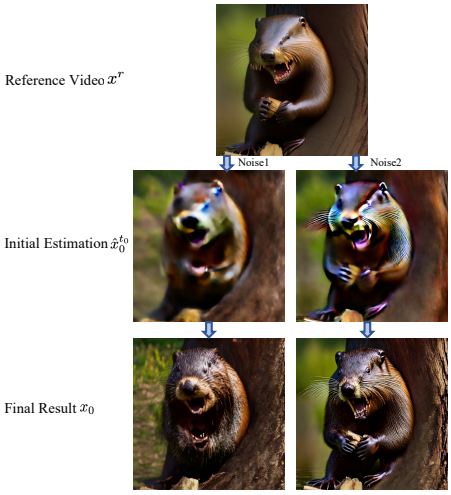

Fig. 2: Decomposition of the video enhancement process based on SDEdit

challenging. Typically, a low-quality video with diverse and semantically accurate motions is generated from the base model, and then an expert model is retrained to implement a noising-denoising process, as pioneered by SDEdit [41], to refine the generated video. The expert model is further trained using high-quality videos with earlier noise scales, aiming to amplify the model's attention to spatio-temporal details. This approach is anticipated to further improve the spatio-temporal continuity and clarity of the video, effectively addressing artifacts in both time and space dimensions.

However, despite the substantial resources invested in retraining, the well-established structures in the original video are often disrupted during the process of quality enhancement. Consequently, we believe that solely utilizing pre-trained Text-to-Video (T2V) diffusion models for content-preserving video enhancement is a worthwhile research direction. This T2V model does not need to generate videos with accurate semantic motions; however, it is required to produce videos with high visual quality. By only using SDEdit for video enhancement based on pre-trained T2V models, it is common for the spatial structure of the original video to be disrupted when the initial denoising step is set sufficiently large to achieve satisfactory quality enhancement, as shown in Fig. 1. For instance, in the "camel" example, while the quality is enhanced, brown camel hair appears to be flying across the sky. In the "rubber duck" example, the duck transforms from a static object into a dynamic creature, and the sunlight is transformed into a swaying tail.

To address this challenge, we propose a novel formulation for video enhancement that ensures both quality improvement and consistency with the content of original video. Specifically, noise is first introduced to disrupt a given video,

which is then gradually eliminated based on a pre-trained model. By leveraging the ability of the pre-trained model to generate high-quality videos, we enhance the quality of the original video. To ensure consistency, we propose an additional loss function that constrains the content loss between the enhancement result and the original video. To solve this optimization problem, we provide a simple but effective solution, **Noise Calibration**. By refining the initial random noise only 1-3 times before adding it to the original video, we can largely preserve the content of original video and significantly improve the enhancement effect.

A significant amount of theoretical analysis and experiments demonstrate that our method can effectively preserve the content of videos before and after enhancement when using pre-trained T2V models for video enhancement. Furthermore, this approach can conveniently serve as a plug-in to enhance the performance of state-of-the-art visual refinement models. We summarize our contributions as follows:

- We introduce a novel formulation for video enhancement based on diffusion models, which focuses not only on improving quality but also on maintaining consistency of content with the original video.
- We propose a concise yet effective content-preserving strategy for video enhancement, called **Noise Calibration**, which only requires calibrating the initial random noise, without any additional fine-tuning or operations.
- Extensive quantitative and qualitative experiments demonstrate that Noise Calibration can be effectively applied to video enhancement and various tasks based on SDEdit, achieving more controllable image/video generation.

2 Related Work

The aim of this study is to investigate how to maintain consistency of content between the enhancement results and the original video while performing video enhancement based on SDEdit. Consequently, we will briefly review relevant domains in this section to facilitate a better understanding.

2.1 Video Diffusion Models

The recent emergence of diffusion models [15, 22, 47, 62, 63, 66, 68, 82] as a type of generative model [14, 17, 30, 31, 45, 48, 53] has significantly advanced the field of T2I generation [6, 15, 16, 22, 25, 33, 46, 47, 49, 56, 60, 62, 63, 65–68, 71, 82]. These models have also demonstrated potential in various tasks, such as image-to-image translation [12, 32, 41, 55, 75, 83, 95], image super-resolution [10, 36, 57, 73, 86, 91], image inpainting [3, 4, 35, 46, 87], and image editing [8, 20, 27, 37–39, 44, 50, 70], among others.

Text-to-video synthesis is a complex and challenging task with significant practical implications, as it aims to generate relevant videos from textual descriptions. Early approaches [5, 61, 69, 77, 78, 80] primarily utilized GANs, which unfortunately resulted in subpar video quality. As a pioneering work introducing diffusion models to the field of video generation, Video Diffusion Models [23]

adopted the 3D U-Net from [13], achieving impressive results in both unconditional and text-conditional video generation tasks. Subsequently, to reduce training costs, a significant number of studies [2,11,19,40,42,54,74,76,79,90,96] have extended pre-trained image diffusion models to video and learned Video Diffusion Models in latent or hybrid-pixel-latent spaces.

Given the complexity of video generation, limited computational resources, and the scarcity of high-quality video data, cascade models [21,24,79,92,93] have emerged as the mainstream paradigm, adopting a divide-and-conquer approach to tackle these challenges. Specifically, a cascade model typically comprises three components: a base model, a frame interpolation model, and a refinement model. Based on the base model trained with a large number of low-resolution videos, we can generate well-structured low-resolution videos. Subsequently, the frame interpolation model enhances the video's continuity by adding frames. Finally, a refinement model is employed to further improve the spatio-temporal continuity and clarity of the video.

2.2 Refinement Models of Video Diffusion Models

As mentioned above, refinement models play a crucial role in determining the final quality of generated videos. In this section, we will review the methods employed by existing refinement models. Refinement models differ from conventional super-resolution models, which solely focus on increasing the resolution. A critical aspect of refinement models is their ability to refine the videos generated by the base model, which might lack sufficient details, by adding appropriate details to enhance the overall quality. There are few existing methods, which can primarily be categorized into two approaches: methods based on SR3 [57], which resemble traditional super-resolution techniques, and methods based on SDEdit [41], which lean towards further generation.

As the first diffusion-based Super Resolution method, SR3 [57] incorporates the low-resolution (LR) image as an additional input to the denoising network, constructing a conditional denoising network. To further enhance visual quality of generated videos and increase spatial resolution, Lavie [79] utilizes SD-x4-Upscaler [54] as a prior and incorporates an additional temporal dimension, enabling temporal processing within the diffusion UNet. Although the additional initial video embedding greatly ensures consistency of content before and after enhancement, it also restricts the refinement capabilities of diffusion models.

SDEdit [41] is a pioneering approach that achieves editing through iterative denoising via a stochastic differential equation (SDE). Initially designed to address the Stroke Painting to Image problem, SDEdit's impressive scalability has since facilitated advancements in various other fields [52, 84, 89]. In the realm of video generation, both Show-1 [92] and Modelscope [93] have successfully implemented SDEdit to enhance the quality of videos produced by their base models. To amplify the refinement model's focus on spatio-temporal details, they specifically train it on low noise scales, using high-resolution videos. Although this method possesses a stronger generative capability and can enrich video details more effectively, the randomness inherent in the generation process may

lead to the final video deviating from the original content or even damaging well-established structures. To address this issue, we propose a training-free and plug-and-play method that aims to enhance the consistency of content between the final video and the original video without compromising the refinement quality. This method is versatile, compatible with both pre-trained T2V models and expert models retrained on a low noise scale.

Despite SDEdit's wide application across various tasks, the trade-off between realism and fidelity often falls short of user expectations in practical applications, as noted in several studies [1, 7, 18, 26, 28, 43, 88, 94]. Despite this, there is a noticeable lack of research in this area. Peng et al. [51] proposed a method that uses source semantics to guide the generation process, aiming to enhance the consistency between the source and target domain content in SDEdit's image translation tasks. However, this method's effectiveness is contingent on the accuracy of semantic maps, and its applicability in the latent space is yet to be assessed. Singh et al. [59] also concentrated on optimizing the subsequent sampling process with the goal of enhancing the realism of the editing results for stroke painting to image. However, a universally applicable method has not been established yet. Our approach, while primarily designed for content-preserving video enhancement, can be easily adapted for other tasks based on SDEdit. Additionally, it is worth noting that our work is partially inspired by ILVR [12], which refines each generation step with low-frequency component of purturbed reference image for controlling the generation of unconditional Diffusion Models.

3 Proposed Methods

3.1 Preliminaries

Diffusion models start from the given image x_0, and then progressively add Gaussian Noise $\epsilon_t \sim \mathcal{N}(0, 1)$. This transformation yields x_t in each timestep t, which can be directly computed as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, \tag{1}$$

where $\bar{\alpha}_t$ represents the diffusion schedule parameters following a given sequence $0 = \bar{\alpha}_T < \bar{\alpha}_{T-1} ... < \bar{\alpha}_1 < \bar{\alpha}_0 = 1$. During inference, diffusion models can synthesize new image by starting with a random noise sample $x_T \sim \mathcal{N}(0, 1)$ and iteratively denoising it. Given a noised image x_t at timestep t, the model predicts the next-step x_{t-1} as follows:

$$\begin{aligned} p_\theta\left(x_{t-1} \mid x_t\right) &= \mathcal{N}\left(\mu_\theta\left(x_t, t\right), \sigma_t \mathbf{I}\right), \\ \mu_\theta\left(x_t, t\right) &= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta\left(x_t, t\right)\right), \end{aligned} \tag{2}$$

where ϵ_θ represents a neural network trained to predict the noise at each step. At each timestep, the noiseless image $\hat{x}_0^t$ can be approximated as:

$$\hat{x}_0^t = \frac{x_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1 - \bar{\alpha}_t} \epsilon_\theta\left(x_t, t\right)}{\sqrt{\bar{\alpha}_t}}. \tag{3}$$

As a pioneering method, SDEdit [41] introduces a reference image x^r to initialize the denoising process at an intermediate step $t_0 \in [0, T]$. This initialization takes the form $x_{t_0} \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_{t_0}} x^r, (1 - \bar{\alpha}_{t_0})\, \mathbb{I}\right)$. The choice of t_0 represents a trade-off between realism ($t_0 \approx T$), understood as producing images in line with the training distribution $p^*(x)$, and faithfulness ($t_0 \approx 0$), emphasizing similarity with the reference image x^r.

3.2 Formulation of Content-Preserving Video Enhancement

As illustrated in Fig. 1, when only using random noise ϵ_{t_0} to perturb the reference video x^r and subsequently applying quality enhancement based on a pre-trained video model $\epsilon_\theta(x_t, t)$, achieving satisfactory enhancement often leads to unintended alterations in content. To tackle this issue, we propose a novel formulation for video enhancement that prioritizes not only visual quality but also imposes constraints on the content loss between the enhanced result x_0 and the reference video x^r, as illustrated below:

$$\begin{array}{ll} \min_{\epsilon_{t_0}} & dist\left(x_0, x^r\right) \\ \text{s.t.} & x_0 \sim P_\theta\left(x \mid x^r, \epsilon_{t_0}\right) \end{array}. \tag{4}$$

Using Noise1 and Noise2 from Fig. 2 as examples, given reference video x^r and pre-trained video model $\epsilon_\theta(x_t, t)$, various initial noises ϵ_{t_0} will generate diverse enhanced videos x_0 based on the training distribution P_θ of the pre-trained video model $\epsilon_\theta(x_t, t)$. The optimization goal is to identify a more appropriate initial noise ϵ_{t_0} that effectively minimizes the content loss, $dist\left(x_0, x^r\right)$, between the reference video x^r and the enhanced video x_0. A smaller value of $dist\left(x_0, x^r\right)$ indicates a higher consistency of content between the two videos.

To derive a specific optimizable form of $dist\left(x_0, x^r\right)$, we decompose the video enhancement process based on SDEdit. Specifically, the reference video x^r is combined with random noise ϵ_{t_0} corresponding to the selected initial denoising step t_0, using the following formula:

$$x_{t_0} = \sqrt{\bar{\alpha}_{t_0}} x^r + \sqrt{1 - \bar{\alpha}_{t_0}} \epsilon_{t_0}, \epsilon_{t_0} \sim \mathcal{N}(0, 1). \tag{5}$$

Based on Eq. (3), the initial estimation $\hat{x}_0^{t_0}$ of the enhancement result at step t_0 can be expressed as:

$$\hat{x}_0^{t_0} = \frac{x_{t_0} - \sqrt{1 - \bar{\alpha}_{t_0}} \epsilon_\theta\left(x_{t_0}, t_0\right)}{\sqrt{\bar{\alpha}_{t_0}}}. \tag{6}$$

In accordance, x_{t_0} is decomposed using the following equation:

$$x_{t_0} = \sqrt{\bar{\alpha}_{t_0}} \hat{x}_0^{t_0} + \sqrt{1 - \bar{\alpha}_{t_0}} \epsilon_\theta(x_{t_0}, t_0). \tag{7}$$

Subsequently, the noise video $x_{t_0 - 1}$ at steps $t_0 - 1$ is obtained as:

$$x_{t_0 - 1} = \sqrt{\bar{\alpha}_{t_0 - 1}} \hat{x}_0^{t_0} + \sqrt{1 - \bar{\alpha}_{t_0 - 1}} \epsilon_{t_0 - 1}, \tag{8}$$

where, if the DDIM [64] sampling method is utilized, ϵ_{t_0-1} should be $\epsilon_\theta(x_{t_0}, t_0)$. Subsequently, noise video x_t undergoes progressive denoising, and the corresponding estimation $\hat{x}_0^t$ for the enhancement result is gradually refined until the final enhancement result x_0 is obtained.

As illustrated in Fig. 2, the primary content loss occurs between the initial estimation $\hat{x}_0^{t_0}$ and the reference video x^r during the enhancement process. Therefore, we propose to measure $dist\,(x_0, x^r)$ by assessing the difference between the low-frequency components of $\hat{x}_0^{t_0}$ and x^r, following the decomposition method outlined in [58]:

$$\begin{aligned} x^r &= f_l^\nu(x^r) + f_h^\nu(x^r), \\ \hat{x}_0^{t_0} &= f_l^\nu(\hat{x}_0^{t_0}) + f_h^\nu(\hat{x}_0^{t_0}), \end{aligned} \tag{9}$$

where f_l^ν denotes the low-frequency component, f_h^ν denotes the high-frequency component, and the threshold frequency ν lies between 0 and 1, defining the boundary between high and low frequencies.

Based on the analysis above, we redefine our formulation as follows:

$$\begin{aligned} &\min_{\epsilon_{t_0}} \; ||f_l^\nu(x^r) - f_l^\nu(\hat{x}_0^{t_0})|| \\ &\text{s.t.} \quad \hat{x}_0^{t_0} \sim \hat{P}_\theta^{t_0}\left(x \mid x^r, \epsilon_{t_0}\right) \end{aligned}, \tag{10}$$

where $\hat{x}_0^{t_0}$ represents the initial estimation of the enhancement result x_0.

3.3 Noise Calibration

In order to solve the formulation defined in the previous section, we propose a simple and effective optimization method called **Noise Calibration**, which essentially obtains a more suitable initial noise through 1-3 iterations. Specifically, by combining Eqs. (5), (7) and (9), we obtain:

$$\sqrt{\bar{\alpha}_{t_0}}(f_l^\nu(x^r) - f_l^\nu(\hat{x}_0^{t_0})) = \sqrt{1-\bar{\alpha}_{t_0}}(\epsilon_\theta(x_{t_0}, t_0) - \epsilon_{t_0}) + \sqrt{\bar{\alpha}_{t_0}}(f_h^\nu(\hat{x}_0^{t_0}) - f_h^\nu(x^r)). \tag{11}$$

This implies that,

$$\min_{\epsilon_{t_0}} ||f_l^\nu(x^r) - f_l^\nu(\hat{x}_0^{t_0})|| \equiv \min_{\epsilon_{t_0}} ||\epsilon_\theta(x_{t_0}, t_0) - \epsilon_{t_0} + \frac{\sqrt{\bar{\alpha}_{t_0}}}{\sqrt{1-\bar{\alpha}_{t_0}}}(f_h^\nu(\hat{x}_0^{t_0}) - f_h^\nu(x^r))||. \tag{12}$$

The goal of ILVR [12] is to generate an image $x \in \{x : \phi(x) = \phi(y)\}$ based on a diffusion model, given a reference image y. Here, $\phi(\cdot)$ denotes a linear low-pass filtering operation, a sequence of downsampling and upsampling. This goal is achieved by ensuring $\phi(x_t) = \phi(y_t)$ in the denoising process through:

$$x_t \leftarrow x_t + \phi\left(y_t\right) - \phi\left(x_t\right). \tag{13}$$

Motivated by this insight, we employ Fixed Point Iteration based on Eq. (12) to optimize the initial noise as:

$$\underline{\epsilon_{t_0}} \leftarrow \epsilon_\theta(\sqrt{\bar{\alpha}_{t_0}}x^r + \sqrt{1-\bar{\alpha}_{t_0}}\underline{\epsilon_{t_0}}, t_0) + \frac{\sqrt{\bar{\alpha}_{t_0}}}{\sqrt{1-\bar{\alpha}_{t_0}}}(f_h^\nu(\hat{x}_0^{t_0}) - f_h^\nu(x^r)). \tag{14}$$

Algorithm 1 SDEdit with Noise Calibration for video enhancement

Input: reference video x^r, initial denoising step t_0, diffusion model $\epsilon_\theta(x_t, t)$, iteration steps N, threshold frequency ν
$\epsilon_{t_0} \sim \mathcal{N}(0,1)$
for $n = 1$ **to** N **do**
  $x_{t_0} = \sqrt{\bar{\alpha}_{t_0}} x^r + \sqrt{1-\bar{\alpha}_{t_0}}\epsilon_{t_0}$
  $\hat{x}_0^{t_0} = (x_{t_0} - \sqrt{1-\bar{\alpha}_{t_0}}\epsilon_\theta(x_{t_0}, t_0))/\sqrt{\bar{\alpha}_{t_0}}$
  $\epsilon_{t_0} = \epsilon_\theta(x_{t_0}, t_0) + \frac{\sqrt{\bar{\alpha}_{t_0}}}{\sqrt{1-\bar{\alpha}_{t_0}}}(f_h^\nu(\hat{x}_0^{t_0}) - f_h^\nu(x^r))$
end for
$x_{t_0} = \sqrt{\bar{\alpha}_{t_0}} x^r + \sqrt{1-\bar{\alpha}_{t_0}}\epsilon_{t_0}$
for $t = t_0$ **to** 1 **do**
  $\epsilon_t \sim \mathcal{N}(0,1)$
  $\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(x_t,t) + \sigma_t\epsilon_t$
end for

Essentially, during each iteration, a replacement on the low-frequency domain resembling LIVR [12] is performed on x_{t_0} as:

$$x_{t_0} \leftarrow x_{t_0} + \sqrt{\bar{\alpha}_{t_0}}(f_l^\nu(x^r) - f_l^\nu(\hat{x}_0^{t_0})). \tag{15}$$

After obtaining the calibrated noise through 1-3 iterations, we re-add noise to the reference video and enhance video quality using pre-trained video models, as shown in Algorithm 1, referred to as NC-SDEdit. As illustrated in Fig. 2, Noise2, being a calibrated version of Noise1, results in a a superior initial estimation $\hat{x}_0^{t_0}$, thereby effectively achieving content-preserving video enhancement.

4 Experiments

4.1 Experiments Setup

Setting up. We conduct our experiments using an open-source T2V diffusion model, VideoCrafter(576×1024) [11], known for its superior visual quality, albeit with limitations in semantic understanding. We use SDEdit as a benchmark, while introducing our approach, NC-SDEdit, which essentially incorporates Noise Calibration into SDEdit. The initial denoising step, iteration steps, and threshold frequency are set to $t_0 = 600, N = 3, \nu = 1.0$ as default.
Dataset and Metrics. We utilize a reference set consisting of 700 videos with a resolution of 320×512, generated by Lavie [79], along with their corresponding texts from EvalCrafter [34] for quantitative evaluation. Firstly, to evaluate the consistency of content, we report the MSE_l (MSE on the low-frequency domain with ν=0.5), MSE and SSIM [81] between the enhancement results and the reference videos. Secondly, to assess the visual quality of the enhanced results, we also report the state-of-the-art video quality assessment metric, DOVER [85]. The image quality metric CLIP-IQA [72] is also used to assist in the evaluation. Finally, spatial frequency (SF) can measure the gradient distribution thus revealing the

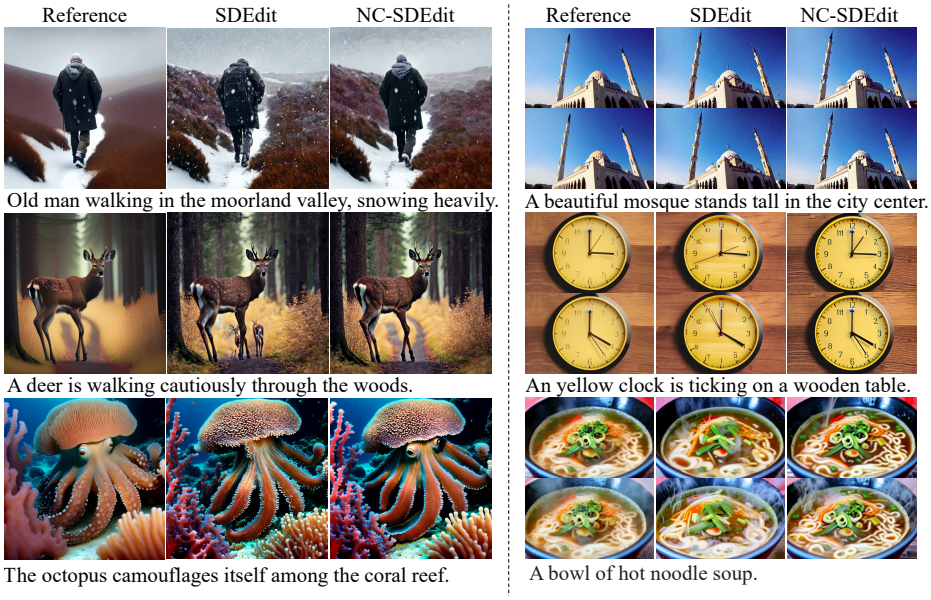

Fig. 3: Visual comparisons of video enhancement based on VideoCrafter [11]

Table 1: Quantitative comparisons based on VideoCrafter [11]

Method	$\text{MSE}_l\downarrow$	MSE↓	SSIM↑	DOVER↑	CLIP-IQA↑	$\text{D}_{\text{SF}}\uparrow$
SDEdit	4.3447	0.7600	0.6464	60.17	**0.4482**	0.0527
Ours(N=1)	<u>2.9201</u>	<u>0.6546</u>	<u>0.6998</u>	<u>60.62</u>	<u>0.4471</u>	0.0531
Ours(N=2)	**2.8039**	**0.6506**	**0.7040**	**62.71**	0.4400	0.0554
Ours(N=3)	2.9540	0.6638	0.6971	62.45	0.4387	<u>0.0584</u>
Ours(N=10)	4.6107	0.7570	0.6209	54.42	0.3873	**0.0741**

detail and texture of the video frame. Therefore, we use $\text{D}_{\text{SF}} = \text{SF}(x) - \text{SF}(x^r)$ to measure whether video details have been enhanced.

4.2 Quantitative and Qualitative Evaluation

Quantitative Evaluation. We evaluate our method against SDEdit, with the quantitative results presented in Tab. 1, by enhancing the videos in the reference set at a resolution of 640×1024. According to the results, our proposed method only requires an additional one to three iterations and significantly outperforms previous approaches in all evaluation metrics, with a slight exception in the image quality assessment metric CLIP-IQA. Additionally, it is observed that when the number of iteration steps becomes too large (e.g., 10), the enhancement effect does not increase but rather decreases. We will discuss this in Sec. 4.3.

Qualitative Evaluation. Fig. 3 presents the visual comparison of video enhancement results with various types and aspect ratios. It can be observed that

Table 2: User study of human preference

Method	Consistency	Visual Quality	Texture
SDEdit	13.89%	26.74%	24.65%
NC-SDEdit	**86.11%**	**73.26%**	**75.35%**

Fig. 4: Comparison with entirely different methods

our method is capable of maintaining the original content during the video enhancement process. In contrast, existing methods either introduce strange noise (as in the 'mosque' case) or object (ike in the 'deer' case), or alter the details of the existing content (e.g., the time in the 'clock' case).

User Study and extra Comparisons. We also conduct a user study with 18 samples and invite 15 volunteers to evaluate the results. As shown in Tab. 2, our method performs much better than SDEdit, especially in terms of consistency. Additionally, we compare our method with a traditional SR method, BasicVSR++ [9], and a SR3-based method, SD-x4-Upscaler [54]. Results are shown in Fig. 4. As discussed in Sec. 2.2, unlike NC-SDEdit, the two other methods merely increase the resolution and fail to enhance the texture.

4.3 Ablation Studies and Analysis

Influence of Iteration Steps N. Our proposed method involves iterating the initial noise multiple times. We provide an intuitive analysis of the effects corresponding to different numbers of iterations using a specific case as an example, depicted in Fig. 5. In the original video, a young man is shown holding his head with one hand while looking at a laptop. However, SDEdit, while enhancing the quality, modifies the content of the original video, such as altering the hand posture and the shape of the coffee cup. In contrast, our method requires only a single iteration to significantly preserve the overall content of the original video while enhancing it. Subsequent iterations, two or three in total, further enhance the preservation of original details, such as the laptop's logo. However, when the number of iteration steps becomes excessive (e.g., 10), the video structure is preserved identically to the reference, but the colors become oversaturated. The reason is that the content loss $||f_l^{\nu}(x^r) - f_l^{\nu}(\hat{x}_0^{t_0})||$ in Eq. (10) emphasizes the consistency of low-frequency information, *i.e.,* video structure, which might have side effects on color. Designing a more appropriate consistency objective could be a direction to mitigate color oversaturation. As a general guideline, we recommend limiting the iteration steps to between 1 and 3.

Fig. 5: Visual comparisons about iteration steps N and threshold frequency ν

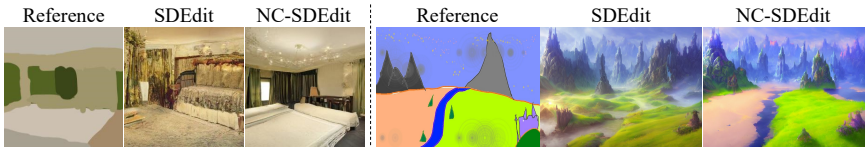

Fig. 6: Visual comparisons of Stroke Painting to Image($t_0 = 800, N = 1, \nu = 0.1$)

Influence of Threshold Frequency ν. We suggest using the Fourier transform to extract the low-frequency components. The threshold frequency ν serves as the boundary between high and low frequencies, influencing the constraint range of the objective function $||f_l^\nu(x^r) - f_l^\nu(\hat{x}_0^{t_0})||$. Using the 'bear' case in Fig. 5 as an example, SDEdit, without any additional consistency preservation, can be likened to having a threshold frequency of $\nu = 0$ resulting in alterations to details such as the teddy bear's collar button. As the threshold frequency increases, the range constrained by the objective function $||f_l^\nu(x^r) - f_l^\nu(\hat{x}_0^{t_0})||$ expands, thereby enhancing the consistency of content between the final enhanced video and the original video. When the threshold frequency reaches 1.0, the texture of the teddy bear's fur in the enhanced video is almost identical to that in the original video. Additionally, the threshold frequency ν enables the application of our method to other tasks based on SDEdit, such as sketch-to-image translation, as depicted in Fig. 6. In addition, as shown in Tab. 3 we conduct an ablation study on 60 videos about initial denoising step t_0 and threshold frequency ν. Firstly, it can be observed that increasing t_0 results in decreased content consistency but enhances details more effectively. Secondly, for a given t_0, a larger ν improves the effectiveness of content-preserving video enhancement. Ultimately, we choose a compromise parameter combination (t_0=600; ν=1.0) as the default hyperparameters.

Convergence of Noise Calibration. To further validate whether Noise Calibration can effectively optimize the objective function (Eq. (10)), We calculate the L2 distance $||f_l^\nu(x^r) - f_l^\nu(\hat{x}_0^{t_0})||_2$ in the latent space between $f_l^\nu(x^r)$ and $f_l^\nu(\hat{x}_0^{t_0})$ on the Lavie700 dataset [34], as the number of iterations N increases for different threshold frequencies ν. As shown in Fig. 7, for different values

Table 3: Ablation Study about t_0 and ν ($L_2^l \downarrow / D_{SF} \uparrow$)

Method	ν=0.0	ν=0.3	ν=0.5	ν=0.7	ν=1.0
t_0=500	2.9677/0.0544	2.2955/0.0535	2.0532/0.0548	**1.9383**/0.0571	2.0706/0.0641
t_0=600	4.2997/0.0561	3.2612/0.0554	2.8896/0.0566	2.7246/0.0589	2.7957/0.0658
t_0=700	6.2563/0.0586	4.7217/0.0574	4.1862/0.0586	3.9240/0.0606	3.9010/**0.0664**

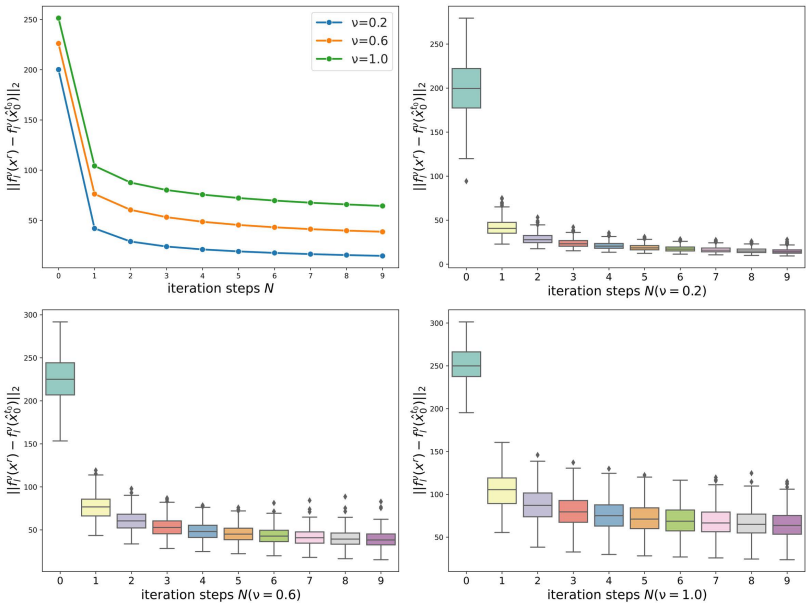

Fig. 7: Demonstration of the Optimization Effect of the Objective Function

of the threshold frequency ν, the average value of the optimization target decreases as the number of iterations N increases. Notably, the most significant effect is observed with just one iteration, indicating that achieving a high level of consistency maintenance can be accomplished with a single iteration.

Evaluation about training and inference costs. Firstly, our method is training-free and incurs no training costs. Secondly, during inference, using the default setting of initial denoising step $t_0 = 600$ and DDIM steps $= 30$, our method requires only 1 to 3 additional refinement steps, resulting in less than 10% additional inference time compared to SDEdit.

4.4 Improving upon State-of-the-Art Visual Refinement Models

MS-Vid2Vid-XL [93] and SDXL-1.0-refiner [52] are two refinement models that fine-tune existing generative models and utilize SDEdit for quality enhancement. However, during the enhancement process, the existing details of the original video/image are often smoothed out, and the originally correct structures tend to

Fig. 8: Visual Demonstration of MS-Vid2Vid-XL [93] with Noise Calibration

be disrupted. Our method, Noise Calibration, can address this issue. To validate the effectiveness of Noise Calibration, we employ the same 700 paired text-video samples from EvalCrafter as the test set for MS-Vid2Vid-XL. Furthermore, we generate corresponding images using SDXL [52] based on these 700 texts as the test set for SDXL-1.0-refiner. Given the default initial denoising step $t_0 = 300$ in SDXL-1.0-refiner, we set the threshold frequency ν to 0.5. The results, as presented in Tab. 4, demonstrate that Noise Calibration improves the performance of existing refinement models across all metrics. Moreover, as illustrated in Figs. 8 and 9, our method yields more consistent enhancement effects intuitively.

5 Limitation, Societal Impact and Acknowledgements

Limitation. Like SDEdit, the enhancement effectiveness of our method is also limited by the performance of the base model.

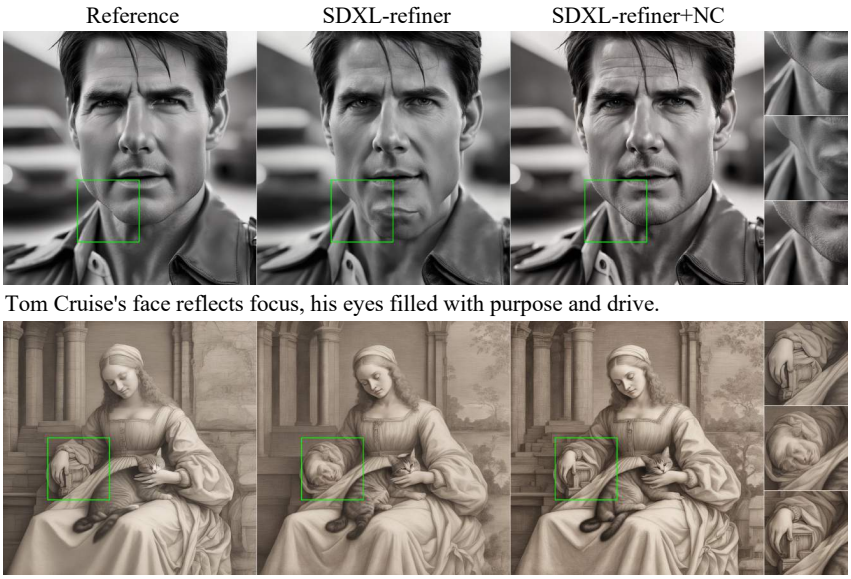

Fig. 9: Visual Demonstration of SDXL-1.0-refiner [52] with Noise Calibration

Table 4: Quantitative Comparisons based on MS-Vid2Vid-XL and SDXL-1.0-refiner

Method	MSE_l↓	MSE↓	SSIM↑	CLIP-IQA↑	DOVER↑	D_{SF}↑
MS-Vid2Vid-XL [93]	3.2214	0.7490	0.7079	0.4232	52.89	0.0478
MS-Vid2Vid-XL+NC	**2.6848**	**0.7120**	**0.7253**	**0.4305**	**57.61**	**0.0517**
SDXL-1.0-refiner [52]	1.2775	0.6933	0.7344	0.8590	-	0.0503
SDXL-1.0-refiner+NC	**0.8750**	**0.5834**	**0.7625**	**0.8734**	-	**0.0530**

Societal Impact. As our method is for improving video quality, it does not introduce additional ethical concerns.

Acknowledgements. This research is supported by National Key R&D Program of China (No. 2018AAA0100300).

6 Conclusion

In this work, we propose a novel formulation for video enhancement that takes into account both visual quality and consistency of content. While using the pre-trained T2V diffusion model for denoising to improve video quality, we introduce **Noise Calibration**, a simple yet effective method for maintaining consistency of content before and after enhancement. Extensive analysis and experiments demonstrate the effectiveness of our approach.

References

1. Ahn, N., Kwon, P., Back, J., Hong, K., Kim, S.: Interactive cartoonization with controllable perceptual factors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16827–16835 (2023)
2. An, J., Zhang, S., Yang, H., Gupta, S., Huang, J.B., Luo, J., Yin, X.: Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. arXiv preprint arXiv:2304.08477 (2023)
3. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023)
4. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
5. Balaji, Y., Min, M.R., Bai, B., Chellappa, R., Graf, H.P.: Conditional gan with discriminative filter generation for text-to-video synthesis. In: IJCAI. vol. 1, p. 2 (2019)
6. Bao, F., Li, C., Zhu, J., Zhang, B.: Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. arXiv preprint arXiv:2201.06503 (2022)
7. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledits++: Limitless image editing using text-to-image models. arXiv preprint arXiv:2311.16711 (2023)
8. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
9. Chan, K.C.K., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment (2021)
10. Chen, C., Zhou, S., Liao, L., Wu, H., Sun, W., Yan, Q., Lin, W.: Iterative token evaluation and refinement for real-world super-resolution. arXiv preprint arXiv:2312.05616 (2023)
11. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
12. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
13. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
14. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine **35**(1), 53–65 (2018)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
16. Dockhorn, T., Vahdat, A., Kreis, K.: Score-based generative modeling with critically-damped langevin diffusion. arXiv preprint arXiv:2112.07068 (2021)
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

18. Hachnochi, R., Zhao, M., Orzech, N., Gal, R., Mahdavi-Amiri, A., Cohen-Or, D., Bermano, A.H.: Cross-domain compositing with pretrained diffusion models. arXiv preprint arXiv:2302.10167 (2023)
19. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)
20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control.(2022). URL https://arxiv. org/abs/2208.01626 (2022)
21. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
23. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models (2022)
24. Hu, Y., Luo, C., Chen, Z.: Make it move: controllable image-to-video generation with text descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18219–18228 (2022)
25. Jiang, Y., Yang, S., Qiu, H., Wu, W., Loy, C.C., Liu, Z.: Text2human: Text-driven controllable human image generation. ACM Transactions on Graphics (TOG) **41**(4), 1–11 (2022)
26. Jiménez, Á.B.: Mixture of diffusers for scene composition and high resolution image generation. arXiv preprint arXiv:2302.02412 (2023)
27. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
28. Kim, H., Lee, G., Choi, Y., Kim, J.H., Zhu, J.Y.: 3d-aware blending with generative nerfs. arXiv preprint arXiv:2302.06608 (2023)
29. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
30. Kingma, D.P., Welling, M., et al.: An introduction to variational autoencoders. Foundations and Trends® in Machine Learning **12**(4), 307–392 (2019)
31. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006)
32. Li, B., Xue, K., Liu, B., Lai, Y.K.: Vqbb: Image-to-image translation with vector quantized brownian bridge. arXiv preprint arXiv:2205.07680 (2022)
33. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
34. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. arXiv preprint arXiv:2310.11440 (2023)
35. Lu, S., Liu, Y., Kong, A.W.K.: Tf-icon: Diffusion-based training-free cross-domain image composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2294–2305 (2023)
36. Luo, F., Xiang, J., Zhang, J., Han, X., Yang, W.: Image super-resolution via latent diffusion: A sampling-space mixture of experts and frequency-augmented decoder approach. arXiv preprint arXiv:2310.12004 (2023)

37. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
38. Ma, Y., He, Y., Wang, H., Wang, A., Qi, C., Cai, C., Li, X., Li, Z., Shum, H.Y., Liu, W., et al.: Follow-your-click: Open-domain regional image animation via short prompts. arXiv preprint arXiv:2403.08268 (2024)
39. Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. arXiv preprint arXiv:2406.01900 (2024)
40. Mei, K., Patel, V.: Vidm: Video implicit diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 9117–9125 (2023)
41. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
42. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2630–2640 (2019)
43. Mishra, S., Saenko, K., Saligrama, V.: Syncdr: Training cross domain retrieval models with synthetic data. arXiv preprint arXiv:2401.00420 (2023)
44. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
45. Ngiam, J., Chen, Z., Koh, P.W., Ng, A.Y.: Learning deep energy models. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 1105–1112 (2011)
46. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
47. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
48. Oussidi, A., Elhassouny, A.: Deep generative models: Survey. In: 2018 International conference on intelligent systems and computer vision (ISCV). pp. 1–8. IEEE (2018)
49. Pandey, K., Mukherjee, A., Rai, P., Kumar, A.: Vaes meet diffusion models: Efficient and high-fidelity generation. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
50. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
51. Peng, D., Hu, P., Ke, Q., Liu, J.: Diffusion-based image translation with label guidance for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 808–820 (2023)
52. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
53. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)

54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
55. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
56. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
57. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
58. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. arXiv preprint arXiv:2309.11497 (2023)
59. Singh, J., Gould, S., Zheng, L.: High-fidelity guided image synthesis with latent diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5997–6006. IEEE (2023)
60. Sinha, A., Song, J., Meng, C., Ermon, S.: D2c: Diffusion-decoding models for few-shot conditional generation. Advances in Neural Information Processing Systems **34**, 12533–12548 (2021)
61. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3626–3636 (2022)
62. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
63. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
64. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
65. Song, Y., Durkan, C., Murray, I., Ermon, S.: Maximum likelihood training of score-based diffusion models. Advances in Neural Information Processing Systems **34**, 1415–1428 (2021)
66. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
67. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems **33**, 12438–12448 (2020)
68. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
69. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
70. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: 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. pp. 1921–1930 (2023)
71. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. Advances in Neural Information Processing Systems **34**, 11287–11302 (2021)

72. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
73. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015 (2023)
74. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
75. Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952 (2022)
76. Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J., Liu, J.: Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. arXiv preprint arXiv:2305.10874 (2023)
77. Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: G3an: Disentangling appearance and motion for video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5264–5273 (2020)
78. Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: Imaginator: Conditional spatio-temporal gan for video generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1160–1169 (2020)
79. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023)
80. Wang, Y., Jiang, L., Loy, C.C.: Styleinv: A temporal style modulated inversion network for unconditional video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22851–22861 (2023)
81. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
82. Watson, D., Chan, W., Ho, J., Norouzi, M.: Learning fast samplers for diffusion models by differentiating through sample quality. arXiv preprint arXiv:2202.05830 (2022)
83. Wolleb, J., Sandkühler, R., Bieder, F., Cattin, P.C.: The swiss army knife for image-to-image translation: Multi-task diffusion models. arXiv preprint arXiv:2204.02641 (2022)
84. Wu, C.H., De la Torre, F.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7378–7387 (2023)
85. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20144–20154 (2023)
86. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. arXiv preprint arXiv:2303.09472 (2023)
87. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)
88. Yang, Z., Chu, T., Lin, X., Gao, E., Liu, D., Yang, J., Wang, C.: Eliminating contextual prior bias for semantic image editing via dual-cycle diffusion. IEEE Transactions on Circuits and Systems for Video Technology (2023)

89. Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22479–22489 (2023)
90. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18456–18466 (2023)
91. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. arXiv preprint arXiv:2307.12348 (2023)
92. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint arXiv:2309.15818 (2023)
93. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
94. Zhang, S., Xiao, S., Huang, W.: Forgedit: Text guided image editing via learning and forgetting. arXiv preprint arXiv:2309.10556 (2023)
95. Zhao, M., Bao, F., Li, C., Zhu, J.: Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. Advances in Neural Information Processing Systems **35**, 3609–3623 (2022)
96. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)