

# DNI: Dilutional Noise Initialization for Diffusion Video Editing

Sunjae Yoon<sup>1</sup>, Gwanhyeong Koo<sup>1</sup>, Ji Woo Hong<sup>1</sup>, and Chang D. Yoo<sup>1</sup>

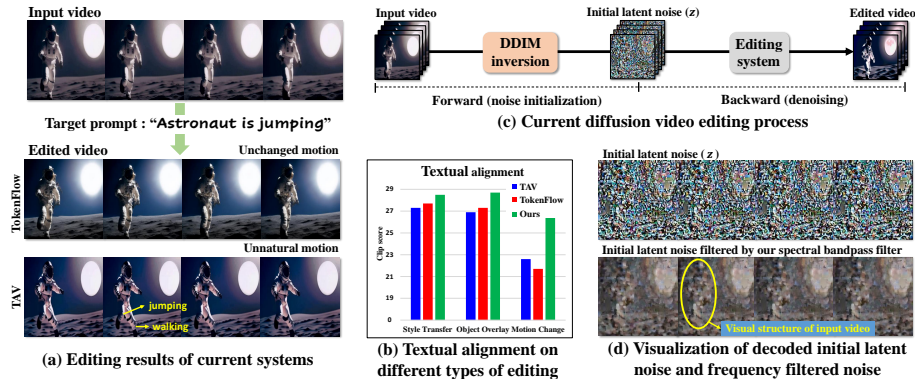
Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea  
{sunjae.yoon, cd\_yoo}@kaist.ac.kr



**Fig. 1:** Edited videos of Dilutional Noise Initialization (DNI) framework. DNI performs text-based rigid and non-rigid edits, enabling effective alteration under high fidelity.

**Abstract.** Text-based diffusion video editing systems have been successful in performing edits with high fidelity and textual alignment. However, this success is limited to rigid-type editing such as style transfer and object overlay, while preserving the original structure of the input video. This limitation stems from an initial latent noise employed in diffusion video editing systems. The diffusion video editing systems prepare initial latent noise to edit by gradually infusing Gaussian noise onto the input video. However, we observed that the visual structure of the input video still persists within this initial latent noise, thereby restricting non-rigid editing such as motion change necessitating structural modifications. To this end, this paper proposes Dilutional Noise Initialization (DNI) framework which enables editing systems to perform precise and dynamic modification including non-rigid editing. DNI introduces a concept of ‘noise dilution’ which adds further noise to the latent noise in the region to be edited to soften the structural rigidity imposed by input video, resulting in more effective edits closer to the target prompt. Extensive experiments demonstrate the effectiveness of the DNI framework.

**Keywords:** Diffusion video editing · Effective editing · Noise dilution



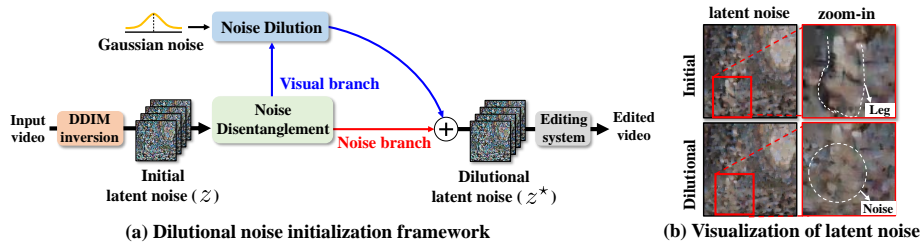
**Fig. 2:** (a) Editing results about motion change of current systems [4, 32]. (b) Categorical analysis of textual alignment with video across different types of editing on DAVIS [19]. (c) Overview of current diffusion video editing process. (d) Visualization of initial latent noise and the latent noise filtered by our designed adaptive spectral filter, where the input video’s visual structure clearly remains in the initial latent noise.

## 1 Introduction

Denoising diffusion models [3, 8, 27, 28] have spurred substantial innovations in the generative capabilities of artificial intelligence. By gradually denoising input Gaussian noise, diffusion models generate various outputs including image [18, 23], audio [13, 34], and video [2, 7], which can be further edited to meet users’ specific needs. We focus here on diffusion video editing which holds immense promise for revolutionizing the entertainment industry.

Video editing systems aim to modify specific attributes in an input video corresponding to users’ requirements from a target textual prompt. Recent video editing systems [4, 15, 20, 31] have succeeded in performing edits with high fidelity to input video and precise textual alignment. However, this success is still restricted to rigid modifications such as style transfer and object overlay by preserving input video structural layouts. Specifically, in Fig. 2 (a), for a target prompt requiring non-rigid modifications (*e.g.* “Astronaut is jumping”), current systems fail to conform and return the original input video under over-fidelity. Otherwise, they often exhibit unnatural motion by blending the original content (*i.e.* walking) with the targeted content (*i.e.* jumping) in the video. In Fig. 2 (b), our categorical analysis of textual alignment with video across different editing types (*i.e.* motion change, style transfer, object overlay) demonstrates that current systems are struggling with the motion change. Therefore the resulting videos about complex non-rigid editing still remain unsatisfactory.

Our investigation revealed that one of the reasons for this unsatisfactory non-rigid editing stems from the initial latent noise fed into the diffusion video editing systems. In a typical process of diffusion video editing, as shown in Fig. 2 (c), the diffusion model infuses a Gaussian noise onto input video to build initial latent noise  $z$  using inverse denoising (*e.g.* DDIM inversion) as a forward process.



**Fig. 3:** (a) Illustration of Dilutional Noise Initialization framework. The noise disentanglement separates the initial latent noise into a visual branch and a noise branch. The visual branch contains a visual noise of input video components and the noise branch contains a Gaussian noise. The noise dilution adds further noise into an editing region of the visual noise, enabling dynamic modifications without being restricted by the input video layout. (b) Visualizations of initial and dilutional latent noises.

In the backward process, the model performs denoising into this  $z$  to generate edited videos conforming to the target prompt. Despite the noise  $z$  assuming a Gaussian noise distribution, in Fig. 2 (d), we observe that this noise still contains the visual structures of the input video. To ascertain their presence, we devise a frequency pass filter referred to as adaptive spectral filter, which captures the clearer structure of the input video within the latent noise. Consequently, the current video editing systems perform edits on top of the input video’s visual structure within the initial latent noise, facilitating rigid editing yet exposing a susceptibility to non-rigid editing that necessitates altering the structure.

To this end, we propose Dilutional Noise Initialization (DNI) that enables video editing systems to perform precise and dynamic modifications encompassing the versatility of non-rigid editing. As shown in Fig. 3 (a), the DNI framework introduces a novel concept of *noise dilution*, which adds further noise into latent noise to ensure the edited video aligns more closely with the input text prompt. Formally, the DNI takes an initial latent noise  $z$  as input and produces a dilutional latent noise  $z^*$  which mitigates structural rigidity imposed by the input video in the area to be edited. To build the  $z^*$ , DNI framework performs two main processes: (1) disentangling initial noise into the visual branch and noise branch and (2) diluting the noise in the visual branch with additional Gaussian noise. For the noise disentanglement, we design a frequency pass filter referred to as *adaptive spectral filter* which effectively isolates the input video components into the visual branch by considering the frequency spectrum of the input video. Subsequently, dilution is carried out within the visual noise by blending Gaussian noise into the targeted editing area guided by the target prompt. Finally, dilutional latent noise  $z^*$  is synthesized by recombining the noises in the two branches. As shown in Fig. 3 (b), the dilutional latent noise preserves the input video’s visual structure while simultaneously reducing the rigidity of this structure in the specific area targeted for editing (*e.g.* the man’s legs). DNI is applied to any diffusion video editing system in a model-agnostic manner, demonstrating effective editability on video editing benchmarks (DAVIS [19], TVGVE [33]).

## 2 Related Works

### 2.1 Diffusion-based generative models

Denoising diffusion models [8, 27] have surpassed erstwhile qualities of generative adversarial networks [5]. Diffusion-based text-to-image (T2I) models [22, 24] have significantly advanced image generation, producing high-fidelity images from text. These models are now extending into text-to-video (T2V) task. Early works [9, 10, 30] in T2V task adapted pre-trained T2I models by incorporating a temporal dimension, where many temporal attentions [7, 26] are also designed to enhance frame consistency. Recently, diffusion models have excelled in various generative works including super-resolution and inpainting [17, 25]. Among these, diffusion video editing presents a new challenge for controlled synthesis across frames obtrusively, discussed in detail below.

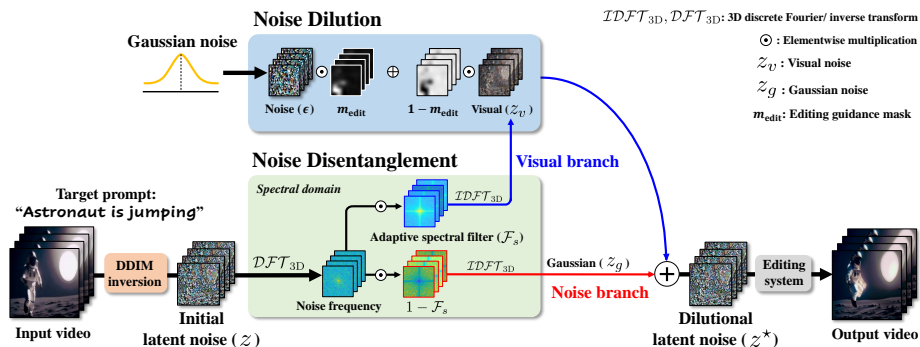
### 2.2 Diffusion video editing

The recent success of text-based image editing [1, 6, 14] bridges to video editing [4, 32, 35]. For video, the challenge lied in seamlessly integrating edited frames under high fidelity to input video. Thus, several technical solutions to enhance temporal consistency are introduced including temporal attention [4, 32] and knowledge injection [15, 20] from input video priors. Previous models sought to preserve the input video’s information to improve editing quality [27], yet paradoxically, they encountered a trade-off, sacrificing versatility in editing. To this end, we present a DNI framework that can enable video editing systems to perform various edits including non-rigid edits, maintaining their video quality.

## 3 Preliminaries

### 3.1 Denoising diffusion probabilistic models

Denoising diffusion probabilistic models (DDPMs) [8] are structured as parameterized Markov chains, methodically restoring noisy data sequences  $\{x_1, \dots, x_T\}$  from an initial  $x_0$ . First, Gaussian noise is progressively added to  $x_T$  through the Markov transition  $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I)$ , following a pre-defined schedule  $\alpha_t$  over steps  $t \in \{1, \dots, T\}$ . This procedure is defined as the *forward process* in diffusion modeling. The *reverse process* is then applied to generate data using diffusion model estimating  $q(x_{t-1}|x_t)$  through trainable Gaussian transitions  $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t))$ , starting from the normal distribution  $p(x_T) = \mathcal{N}(x_T; 0, I)$ . The model is trained to maximize log-likelihood  $\log(p_\theta(x_0))$  over  $\theta$ , where variational inference maximizes the lower bound of  $\log(p_\theta(x_0))$ , yielding a closed-form KL divergence between distributions  $p_\theta$  and  $q$ . This process is summarized as training a denoising network  $\epsilon_\theta(x_t, t)$  to predict noise  $\epsilon \sim \mathcal{N}(0, I)$  as  $\mathbb{E}_{x, \epsilon \sim \mathcal{N}(0, 1), t \sim \mathcal{U}\{1, T\}}[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2]$ , where  $\mathcal{U}\{1, T\}$  is discrete uniform distribution from 1 to  $T$  for robust training on each step  $t$ .



**Fig. 4:** Illustration of Dilutional Noise Initialization (DNI) framework, which refines initial latent noise  $z$  into dilutional latent noise  $z^*$ , enabling editing systems to perform effective editing including non-rigid editing. DNI contains two main modules: (1) Noise Disentanglement which separates the noise  $z$  into Gaussian noise  $z_g$  and visual noise  $z_v$  containing input video components and (2) Noise Dilution which adds a Gaussian noise  $\epsilon$  on the  $z_v$  to mitigate restrictions of the input video structure near the editing region. The noises  $z_v$  and  $z_g$  are recombined to build  $z^*$  for an input of video editing.

### 3.2 Denoising diffusion implicit model and Inversion

Denoising diffusion implicit model (DDIM) [27] accelerates diffusion reverse process, sampling with fewer steps as  $x_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} x_t + \left( \sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \epsilon$ . We can also build an inverse process of this acting as the forward process as  $x_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} x_t + \left( \sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \epsilon$ , referred to as DDIM inversion process. In diffusion editing, DDIM inversion enhances fidelity to the input video.

### 3.3 Text-conditioned diffusion model

The text-conditioned diffusion model generates the output data  $x_0$  conditioned on a text prompt. The training objective incorporates textual condition under latent space as  $\mathbb{E}_{z, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, \mathbf{c})\|_2^2]$ , where  $z_t$  is a latent noise encoding of  $x_t$  using VQ-VAE [29] and  $\mathbf{c}$  is target prompt CLIP [21] embedding. Video editing takes  $z_t$  as input video latent noise and  $\mathbf{c}$  for a conditional target prompt.

## 4 Dilutional Noise Initialization

Dilutional Noise Initialization (DNI) framework aims to enable video editing systems to perform effective editing including non-rigid modifications. Fig. 4 illustrates the overall process of DNI framework, where it takes initial latent noise  $z$  based on  $T$  step inverse denoising (*i.e.* DDIM inversion) of the input video and synthesizes dilutional latent noise  $z^*$  to mitigate constraints from the visual structure of the input video in the area to be edited. The DNI framework

consists of two primary components: (1) Noise Disentanglement and (2) Noise Dilution. The noise disentanglement separates initial latent noise  $z$  into visual noise  $z_v$  and Gaussian noise  $z_g$  using our designed adaptive spectral filter in the 3-dimensional frequency domain. Noise dilution specifies editing region within the visual noise using a target prompt and blends additional Gaussian noise, thereby diminishing the input video’s structural influence in the region to be edited. Finally the dilutional latent noise  $z^*$  is synthesized by merging the two noises from the visual branch and the noise branch.

#### 4.1 Noise Disentanglement

Noise disentanglement aims to extract the inherent input video components from the initial latent noise  $z$ . To conduct this, we investigated the spectral characteristics of the latent noise across spatial and temporal domains. Fig. 5 presents a spatial (top) and temporal (bottom) frequency of (a) initial latent noise  $z$ , (b) input video latent feature  $z_0$  prior to noise addition, and (c) white Gaussian noise  $\epsilon \sim N(0, I)$  (*i.e.* Isotropic Gaussian). The low frequency of  $z$  shows similar distributions (region in red) with  $z_0$  in both spatial and temporal domains. Based on this observation, we may apply a low-frequency pass filter (LPF) to acquire components of the input video. However, the LPF causes the loss of high-frequency components in the input video and also becomes heuristic to correspond to different frequencies for each

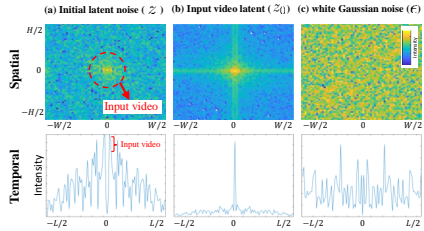
input video. Therefore, we introduce an adaptive spectral filter (ASF) that can adaptively respond to the frequency changes in the input video. (Sec. 5.4 provides detailed analysis of adaptive spectral filter with visualization in Fig. 9.) The ASF builds a frequency pass filter based on the frequency spectrum of the input video latent  $z_0$ , such that it appropriately captures the frequency range of each input video. Formally, the adaptive spectral filter  $\mathcal{F}_s$  is defined as below:

$$\mathcal{F}_s = \text{Norm}_{\min\text{-max}}(\mathcal{DFT}_{3D}(z_0)) \in \mathbb{R}^{W \times H \times L \times C}, \quad (1)$$

where  $W, H, L, C$  are the width, height, length, and channel of  $z_0$ . The  $\mathcal{DFT}_{3D}(\cdot)$  is 3-dimensional discrete Fourier transform, and  $\text{Norm}_{\min\text{-max}}$  is the min-max normalization for the scaling between 0 to 1. Thus, employing the  $\mathcal{F}_s$ , we separate initial latent noise  $z$  into visual noise  $z_v$  and Gaussian noise  $z_g$  as given below:

$$z_v = \text{IDFT}_{3D}(\mathcal{F}_s \odot \mathcal{DFT}_{3D}(z)), \quad z_g = \text{IDFT}_{3D}((1 - \mathcal{F}_s) \odot \mathcal{DFT}_{3D}(z)), \quad (2)$$

where the  $\text{IDFT}_{3D}$  is inverse  $\mathcal{DFT}_{3D}$  and  $\odot$  is the elementwise multiplication.

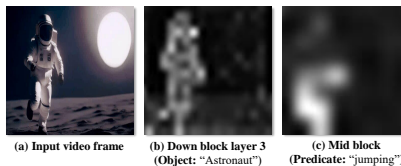


**Fig. 5:** Discrete Fourier transform (DFT) of (a) initial latent noise  $z$ , (b) video latent feature  $z_0$ , and (c) white Gaussian noise  $\epsilon$ . A similar distribution between  $z$  and  $z_0$  (red circle) shows that  $z$  contains the input video components (top: spatial domain 2D-DFT, bottom: temporal domain 1D-DFT).

## 4.2 Noise Dilution

Noise dilution aims to enhance the flexibility of modifications by mitigating the constraints imposed by the input video’s visual structure in editing region. Thus, by taking visual noise  $z_v$  as input, the noise dilution specifies the editing region on the  $z_v$  using the target prompt and blends additional Gaussian noise there. We describe this process as editing guidance and noise blending in the following.

**Editing guidance.** To provide guidance about the editing region, we first select the editing reference words (*e.g.* ‘Astronaut’, ‘jumping’) in the textual prompts (*e.g.* Astronaut is jumping) and obtain the guidance mask  $m_{\text{edit}}$  about these words. Our initial choice for the mask generation was to use pre-trained segmentation model (*e.g.* SAM [11]), but this was not appropriate for specifying the region for non-rigid editing due to noun-based detection. (*i.e.* Non-rigid editing mainly requires motion/pose modification by the predicate in a target prompt.) Therefore we use a cross-attention map between the reference words and video frames in the diffusion model. To be specific, as shown in Fig. 6 (b), when the reference word is a noun or adjective, the attention map  $m_{\text{rgd}}$  from the down-block of the UNet is used to provide the clear boundary of the editing target for rigid modification. In Fig. 6 (c), when the reference word is a predicate, blurry attention map  $m_{\text{non-rgd}}$  of mid-block is employed to encompass areas of non-rigid modification. All attention maps of  $m_{\text{rgd}}$ ,  $m_{\text{non-rgd}}$  are resized to the latent spatial dimension (*i.e.*  $W \times H$ ) and added together to form an editing guidance mask  $m_{\text{edit}}$ <sup>1</sup> as given below:



**Fig. 6:** Cross-attention map between frame and reference words.

$$m_{\text{edit}} = \alpha \times m_{\text{rgd}} + \beta \times m_{\text{non-rgd}} \in \mathbb{R}^{W \times H \times L}, \quad (3)$$

where  $\alpha$  and  $\beta$  are hyperparameters between 0 and 1 that modulate the mask’s intensity. If closer to 1, they increase Gaussian noise and suppress the input video’s influence through noise blending. The  $\alpha$  enables effective modifications within the visual structure, while  $\beta$  supports editing beyond those structures.

**Noise blending.** To the specified editing region by the mask  $m_{\text{edit}}$ , we blend Gaussian noise to diminish the input video structural influence. To conduct this, the visual noise  $z_v$  and white Gaussian noise  $\epsilon$  are blended based on the  $m_{\text{edit}}$  with channel-wise broadcasting. After that, the separated Gaussian noise  $z_g$  of the initial noise is also combined to build dilutional latent noise as given below:

$$z^* = z_g + m_{\text{edit}} \times \epsilon + (1 - m_{\text{edit}}) \times z_v \in \mathbb{R}^{W \times H \times L \times C}. \quad (4)$$

The dilutional noise  $z^*$  is then used for the denoising of video editing systems.

<sup>1</sup> Multiple attention maps of  $m_{\text{rgd}}$  and  $m_{\text{non-rgd}}$  by multiple reference words are mean-pooled before the Eq. (3) and  $m_{\text{edit}}$  is scaled between 0 to 1 after the Eq. (3)



### 4.3 Plug-and-play DNI framework

Recent video editing systems predominantly are largely grouped into two distinct operational approaches: (1) tuning-based [15, 32] and (2) tuning-free methods [4, 20]. The DNI framework can be applied to both methods, offering a model-agnostic approach that enhances editing versatility. In an inference time (*i.e.* denoising) of a model, DNI injects dilutional latent noise  $z^*$  in the model as input instead of initial latent noise  $z$  as given below:

$$\mathcal{V}_{\text{edit}} = \text{Denoise}(z^*, \mathcal{T}), \quad (5)$$

where the  $\mathcal{T}$  is target prompt and  $\mathcal{V}_{\text{edit}}$  is the edited video.

## 5 Experiment

### 5.1 Experimental Settings

**Implementation Details.** The VQ-VAE [29] is used for patch-wise frame encoding, and CLIP model (ViT-L/14) [21] for text embedding. We follow original settings of the baselines for video diffusion models (*i.e.* Stable Diffusion v1.5 and v2.1). Experiment is performed on NVIDIA A100 GPU, where the  $W, H = 64$  is used, rigid editing uses  $0.3 < \alpha < 0.7$ ,  $\beta = 0$ , and non-rigid editing use  $\alpha < 0.4$ ,  $\beta > 0.6$ . For the tuning-free model, DNI is applied to all injections of latents tuned by a source prompt. Empirically, we found leveraging the noise branch in Eq. (4) also enhances effectiveness for some edits, so we utilize this by flexibly multiplying  $0 < \gamma < 2.0$  to noise branch and  $2 - \gamma$  to visual branch.

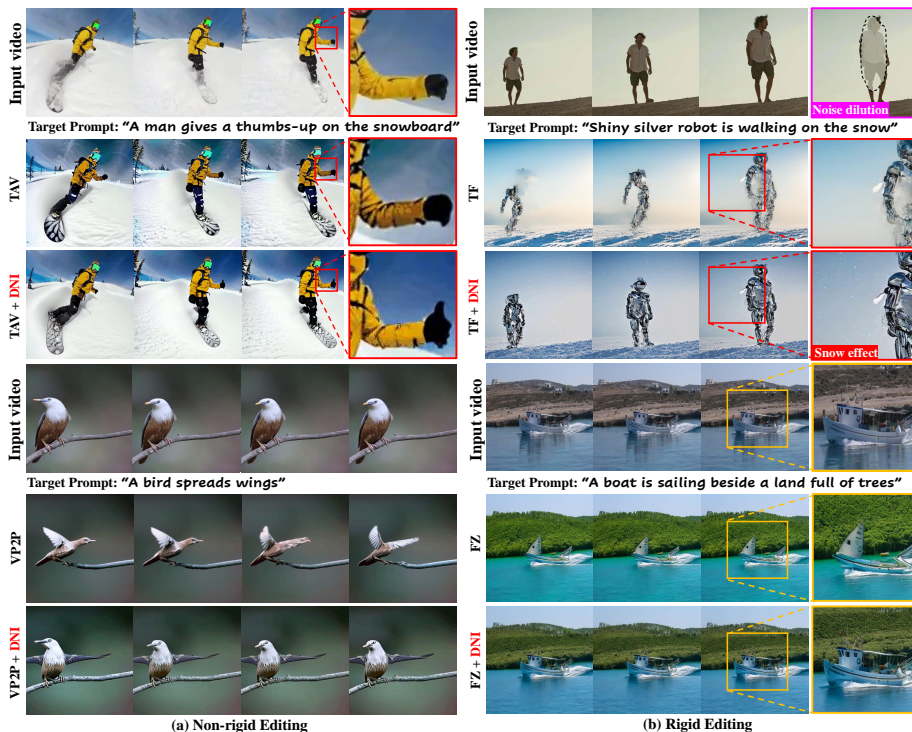
**Dataset and Baselines.** We validate videos on DAVIS [19] and LOVEU-TGVE [33], which are video editing challenge dataset<sup>2</sup> comprising 32 to 128 frames of each. DNI framework is validated about non-rigid/rigid editing on recent editing systems including Tune-A-Video (TAV) [32], Video-P2P (VP2P) [15], FateZero (FZ) [20] and TokenFlow (TF) [4] on their public codes.

### 5.2 Evaluation Metric

Editing results are validated using four criteria: (1) textual alignment, (2) input fidelity, (3) frame consistency, and (4) human preference. The textual alignment measures the semantic alignment between a target prompt and an edited video using the CLIP score [21] and PickScore [12]. The PickScore approximates human preferences by a large-scale trained model. The fidelity measures the preservation of original content in the unedited region using learned perceptual image patch similarity (LPIPS), and structural similarity index measure (SSIM). The frame consistency measures image CLIP scores between sequential frames and measures fréchet video distance (FVD) to evaluate the naturalness of videos. For the human evaluation, we investigate the preferences of edited videos according to the target prompt between the editing models and the models with DNI.

<sup>2</sup> <https://sites.google.com/view/loveucvpr23/track4>





**Fig. 7:** Qualitative results of applying DNI on recent editing systems according to (a) non-rigid editing (motion change) and (b) rigid editing (style transfer, object overlay). TAV: Tune-A-Video, VP2P: Video-P2P, FZ: FateZero, TF: TokenFlow. The yellow box shows a zoomed view of fidelity on unedited regions and the red box shows editing effects. Editing reference words are {gives, thumbs-up}, {spreads, wings}, {shiny, silver, robot, snow}, {land, trees} respectively for each sample. The noise dilution in the top of the right (*i.e.* region in black dotted line) is visualized by overlapping together with the input frame to show that the editing based on dilution is seamlessly connected with the surroundings, even to naturally extend the scope of dilution.

### 5.3 Experimental Results

**Qualitative Results.** Fig. 7 presents qualitative results of DNI framework incorporating with recent editing systems [4, 15, 20, 32]. To validate qualitative impact of DNI framework, we conduct case studies in terms of two distinct editing categories: (a) non-rigid editing and (b) rigid editing. For non-rigid editing, current editing systems fail to synchronize with the intended target prompt, resulting in the reconstruction of original input videos or the improper fusion of original contents (*e.g.* trees) with the desired alterations (*e.g.* wings). However, these models using the DNI effectively perform non-rigid editing on humans and objects. Remarkably, motion editing for actions such as a thumbs-up (*i.e.* red box) is selectively performed based on the visibility of the skier’s hand. This

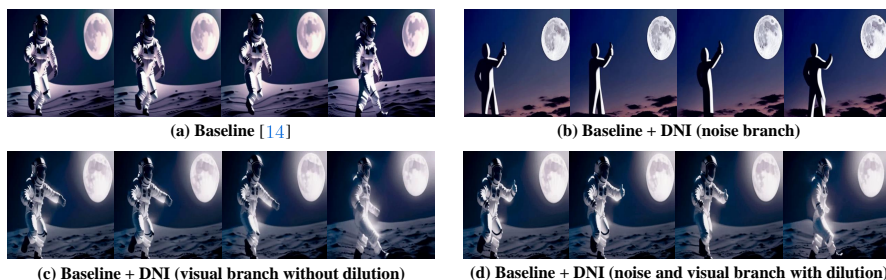
**Table 1:** Evaluations of edited videos from DAVIS and TGVE for non-rigid/rigid type editing in terms of textual alignment, fidelity, consistency, and human preference. CLIP\*: text-video clip, CLIP†: image-image clip, P-Score: PickScore, PF: preference

	Textual Alignment		Fidelity		Consistency		Human
	CLIP* ↑	P-Score ↑	LPIPS ↓	SSIM ↑	CLIP† ↑	FVD ↓	PF ↑
TAV [32]	22.6/27.1	19.5/20.2	0.193/0.181	0.621/0.653	0.921/0.952	3481/3392	0.14
TAV+DNI	27.6/28.5	20.6/20.9	0.168/0.161	0.706/0.711	0.952/0.961	3270/3151	0.86
FZ [20]	21.2/26.1	19.4/20.1	0.173/0.165	0.636/0.643	0.958/0.963	3319/3106	0.34
FZ+DNI	26.1/28.7	20.1/21.2	0.168/0.157	0.672/0.687	0.965/0.968	3209/3071	0.66
VP2P [15]	22.5/27.2	19.6/20.0	0.181/0.172	0.645/0.677	0.954/0.958	3231/3095	0.38
VP2P+DNI	<b>27.9/29.3</b>	<b>20.9/21.3</b>	0.161/0.158	0.717/0.719	0.961/0.964	3135/2953	0.62
TF [4]	21.7/27.4	19.4/20.1	0.160/0.157	0.653/0.677	0.971/0.974	3152/3043	0.41
TF+DNI	25.9/ <b>29.6</b>	20.7/ <b>21.5</b>	<b>0.143/0.151</b>	<b>0.731/0.733</b>	<b>0.980/0.977</b>	<b>3103/2912</b>	0.59

precision is attributed to the sensible application of noise dilution, which is selectively applied to the visible editing region (*i.e.* the hand) throughout the video frames. For rigid editing, including object overlay at the top and style transfer at the bottom, both current editing systems and those enhanced with DNI achieve qualitatively appropriate modifications. However, upon closer comparison, solely the models utilizing the DNI framework preserve a superior fidelity within the unedited regions (*i.e.* yellow box) of the video. It is considered that selective dilution by editing the guidance mask improves fidelity in the model. At the top, we adapted the model to transform a man into a robot walking in the snow. Intriguingly, the model with the DNI framework not only alters the human’s appearance but also adds a snowing effects in the sky. For this sample, we also visualize the diluted region on top of the input frame (*i.e.* frame outlined by pink color) indicated by a black-dotted line.<sup>3</sup>, where it shows that the scope of modification surpasses the initially established diluted perimeters. This indicates that the editing based on the diluted latent noise is seamlessly connected with the surrounding area and naturally extends the editing effect to the surroundings, enhancing the overall effectiveness of the editing.

**Quantitative Results.** Tab. 1 provides evaluations of non-rigid and rigid editing on DAVIS and TGVE videos using recent editing systems with the DNI framework. The assessments covers textual alignment, fidelity, consistency, and human evaluation. The effectiveness of the DNI framework is validated across all the video editing systems, with a notable enhancement in textual alignment particularly observed within the realm of non-rigid editing. For fidelity, it measures the preservation of unedited areas in the video after masking the same specified regions for editing. The fidelity is lower in the tuning-based models (*i.e.* TAV, Video-P2P) compared to tuning-free models (*i.e.* FateZero, TokenFlow).

<sup>3</sup> Although dilution is applied over the initial latent noise, we marked an area over the actual frame to help qualitatively understand the dilution effect.

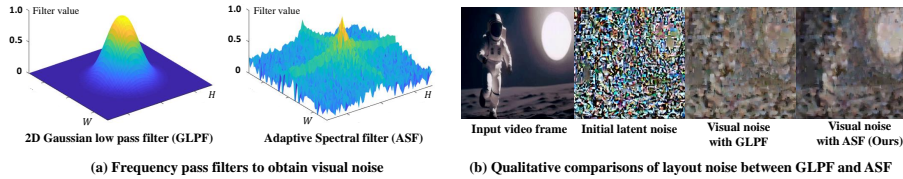


**Fig. 8:** Ablation studies about visual and noise branches in DNI. The input video is shown in Fig. 1. The target prompt is “Astronaut gives a thumbs-up under the moon”.

Tuning-based models train the diffusion model based on the input video, which sometimes leads to overfitting. Therefore, in the inference of editing, overly adhering to the trained video results in unnatural edits. In these models, dilutional noise identifies areas for additional noise to obscure and reduces over-reliance on the initial visual structure. This allows for better adaptation to the target prompt, improving naturalness, fidelity, and consistency. Tuning-free models excel in consistency but are less responsive to target prompts, making them suitable only for rigid editing due to their dependence on input video structure in the initial latent noise. The dilutional noise adaptively reduces this reliance and enhances the effectiveness of editing.

#### 5.4 Ablation Study

**Ablative results about noise and visual branch.** To investigate the effectiveness of the visual and noise branches in the editing of DNI framework, in Fig. 8, we perform ablation studies of these two branches about editing a video based on a target prompt “Astronaut gives a thumbs-up under the moon”. The target prompt demands a complex blend of rigid and non-rigid editing, necessitating the adjustment of the hand’s pose and reimaging the pale glow planet behind as moon. Fig. 8 (a) shows the editing results for the baseline [15]. In the edited video, the planet in the sky is changed into the moon, but the hand gesture fails to transition from the original motion to the intended thumbs-up, displaying an unnatural pose between the thumbs-up and its original stance. Fig. 8 (b) and (c) show the results of integrating the DNI framework using only one of the branches, either the visual branch or the noise branch. The results of (b) show the edited frames using only the noise branch. While effective changes are shown in the results, they significantly deviate from the input video, especially in the original astronaut. This indicates that the input video’s visual structure within the initial latent noise plays a role in maintaining fidelity to the input video. The outcomes from (c) illustrate the results of employing the visual branch without dilation. Unlike (b), these results maintain a strong fidelity to the original video, yet they fall short in executing non-rigid editing for



**Fig. 9:** Frequency pass filters for extracting visual noise from initial latent noise. (top-left: Gaussian low pass filter, top-right: our proposed adaptive spectral filter, bottom: frequency-filtered results using GLPF and ASF).

a thumbs-up, resulting in edits that display awkward motion through unnatural elongation of the arm. The results of (d) show the edited video combining the two branches with applying dilation to the visual branch. It maintains fidelity to the astronaut well, while successfully applying non-rigid editing about thumbs-up. Notably, when the arm is shaded by the torso, (*i.e.* fourth frame) the thumbs-up also disappears, resulting in a natural movement. This denotes that dilation is discerningly applied based on the motions displayed within the video frames.

#### Effectiveness of adaptive spectral filter.

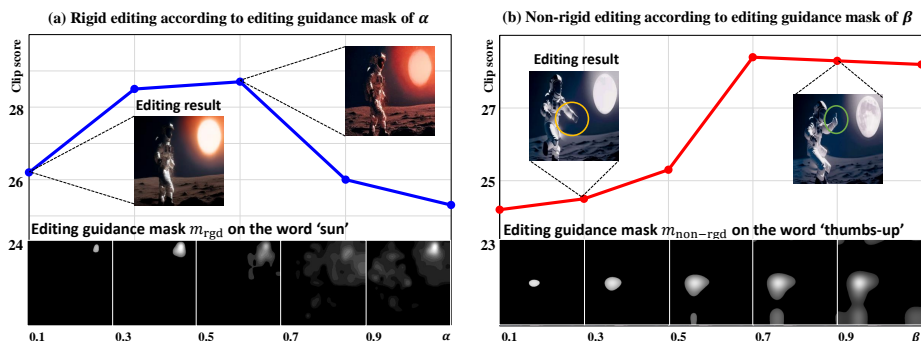
The initial latent noise fed into video editing systems contains the inherent visual structure of the input video. Within the DNI framework, noise disentanglement aims to isolate this input video structure from the initial latent noise. Fig. 9 shows the adaptive spectral filter (ASF) we designed for this purpose.<sup>4</sup> As shown in Fig. 5, the input video contains multiple frequencies ranging from low to high, and some of these remain in the initial latent noise. To appropriately capture these, we use the frequencies of the input video as a frequency filter by scaling between 0 to 1. To demonstrate the effectiveness of the ASF, we compare it with a Gaussian low pass filter (GLPF) in the spatial domain. Fig. 9 (b) shows the input video frame, its initial latent noise, and frequency-filtered visual noise using the GLPF and ASF. Qualitatively, the visual noise obtained using ASF more clearly reveals the input video’s contents. To quantitatively measure this, in Tab. 2, we conducted comparisons in terms of the peak signal-to-noise ratio (PSNR) with the input video and filtered videos using the ASF and variants of GLPF<sup>5</sup> by adjusting  $\sigma$  to leverage filtering frequency band. This shows that ASF extracts the input video structure more effectively than all variations of the GLPF.

**Table 2:** Quantitative validation to assess the involvement of input video’s component within the visual noise, filtered by different frequency pass filters.

Frequency pass filter type	PSNR
Gaussian low pass filter ( $\sigma = 1$ )	6.2
Gaussian low pass filter ( $\sigma = 3$ )	9.1
Gaussian low pass filter ( $\sigma = 5$ )	13.2
Gaussian low pass filter ( $\sigma = 10$ )	11.3
adaptive spectral filter (Ours)	<b>18.4</b>

<sup>4</sup> Although ASF is a 3D filter, for visual clarity, we show a 2D filter in spatial domain.

<sup>5</sup> It follows the function of  $G(x, y) = e^{-(x^2+y^2)/2\sigma^2}$  to scale from 0 to 1.



**Fig. 10:** Sensitivity analysis about textual alignment and edited video according to editing guidance mask modulated by  $\alpha$  and  $\beta$  in Eq. (3). For the setting of rigid editing in (a), the target prompt uses “Astronaut walks through red sunset”, the parameter  $\beta = 0$  is fixed, and the reference word for editing is specified as ‘red’ and ‘sunset’, where the mask above visualizes the attention map of ‘sunset’. For non-rigid editing of (b), the target prompt uses “Astronaut gives a thumbs-up under the moon”, the  $\alpha = 0.2$  is fixed, the reference word for editing is specified as ‘gives’, ‘thumbs-up’, and ‘Astronaut’ for (b), where the mask visualizes attention map of ‘thumbs-up’.

**Sensitivity analysis on editing guidance mask.** Fig. 10 presents a sensitivity analysis of the editing results and their textual alignment in response to changes in the editing guidance mask variation. Fig. 10 (a) shows the editing results for rigid editing, where ‘red’ and ‘sunset’ are selected as reference words from the target prompt to obtain the editing guidance mask. Based on our designed Eq. (3), the influence of rigid editing mask  $m_{\text{rgd}}$  can be adjusted according to variation of mask scaler  $\alpha$ . Therefore, we investigate the results of rigid editing according to variation of  $\alpha$ , and below, the mask  $m_{\text{rgd}}$  is visualized together. (*i.e.* To enhance the visibility of changes in the attention map, we visualized the map for a single word ‘sunset’.) For rigid editing, a 16x16 attention mask is obtained from the down-block of the UNet, which is then resized to 64x64. It can be observed that as  $\alpha$  increases, the influence of the mask becomes more intensive amplified by the  $\alpha$ . Thus, the edited area also expands, and simultaneously, more effects are incorporated into the video. This leads to an improvement in textual alignment, and it can also be observed that alignment starts to decline at points where  $\alpha$  exceeds 0.7, indicating that excessive effects have been introduced. Fig. 10 (b) displays the edited videos and textual alignment for non-rigid editing according to mask scaler  $\beta$ . The reference words selected were ‘gives’, ‘thumbs-up,’ and ‘Astronaut’. The mask  $m_{\text{rgd}}$  for ‘Astronaut’ utilizes a 16x16 attention map for the rigid editing,<sup>6</sup> and the masks  $m_{\text{non-rgd}}$  for ‘gives’ and ‘thumbs-up’ employs an 8x8 attention map provided by

<sup>6</sup> Natural language toolkit [16] is used to automatically classify part of speech about words in target prompt, where DNI uses the rigid editing mask for words about objectives and adjectives and the non-rigid editing mask for words in predicate.

the mid-block of the UNet to enhance non-rigid editing. By fixing  $\alpha = 0.2$ , we investigate the variation of non-rigid editing according to the mask scaler  $\beta$ . As the  $\beta$  increases, it is observed that the highlighted area for non-rigid editing expands, and simultaneously, the thumbs-up, which was not edited (*i.e.* yellow circle) with lower  $\beta$ , is gradually being synthesized. (*i.e.* green circle). This indicates that the expansion of dilution by the mask area mitigates the constraints imposed by the input video, allowing the synthesis effect to be free from the original motion of an object (*i.e.* walking motion of arms and hands). Therefore,  $\alpha$  and  $\beta$  influence editing effectiveness and also hold editing robustness regions (*i.e.*  $0.3 < \alpha < 0.7$  for rigid editing and  $\beta > 0.6$  for non-rigid editing) to properly synthesize desired attributes to conform to the target prompt.

### Image editing with DNI framework.

The DNI framework is structurally designed for flexible adaptability within diffusion-based editing systems, such that we apply our work to diffusion-based image editing. Fig. 11 illustrates the enhanced capabilities of the editing system [6] when incorporated with the DNI framework, showcasing the successful application of non-rigid editing that was previously unattainable by the baseline model. To be specific, the top of Fig. 11 shows non-rigid editing about the pose change, where the current image editing model outputs an image similar to the input image, unable to change the pose according to the target prompt. However,

when integrated with the DNI framework, it demonstrates the ability to perform precise editing. The bottom of Fig. 11 represents non-rigid editing that changes the view. The model, in attempting to switch to a zoomed-out view, is constrained by the layout of the input image and fails to transform correctly. The results of the incorporation with the DNI framework show that editing can be performed more freely, not restricted by the layout.



**Fig. 11:** Application DNI framework into image editing system [6] in terms of non-rigid editing.

## 6 Conclusion

This paper introduces a diffusion-based video editing framework, termed Dilutional Noise Initialization (DNI), designed to facilitate intricate non-rigid modifications of subjects or objects within video. We introduce a novel concept of ‘noise dilution’ which adds Gaussian noise into initial latent noise to alleviate the restrictive influences imposed by the input video’s visual structure on the specified editing regions. DNI can be easily applied to any diffusion-based editing system in a model-agnostic manner and enhances them to perform non-rigid editing. Extensive experiments validate its editability and visual effectiveness.



## Acknowledgements

This work was partly supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments) and partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C2012706).

## References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800 (2022)
2. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. arXiv preprint arXiv:2401.09047 (2024)
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* **34**, 8780–8794 (2021)
4. Geyer, et al, M.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
5. 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)
6. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
7. 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)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020)
9. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)
10. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
12. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. arXiv preprint arXiv:2305.01569 (2023)
13. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)
14. Koo, G., Yoon, S., Yoo, C.D.: Wavelet-guided acceleration of text inversion in diffusion-based image editing. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 4380–4384. IEEE (2024)



15. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv preprint arXiv:2303.04761 (2023)
16. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002)
17. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
18. 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)
19. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
20. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
22. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
23. 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)
24. 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)
25. 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* (2022)
26. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
27. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
28. 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)
29. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. *Advances in neural information processing systems* **30** (2017)
30. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: Nüwa: Visual synthesis pre-training for neural visual world creation. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*. pp. 720–736. Springer (2022)
31. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-

- to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
32. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565 (2022)
  33. Wu, J.Z., Li, X., Gao, D., Dong, Z., Bai, J., Singh, A., Xiang, X., Li, Y., Huang, Z., Sun, Y., et al.: Cvpr 2023 text guided video editing competition. arXiv preprint arXiv:2310.16003 (2023)
  34. Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., Yu, D.: Diffsound: Discrete diffusion model for text-to-sound generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)
  35. Yoon, S., Koo, G., Kim, G., Yoo, C.D.: Frag: Frequency adapting group for diffusion video editing. arXiv preprint arXiv:2406.06044 (2024)