FRAG: Frequency Adapting Group for Diffusion Video Editing

Sunjae Yoon [1] **Gwanhyeong Koo** [1] **Geonwoo Kim** [1] **Chang D. Yoo** [1]

Abstract

In video editing, the hallmark of a quality edit lies in its consistent and unobtrusive adjustment. Modification, when integrated, must be smooth and subtle, preserving the natural flow and aligning seamlessly with the original vision. Therefore, our primary focus is on overcoming the current challenges in high quality edit to ensure that each edit enhances the final product without disrupting its intended essence. However, quality deterioration such as blurring and flickering is routinely observed in recent diffusion video editing systems. We confirm that this deterioration often stems from high-frequency leak: the diffusion model fails to accurately synthesize high-frequency components during denoising process. To this end, we devise Frequency Adapting Group (FRAG) which enhances the video quality in terms of consistency and fidelity by introducing a novel receptive field branch to preserve high-frequency components during the denoising process. FRAG is performed in a model-agnostic manner without additional training and validates the effectiveness on video editing benchmarks (i.e., TGVE, DAVIS).

1. Introduction

Denoising diffusion models (Dhariwal & Nichol, 2021; Song et al., 2020b;a; Ho et al., 2020) have significantly advanced the generative capabilities of artificial intelligence, leading to groundbreaking achievements in image, speech, and video generation. We focus here on video editing based on diffusion which holds immense promise for revolutionizing the entertainment industry. Video editing systems (Bar-Tal et al., 2022; Wu et al., 2023a; Geyer et al., 2023) are designed to work with both the input video and a target text prompt that outlines the user's desired modifications. The systems incorporate these modifications into the video, ensuring that the edits are seamless and unobtrusive. This process is then carefully managed to produce a coherent final output that maintains a natural flow, whilst aligning closely with the original input video.

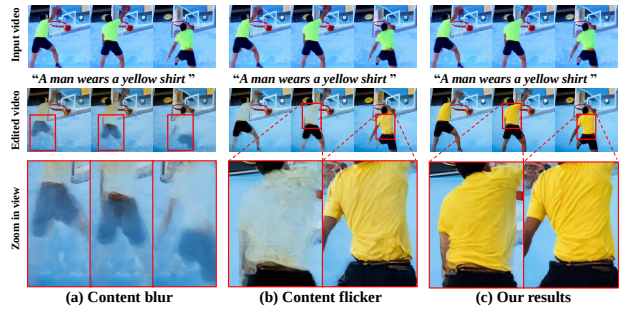


Figure 1. Illustration of video quality deterioration represented into two distinct categories: (a) content blur and (b) content flicker. For the comparison, we present our results in (c).

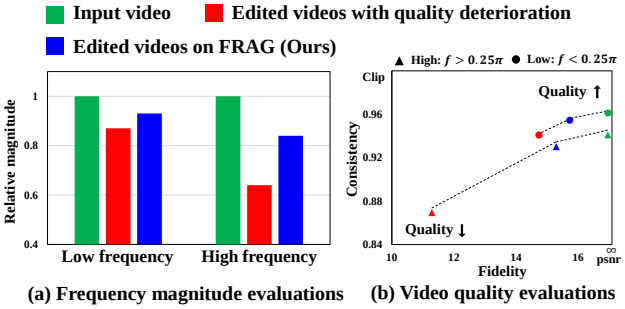


Figure 2. (a) Frequency magnitude evaluations of videos according to low and high frequencies. (b) Video quality evaluations about frame consistency and fidelity according to low and high-frequency components in videos. Normalized frequency $0 < f < \pi$, low frequency: $f < 0.25\pi$, high frequency: $f > 0.25\pi$.

Recent advancements (Wu et al., 2023a; Liu et al., 2023; Geyer et al., 2023) in video editing systems have aimed at preserving the temporal consistency across edited frames. However, a significant challenge persists as these systems often struggle with maintaining the quality of various attributes, including the color and shape of objects. This inconsistency manifests not just over time but also across the spatial dimensions of the video, leading to a deterioration in the overall quality of the edits. To be specific, Figure 1 illustrates two distinct types of quality deterioration: (1) content blur and (2) content flicker. As shown in Figure 1

[1]Korea Advanced Institute of Science and Technology (KAIST), South Korea. Correspondence to: Chang D. Yoo <cd_yoo@kaist.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

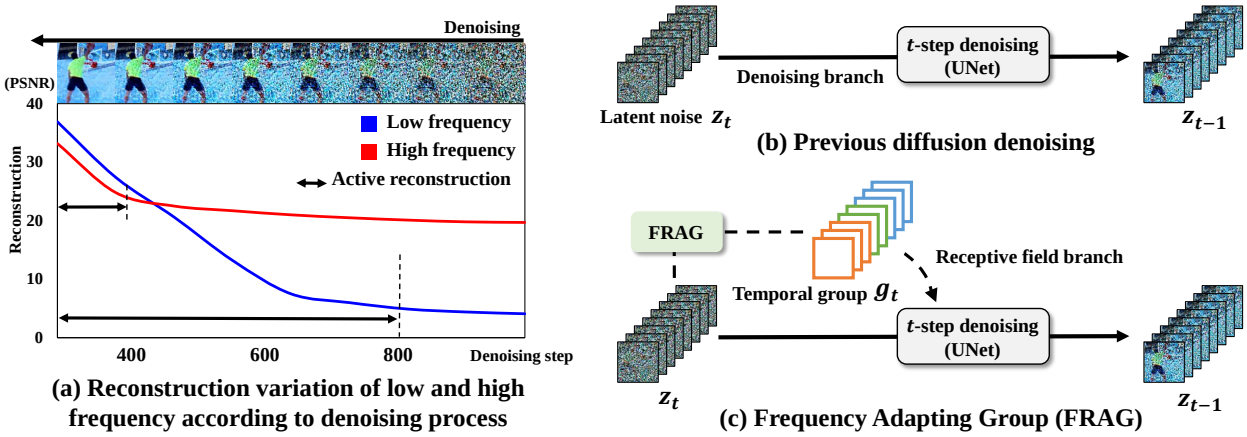


Figure 3. (a) shows experimental observations about latent noise reconstruction in terms of low and high frequencies, where high-frequency components are synthesized later in denoising than low frequencies. (b) illustrates previous video diffusion denoising and (c) illustrates our proposed denoising with the receptive field branch using Frequency Adapting Group (FRAG) to enhance the quality of editing.

(a), the content blur denotes that attributes (*e.g.*, color and shape) synthesized for a content (*e.g.*, shirt) are irregularly mixed with other unintended contents (*e.g.*, background) in the entire video, which makes the content faint or unclear. Figure 1 (b) also shows another case of quality deterioration as a content flicker. This indicates a disruption of visual continuity in a synthesized attribute at a certain moment, causing the attribute (*e.g.*, shirt color) to display contrasting characteristics (*e.g.*, light and dark) at different times.

Our observation suggests that blurring and flickering are due to a high-frequency leak in the diffusion denoising process. The high-frequency leak denotes a shortfall of the video diffusion model's ability to accurately synthesize high-frequency attributes during the denoising process, leading to the lack of high-frequency components. Figure 2 presents experimental evidence about the high-frequency leak of current diffusion video editing systems (Geyer et al., 2023; Wu et al., 2023a; Khachatryan et al., 2023). In Figure 2 (a), we collected edited videos exhibiting quality deterioration and measured the average magnitude of frequency by converting them into low and high frequencies using a spatial frequency filter. These videos show a deficiency in high-frequency components. Furthermore, in Figure 2 (b), qualities related to high frequency such as frame consistency and fidelity are degraded compared to those of the input video.[1] We further investigate to identify denoising dynamics in latent noise in terms of low and high frequency. Figure 3 (a) shows the reconstruction[2] of low and high-frequency latent noises according to the denoising step. Notably, low-frequency components are reconstructed in the early denoising steps, while high frequencies tend to be reconstructed later. This denotes that it is crucial to properly capture and preserve synthesized attributes at each frequency generation.

To achieve this, we devise Frequency Adpating Group (FRAG) for diffusion video editing, which enhances the quality of edited videos by effectively preserving high-frequency components. As shown in Figure 3 (c), FRAG has an auxiliary branch for denoising process on top of the original denoising branch in Figure 3 (b). This branch is defined as receptive field branch which guides denoising UNet (Ronneberger et al., 2015) to properly synthesize the frequency components during the denoising process. To be specific, this receptive field decides the frame-level operating range for the quality enhancement modules (*e.g.*, attention, propagation) within the UNet. Previously, this field has employed fixed sliding windows or the entire video length (Geyer et al., 2023; Wu et al., 2023a), where both inevitably lead to a high-frequency leak problem.[3] Thus, we devise a dynamic receptive field referred to as temporal group g_t, which adaptively refines the field according to synthesized frequency variations in each denoising step t. Following the frequency characteristics of denoising in Figure 3 (a), g_t builds large receptive fields in early denoising to facilitate the generation of low frequencies. As the denoising progresses, g_t shifts to forming numerous smaller fields for high frequencies. FRAG works in a model-agnostic manner without additional training and validates its effectiveness of quality on video editing benchmarks (*i.e.*, TGVE, DAVIS).

[1] For consistency, due to the lack of supervision of editing, we applied editing (*e.g.*, style transfer) conforming to the consistency patterns in the input video, and measured clip score of input and output videos. For fidelity, we measure peak signal-to-noise (psnr) about the unedited region between input and output videos. Please note the difference of psnr between low and high frequencies in the samples (*i.e.*, red circle and triangle) exhibiting the deterioration.

[2] We measure the psnr between the input video and video decoded by each step latent noise.

[3] A fixed small receptive field leads to flicker and a wide field leads to blur. Please see more details of this in Appendix D.

2. Related Work

2.1. Diffusion-based Video Editing

Video editing aims to edit the input video as seamlessly and unobtrusively as possible incorporating the target text descriptions. The pre-trained text-to-image diffusion models (Rombach et al., 2022; Ramesh et al., 2022) have presented an effective solution for generative editing, where earlier works (Kim et al., 2022; Hertz et al., 2022) in image editing laid the foundation for the development of controlled synthesis of visual information. Extending the work in image, diffusion-based video editing (Molad et al., 2023; Wu et al., 2023a) has been attempted based on the video diffusion models (Ho et al., 2022; Hong et al., 2022). To achieve accurate editing outcomes aligned with the target text, it is crucial to have controlled synthesis capabilities. Thus, there have been lines of works (Zhang et al., 2023; Liu et al., 2023) to improve text-conditioned editing controllability. To enhance the efficiency of diffusion editing, zero-shot frameworks (Khachatryan et al., 2023; Qi et al., 2023) have been proposed. These frameworks remove the process of training with the input video. In particular, maintaining video quality across resulting frames is another crucial issue for video editing. We further elaborate on this in the following.

2.2. Diffusion Video Editing Quality Enhancement

The quality of edited video is evaluated through two standards: (1) frame consistency and (2) fidelity. Frame consistency refers to the uniformity and coherence of consecutive frames in a video, while fidelity refers to the degree to which the edited video maintains the integrity and quality of the original content that is not meant to be altered. There are three popular choices for diffusion video editing to video quality enhancement: (1) attention, (2) propagation, and (3) prior guidance. Attention-based approach (Wu et al., 2023a; Liu et al., 2023) is a method to highlight the visual commonality across frames based on their feature similarities. It is effective in maintaining consistency based on contextual understanding of video scenes. The propagation-based approach (Khachatryan et al., 2023; Geyer et al., 2023) selects a pivotal key frame and shares its visual attribute with the attributes in frames within a given temporal receptive field of propagation. This ensures a highly consistent video at a visual attribute level. The prior guidance methods (Cong et al., 2023; Chai et al., 2023) perform editing following precomputed prior observations (*e.g.*, optical flow, object mask), which effectively enhances fidelity to the input video. Although all of these approaches have pursued quality enhancement, they are still vulnerable to quality deterioration due to high-frequency leaks. Thus, we design an adaptive receptive field branch to guide quality enhancement modules to have robustness on the frequency variation in denoising.

3. Preliminaries

3.1. Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) are parameterized Markov chains to sequentially reconstruct a noisy data $\{x_1, \cdots, x_T\}$ based on initial raw data x_0. To construct this, Gaussian noise is gradually added upto x_T via the Markov transition $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$ utilizing a predefined schedule α_t across steps $t \in \{1, \cdots, T\}$. This procedure is termed as *forward process* in the diffusion model. The counterpart to this, known as the *reverse process*, involves the diffusion model estimating $q(x_{t-1}|x_t)$ through trainable Gaussian transitions $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t))$, beginning from the normal distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$. The training objective of diffusion model is to maximize log-likelihood $\log(p_\theta(x_0))$ updating parameters θ. To this, we can apply variational inference of maximizing the variational lower bound about $\log(p_\theta(x_0))$, which builds a closed form of KL divergence[4] between two distributions p_θ and q. This whole process is summarized as introducing a denoising network $\epsilon_\theta(x_t, t)$ to predict noise $\epsilon \sim \mathcal{N}(0, I)$ as given below:

$$\mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t\sim\mathcal{U}\{1,T\}}[||\epsilon - \epsilon_\theta(x_t, t)||_2^2]. \quad (1)$$

where $\mathcal{U}\{1, T\}$ is discrete uniform distribution between 1 and T for training robustness on each step t.

3.2. Denoising Diffusion Implicit Model

Denoising diffusion implicit model (DDIM) (Song et al., 2020a) accelerates the reverse process of DDPM, which samples noisy data with a smaller number of T as below:

$$x_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}x_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right)\epsilon \quad (2)$$

We can also inverse this process to compute latent noise 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. For diffusion editing, noise initialization with this enhances fidelity to the input video.

3.3. Text-conditioned Diffusion Model

The text-conditioned diffusion model reconstructs the output data x_0 from random noise conditioned on a text prompt $\mathcal{T}$. The training objective also incorporates text condition under latent space for semantic interaction as $\mathbb{E}_{z,\epsilon,t}[||\epsilon - \epsilon_\theta(z_t, t, \mathbf{c})||_2^2]$, where $z_t = E(x_t)$ is a latent noise encoding (*e.g.*, VQ-VAE (Van Den Oord et al., 2017)) and $\mathbf{c} = \psi(\mathcal{T})$ is textual embedding (*e.g.*, CLIP (Radford et al., 2021)). Diffusion video editing takes z_t as the encoding of video data, and $\mathbf{c}$ for encoding the target text prompt.

[4] See the detailed proof in Appendix B.

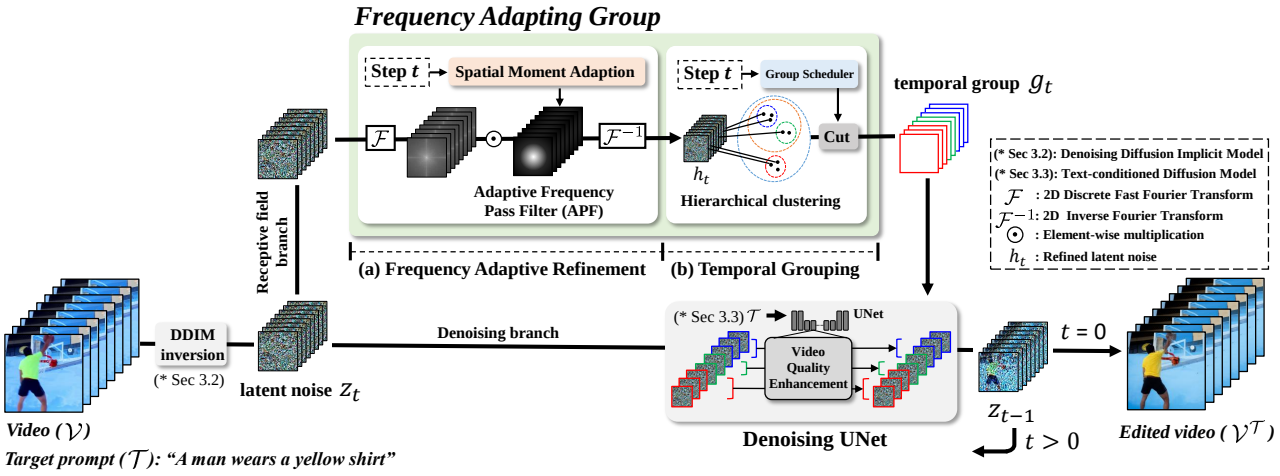


Figure 4. Illustration of Frequency Adapting Group (FRAG). FRAG takes t step latent noise z_t and produces receptive field g_t referred to as temporal group. The g_t guides denoising UNet to adaptively synthesize the frequency components according to frequency variations of latent noise during the denoising process. FRAG contains (a) frequency adaptive refinement that enhances the visual quality of attributes within latent noise and (b) temporal grouping that clusters latent noise frames to build g_t.

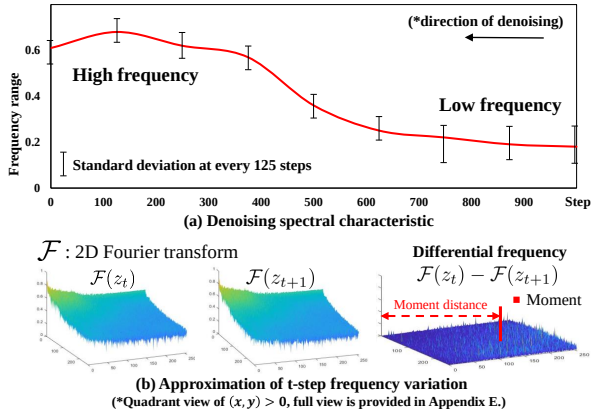


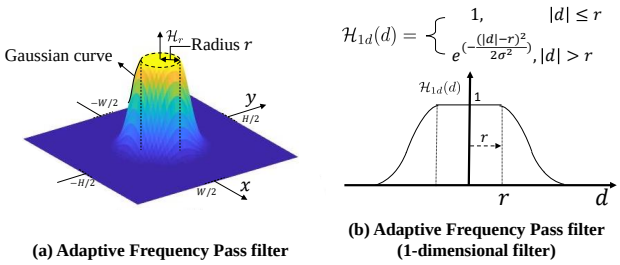
Figure 5. (a) Denoising spectral characteristic: Average frequency variation according to denoising from 1000 to 0 step on 800 videos in TGVE (Wu et al., 2023b) and UCF-101 (Soomro et al., 2012). (b) t step frequency variation: It is approximated by a normalized distance of moment in the differential frequency distribution.

3.4. Denoising Spectral Characteristic

Frequency analysis of latent space revealed the process of denoising is patterned by the spatial frequencies within the latent noises. For clarity, we term this pattern as 'denoising spectral characteristics'. This concept encapsulates the sequential synthesis of spatial frequency of latent noise in the denoising process: Low frequency is synthesized in the early stages, followed by the synthesis of high frequency in the subsequent phases. Figure 5 (a) demonstrates experimental observations of the denoising spectral characteristics, which measures the frequency variation of synthesized latent noise according to denoising. This shows a gradual increase in the frequency as the denoising process advances. To estimate the frequency variation, in Figure 5 (b), we transform each latent noise into spatial frequency and calculate each step differential frequency distribution by subtracting the previous step frequency. Based on differential distribution, we measure a normalized distance of the moment in it (Please see details in Sec 4.1) and approximate the distance as frequency variation. Leveraging this spectral characteristic, we present the Frequency Adapting Group in the following.

4. Frequency Adapting Group

Diffusion video editing system takes inputs of video $\mathcal{V}$ and target prompt $\mathcal{T}$, where it produces edited video $\mathcal{V}^{\mathcal{T}}$ conforming to the meaning of $\mathcal{T}$. Figure 4 shows the application of Frequency Adapting Group (FRAG) into the general diffusion video editing system, which allows both supervised (*i.e.*, tuning) and unsupervised (*i.e.*, tuning-free) models. FRAG aims to enhance the quality of edited videos by effectively preserving high-frequency components. At each denoising step t, FRAG takes an input latent noise z_t and produces a receptive field g_t referred to as a temporal group. This temporal group guides the quality enhancement module (*e.g.*, attention, propagation) in denoising UNet to preserve the frequencies dynamically synthesized during the denoising process. To perform this, FRAG comprises two main modules: (1) Frequency Adaptive Refinement (Sec 4.1) and (2) Temporal Grouping (Sec 4.2). Frequency adaptive refinement refines the visual quality of synthesized attributes within latent noise by applying our adaptive frequency pass filter. Based on this refinement, temporal grouping clusters frames with similar latent noise into temporal groups based on shared content. Finally, these groups are provided as receptive fields for quality enhancement of denoising UNet.



(a) Adaptive Frequency Pass filter

(b) Adaptive Frequency Pass filter (1-dimensional filter)

Figure 6. Visualization of Adaptive Frequency Pass filter (APF). (a) shows APF in 3D view, where x, y are spatial axes and $\mathcal{H}_r$ is the filter value. (b) shows a 1-dimensional APF $\mathcal{H}_{1d}$ of 2D view.

4.1. Frequency Adaptive Refinement

Frequency Adaptive Refinement aims to enhance the visual quality of synthesized attributes within the latent noise of each step. According to the denoising spectral characteristic, attributes are progressively synthesized from low to high frequency, such that we have devised an adaptive frequency pass filter (APF) capable of progressively passing the frequency ranging from low to high corresponding to the currently synthesized frequency. Thus, APF reveals the synthesized attributes better within latent noise by keeping track of frequency synthesizing trends. To be specific, as shown in Figure 6 (a), the APF is a 2-dimensional low-frequency pass filter denoting $\mathcal{H}_r \in \mathbb{R}^{L\times W\times H}$, where W, H are width and height of the filter. The L is the number of filters corresponding to frame length. It is a cylindrical structure of radius r (*i.e.*, $0 < r < \sqrt{(W/2)^2 + (H/2)^2}$) whose edges follow a Gaussian curve for smoothing effect (Figure 6 (b) presents a 1-dimensional APF to enhance the understanding of its shape). Formally, the $\mathcal{H}_r$ can be defined as below:

$$\mathcal{H}_r^i = \begin{cases} 1 & d \le r \\ e^{-(d-r)^2/2\sigma^2} & d > r, \end{cases} \tag{3}$$

where σ is coefficient for scaling the Gaussian curve. The superscript i denotes the i-th filter. (We omit this in the following for the simplicity.) The d is a distance of each point (x, y) in 2D frequency domain from the center point, satisfying $d(x, y) = \sqrt{x^2 + y^2}$. We multiply this $\mathcal{H}_r$ into the latent noise frequency and convert it back to the real domain, which preserves the components inside the radius r of the latent noise frequency as given below:

$$h_t = \mathcal{F}^{-1}(\mathcal{H}_r \odot \mathcal{F}(z_t)) \in \mathbb{R}^{L\times W\times H\times C}, \tag{4}$$

where z_t is t step latent noise and C is the channel. $\mathcal{F}, \mathcal{F}^{-1}$ are discrete time fast Fourier transform and inverse transform. $\odot$ is element-wise multiplication with broadcasting[5].

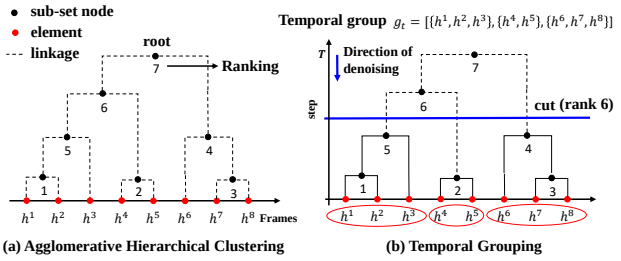
[5]Since in discrete time, $\mathcal{F}(z_t) \in \mathbb{R}^{L\times W\times H\times C}$ builds same dimension of latent noise $z_t \in \mathbb{R}^{L\times W\times H\times C}$. Thus, spatial frequency filter $\mathcal{H}_r \in \mathbb{R}^{L\times W\times H}$ is broadcasting to the channel axis.

Therefore, h_t is refined latent noise by $\mathcal{H}_r$. By expanding r, the $\mathcal{H}_r$ encompasses the generated frequency.[6] To achieve this, we introduce a spatial moment adaption below.

Spatial Moment Adaption. The spatial moment adaption adjusts the radius r of the adaptive filter $\mathcal{H}_r$ to include the frequency generated at each denoising step. To identify the generated frequency at t step, as shown in Figure 5 (b), we obtain differential frequency $\mathcal{Z}_t$ by subtracting the previous step frequency, from the t step as $\mathcal{Z}_t = \mathcal{F}(z_t) - \mathcal{F}(z_{t+1})$. Thus, the $\mathcal{Z}_t$ contains spatial frequencies generated during the t step denoising process, to cover these frequencies by radius r in APF, we calculate a point of the spatial moments about $\mathcal{Z}_t$ as M_x, M_y on a space $(x, y) > 0$ satisfying below:

$$M_x = \frac{\sum_{x,y} x\mathcal{Z}_t(x,y)}{\sum_{x,y} \mathcal{Z}_t(x,y)}, M_y = \frac{\sum_{x,y} y\mathcal{Z}_t(x,y)}{\sum_{x,y} \mathcal{Z}_t(x,y)}. \tag{5}$$

Finally, the radius for $\mathcal{H}_r$ is defined as $r = d(M_x, M_y) + d_0$ with margin d_0. Therefore r is adaptively updated following the t step synthesized frequency distribution.



(a) Agglomerative Hierarchical Clustering

(b) Temporal Grouping

Figure 7. (a) shows the agglomerative hierarchical clustering and (b) shows temporal grouping by group scheduler to cut the tree (a).

4.2. Temporal Grouping

Temporal grouping aims to provide a receptive field composed of multiple frame groups, where each group contains latent noise features containing similar attributes. Based on the refined latent noise $h_t \in \mathbb{R}^{L\times W\times H\times C}$, the temporal grouping measures their frame distances between consecutive frames and clusters frames based on the distances, producing multiple frame groups g_t. We first define frame distance between i-th and j-th frame latent noises by applying Euclidean distance as given below:

$$D(h^i, h^j) = ||h^i - h^j|| \in \mathbb{R}, \tag{6}$$

where h^i is i-th refined latent noise feature.[7] Based on our defined frame distance, we apply hierarchical clustering (HC) (Jain & Dubes, 1988) with agglomeration (*i.e.*, bottom-up). As shown in Figure 7 (a), the HC constructs a binary

[6]High frequencies can also be distinguished as they are much lower than the Gaussian noise frequencies within the latent noise.

[7]we skip subscript step t for the simplicity.

merge tree, initiating with individual elements (*i.e.*, single frame latent noise, h^i). It progressively merges the nearest sub-sets (*i.e.*, multi frames) in pairs, advancing towards the root of the tree which ultimately encompasses all elements. For each frame to cluster the closest pair of sub-sets, we extend the frame distance $D(h^i, h^j)$ into a sub-set distance between any two sub-sets of frames X and Y as $\Delta(X, Y) = \min_{h^i \in X, h^j \in Y} D(h^i, h^j)$ with minimum distance linkage. Therefore, the HC algorithm[8] builds a tree that links each h^i in order of minimum distance about $\Delta(X, Y)$, where the number below the linkage in Figure 7 (a) denotes the ranking of linkages. To build temporal groups g_t using this tree, as shown in Figure 7 (b), we cut one of the linkages in the tree. For the selection of linkage to cut, we design a 'group scheduler' which selects the ranking of linkage from high to low according to denoising. This is because, according to the denoising spectral characteristic, the attributes of high-frequency components are synthesized in the latter step of denoising (*e.g.*, $t < 400$), so the receptive field also needs to be narrow for appropriately clustering high-frequency components such as fine-grained attributes. Thus, the group scheduler cuts off high rankings in the beginning of denoising to form a small number of temporal groups with a wide range and cuts out low rankings in the later stages to form a large number of temporal groups with a short range. Formally, we choose a logistic curve for the group scheduler to stably decide the number of rankings according to each step t as given below:

$$n_{\text{cut}} = \lceil n_{\text{root}} \times (\alpha \log(T - t) + 1) \rceil \,, \tag{7}$$

where $n_{\text{cut}}, n_{\text{root}}$ are the integer numbers of rank to cut and the root rank. $T = 1000$ is a maximum step and $\alpha = -1/\log(T - 1)$ is scaler to fit output range from 1 to n_{root}. After cutting, we construct a temporal group (*e.g.*, $g_t = [\{h^1, h^2, h^3\}, \{h^4, h^5\}, \{h^6, h^7, h^8\}]$) by remained subsets.

4.3. Plug-and-Play FRAG

We integrate temporal group g_t into video editing systems by applying it into a quality enhancement module (*e.g.*, attention, propagation) of video diffusion UNet. In general, when the quality module is defined as $f : \mathbb{R}^{l \times d} \rightarrow \mathbb{R}^{l \times d}$, l is the range of frames (*e.g.*, $l = L$) to be performed of enhancement and d is the feature dimension (*e.g.*, $d = W \times H \times C$). We can simply update the l as g_t as below:

$$f : \mathbb{R}^{g_t^i \times d} \rightarrow \mathbb{R}^{g_t^i \times d}, \tag{8}$$

where the g_t^i is the i-th group of temporal group. Therefore, the diffusion model with FRAG adaptively denoises latent noise within the temporal groups designed for preserving synthesized frequencies throughout the denoising process.

[8]Please refer algorithm of agglomerative HC in Appendix C.

5. Experiment

5.1. Experimental Settings

Implementation Details. Diffusion video editing systems that apply FRAG use CLIP (Radford et al., 2021) for the text encoder and VQ-VAE (Van Den Oord et al., 2017) for the patch-wise image frames encoder. We use a pre-trained Stable Diffusion v2.1 (Rombach et al., 2022) for knowledge of editing. The experimental settings are $W = H = 64, L = 48, C = 4, \sigma = 0.25, d_0 = 6$ on NVIDIA A100 GPU. More details are also available in Appendix A.

Data and Baseline. We validated videos using the TGVE and DAVIS datasets, both of which are video editing challenge datasets[9] containing 32 to 128 frames each. FRAG is validated on recent editing systems including TokenFlow (Geyer et al., 2023), FLATTEN (Cong et al., 2023), Tune-A-Video (TAV) (Wu et al., 2023a), FateZero (Qi et al., 2023) on their public codes and papers.

5.2. Evaluation Metrics

Editing is measured based on the following five qualities: (1) frame consistency, (2), fidelity to input video, (3) spectral analysis, (4) textual alignment, (5) human preference. The frame consistency measures image CLIP scores between sequential frames and measures Fréchet Video Distance (FVD) to evaluate the naturalness of videos. The fidelity measures the preservation of original content in the unedited region using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The spectral analysis measures consistency and fidelity in terms of low and high frequency. The textual alignment assesses semantic coherence between a target prompt and an edited video, utilizing CLIP score. For human preference, we analyze the preferences for edited videos based on the target prompt. The details about capturing unedited regions for fidelity and human evaluation are provided in Appendix A.

5.3. Experimental Results

Qualitative Comparisons. To demonstrate the effectiveness of our proposed FRAG, as shown in Figure 8, we apply FRAG to the recent four video editing systems (*i.e.*, TokenFlow, FLATTEN, FateZero, TAV) and qualitatively evaluate the edit results. The top left shows results pertaining to TokenFlow which relies on a fixed-size receptive field for quality enhancement module (*i.e.*, propagation) in all denoising steps. This results in a content blur (*i.e.*, yellow box) in terms of the colors and edges. However, TokenFlow with FRAG solves this blurring by dynamically configuring the receptive field during each denoising process, and it can be seen that the attributes become clearer. The frames at

[9]https://sites.google.com/view/loveucvpr23/track4

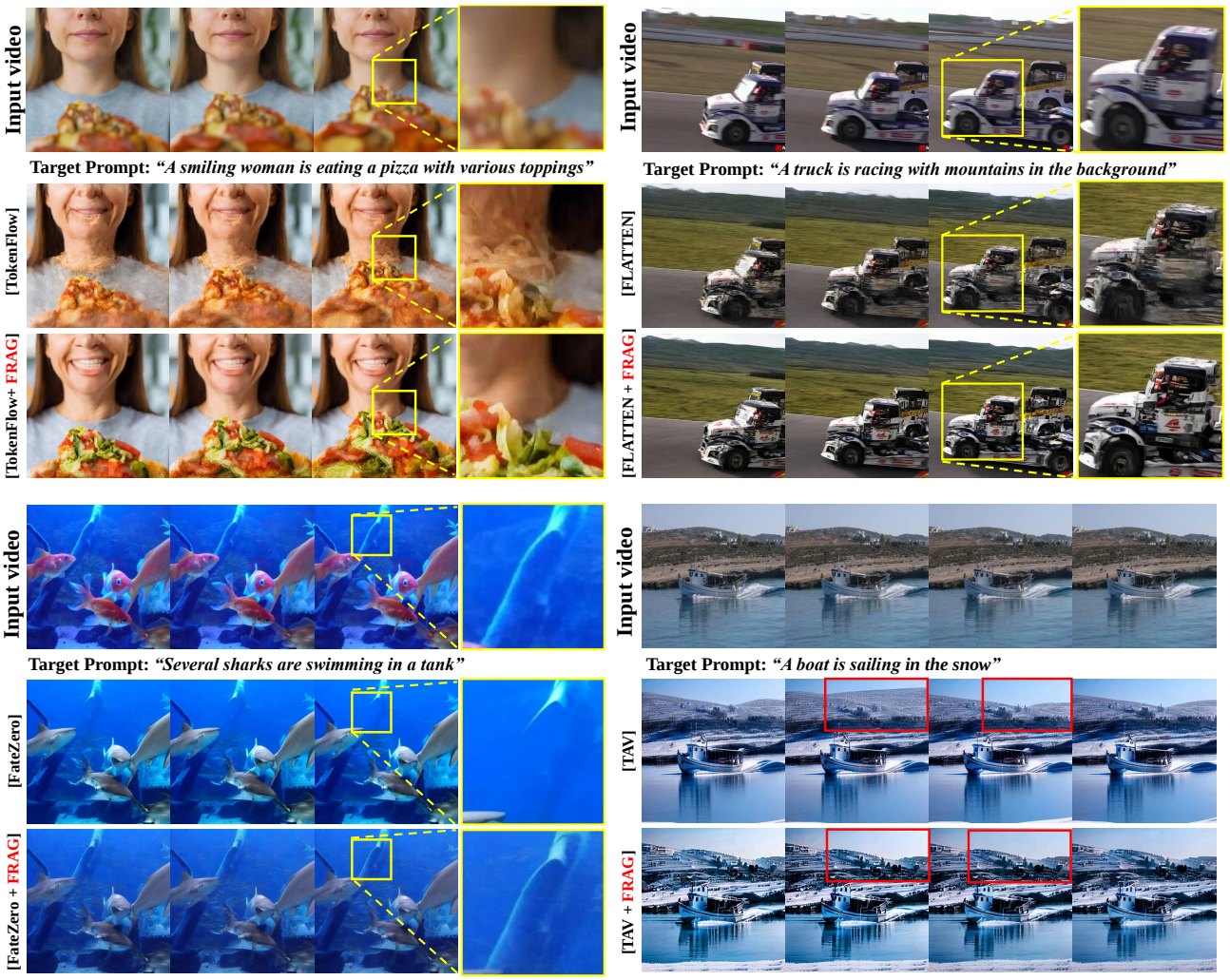


Figure 8. Qualitative result about applying FRAG on recent editing systems (TokenFlow: propagation-based model, FLATTEN: optical flow guidance, FateZero: zero-shot model, Tune-A-Video: attention based model), (Red box: content flicker, yellow box: content blur).

the top right show the results about FLATTEN[10], where it uses auxiliary guidance (*i.e.*, optical flow) to preserve consistency using pre-trained model (Teed & Deng, 2020). The editing outcomes demonstrate high consistency and fidelity, yet they still reveal instances of content blur (*i.e.*, yellow box). The model, when integrated with FRAG, effectively mitigates the blur, delivering clearer attributes while preserving a high degree of fidelity to the input video. The frames at the bottom left show results about FateZero, which performs video editing in zero-shot approach. In this model, blurring is also observed in the editing results. The application of FRAG notably enhanced the blurring, concurrently, improving the fidelity (*i.e.*, Tree of background correctly is preserved). We think that building a receptive field under consideration of frequency in latent noise can lead to consistent alterations of the same attribute during the attribute synthesizing process, ensuring uniformity in changes. The frames displayed on the lower right are the outcomes of TAV, a tuning-based attention approach that has served as a foundational baseline for numerous editing systems. TAV employs sliding window attention for quality enhancement, however, it exhibits severe temporal flickering. When integrated with FRAG, TAV improves this flickering issue and preserves high frequencies such as fine-grained details.

Quantitative Results. Table 1 presents evaluations of edited videos on DAVIS and TGVE of recent editing systems with FRAG in five criteria (*i.e.*, consistency, fidelity, spectral analysis, alignment, human evaluation). The baselines include different types of quality enhancement modules (*i.e.*, TAV: attention, TokenFlow: propagation, FLATTEN: optical flow), and the effectiveness of FRAG is confirmed in all the models. The consistency and fidelity are

[10]We reproduce it based on their paper.

Table 1. Quantitative results of edited videos on DAVIS and TGVE based on editing systems with FRAG about consistency (frame consistency), fidelity (fidelity to input video), spectral analysis (consistency and fidelity of low/high normalized frequency f, low: f <0.25π, high: f >0.25π), alignment (textual alignment), and human (preference). CLIP*: text-video clip, CLIP†: image-image clip.

	Consistency		Fidelity		Spectral Analysis		Alignment	Human
	CLIP† ↑	FVD ↓	PSNR ↑	SSIM ↑	CLIP† ↑	PSNR ↑	CLIP* ↑	Preference ↑
TAV	0.932	3452	13.7	0.647	0.961 / 0.872	14.1 / 11.0	25.2	0.27
TAV + FRAG	**0.951**	**3251**	**14.9**	**0.687**	**0.965 / 0.913**	**15.2 / 13.2**	**25.7**	0.73
FateZero	0.945	3241	13.9	0.651	0.969 / 0.893	14.3 / 12.6	24.5	0.39
FateZero + FRAG	**0.956**	**3119**	**15.3**	**0.694**	**0.971 / 0.911**	**15.6 / 13.9**	**25.1**	0.61
FLATTEN	0.962	3002	14.2	0.672	0.968 / 0.911	14.6 / 12.7	25.4	0.41
FLATTEN + FRAG	**0.970**	**2951**	**15.3**	**0.702**	**0.971 / 0.928**	**16.0 / 14.8**	**25.6**	0.59
TokenFlow	0.968	2984	15.1	0.691	0.972 / 0.931	15.1 / 13.1	25.8	0.43
TokenFlow + FRAG	**0.978**	**2841**	**18.2**	**0.736**	**0.981 / 0.954**	**18.2 / 16.3**	**26.4**	0.57

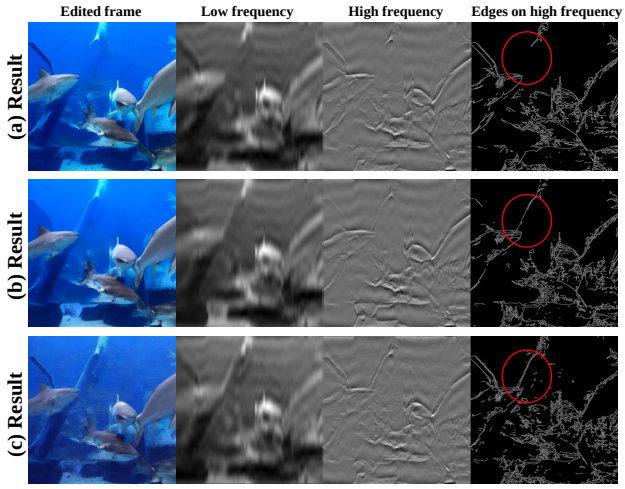


Figure 9. Ablation studies on main modules in FRAG about low and high frequency and edges on high frequency. (a): baseline (FateZero), (b): baseline + FRAG (temporal grouping), (c): baseline + FRAG (frequency adaptive refinement + temporal grouping). The input video is presented in Figure 8.

effectively enhanced in the models with FRAG. In spectral analysis, we separate video into high-frequency and low-frequency components using a frequency filter, where FRAG significantly improves video quality of high frequency.

5.4. Ablation Study

Figure 9 presents ablative studies on FRAG in terms of low and high frequencies in edited videos. The results (a) are the editing with baseline (FateZero), where quality deterioration, such as blurring due to mixing with trees in the background, is identified. The results (c) confirm that combining FRAG with the model effectively mitigates this deterioration in both low and high frequencies. Especially, frequency adaptive refinement plays a key role in this mitigation, improving the high-frequency details with more edges

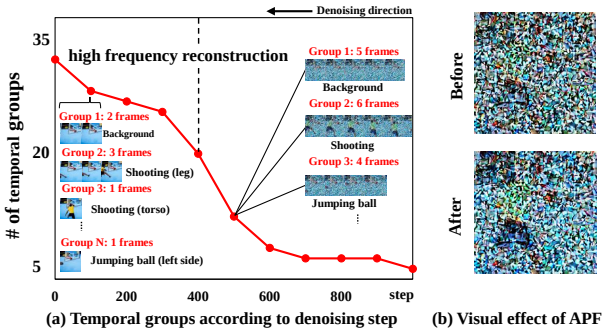


Figure 10. (a) shows variation in the number of temporal groups according to the denoising step. The groups are progressively abundant and fine-grained to contain high-frequency details. (b) shows visual effectiveness before/after applying the adaptive frequency pass filter (APF) on decoded latent noise at step 781.

about the original shape. Figure 10 (a) shows variations in temporal groups during the denoising process. At the start of denoising, there are fewer groups with many frames in each. Interestingly, frames are grouped by similar scenes (*e.g.*, dunk-shoot scenes). When high frequencies appear (*i.e.*, 400 step), group numbers surge, by clustering frames of similar fine details (*e.g.*, scene of the appearance of legs or torso). Figure 10 (b) qualitatively shows the effect of applying APF to latent noise, enhancing visual quality by extracting low-frequency attributes in Gaussian noise.

6. Discussion

Limitations. Our model can be applied to several consistency modules but has a certain degree of sensitivity across modules. Empirical studies show FRAG excels in propagation methods. It's also effective with prior-guidance methods like optical flow, but less so compared to the former. We consider this is because consistency is enforced by the guidance prior. In this way, our proposed adjusting the receptive field also has sensitive effectiveness according to

its application and there are still many additional improvements needed to enhance video quality comprehensively. Moreover, although this paper employs frequency characteristics for temporal grouping and some works (Si et al., 2023; Huang et al., 2024; He et al., 2024) are also concerned about the frequency for the diffusion model, it is also important to understand the scene characteristics of the image/video to achieve robustness even for videos with longer and more diverse scenes. To the best of our knowledge, employing scene knowledge for the diffusion model has never been studied before. To this end, utilizing video search technology (Yoon et al., 2022a; 2023c) appears promising by constructing scene-aware temporal grouping. Additionally, employing recent weakly-supervised (Yoon et al., 2023e; Ma et al., 2020) and unsupervised (Luo et al., 2024) methods can reduce the training resource and enhance the effectiveness of the approach.

Future work. Video editing has recently surged in popularity, yet numerous unresolved issues remain. We briefly introduce the various methods we are considering for future work to address the issues. Video editing performance remains highly sensitive to prompts. To address this, it is essential to further incorporate prompt optimization and tuning methods (Yoon et al., 2023a), similar to those used in image editing (Kawar et al., 2023). Current video editing technology simply relies on human's intuitive decisions about the success of editing, but it needs to be integrated with more detailed automatic control of the editing effect. For this purpose, integrating calibration systems (Yoon et al., 2023b) seems novel for fine-grained controllability of editing. Furthermore, current video editing is slow. The zero-shot (Geyer et al., 2023; Qi et al., 2023) can solve this but it offers limited editability. The tuning method provides high editability but requires a significant amount of time. Thus, it is essential to enhance the video tuning method to address these challenges similar to the works (Koo et al., 2024) in the image. Finally, there is still a lack of research on the vision-language model applying video editing/generation technology. We believe that the convergence of generative technologies will bring explosive innovation to image/video high-level tasks including dialogue (Yoon et al., 2022b), and commonsense reasoning (Liang et al., 2022).

7. Conclusion

This paper proposes Frequency Adaptive Group (FRAG) which enhances the video quality of diffusion video editing systems in a model-agnostic manner. We found the spectral characteristics of latent noise that low-frequency attributes emerge in the early stages, followed by the synthesis of higher-frequency attributes. Based on this characteristic, FRAG enhances the video quality according to the frequency variation of synthesized latent noise.

Impact Statement

Visual generative models are associated with ethical challenges, including the creation of unauthorized counterfeit content, risks to privacy, and issues of fairness. Our work is built on these generative models, inheriting their vulnerabilities. Therefore, it's crucial to tackle these issues through a combination of robust regulations and technical safeguards. We think that researchers assume responsibility for these issues, actively working to implement technical safeguards. We are also considering measures like adopting learning-based digital forensics and implementing digital watermarking. These actions are designed to guide the ethical use of visual generative models, ensuring their responsible and positive application.

Acknowledgements

This work was 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 Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics).

References

Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., and Dekel, T. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pp. 707–723. Springer, 2022.

Chai, W., Guo, X., Wang, G., and Lu, Y. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23040–23050, 2023.

Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.-M., Rosenhahn, B., Xiang, T., and He, S. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Geyer, M., Bar-Tal, O., Bagon, S., and Dekel, T. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

He, F., Li, G., Zhang, M., Yan, L., Si, L., and Li, F. Freestyle:

Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*, 2024.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Huang, L., Fang, R., Zhang, A., Song, G., Liu, S., Liu, Y., and Li, H. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. *arXiv preprint arXiv:2403.12963*, 2024.

Jain, A. K. and Dubes, R. C. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and 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.

Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

Kim, G., Kwon, T., and Ye, J. C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022.

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.

Koo, G., Yoon, S., and Yoo, C. D. Wavelet-guided acceleration of text inversion in diffusion-based image editing. *arXiv preprint arXiv:2401.09794*, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liang, C., Wang, W., Zhou, T., and Yang, Y. Visual abductive reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15565–15575, 2022.

Liu, S., Zhang, Y., Li, W., Lin, Z., and Jia, J. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023.

Luo, D., Huang, J., Gong, S., Jin, H., and Liu, Y. Zero-shot video moment retrieval from frozen vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5464–5473, 2024.

Ma, M., Yoon, S., Kim, J., Lee, Y., Kang, S., and Yoo, C. D. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pp. 156–171. Springer, 2020.

Molad, E., Horwitz, E., Valevski, D., Acha, A. R., Matias, Y., Pritch, Y., Leviathan, Y., and Hoshen, Y. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., and Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

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.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and 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.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Si, C., Huang, Z., Jiang, Y., and Liu, Z. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Teed, Z. and Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 402–419. Springer, 2020.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and 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, 2023a.

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*, 2023b.

Yoon, E., Yoon, H. S., Harvill, J., Hasegawa-Johnson, M., and Yoo, C.-D. Information-theoretic adversarial prompt tuning for enhanced non-native speech recognition. In *The 61st Annual Meeting of the Association for Computational Linguistics*. The 61st Annual Meeting of the Association for Computational Linguistics, 2023a.

Yoon, H. S., Tee, J. T. J., Yoon, E., Yoon, S., Kim, G., Li, Y., and Yoo, C. D. Esd: Expected squared difference as a tuning-free trainable calibration measure. *arXiv preprint arXiv:2303.02472*, 2023b.

Yoon, S., Hong, J. W., Yoon, E., Kim, D., Kim, J., Yoon, H. S., and Yoo, C. D. Selective query-guided debiasing for video corpus moment retrieval. In *European Conference on Computer Vision*, pp. 185–200. Springer, 2022a.

Yoon, S., Yoon, E., Yoon, H. S., Kim, J., and Yoo, C. D. Information-theoretic text hallucination reduction for video-grounded dialogue. *arXiv preprint arXiv:2212.05765*, 2022b.

Yoon, S., Hong, J. W., Eom, S., Yoon, H. S., Yoon, E., Kim, D., Kim, J., Kim, C., and Yoo, C. D. Counterfactual two-stage debiasing for video corpus moment retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023c.

Yoon, S., Kim, D., Yoon, E., Yoon, H., Kim, J., and Yoo, C. HEAR: Hearing enhanced audio response for video-grounded dialogue. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11911–11924, Singapore, December 2023d. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.797. URL https://aclanthology.org/2023.findings-emnlp.797.

Yoon, S., Koo, G., Kim, D., and Yoo, C. D. Scanet: Scene complexity aware network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13576–13586, 2023e.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

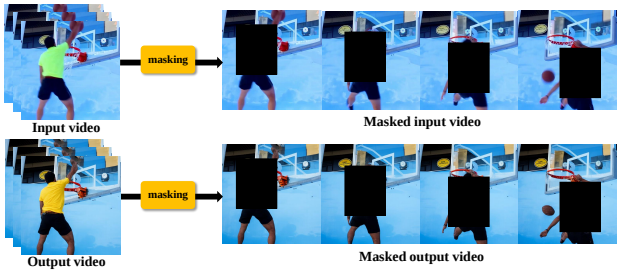


Figure 11. Masked input and output videos for fidelity metric assessment, applying the same masks in the edited regions.

Appendix

A. Details of Implementation and Evaluations

Implementation Details. For video encoding, all the baselines utilize VQ-VAE (Van Den Oord et al., 2017), which provides patch-wise encoding of each frame. For text encoding the CLIP model (ViT-L/14) (Radford et al., 2021) is used for conditional text input. For Equation (6), to compute frame distance we apply mean-pooing along the spatial domain and then perform Euclidean distance for the computational efficiency. We set the margin $d_0 = 6$ for the Sptial Moment Adaption, which was the most effective in our framework, where the Gaussian curve in Equation (3) also smoothly includes the frequency near the moment (M_x,M_y) in the 2-dimensional frequency domain. We adjusted the minimum size of the temporal group along the frame axis to a range between 1 and 4, according to video quality enhancement modules (*e.g.*, propagation, attention).

Fidelity Evaluation Details. To assess fidelity to the input video, we applied the same zero mask to the edited areas in both the input and output videos, as illustrated in Figure 11. By masking the area for editing, we can evaluate the preservation of unedited content between the input and output videos, yielding PSNR and SSIM scores that reflect their similarity and consistency. We utilize automatic detectors (Kirillov et al., 2023) to specify the areas of square mask for editing in both input and output videos. For some edits (*i.e.*, background change) that are not proper by automatic detectors, we specify the edited region.

Human Evaluation Details. Human evaluation is conducted to assess preferences for the edited outcomes based on a specified target prompt. Motivated the format of human evaluation in the work (Yoon et al., 2023d), we conducted a survey comparing preferences for outputs from existing editing systems and the FRAG framework under consideration of semantic alignment, and video quality. A survey was conducted with 36 participants from varied academic fields such as engineering, literature, and art.

B. Closed Form of KL Divergence

The reverse process of DDPM is to approximate $q(x_{t-1}|x_t)$ based on learnable Gaussian transitions $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t))$. We first define whole T step transitions, by sequentially constructing them as $p_\theta(X) = p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$. where staring normal distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$, this considers transitions of $X = x_{0:T}$. For the training objective of $p_\theta(X)$, we should maximize log-likelihood $\log(p_\theta(X))$. Otherwise, we can also apply variational inference by maximizing the variational lower bound $-\mathcal{L}_{vlb}$ as given below:

$$\begin{aligned} -\mathcal{L}_{vlb} &= \log p_\theta(X) - D_{\mathrm{KL}}(q(Z|X)||p_\theta(Z|X)) \\ &\leq \log p_\theta(X), \end{aligned} \tag{9}$$

where D_{KL} is the Kullback-Leibler divergence and the Z is the latent variable using reparametrization trick by the variational auto-encoder. The q can be any distributions that we can approximate. We inverse the inequality condition as $-\log p_\theta(X) \leq \mathcal{L}_{vlb}$. Here, the $-\log p_\theta(X)$ is conditioned by $\mathcal{L}_{vlb}$ by expanding it as $\mathcal{L}_{vlb} = \mathcal{L}_T + \mathcal{L}_{T-1} + \cdots + \mathcal{L}_0$, where they are defined with $1 \leq t \leq T$ as given below:

$$\begin{aligned} \mathcal{L}_T &= D_{\mathrm{KL}}(q(x_T|x_0)||p_\theta(x_T)), \\ \mathcal{L}_t &= D_{\mathrm{KL}}(q(x_t|x_{t+1}, x_0)||p_\theta(x_t|x_{t+1})), \\ \mathcal{L}_0 &= -\log p_\theta(x_0|x_1). \end{aligned} \tag{10}$$

These terms make the closed form of KL divergence under step t with a range of $0 \leq t \leq T$.

C. Agglomerative Hierarchical Clustering

Algorithm 1 Agglomerative Hierarchical Clustering

Input: data x_i
Initialize for each data element $x_i \in X$ its cluster singleton $G_i = \{x_i\}$ in a list
while there remain two elements in the list **do**
  Choose G_i and G_j so that $\Delta(G_i, G_j)$ is minimized among all pairs,
  Merge $G_{i,j} = G_i \cup G_j$,
  Add $G_{i,j}$ to the list,
  Remove G_i and G_j from the list.
end while
Return the remaining group in the list ($G_{\text{root}} = X$) as the dendrogram root.

D. Ablation Studies on Fixed Receptive Field

Figure 12 (a) presents a conceptual illustration of uniform sliding windows for receptive field for quality enhancement module (*e.g.*, temporal attention, propagation) of current systems (Wu et al., 2023a; Geyer et al., 2023; Liu et al., 2023).

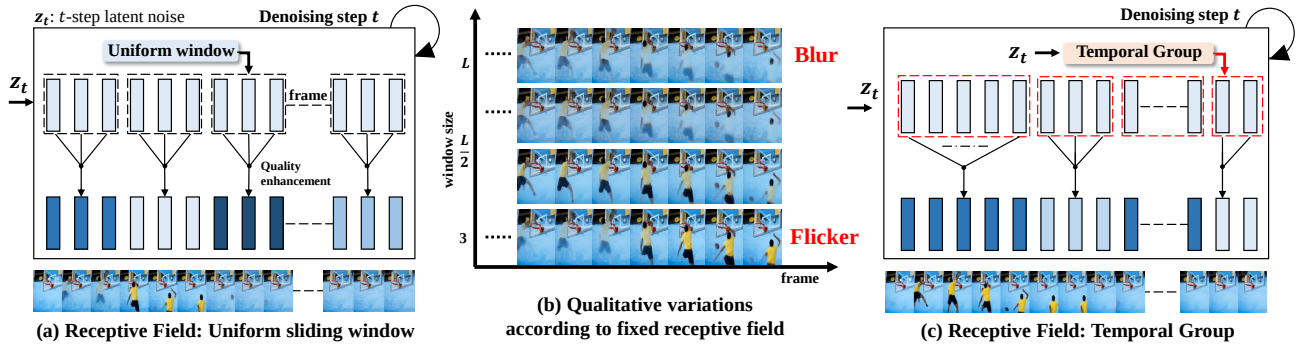


Figure 12. (a) Illustration of fixed receptive field using a uniform sliding window, (b) qualitative results about ablation study about video results according to window size. (c) Illustration of dynamic receptive field using temporal group.

Although this structure offers an intuitive understanding of the frame-level enhancement, it inevitably causes content blur or content flicker. As shown in Figure 12 (b), when applying the windows with a relatively large size (*e.g.*, frame length L), they make a blurred video. latent noise interactions for quality enhancement in long-range receptive fields often overlook the high-frequency attributes synthesized in individual frames, leading to a blurring of content. Conversely, if the receptive field is reduced (*i.e.*, window size of $2 \sim 4$), high-frequency components are generated, but these are ununiformly synthesized across each field. Conversely, when the receptive field is reduced, high-frequency components are generated, but the low-frequency components are not positioned uniformly, so the high-frequency components generated above them are also not uniform. Thus, this makes a content flicker. Figure 12 (c) shows our proposed dynamic receptive field using a temporal group. The temporal group (*i.e.*, red box) adaptively designs the receptive field according to the variational frequency synthesis during the denoising step.

E. More Qualitative Results

All the video samples are based on DAVIS, TGVE, and copyright-free videos at https://www.pexels.com. Here we present qualitative results about (1) the frequency distribution of decoded latent noise, (2) more comparison results, and (3) more results with FRAG.

Frequency Distribution of Decoded Latent Noise.

More Comparison Results.

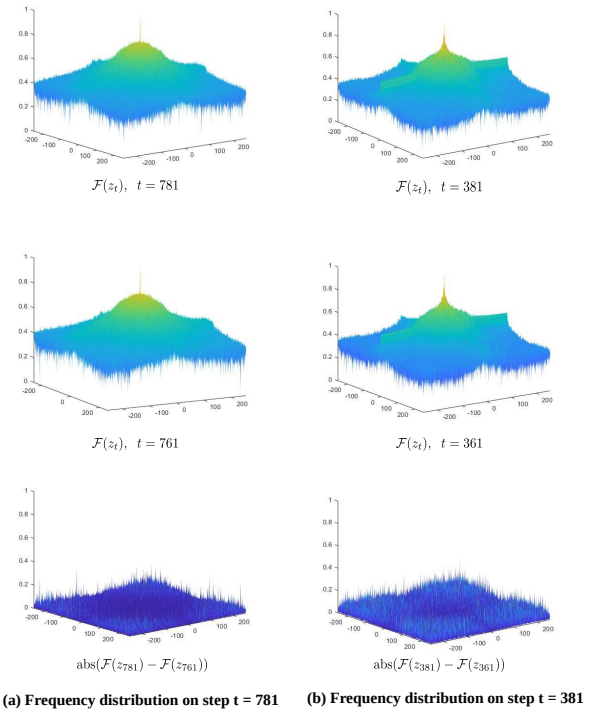
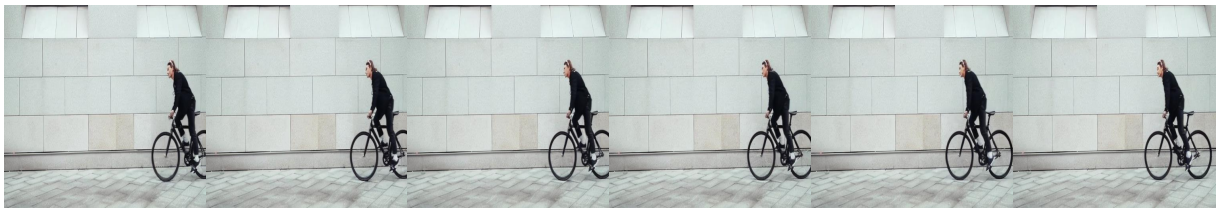


Figure 13. Illustration of a frequency distribution (full view on discrete frequency domain) on a decoded image about latent noise at step $t = 781$, and $t = 381$. (a) shows the differential frequency in the early stage of denoising (*i.e.*, $t = 781$), where there is less information in the region of high frequency ($f > 0.25\pi$) in the difference. (b) shows the differential frequency in the latter stage of denoising (*i.e.*, $t = 381$), where there is dense information in the high-frequency region. We used DDIM for denoising, such that the unit step is 20 for the sampling.

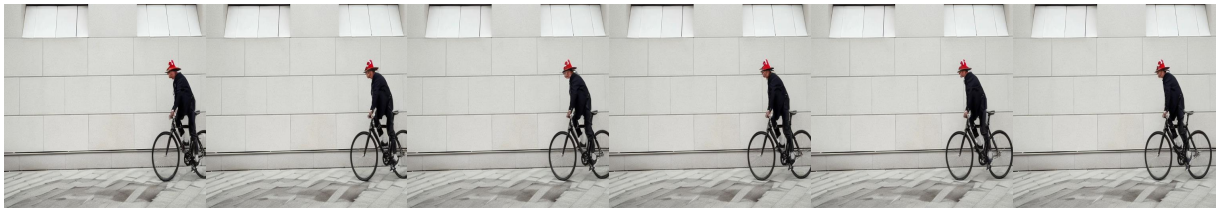
Input video



Target prompt: ***"A man wearing a hat rides a bicycle"***

[TokenFlow]

[FRAG]



Input video

Target prompt: ***"A marble sculpture of a man is running, Venus de Milo style"***

[FLATTEN]

[FRAG]

Input video

Target prompt: ***"A man wears a yellow shirt"***

[TokenFlow]

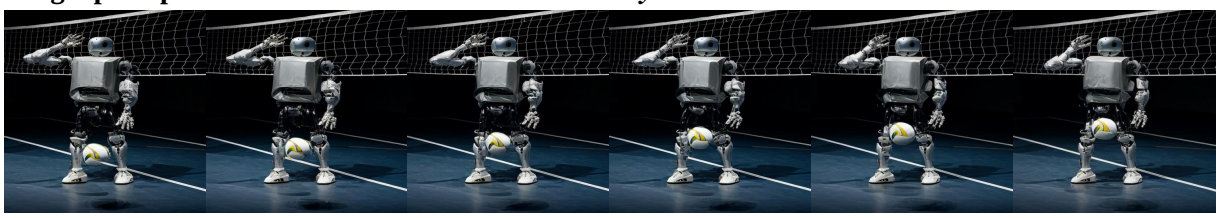
[FRAG]

More Results with FRAG.

Input video

Target prompt: ***"A silver robotic man bounces a volleyball"***

[FRAG]



Input video

Target prompt: ***"A man is riding a snowboard, cartoon style"***

[FRAG]

Input video

Target prompt: ***"A robotic bird on a tree branch"***

[FRAG]

Target prompt: ***"A origami bird on a tree branch"***

[FRAG]