Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution

Shangchen Zhou*   Peiqing Yang*   Jianyi Wang   Yihang Luo   Chen Change Loy
S-Lab, Nanyang Technological University
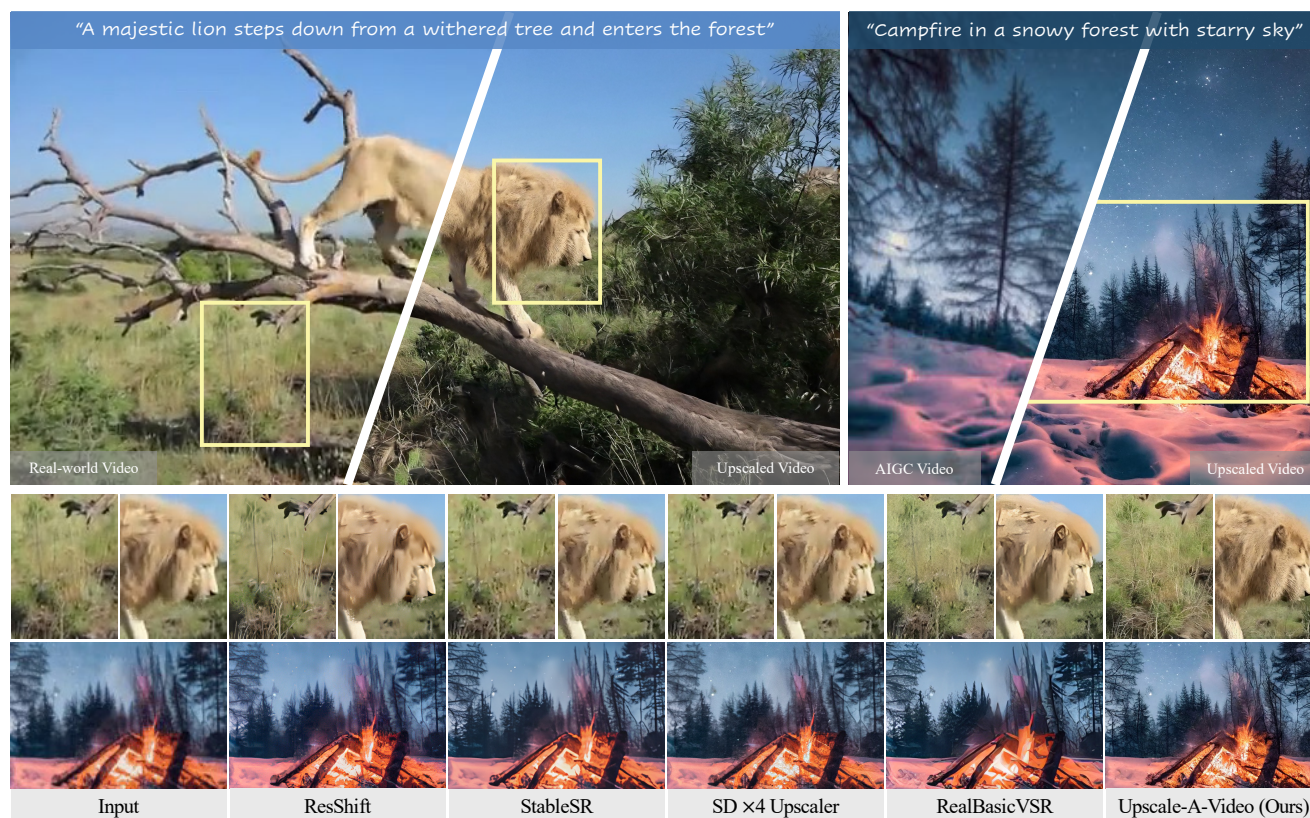https://shangchenzhou.com/projects/upscale-a-video

Figure 1. Video super-resolution comparisons on both real-world and AI-generated videos. Our proposed Upscale-A-Video showcases excellent upscaling capabilities. By using appropriate text prompts, it achieves impressive results characterized by finer details and heightened visual realism. **(Zoom-in for best view)**

Abstract

*Text-based diffusion models have exhibited remarkable success in generation and editing, showing great promise for enhancing visual content with their generative prior. However, applying these models to video super-resolution remains challenging due to the high demands for output fidelity and temporal consistency, which is complicated by the inherent randomness in diffusion models. Our study introduces **Upscale-A-Video**, a text-guided latent diffusion framework for video upscaling. This framework ensures temporal coherence through two key mechanisms: locally, it integrates temporal layers into U-Net and VAE-Decoder, maintaining consistency within short sequences; globally, without training, a flow-guided recurrent latent propagation module is introduced to enhance overall video stability by propagating and fusing latent across the entire sequences. Thanks to the diffusion paradigm, our model also offers greater flexibility by allowing text prompts to guide texture creation and adjustable noise levels to balance restoration and generation, enabling a trade-off between fidelity and*

*Equal contribution.

*quality. Extensive experiments show that **Upscale-A-Video** surpasses existing methods in both synthetic and real-world benchmarks, as well as in AI-generated videos, showcasing impressive visual realism and temporal consistency.*

1. Introduction

Video super-resolution (VSR) in real-world scenarios is a challenging task that aims at enhancing the quality of low-quality videos to produce high-quality results. Unlike previous works that mainly focus on either synthetic degradations [8, 9, 64] or specific camera-related degradations [80], this task is more demanding due to the need for addressing complex and unknown degradations commonly found in low-quality videos, such as downsampling, noise, blur, flickering, and video compression. In addition, maintaining the visual fidelity and ensuring temporal coherence are crucial for the perceptual quality of the output video. Although recent convolutional neural network (CNN)-based networks [10, 73] have shown success in mitigating many forms of degradation, they still fall short in producing realistic textures and details due to their limited generative capabilities, often resulting in over-smoothing, which can be observed in RealBasicVSR results shown in Fig. 1.

Diffusion models [24] have exhibited impressive proficiency in generating high-quality images [49, 50, 52, 85] and videos [6, 11, 17, 25, 87]. Harnessing their generative potential, diffusion-based models have been proposed for image restoration, including both training from scratch [50, 51, 83] and fine-tuning from pretrained Stable Diffusion models [38, 62]. These methods have effectively mitigated the over-smoothing issue often observed in CNN-based models, yielding results with more realistic fine-grained details. However, adapting these diffusion priors to VSR remains a non-trivial challenge. This difficulty stems from the inherent randomness in diffusion sampling, which inevitably introduces unexpected temporal discontinuities in resulting videos. This issue is more pronounced in latent diffusion, where the VAE decoder further introduces flickering in low-level texture details.

Recent efforts have been made to adapt image diffusion models for video tasks by introducing strategies for temporal consistency, which include: 1) Fine-tuning video models with temporal layers such as 3D convolution [6, 66] and temporal attention [6, 11, 66, 87]; 2) Employing zero-shot mechanisms like cross-frame attention [71, 78] and flow-guided attention [14, 18] in the pretrained models. Although these solutions significantly improve video stability, there remain two primary issues: i) The current methods, operating in the U-Net feature or latent space, struggle to maintain low-level consistency, with issues like texture flickering still present. ii) The existing temporal layers and attention mechanisms can only impose constraints on short, local input sequences, limiting their capacity to ensure global temporal consistency in longer videos.

To tackle these issues, we adopt a local-global strategy for maintaining temporal consistency in video reconstruction, focusing on both fine-grained textures and overall consistency. On a local video clip, we explore finetuning a pretrained image ×4 upscaling model [2] with additional temporal layers on the video data. Specifically, within a latent diffusion framework, we first finetune the U-Net with integrated 3D convolutions and temporal attention layers, and then tune the VAE-Decoder with video-conditioned inputs and 3D convolutions. The former significantly achieves structure stability in local sequences, while the latter further improves low-level consistency, reducing texture flickering. On a more global scale, we introduce a novel, training-free flow-guided recurrent latent propagation module. Spanning short video segments, it bidirectionally conducts frame-by-frame propagation and latent fusion during inference. Leveraging latent fusion in a long-term sequence, this module encourages overall stability for long videos.

We further investigate the generative potential of diffusion models in the VSR task. Following the text-guided fashion, our model can leverage text prompts as an optional condition to steer the model towards producing more realistic and high-quality details, as illustrated in Fig. 1. Furthermore, we enhance the robustness of our model against heavy or unseen degradation by injecting noise into input to dilute its degradation. By adjusting the level of added noise, we can modulate the balance between restoring and generating within the diffusion model. Lower noise levels prioritize the model's restoration capabilities, while higher levels encourage generation of more refined details, thus achieving a trade-off between fidelity and quality of the output.

To summarize, the main contribution of our study is a practical and robust approach to real-world VSR. In particular, we conduct an exhaustive exploration of integrating a local-global temporal strategy within a latent diffusion framework, enhancing temporal coherence and generation quality. We further examine ways to improve versatility by allowing text prompts to guide texture creation, and offer control over noise levels to balance restoration and generation, thereby achieving a trade-off between fidelity and quality. Thanks to our design and pretrained generative prior, our model achieves state-of-the-art performance on existing benchmarks, showing remarkable visual realism and temporal consistency.

2. Related Work

Video Super-Resolution. VSR aims to restore a sequence of high-resolution (HR) video frames from its degraded low-resolution (LR) counterparts. Most existing approaches [7–9, 28–31, 36, 37, 64, 75] assume a pre-defined degradation process [39, 45, 75, 82], and their performance

deteriorates significantly in real-world scenarios due to the limited generalizability. To tackle real-world VSR, recent works go beyond the traditional paradigms by assuming inputs with unknown degradations. Due to the lack of real-world paired data for training, Yang *et al.* [80] propose to collect HR-LR data pairs with iPhone cameras. While the VSR model trained on such data can be effective to videos captured by similar mobile cameras, its generalization capability to other devices is in doubt. Rather than relying on labor-intensive pair data collection, recent studies [10, 73] have shifted towards employing diverse degradations for data augmentation during training, demonstrating better performance in handling real-world cases. Nonetheless, it is still challenging for existing CNN-based approaches [10, 73] to generate photo-realistic textures due to the absence of generative prior. In this study, we set our sights on exploiting robust and extensive generative prior encapsulated in a pretrained image diffusion model, *i.e.*, Stable Diffusion (SD) ×4 upscaler [2]. By integrating this strong diffusion prior, our approach circumvents the need for exhaustive training from scratch and exhibits improved performance in producing detailed textures.

Diffusion Models for Video Tasks. Following the success of text-to-image diffusion models [4, 12, 16, 20, 22, 44, 85], recent studies have ventured into the applications of video diffusion models [15, 25–27, 40–42]. Instead of training from scratch, some methods [14, 19, 48, 71, 78] focus on video generation using off-the-shelf image diffusion models [50, 85] in a zero-shot manner. To keep temporal consistency, cross attention [60] between neighboring frames and optical flow [58] warping are popular solutions adopted in these methods. While being efficient, these methods generally suffer from limited generalizability and the selection of hyperparameters can be tricky. Most recently, Blattmann *et al.* [6] propose to extend pretrained image diffusion models to the video domain by introducing an additional temporal dimension and fine-tuning the temporal layers only to speed up the training process. Subsequent works [21, 74] following such paradigm are capable of generating impressive video sequences. Inspired by these works, we employ a pretrained image ×4 upscaling model [2] as generative prior and propose a novel local-global temporal strategy, resulting in temporally coherent outputs with faithful details.

Diffusion Models for Restoration Tasks. With notable advancements in diffusion models [24, 46, 55, 77], numerous diffusion-based works have been proposed for image restoration. A straightforward way is to train a diffusion model conditioned on LR images from scratch [50, 51, 53, 72, 83], which, however, demands significant computational resources for training. To avoid such heavy training costs, another prevalent approach [12, 13, 32, 56, 67, 69, 76] is to incorporate constraints into the reverse diffusion process of a pretrained diffusion model. Though efficient, designing these constraints relies on pre-defined image degra-

dation processes or pretrained SR models as a priori, leading to limited generalizability and inferior performance. Recent works [62, 79] go further by fine-tuning directly on a frozen pretrained diffusion model with several additional trainable layers, demonstrating impressive performance. Inspired by these recent advances, we focus on exploiting the effective diffusion prior for real-world VSR, which remains under-explored and challenging due to the temporal discontinuities caused by inherent randomness during diffusion sampling. With the proposed local-global temporal strategy, our approach is capable of generating temporally coherent results for real-world VSR.

3. Methodology

Our objective is to develop a text-guided diffusion framework tailored for real-world VSR. The diffusion denoising process, characterized by its inherent stochastic nature, poses significant challenges when applied to video tasks. These challenges include temporal instability and the emergence of flickering artifacts, which are particularly prominent in VSR tasks involving lengthy video sequences. The complexity of these tasks lies not just in achieving temporal consistency within localized segments, but also in preserving coherence throughout the entire video.

As illustrated in Fig. 2, our framework incorporates both local and global modules in the latent diffusion model (LDM) to preserve temporal consistency both within and across video segments. Within each diffusion time step ($t = 1, 2 \dots T$), the video is split into segments and processed with a U-Net that includes temporal layers to ensure consistency within each segment. If the current time step falls within the user-specified global refinement steps (T^*), a recurrent latent propagation module is employed to improve consistency across segments during inference. Lastly, a finetuned VAE-Decoder is used to reduce remaining flickering artifacts. Thanks to the diffusion paradigm, our model exhibits remarkable versatility: 1) Input text prompts can enhance video quality further with improved realism and details. 2) User-specified noise levels provide a control over the trade-off between quality and fidelity.

3.1. Preliminary: Diffusion Models

Pretrained Stable Diffusion Image ×4 Upscaler. Our Upscale-A-Video is built upon the pretrained text-guided SD ×4 Upscaler [2]. It employs the LDM framework [50], which uses an autoencoder to transform an image into a latent with an encoder $\mathcal{E}$ (×4 downsampling), and reconstructs it with a decoder $\mathcal{D}$. Conditioned on the low-resolution images x, it learns to generate the high-quality counterparts via iterative denoising in the latent space. To achieve this, given latent samples z $\sim p_{data}$, diffused latents $z_t = \alpha_t z + \sigma_t \epsilon$ is generated by introducing Gaussian noise to the latents z at each diffusion step t; where
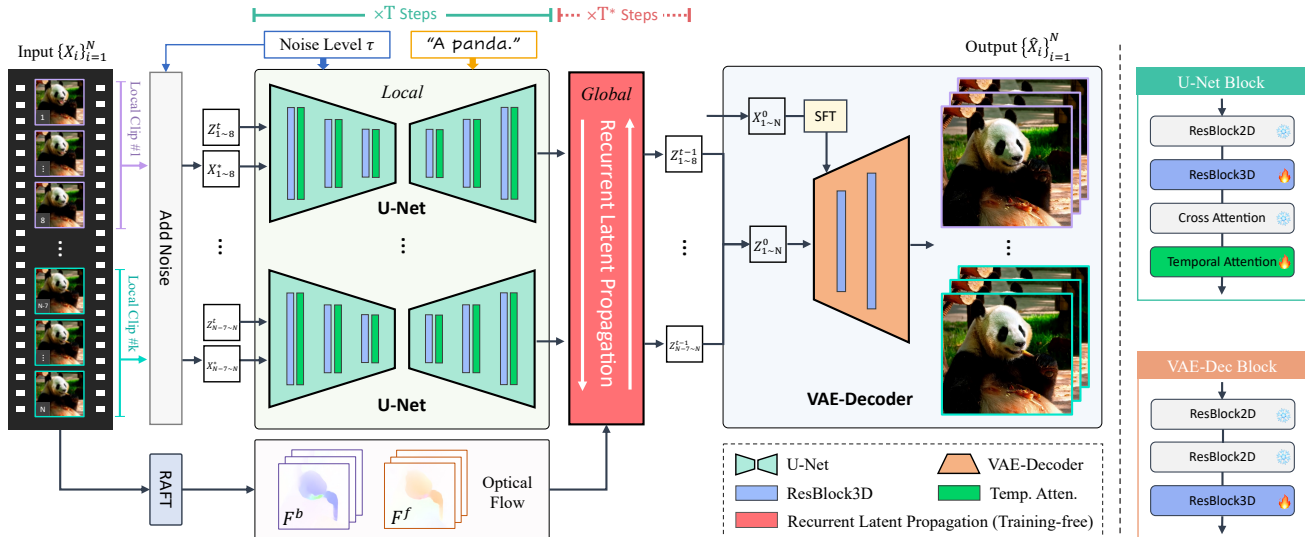
Figure 2. An overview of Upscale-A-Video. Upscale-A-Video processes long videos using both local and global strategies to maintain temporal coherence. It divides the video into segments and processes them using a U-Net with temporal layers for intra-segment consistency. During user-specified diffusion steps for global refinement, a recurrent latent propagation module is used to enhance inter-segment consistency. Finally, a finetuned VAE-Decoder reduces remaining flickering artifacts for low-level consistency. Our model also allows users to guide texture creation with text prompts and adjust noise levels to balance the effect of restoration and generation.

$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, α_t and σ_t define a noise schedule. To enhance the ability of our model to generate fine details, it also applies random noise to input images, *i.e.*, the diffused images $\mathbf{x}_\tau = \alpha_\tau \mathbf{x} + \sigma_\tau \epsilon$, where τ is the noise level that corresponds to the early steps in noise schedule.

Adopting v-prediction perameterization [54], the U-Net denoiser f_θ is trained to make predictions of $\mathbf{v}_t \equiv \alpha_t \epsilon - \sigma_t \mathbf{x}$. The optimization objective for LDM is as follows:

$$\mathbb{E}_{\mathbf{z},\mathbf{x},\mathbf{c},t,\epsilon} \left[\| \mathbf{v} - f_\theta(\mathbf{z}_t, \mathbf{x}_\tau; \mathbf{c}, t) \|_2^2 \right], \quad (1)$$

where c serves as an optional set of conditions, including text prompts and noise levels of the diffused image. During inference, the model has the flexibility to involve different text prompts and noise levels to diffusion sampling of x_0, and finally decode it to produce the $\times 4$ upscaled images.

Inflated 2D Convolution. When adapting pretrained 2D diffusion models to video tasks, it is common to inflate their 2D convolutions into the 3D convolutions [6, 17, 71, 78]. This modification allows the diffusion models to smoothly integrate new temporal layers, enabling them to capture and encode temporal information within the pretrained model.

In this study, we explore a new avenue to create a text-guided video diffusion model for VSR tasks, starting with the pretrained SD $\times 4$ Upscaler [2]. To process video data, we first modify its network structure by inflating the 2D convolutions into 3D convolutions, and then initialize our network with this upscaler to inherit its enhancement capabilities. Our goal is to transfer the knowledge learned from image upscaling to video enhancement, enabling more efficient training. In the following, we we will describe our

local-global strategy within the LDM framework to achieve temporal coherence while harnessing the capability of diffusion for VSR tasks.

3.2. Local Consistency within Video Segments

To apply the pretrained text-to-image SD model to video-related tasks, existing video diffusion models typically employ techniques such as 3D convolutions [6, 66], temporal attention [6, 17, 66, 68]sx, and cross-frame attention [71] to ensure temporal consistency.

Finetuning Temporal U-Net. Following the existing methods [6, 66], we introduce additional temporal layers into the pretrained image model and learn the local consistency constraint within video segments. As illustrated in Fig. 2, in the modified temporal U-Net, we opt for temporal attention and 3D residual blocks based on 3D convolutions to serve as our temporal layers, and insert them within the pretrained spatial layers. The temporal attention layer performs self-attention along the temporal dimension and focuses on all local frames. Additionally, we add Rotary Position Embedding (RoPE) [59] into the temporal layers to provide the model with positional information for time.

Our temporal layers are trained using the same noise schedule as the pretrained image model. Importantly, we keep the pretrained spatial layers fixed during training and only optimize the inserted temporal layers using Eq. (1). An essential benefit of this strategy is that it allows us to leverage the pretrained spatial layers that were learned from a huge, high-quality image dataset. This enables us to concentrate our training efforts on refining the temporal layers.

Finetuning Temporal VAE-Decoder. Even after finetuning the U-Net on video data, the VAE-Decoder within the LDM framework, which is trained on images only, still tends to produce flickering artifacts when decoding a latent sequence. To mitigate this issue, we introduce additional temporal 3D residual blocks into the VAE-Decoder to enhance low-level consistency. Furthermore, the diffusion denoising process in the U-Net often introduces color shifts, a problem also encountered by other diffusion-based restoration networks [62, 83]. To tackle this, we condition the input videos with a Spatial Feature Transform (SFT) layer [63], which employs the inputs to transform the features of the first layer of the VAE-Decoder. This allows the input videos to provide low-frequency information, such as color, to strengthen the color fidelity of the output results.

Similar to the training of the Temporal U-Net, we keep the pretrained spatial layers unchanged and only train the newly added temporal layers. These temporal layers are trained on video data using a hybrid loss consisting of L1 loss, LPIPS perceptual loss [86], and an adversarial loss employing a temporal PatchGAN [89] discriminator. As demonstrated by the ablation study in Fig. 6 and Table 2, this step is crucial for achieving favorable results.

3.3. Global Consistency cross Video Segments

The trained temporal layers within the LDM are limited to processing local sequences (*e.g.*, eight frames in our U-Net setting), making it impossible to enforce global consistency constraints across video segments. Prior studies [9, 10, 88, 89] have already showcased the benefits of flow-guided long-term propagation in enhancing temporal consistency for video restoration tasks. However, their performance gains are often observed when dealing with long video sequences that provide long-term information. Unfortunately, these methods are not well-suited for diffusion models due to memory constraints, which typically restrict them to processing short video clips.

Training-Free Recurrent Latent Propagation. We introduce a *training-free* flow-guided *recurrent* propagation module within the latent space. This module ensures global temporal coherence for long input videos, involving *bidirectional* propagation in the forward and backward directions. Here, we elaborate on the forward propagation, and backward propagation follows the same process.

Given an input low-resolution video, we first adopt RAFT [58] to estimate optical flow, with its resolution exactly matching the latent resolution, thus no need for resizing. We then check the validity of the estimated flow by evaluating forward-backward consistency error [43]:

$$E_{i-1\to i}(p) = \left\| f_{i-1\to i}(p) + f_{i\to i-1}\big(p + f_{i-1\to i}(p)\big) \right\|_2^2, \tag{2}$$

where p denotes the position of the last frame latent, $f_{i-1\to i}$ and $f_{i\to i-1}$ are forward and backward flow respectively. As
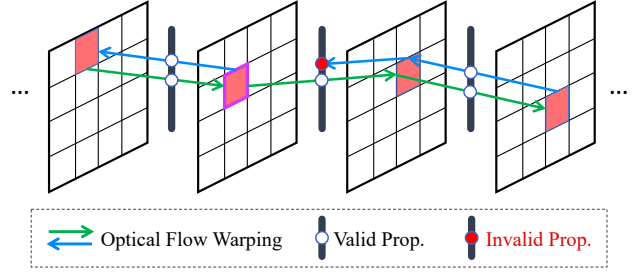

Figure 3. An illustration of flow-guided recurrent latent propagation. Without requiring any learning, this module can achieve coherence across video segments via long-term latent propagation and aggregation. It relies on optical flow validity determined by forward-backward consistency error [43]. Only latent positions with low consistency errors will be propagated, while those with high errors, marked with a red dot, are not.

shown in Fig. 3, only latents with small consistency error will be propagated, which can be thought of as an occlusion mask, $M : E_{i-1\to i}(p) < \delta$, where δ is a threshold. Let z_i be the latent feature for the i-th frame at diffusion step t, we update the predicted $\hat{z}_0$ as $\widetilde{z}_0$ The process of *recurrent* latent propagation and aggregation is expressed as:

$$\widetilde{z}_0^i = \left[\mathcal{W}(\widetilde{z}_0^{i-1}, f_{i\to i-1}) * \beta + \hat{z}_0^i * (1-\beta) \right] * M \\ + \hat{z}_0^i * (1-M), \tag{3}$$

where $\mathcal{W}(\cdot)$ denotes warping operation (with `nearest` mode), and $\beta \in [0, 1]$ serves as a fusion weight for aggregating the warped latent from the previous $(i-1)^{th}$ frame to the current i^{th} frame, where smaller values tend to preserve current information, while larger values favor the propagated information. We set β to 0.5 by default.

It is not necessary to apply this module at each diffusion step during the inference process. Instead, we can choose T^* steps for latent propagation and aggregation. When dealing with minor video jitter, one can opt to integrate this module early in the diffusion denoising process, while for severe video jitter, such as AIGC videos, it is preferable to execute this module later in the denoising process.

3.4. Inference with Additional Conditions

We can further adjust additional conditions of *text prompts* and *noise levels* in Upscale-A-Video to influence the diffusion denoising process. Text prompts can guide the generation of texture details, such as animal fur, as shown in Fig. 7. Besides, adjusting the noise level allows us to balance the model's restoration and generation abilities, with smaller values favoring restoration and larger values promoting the generation of more details (see comparisons in Fig. 7). We also adopt Classifier-Free Guidance (CFG) [23] during inference, which can significantly enhance the impact of both text prompts and noise levels, helping produce high-quality videos with finer details.

Table 1. Quantitative comparisons on different VSR benchmarks from diverse sources, *i.e.*, synthetic (SPMCS, UDM10, REDS30, YouHQ40), real (VideoLQ), and AIGC (AIGC30) data. The best and second performances are marked in red and orange, respectively. E^*_{warp} denotes E_{warp} ($\times 10^{-3}$).

Datasets	Metrics	Real-ESRGAN [65]	SD ×4 Upscaler [2]	ResShift [83]	StableSR [62]	RealVSR [80]	DBVSR [47]	RealBasicVSR [10]	Ours
SPMCS	PSNR ↑	22.89	23.19	23.27	22.71	23.88	24.28	24.51	25.32
	SSIM ↑	0.669	0.631	0.667	0.657	0.681	0.726	0.717	0.741
	LPIPS ↓	0.238	0.304	0.257	0.231	0.437	0.302	0.198	0.222
	E^*_{warp} ↓	1.364	5.008	4.942	4.815	0.294	1.360	0.559	0.367
UDM10	PSNR ↑	27.13	28.07	27.62	26.45	27.38	29.60	29.11	30.79
	SSIM ↑	0.843	0.811	0.827	0.825	0.825	0.880	0.876	0.878
	LPIPS ↓	0.190	0.186	0.222	0.181	0.278	0.155	0.172	0.133
	E^*_{warp} ↓	1.462	1.710	2.196	2.797	0.531	1.943	0.602	0.446
REDS30	PSNR ↑	22.40	22.98	23.00	23.72	23.05	24.37	23.91	24.41
	SSIM ↑	0.591	0.572	0.580	0.635	0.603	0.633	0.636	0.631
	LPIPS ↓	0.303	0.399	0.369	0.352	0.658	0.588	0.249	0.335
	E^*_{warp} ↓	3.658	3.753	4.131	1.645	0.378	9.659	1.557	1.278
YouHQ40	PSNR ↑	24.37	19.71	23.77	24.53	24.19	25.37	24.09	25.83
	SSIM ↑	0.710	0.579	0.654	0.711	0.695	0.719	0.689	0.733
	LPIPS ↓	0.272	0.442	0.376	0.271	0.484	0.430	0.306	0.268
	E^*_{warp} ↓	1.856	3.399	4.426	1.529	0.485	1.149	1.052	0.737
VideoLQ	CLIP-IQA ↑	0.360	0.158	0.430	0.344	0.211	0.274	0.387	0.530
	MUSIQ ↑	49.48	26.21	40.95	44.23	24.52	29.15	55.33	57.99
	DOVER ↑	7.161	2.884	4.679	6.783	2.531	3.628	7.562	7.811
AIGC30	CLIP-IQA ↑	0.430	0.329	0.569	0.467	0.276	0.290	0.565	0.674
	MUSIQ ↑	47.09	35.30	43.32	44.93	24.39	27.22	58.87	57.66
	DOVER ↑	9.710	5.646	7.042	9.668	3.285	3.523	10.68	11.67

4. Experiments

4.1. Datasets and Implementation

Training Datasets. We train our Upscale-A-Video using the following datasets: 1) The subset of WebVid10M [5] contains around 335K video-text pairs with each resolution around 336×596, which is commonly used in training video diffusion models [6, 11, 17, 84]; 2) YouHQ dataset. Due to the lack of high-quality video data for training, we additionally collect a large-scale high-definition (1080×1920) dataset from YouTube, containing around 37K video clips with diverse scenarios, *i.e.*, street view, landscape, animal, human face, static object, nighttime scene, and among others. Training on such high-quality data further enhances the generation ability of our model for real-world VSR. Following the degradation pipeline of RealBasicVSR [10], we generate the LQ-HQ video pairs for training.

Testing Datasets. We construct four synthetic testing datasets (*i.e.*, SPMCS [81], UDM10 [57], REDS30 [45], and YouHQ40), which follow the same degradation pipeline in training to generate LQ videos. We split YouHQ40 test set from the proposed YouHQ dataset, containing 40 videos. Additionally, we evaluate the models on a real-world dataset (*i.e.*, VideoLQ [10]) and an AIGC dataset (*i.e.*, AIGC30) that collects 30 AI-generated videos by popular text-to-video generation models [1, 3, 6, 17, 25, 41, 68, 84].

Training Details. Our Upscale-A-Video is trained on 32 NVIDIA A100-80G GPUs with a batch size of 384. The training data is cropped to 80×80 with a length of 8. The learning rate is set to 1×10^{-4} using the Adam [34]

optimizer. We first train the U-Net model on both Web-Vid10M [5] and YouHQ for 70K iterations. Then we train another 10K iterations on YouHQ only. Since there are no text prompts for YouHQ, we use the null prompts during training. In this way, our model can handle VSR with either LQ inputs and prompts or LQ inputs only, leading to a more flexible use in practice. As for VAE-Decoder fine-tuning, we follow StableSR [62] to first generate 100K synthetic LQ-HQ video pairs on WebVid10M [5] and YouHQ, and the finetuned U-Net model is adopted to generate the corresponding latent codes for the LQ videos.

Evaluation Metrics. We adopt different metrics to evaluate both the frame quality as well as temporal coherency of the generated results. For synthetic datasets with LQ-HQ pairs, we employ PSNR, SSIM, LPIPS [86], and flow warping error [35] E^*_{warp} for evaluation. For real-world and AIGC test data, since no ground-truth videos are available, we conduct our evaluation on commonly used non-reference metrics, *i.e.*, CLIP-IQA [61], MUSIQ [33], and DOVER [70].

4.2. Comparisons

We compare Upscale-A-Video with several state-of-the-art methods, including Real-ESRGAN [65], SD ×4 Upscaler [2], ResShift [83], StableSR [62], RealVSR [80], DBVSR [47], and RealBasicVSR [10].

Quantitative Evaluation. As shown in Table 1, our approach achieves the highest PSNR across all four synthetic datasets, suggesting its outstanding reconstruction ability. Moreover, our approach achieves the lowest LPIPS scores on both UDM10 and YouHQ40, indicating the high percep-
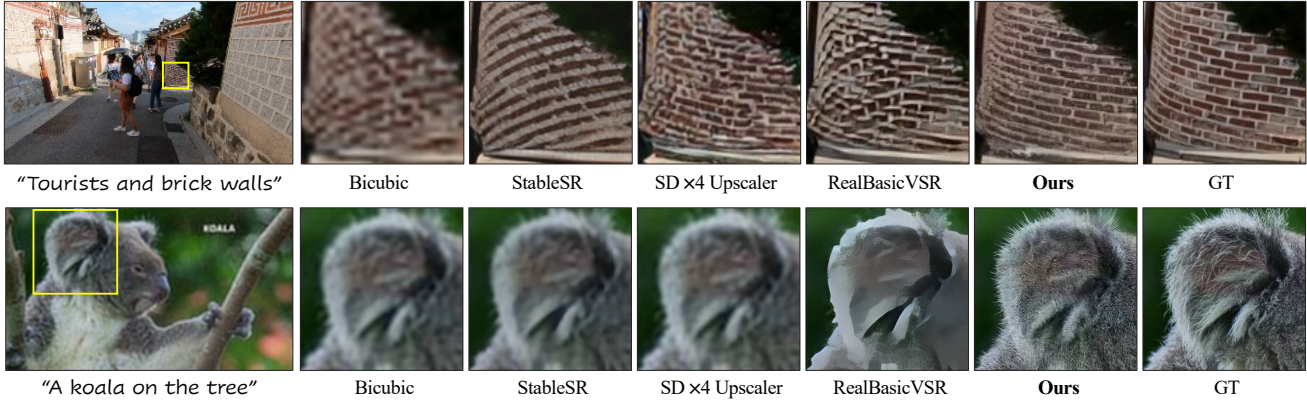
Figure 4. Qualitative comparisons on synthetic low-quality videos from REDS30 [45] and YouHQ40 datasets. Among the tested methods, only our Upscale-A-Video can recover the accurate wall structure and produce detailed koala fur. (**Zoom-in for best view**)
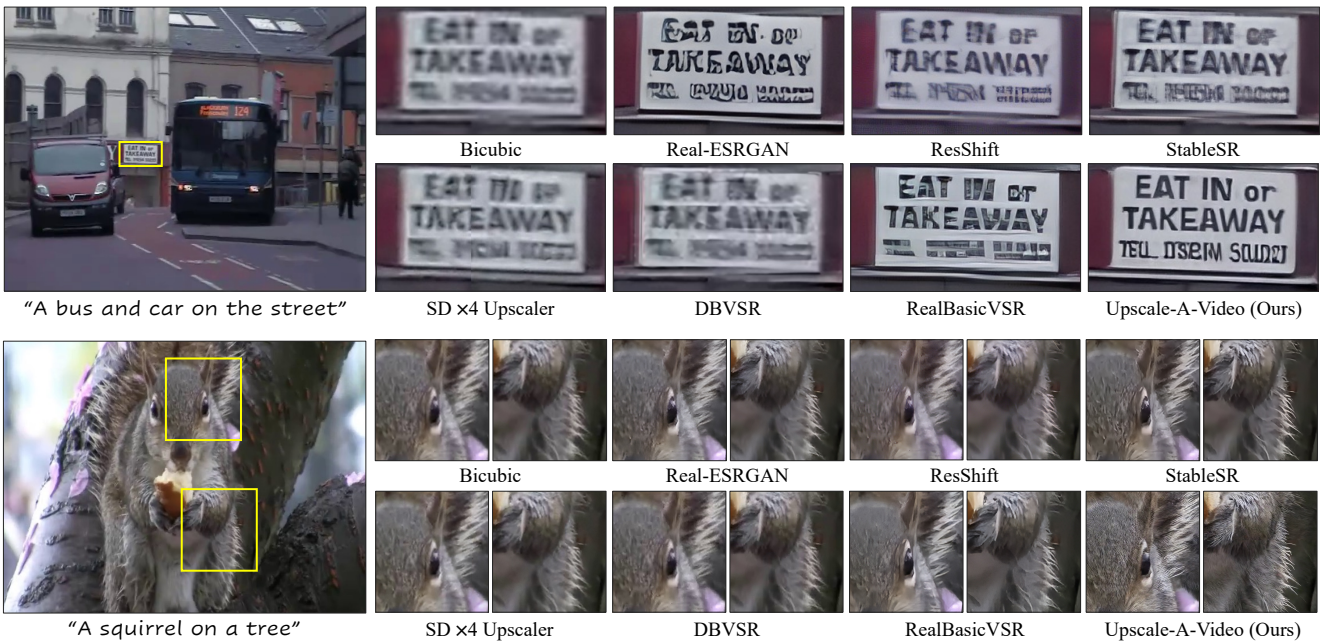

Figure 5. Qualitative comparisons on real-world test videos in VideoLQ [10] dataset. Our Upscale-A-Video effectively leverages the advantages of the diffusion paradigm in generating high-quality results. When compared to existing methods, it notably excels in its restoration capabilities, successfully recovering the billboard word "EAT IN or TAKEAWAY". In particular, when guided by text prompts, Upscale-A-Video showcases promising enhanced results with more details and heightened realism. (**Zoom-in for best view**)

tual quality of our generated results. In addition, our approach further obtains the highest CLIP-IQA and DOVER scores on both the real dataset [10] and AIGC videos. The superiority across datasets from various sources demonstrates the effectiveness of our approach.

Qualitative Evaluation. We present visual results on both synthetic [45, 57, 81] and real-world [10] videos in Fig. 4 and Fig. 5, respectively. It is observed that our Upscale-A-Video significantly outperforms existing CNN- and diffusion-based approaches in both artifact removal and details generation. Specifically, Upscale-A-Video is capable of generating more natural details in Fig. 4, and successfully recovers the words "EAT IN or TAKEAWAY" on the billboard in the first example of Fig. 5, while other methods generate blurry or distorted results. Note that SD ×4 Upscaler shows less effectiveness for the real-world VSR since it does not consider mixed degradation in its training data.

Temporal Consistency. Benefiting from our local-global temporal strategy, Upscale-A-Video achieves the best optical flow error score on UDM10 and the second-best scores on REDS30, SPMCS and YouHQ40, remarkably outperforming other diffusion-based approaches, and even beats the strong CNN-based VSR methods, i.e., RealBasicVSR and DBVSR. We also visualize the temporal profile in
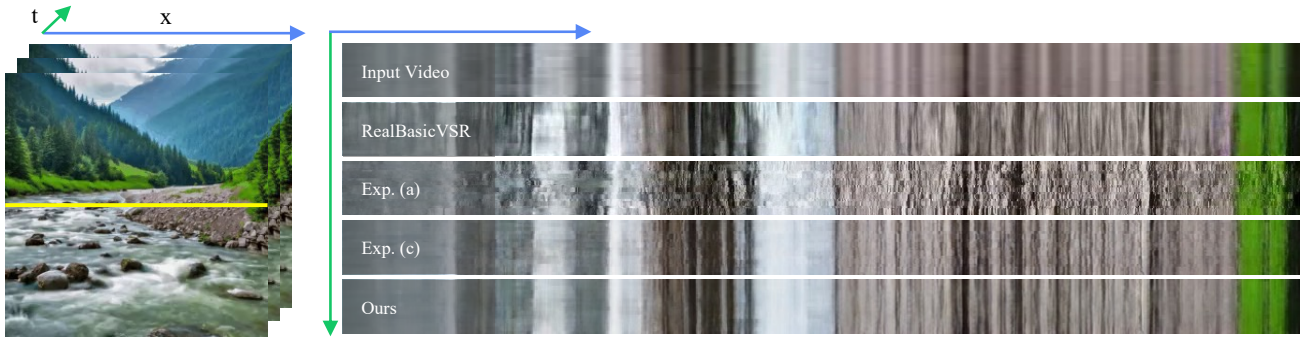
Figure 6. Comparison of temporal profile. We examine a row and track changes over time. The profile from existing methods and Exp. (a) and (c) (w/o latent propagation module) exhibit noise, suggesting the presence of flickering artifacts. Even with the finetuned decoder, the profile of Exp. (c) still displays some discontinuities. Thanks to long-term latent propagation and aggregation, the profile of our Upscale-A-Video exhibits a more seamless and smoother transition.

Table 2. Ablation study of finetuned VAE-Decoder and propagation module on YouHQ40.

Exp.	ft-VAE-Dec.	Latent Prop.	PSNR↑	SSIM↑	E^*_{warp} ↓
(a)			23.82	0.6385	2.398
(b)		✓	25.47	0.7215	1.815
(c)	✓		25.75	**0.7328**	0.842
(d)	✓	✓	**25.83**	0.7326	**0.737**

Fig. 6. It is observed that our approach achieves superior performance with a more seamless and smoother transition.

4.3. Ablation Study

Effectiveness of Finetuned VAE-Decoder. We first investigate the significance of the fine-tuned VAE-Decoder. As shown in Table 2, replacing our finetuned VAE-Decoder with the original decoder leads to worse PSNR, SSIM, and E^*_{warp}. Particularly, the increase of E^*_{warp} from 0.737 to 1.815 indicates a significant deterioration of temporal coherency. The comparisons in Fig. 6 also suggest inferior temporal consistency without the finetuned VAE-Decoder.

Effectiveness of Propagation Module. In addition to fine-tuning VAE-Decoder, our proposed flow-guided recurrent latent propagation module further enhances the stability of long videos. As shown in Table 2, adopting the propagation module can further reduce E^*_{warp} error, effectively improving temporal consistency while maintaining high PSNR. Similar phenomena can also be observed in the temporal profile in Fig. 6, showing a more seamless transition.

Text Prompt. Upscale-A-Video is trained on video data with either labeled prompts or null prompts and thus can handle both scenarios. We examine the use of classifier-free guidance [23] to improve the visual quality during sampling. As shown in Figure 7, compared with the null prompt, adopting proper text prompts can significantly enhance the perceptual quality with more faithful details.

Noise Level. It is observed that the level of noise added to the input can affect the performance of our approach. As shown in Fig. 7, with low noise level, the generated results tend to be suboptimal with blurry details. However, a noise
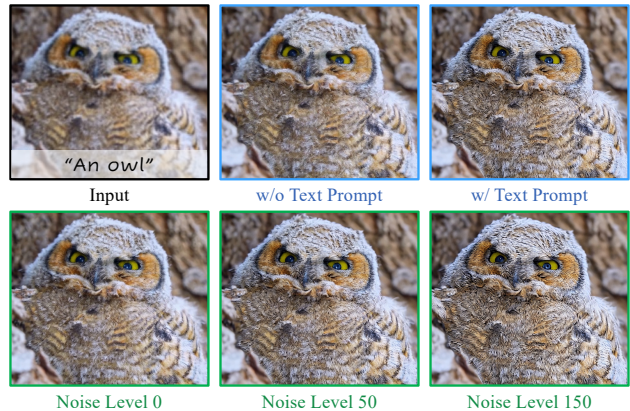

Figure 7. Comparison of various conditions of text prompts and different levels of noise. (**Zoom-in for best view**)

level that is too large may result in oversharpening.

5. Conclusion

While diffusion models have achieved impressive performance on a wide range of image tasks, their applications on video tasks, particularly real-world VSR remain challenging and understudied. In this paper, we present Upscale-A-Video, a novel approach to exploit image diffusion prior for real-world VSR while avoiding temporal discontinuities drawn from the inherent randomness during sampling process. Specifically, we enhance the temporal coherence by proposing a novel local-global temporal strategy within the latent diffusion framework. We additionally devote our efforts to achieving a trade-off between fidelity and quality by enabling texture creation with text prompts and noise level control, further facilitating the practical use in real-world scenarios. We believe that our exploration would lay a good foundation for future works.

References

[1] Pika Labs. https://www.pika.art/, 2023. 6

[2] Stable Diffusion x4 Upscaler. https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler, 2023. 2, 3, 4, 6

[3] Zeroscope V2 XL. https://replicate.com/anotherjesse/zeroscope-v2-xl, 2023. 6

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 3

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 6

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your Latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2, 3, 4, 6

[7] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 2

[8] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In *CVPR*, 2021. 2

[9] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2022. 2, 5

[10] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, 2022. 2, 3, 5, 6, 7

[11] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2, 6

[12] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021. 3

[13] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *NeurIPS*, 2022. 3

[14] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023. 2, 3

[15] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023. 3

[16] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 3

[17] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve Your Own Correlation: A noise prior for video diffusion models. In *ICCV*, 2023. 2, 4, 6

[18] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2

[19] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 3

[20] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 3

[21] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPSW*, 2022. 5, 8

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[25] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3, 6

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022.

[27] Yaosi Hu, Zhenzhong Chen, and Chong Luo. LaMD: Latent motion diffusion for video generation. *arXiv preprint arXiv:2304.11603*, 2023. 3

[28] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *ECCV*, 2020. 2

[29] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *CVPR*, 2020.

[30] Takashi Isobe, Fang Zhu, Xu Jia, and Shengjin Wang. Revisiting temporal modeling for video super-resolution. *BMVC*, 2020.

[31] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, 2018. 2

[32] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022. 3

[33] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, 2021. 6

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[35] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 6

[36] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 2

[37] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *NeurIPS*, 2022. 2

[38] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2

[39] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. In *IEEE TPAMI*, 2013. 2

[40] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. VDT: An empirical study on video diffusion with transformers. *arXiv preprint arXiv:2305.13311*, 2023. 3

[41] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tie-niu Tan. VideoFusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 6

[42] Kangfu Mei and Vishal Patel. VIDM: Video implicit diffusion models. In *AAAI*, 2023. 3

[43] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018. 5

[44] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[45] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, 2019. 2, 6, 7

[46] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 3

[47] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *ICCV*, 2021. 6

[48] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. FateZero: Fusing attentions for zero-shot text-based video editing. *ICCV*, 2023. 3

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[51] Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *arXiv preprint arXiv:2302.07864*, 2023. 2, 3

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[53] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2022. 3

[54] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 4

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[56] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*, 2023. 3

[57] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 6, 7

[58] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 3, 5

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3

[61] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6

[62] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 2, 3, 5, 6

[63] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 5

[64] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 2

[65] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 6

[66] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2, 4

[67] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *ICLR*, 2022. 3

[68] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. LaVie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 4, 6

[69] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *ICLR*, 2023. 3

[70] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 6

[71] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 2, 3, 4

[72] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. DiffIR: Efficient diffusion model for image restoration. *ICCV*, 2023. 3

[73] Liangbin Xie, Xintao Wang, Shuwei Shi, Jinjin Gu, Chao Dong, and Ying Shan. Mitigating artifacts in real-world video super-resolution models. In *AAAI*, 2023. 2, 3

[74] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. SimDA: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 3

[75] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. In *ICCV*, 2019. 2

[76] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. In *NeurIPS*, 2023. 3

[77] S. Yang, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[78] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender A Video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, 2023. 2, 3, 4

[79] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 3

[80] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *ICCV*, 2021. 2, 3, 6

[81] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019. 6, 7

[82] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019. 2

[83] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2023. 2, 3, 5, 6

[84] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 6

[85] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3

[86] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6

[87] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. MagicVideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2

[88] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*, 2019. 5

[89] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In *ICCV*, 2023. 5