This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PatchVSR: Breaking Video Diffusion Resolution Limits with Patch-wise Video Super-Resolution

Shian Du[1*‡], Menghan Xia[2†‡], Chang Liu[1], Xintao Wang[2], Jing Wang[3], Pengfei Wan[2],
Di Zhang[2], Xiangyang Ji[1†]
[1]Tsinghua University, [2]Kuaishou Technology, [3]Beijing Institute of Technology

Abstract

Pre-trained video generation models hold great potential for generative video super-resolution (VSR). However, adapting them for full-size VSR, as most existing methods do, suffers from unnecessary intensive full-attention computation and fixed output resolution. To overcome these limitations, we make the first exploration into utilizing video diffusion priors for patch-wise VSR. This is non-trivial because pre-trained video diffusion models are not native for patch-level detail generation. To mitigate this challenge, we propose an innovative approach, called PatchVSR, which integrates a dual-stream adapter for conditional guidance. The patch branch extracts features from input patches to maintain content fidelity while the global branch extracts context features from the resized full video to bridge the generation gap caused by incomplete semantics of patches. Particularly, we also inject the patch's location information into the model to better contextualize patch synthesis within the global video frame. Experiments demonstrate that our method can synthesize high-fidelity, high-resolution details at the patch level. A tailor-made multi-patch joint modulation is proposed to ensure visual consistency across individually enhanced patches. Due to the flexibility of our patch-based paradigm, we can achieve highly competitive 4K VSR based on a 512×512 resolution base model, with extremely high efficiency.

1. Introduction

Video super-resolution (VSR) has long been a challenging research problem in the field of computer vision. With the advent of deep learning, significant advancements have been made in end-to-end VSR models trained on curated datasets. However, due to limited model capacity and data coverage, these models typically achieve sharp object structures or coarse textures but struggle to synthesize faithful details. Recently, the emergence of diffusion generative models has enabled open-domain video generation of remarkable quality. The generative capabilities of these foundation models present new opportunities for video super-resolution techniques [16, 55]. Nevertheless, existing base models are generally constrained to generating videos of fixed sizes due to their inherent architecture. This limitation is inherited in their downstream super-resolution models, making it challenging for them to support arbitrary resolution outputs.

This study is motivated by the insight that feature attention in super-resolution models is more locally focused compared to generation models. This localization is reasonable, as super-resolution tasks have a low-resolution video as a reference, allowing details to be generated based on neighborhood semantics rather than requiring intensive global coherence considerations. To leverage this insight, we propose a patch-based strategy for video super-resolution. The concept is straightforward and intuitive: we split the input video into patches sized to be compatible with the pre-trained base model, enhance each patch individually, and then combine them to produce the final result. This approach enables the achievement of super-resolution results at arbitrary resolutions, independent of the base model's inherent resolution constraints.

The key challenge lies in that video generation-based models are trained on complete video frames and face noticeable performance degradation for patch-level generation. To adapt a pre-trained video generation model for generating patch-level details in super-resolution, we propose an innovative approach, called *PatchVSR*, which integrates a dual-stream adapter for conditional guidance. The local branch extracts features from the input patches to guide the base model for detail synthesis adhering to the input, while the global branch captures contextual information from the entire input frame to bridge the generation gap caused by incomplete semantics of patches. To make the context guid-

∗ Work done during an internship at KwaiVGI, Kuaishou Technology.
‡ Equal contribution. † Corresponding author.

ance more customized, the global branch also incorporates a binary mask indicating the target patch position as input. Additionally, we employ LoRA technology to fine-tune the base model, further improving its adaptability for generating patch-level details.

Experimental results demonstrate that our method effectively adapts the generative ability of the base model to produce faithful details at the patch level. Specifically, to alleviate the boundary inconsistency across patches, we propose a training-free multi-patch joint modulation scheme, where auxiliary bridging patches are created to allow feature interaction across patches during generation. Thanks to the flexibility of PatchVSR, our method, based on a pre-trained 512×512 resolution video base model, supports super-resolution video generation up to 4K resolution, while significantly outperforming current state-of-the-art methods in terms of fidelity and detail realism. Notably, compared to existing methods that operate on full-frame videos, our approach processes relatively smaller patches, without full-range attention computation, demonstrating a significant efficiency advantage.

Our contributions are summarized as follows:

- We introduce an innovative approach to overcome the resolution limitations of using pre-trained video diffusion models for VSR. Our method achieves any-resolution VSR output based on a fixed-resolution base model.
- We propose a novel dual-branch adapter paradigm that effectively harnesses the capabilities of pre-trained video models for patch-level detail generation.
- Through extensive experimentation, we demonstrate the significant superiority of our method in terms of faithful detail generation and computational efficiency.

2. Related Work

2.1. Text-to-video Generation

With the widespread use of pre-trained text-to-image models [10, 13, 18, 32, 39, 52] and the preparation of large-scale text-video pairing datasets [9, 34], substantial efforts have been made to train a large-scale text-to-video model. Instead of training from scratch, some methods [2, 41, 46, 51] focus on inserting temporal modules and fine-tuning them or all parameters on video data. To avoid catastrophic forgetting and degradation on visual quality, they often adopt image-video joint training strategy [7, 8, 47]. Recently, with Diffusion Transformer (DiT) [36] being proposed, it has replaced the traditional 3D UNet architecture used by previous approaches with better generation quality and scalability, and has become the dominant framework for text-to-video generation [28, 30, 49]. Although achieving high visual quality, extending them to higher-resolution is non-trivial since it requires substantial training resource and high-quality high-resolution video dataset, which is unavailable by this moment. To bypass this difficulty, current methods often adopt a cascade framework [15, 44, 54]: the low-resolution video is first generated by the base model using text prompt, and the refinement model outputs high-resolution video through a noising-denoising process [31] by taking the low-resolution video as condition. However, existing methods are limited to generate full video at once, which still demands a high training overhead. At the same time, the low inference efficiency and high memory usage also hinder its downstream applications.

2.2. Video Super-Resolution

Video super-resolution aims at restoring high-resolution (HR) video frames from its degraded low-resolution (LR) counterparts and enhancing its visual quality. Traditional methods often assume a fixed degradation process and their performance deteriorates significantly when handling AI-generated and real-world videos. Due to the lack of real-world LR-HR video data pair for training, RealVSR [48] collects LR-HR video data pair by utilizing the multi-camera system of iPhone. Nevertheless, the sequences are captured by a single device, which only contains camera-specific degradations. To avoid burdensome data collection, other methods [6, 22, 55] use stochastic combinations of multi-order degradation processes to simulate real-world degradation processes. While performing more robust when handling real-world cases, due to the absence of deep generative prior, it is challenging for traditional CNN-based [4–6, 21, 43] and transformer-based [3, 14, 25, 40] methods to generate photo-realistic details and textures. By utilizing a pre-trained image diffusion model like Stable Diffusion (SD) $\times 4$ Upscaler, Lavie-SR [44] and Upscale-A-Video [55] circumvent the need for exhaustive training from scratch and exhibit improved performance in fixed spatial scale within a short sequence. To improve temporal consistency for long sequence, I2VGen-XL [54] and VEnhancer [16] utilize a pre-trained low-resolution video generative model as the base model, and directly fine-tune the model parameters using high-resolution video dataset following noising-denoising manner [31]. However, the aforementioned methods are all limited to generate full video at once, which requires expanding the training resolution of pre-trained model, bringing low training efficiency and slow convergence speed. Meanwhile, high resolution pre-trained video generative models are currently unavailable to access, hindering the scalability of video super-resolution methods with deep generative prior. To tackle this problem, we make the first exploration of utilizing pre-trained video generative model for patch-wise video super-resolution. By maintaining the training resolution of the pre-trained model, our approach not only avoids significant training overhead, making it possible to generate 4K resolution videos using a 512×512 resolution pre-trained model. At the same time,

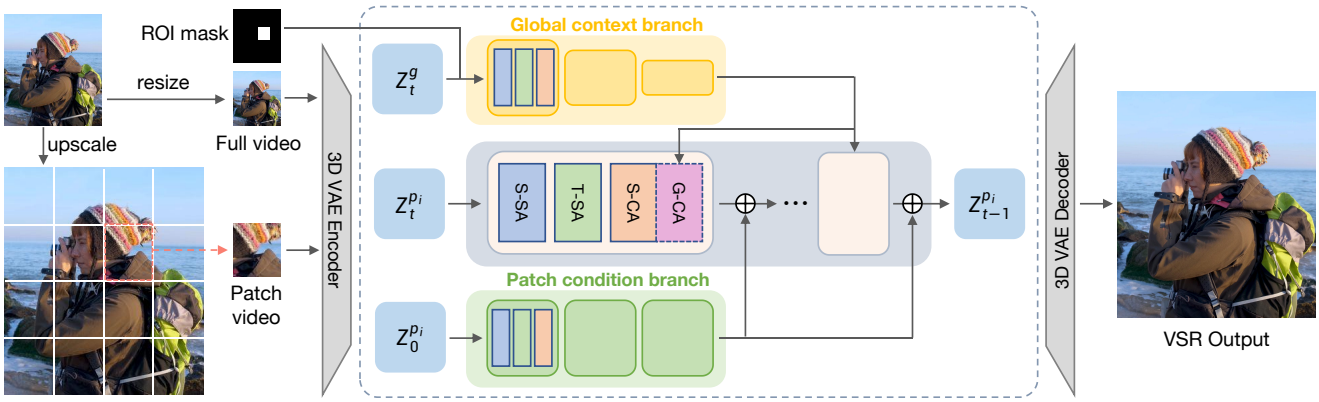


Figure 1. Flowchart of our *PatchVSR*. Building upon a pre-trained latent T2V model, we incorporate a patch condition branch and a global context branch. These branches extract features from partitioned video patches and the resized full video together with a binary mask that indicates the location of the ROI patch, respectively. Particularly, local patch features are added to the output of each block, while the global context feature is fused with the backbone feature through newly introduced cross-attention modules (*G-CA*). For simplicity, we have omitted other conditional inputs such as text prompts and time steps from this diagram. The processed patches are fused via a joint modulation scheme to produce a coherent super-resolution video.

our method significantly improves inference efficiency and reduces memory usage, facilitating downstream applications of video super-resolution methods.

3. Method

We aim to develop a video super-resolution (VSR) framework designed to restore degraded videos of arbitrary resolutions by generating visually coherent and structurally faithful details. Our primary innovation involves strategically adapting pre-trained text-to-video (T2V) diffusion models—powerful generative tools typically constrained to fixed output size (denoted as $s = h \times w$)—through a curated conditioning paradigm. To bypass their inherent token-length limitations, we introduce a patch-based processing strategy: the input video is decomposed into s-sized patches, each independently enhanced using the diffusion model, and then reintegrated into a unified high-resolution output through optimized spatial-temporal fusion. This approach is grounded in the distinct characteristics of VSR compared to video generation. While video synthesis demands global coherence across full-frame contexts, super-resolution prioritizes localized detail restoration atop existing structural content.

Specifically, we augment the base model with two conditional branches: a patch branch that extracts features from input patches, and a global branch that extracts context from the full video. This global context mainly aids in generating patch-level details from the T2V model that was originally trained on full-frame videos. An overview of the model architecture is illustrated in Fig. 1. Given a low-resolution image $\mathbf{V}_l \in \mathbb{R}^{F\times H\times W\times 3}$ and the target upscale factor k, our process begins by upscaling $\mathbf{V}_l$ to the desired resolution $\mathbf{V}_l^{\uparrow} \in \mathbb{R}^{F\times \lfloor H\cdot k\rfloor \times \lfloor W\cdot k\rfloor \times 3}$ using bicubic interpolation per frame. We then partition $\mathbf{V}_l^{\uparrow}$ into patches $\{\mathbf{P}_i \in \mathbb{R}^{F\times h\times w\times 3}\}_{i=1}^{N}$, where the patch dimensions (h, w) match the generation size of the pre-trained T2V model. Additionally, we resize the full video to $\mathbf{G} \in \mathbb{R}^{F\times h\times w\times 3}$ to serve as global guidance. Finally, both $\mathbf{P}_i$ and $\mathbf{G}$ are input into the VSR model for patch-level detail enhancement, and multi-patch joint modulation is involved to ensure coherence across patches. This formulation ensures compatibility with the pre-trained T2V model while allowing for flexible processing of videos of arbitrary sizes.

3.1. Preliminary: Video Diffusion Base Model

Our model is implemented on top of a pre-trained T2V diffusion model, which adopts the Rectified Flow framework [11] for the noise schedule and denoising process. The forward process is defined as straight paths between data distribution and a standard normal distribution, i.e.

$$z_t = (1-t)z_0 + t\epsilon, \tag{1}$$

where $\epsilon \in \mathcal{N}(0, I)$ and t denotes the iterative timestep. To solve the denoising processing, we define a mapping between samples z_1 from a noise distribution p_1 to samples z_0 from a data distribution p_0 in terms of an ordinary differential equation (ODE), namely:

$$dz_t = v_\Theta(z_t, t)dt, \tag{2}$$

where the velocity v is parameterized by the weights Θ of a neural network. For inference, we employ Euler discretization for Eq. 2 and perform discretization over the timestep interval at $[0, 1]$, starting at $t = 1$. We then processed with iterative sampling with: $z_{t_i} = z_{t_j} + v_\Theta(z_{t_j}, t_j) * (t_j - t_i)$.

For consideration of computing efficiency, the mapping is formulated in a latent space constructed with a

3D Variational Auto-Encoder (VAE) [24]. The core is a Transformer-based diffusion model (DiT) [35], where each block is instantiated as a sequence of spatial attention, temporal attention, and cross-attention modules. Text prompts and other micro parameters (like time step, aspect ratio, etc) are used as generative conditions. Due to the full-attention property of Transformer, our model only support the generation of a fixed number of tokens as it was trained with, i.e. 512×512 of diverse aspect-ratio. This is a common characteristic of almost all existing diffusion models for image or video generation [7, 8, 37, 49], and extending pre-trained diffusion models for higher resolution generation remains an open research topic [15, 17]. For VSR as a downstream task, we propose a patch-wise processing strategy to break this resolution limit.

3.2. Patch Video Diffusion with Global Guidance

To adapt a pre-trained T2V model for patch-level VSR, we must consider two key aspects: (i) enabling the input video to guide the base model's denoising process and (ii) bridging the gap between the base model's capabilities in generating full-frame content and patch-level content synthesis requirements. We achieve this through a dual-branch adapter, as detailed below.

Patch Condition Branch. There are generally two approaches to incorporating visual conditions into pre-trained generative models: concatenating with noisy data in the noise space [39, 55], or fusing with the backbone feature through an adapter branch [16, 42]. The former involves modifying the first layers' channel count and necessitates base model fine-tuning, which is typically more data and computationally intensive. In contrast, the latter is more lightweight, as only a newly added branch requires training for feature adaptation. We adopt a hybrid solution that combines elements of both approaches. First, we employ an adapter comprising several Transformer blocks to extract features from the input video patch $\mathbf{P}_i$ and inject them into the base model's blocks. Additionally, considering that the base model was trained for full video generation, we need to adapt its embedded knowledge for patch-level generation. To this end, we allow the base model to be fine-tuned using a Low-Rank Adaptation (LoRA) [19]. An ablation study in Sec. 4.3 demonstrates the advantages of this design.

Global Context Branch. To further guide content generation in a natural manner, we incorporate a global branch that extracts context from the full input video. This branch utilizes a Transformer-based encoder architecture that reduces the number of tokens to one quarter. The resulting context features are fused with each block of the base model through cross-attention $\{\mathbf{Q}, \mathbf{K}_g, \mathbf{V}_g\}$, where the $\mathbf{Q}$ is shared with the existing text-prompt cross-attention $\{\mathbf{Q}, \mathbf{K}_t, \mathbf{V}_t\}$. Specifically, the input is constructed by concatenating the resized full video $\mathbf{V}_g$ (to accommodate a fixed number of tokens) with a binary map $\mathbf{M}_i$ indicating the target patch's $\mathbf{P}_i$ location within it. Our experiments show that the global context branch not only enhances the vividness of generated patch-level details but also improves consistency across patches.

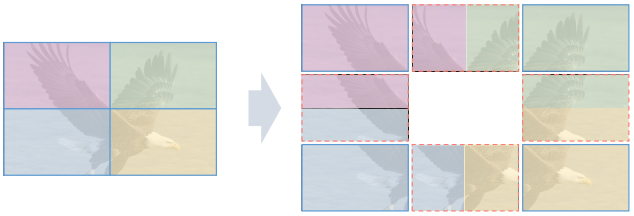


Figure 2. Patch Partition Visualization. The input video is divided into non-overlapping segments, as the solid blue boxes mark. For joint modulation, auxiliary patches are created, indicated by the red dashed boxes, resulting in an overlapping ratio of 50%.

3.3. Multi-Patch Joint Modulation

While the global guidance shared by patches is critical for maintaining color consistency, tonal uniformity, and semantic coherence of details, naive stitching of independently enhanced patches often introduces visible seams. These boundary artifacts primarily stem from the stochastic nature of the generative process when hallucinating textures on degraded or underdetermined regions. This is a fundamental challenge in ill-posed inverse problems where multiple high-resolution solutions may satisfy the constraints of a single low-resolution input.

To mitigate this problem, we introduce inter-patch modulation during the detail generation process. Specifically, a training-free inference strategy called multi-diffusion [1] is adapted to solve our problem. This approach involves splitting the original video into overlapping patches and updating the overlapping regions by averaging contributions from all relevant patches at each denoising step. As illustrated in Fig. 2, instead of directly splitting the video into overlapping patches, we first partition it into non-overlapping segments, and then construct auxiliary patches by combining the halves of adjacent patches. This creates a new set of patch groups with an overlapping ratio of 50%. Our experiment shows that directly taking the average in overlap regions suffers from artifacts of black holes or seamlines, as evidenced in Fig. 5 (middle). Instead, when fusing base patches and its related auxiliary patches, we manage to gradually reduce its influence of auxiliary patches from its center splitting line to two sides by assigning spatial weight maps to them. This method effectively reduces visible seams between patches and promotes overall consistency. A comparative result demonstrating the effectiveness of this approach is presented in Fig. 5.



Figure 3. Qualitative comparisons on SynVideo30 (test). These videos $\times 4$ super-resolution results. **Zoom in for best view.**

4. Experiments

To verify the superiority of PatchVSR, we compare it with SOTA methods, including VEnhancer [16], Upscale-A-Video [55], LaVie-SR [44] and RealBasicVSR [6]. We also make ablation study to analyze the effectiveness of our proposed designs. Besides, some extended analysis is conducted to demonstrate the capability of our method.

4.1. Implementation Details

Training Datasets We train our model using 460K self-collected high-quality video-text pairs, with each resolution from 1024×1024 to 2K. The text prompts are all captioned using LLAVA captioner [26], and encoded by T5 text encoder [38] with no more than 256 tokens.

Testing Datasets. We evaluate the VSR methods on two kinds of dataset, i.e. downscale synthetic videos and AIGC videos. For synthetic videos, since our employed base model is fixed to generate 77-frames video, we utilize a public dataset: REDS30 [33] for fair comparisons, which contains 100 frames. To evaluate higher resolution, we self-collect 30 high-quality 2K resolution videos, each containing more than 150 frames. Additionally, we collect and test on an AIGC dataset, which comprises 30 AI-generated videos from public text-to-video generative models [50].

Training Details. Our model is trained on 16 NVIDIA H800-80G GPUs with a total batch size of 64. AdamW [29] is used as the optimizer with a learning rate of 10^{-5}. The text prompt is randomly replaced by a null prompt with 10% probability. To obtain the LR-HR training data pair, the HR video is first down-sampled using bilinear interpolation with down scale s and up-sampled back to the original resolution. To match the size of the pre-trained model, we randomly crop a 512×512 area from the up-sampled LR video with aspect ratios from 0.5 to 2.0. To enhance the robustness of our model to different degradation scenarios, we utilize the noise augmentation technique by injecting noise into the input latent using a diffuse process. The noise timestep is randomly selected from 200 to 300, aiming at preserving the structure of the input video. We have additionally encoded noise timestep, downscale factor, and crop location as conditions for the model. We have replaced the text prompt with a fixed prompt for the patch condition branch, with the base model and global context branch using text prompts. To provide initialization weight, we followed the structure of the pre-trained model and sampled the layer index from it at equal intervals.

Inference Details. We perform 50 PNDM [27] sampling steps with classifier-free guidance scale 5.0. We have also used timestep shift [12] with shift value 5.0.

Evaluation Metrics. To evaluate the fidelity of our method on synthesized videos, we utilize PSNR, SSIM, and LPIPS [53]. For visual quality, we conduct our evaluation of commonly used quality assessment metrics MUSIQ [23] and DOVER [45]. Furthermore, we also include *aesthetic score* from a video generation benchmark VBench [20] for more comprehensive comparisons.

4.2. Comparison with SOTA methods

Qualitative Results. The visual results of synthetic and AIGC videos are presented in Fig. 3 and Fig. 4 respectively. It is observed that our method significantly outperforms existing CNN-based and diffusion-based models in both details generation and temporal consistency (Readers are recommended to check our provided video results). Specif-



Figure 4. Qualitative comparisons on VideoGen30 (test). These videos are $\times 4$ super-resolution results. **Zoom in for best view.**

Datasets	Metrics	RealBasicVSR [6]	LaVie-SR [44]	Upscale-A-Video [55]	VEnhancer [16]	Ours
REDS30	PSNR ↑	29.698	28.735	31.257	28.544	30.111
	SSIM ↑	0.656	0.622	0.661	0.611	0.647
	LPIPS ↓	0.334	0.407	0.328	0.421	0.326
	DOVER ↑	0.305	0.318	0.310	0.344	0.323
	MUSIQ ↑	40.587	43.577	42.575	44.533	44.956
	Aesthetics ↑	0.422	0.486	0.471	0.482	0.493
SynVideo30	PSNR ↑	33.507	31.433	33.432	28.856	30.857
	SSIM ↑	0.776	0.716	0.728	0.697	0.732
	LPIPS ↓	0.185	0.194	0.205	0.199	0.183
	DOVER ↑	0.584	0.612	0.529	0.550	0.597
	MUSIQ ↑	49.557	48.146	49.839	43.538	50.695
	Aesthetics ↑	0.496	0.509	0.494	0.503	0.520
VideoGen30	DOVER ↑	0.588	0.609	0.510	0.512	0.590
	MUSIQ ↑	47.564	48.329	46.899	46.930	50.559
	Aesthetics ↑	0.547	0.600	0.574	0.551	0.602

Table 1. Quantitative comparisons on different VSR benchmarks from synthetic (REDS30, SynVideo30) and AIGC (VideoGen30) data. Red and blue numbers indicate the best and second best results respectively.

ically, our method not only successfully recovers the details of the red silk scarf in Fig. 3, but also adds more details to the "Angkor Wat" as shown in Fig. 4, which others tend to generate blurry or unnatural results. In Fig. 5, we also illustrate full VSR results achieved through the multi-patch joint modulation scheme. It can be seen that creating auxiliary patches greatly alleviates the seam problem and significantly reduces the color difference between different patches, while further improving the contrast of the results. To show the advantage of our multi-patch joint modulation, we compare with a baseline method: stitching the patches in latent space and decoding the big latent together, which suffers obvious seamline artifacts.

Quantitative Results. The quantitative comparison is shown in Tab. 1. It can be seen that the proposed method achieves the highest LPIPS score in both REDS30 and SynVideo30, indicating the high fidelity of our method for synthetic videos. However, our method show inferior performance in PSNR and SSIM. This can be explained by the fact that our method tends to generate rich and vivid details, which are not favorable to metrics that capture pixel-wise deviations (i.e. PSRN and SSIM). Instead, as a perceptual-level fidelity measure, LPIPS might be more informative to reflect the content fidelity for generative tasks. We evaluate the visual quality of generated high-resolution videos through comparisons with competing methods on MUSIQ

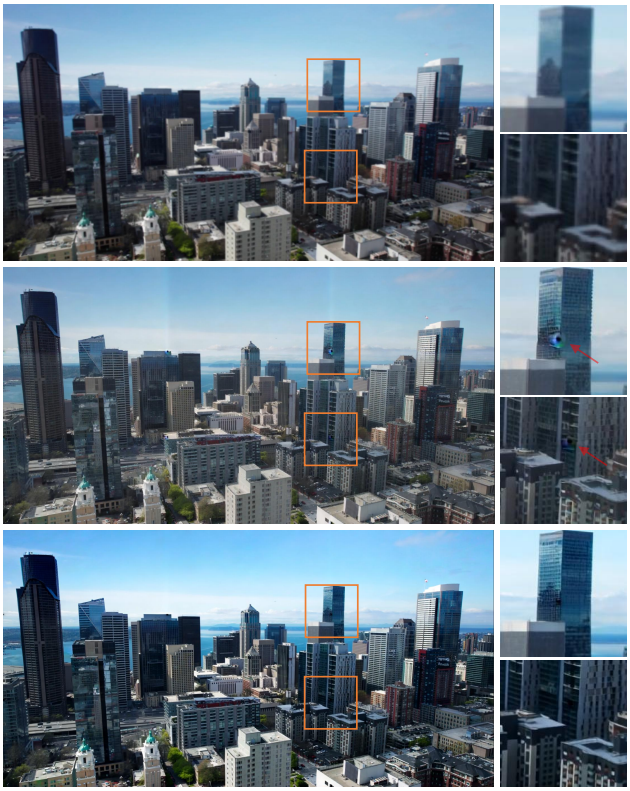


Figure 5. Visual evaluation on multi-patch joint modulation. Here the super-resolved videos of 2K resolution are used for comparison, where $3 \times 3 = 9$ patches are involved. The top, middle, bottom denote input video, result of stitching latent of patches, and result of ours respectively.

and DOVER metrics, demonstrating our approach's superiority. This performance reflects contributions from both the base model's capabilities and the effectiveness of our method in leveraging full-frame generation for patch-level detail synthesis. Interestingly, our approach even preserves subjective properties like aesthetic quality from the base model, as evidenced by the aesthetic scores. This highlights our framework's ability to inherit and maintain critical perceptual characteristics during the enhancement process.

4.3. Ablation Study

We conduct ablation studies on key designs. Considering the limited space, we provide a quantitative evaluation here and the typical visual results are compared in the supplementary material.

Global Branch. To validate the effectiveness of the global branch, We build a baseline by removing the global branch from our method. As shown in Tab. 2, removing the global branch results in significant performance deterioration, leading to worse DOVER, MUSIQ, and *aesthetic score*. Besides, to check the necessity of using the patch location embeddings, we build a baseline by removing the patch position information from the global branch input. As shown in Tab. 2, after incorporating the location embedding, we obtain an increase of DOVER from 0.574 to 0.590, which demonstrates the effectiveness of utilizing the location mask as the patch position information.

Component	DOVER ↑	MUSIQ ↑	Aesthetics ↑
w/o global branch	0.502	46.074	0.589
w/o LoRA	0.582	50.496	0.600
w/o location embed	0.574	50.084	0.601
w/o fixed prompt	0.562	48.133	0.597
Ours	**0.590**	**50.559**	**0.602**

Table 2. Ablation study of different model components on VideoGen30 test dataset. The **bold** numbers indicate the best.

Text Prompts. Text prompts are used in the base model and the conditional global branch. As shown in Tab. 2, we observe that replacing the text prompt with a fixed prompt leads to superior results. We conjecture that the use of text prompts in patch condition branch brings about a mismatch between the video patch and the global semantic information, which in turn destroys the dynamic visual prior in the pre-trained model aligned to the text during training.

Base model finetuning. Since the data distribution of local patches deviates from the full view of the range of common videos, we propose to fine-tune the base model instead of freezing it. It can be seen in Tab. 2 that adding LoRA results in improved performance across all metrics, which demonstrates the effectiveness of utilizing LoRA to adapt to the distribution of local video patches.

4.4. Analysis and Discussion

We would like to further study some characteristics of our proposed method, including detail generation magnitude, robustness to patch semantics, and computation efficiency.

Detail generation magnitude. The input and output sizes of our VSR model are the same and our model can synthesize different level of details. In our method, a scale factor is used to guide the model to generate reasonable details that should be presented on the input video if it is captured normally by the camera at that resolution/size. So, the scale factor actually plays the role of affecting the magnitude of details. Here we illustrate the effects of using different scale factors as conditions. In Fig. 6, increasing the scale factor progressively adds more high-resolution details to the results, which demonstrates the flexibility of our method to generate results with different detail magnitudes.

Robustness to Patch Semantics. As stated above, our approach will partition the input video into patches in a content-agnostic manner. So, some patches can contain incomplete or even unclear semantics, which are more challenging for the model with the generative reasonable details.

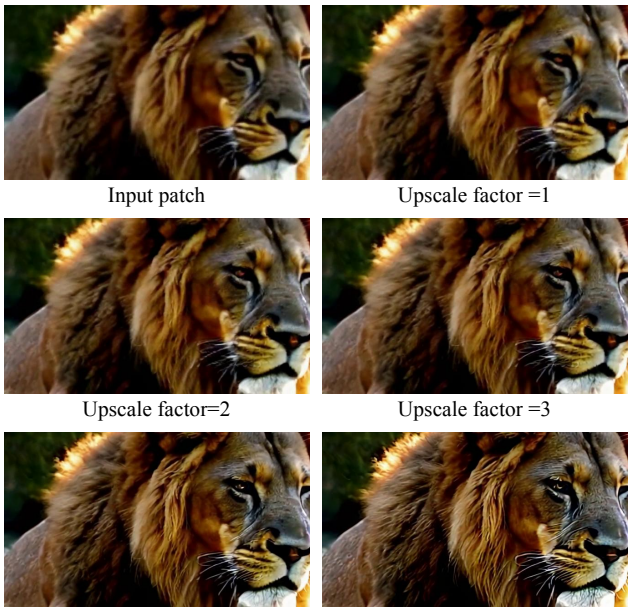


Figure 6. Comparison of different up-sampling scales. Note that the upscale factor is uniformly randomly sampled within the range of 1 to 4 during the training phase.

Methods	Time (s) ↓	Memory (G) ↓
LaVie-SR	2261	68
Upscale-A-Video	2743	47
VEnhancer	1562	62
Ours	**680**	**40**

Table 3. Computation efficiency for 2K video generation.

It is interesting to study whether the performance varies across patches with diverse semantic quality. It can be seen in Fig. 7 that our model can successfully handle the video patches with incomplete semantics. As the global semantic tokens are extracted and input to the base model using global cross-attention modules, the model learns to utilize global semantic information to complete the semantics in the local video patch, which improves the robustness of our model against incomplete semantics.

Computation efficiency. Our method is more efficient than existing pre-trained T2V-based VSR methods because the self-attention is conducted within patches. Here presents a simple deduce of the computation complexity against patch size: given a video of n tokens and each patch of m tokens, the patch number is $k = \frac{n}{m}$ and the inter-token computation times are $k * (\frac{n}{k})^2 = n * m$. Since the base model we adopted uses fixed 512×512 resolution, our patch size follow this size in our experiments. The inference time and memory usage of generating a 2K resolution video is illustrated in Tab. 3. It can be seen that, after dividing the input video into patches, the computation burden is significantly reduced, further promoting the applications of generative VSR models. Thanks to the computation efficiency and the inference flexibility, our method even support 4K video generation, which are shown in the supplementary materials.



Figure 7. Robustness against patch semantic incompleteness.

Limitation. Our method has limitations in two aspects. Firstly, it experiences noticeable performance degradation when applied to real-world low-resolution images, due to the lack of specialized data augmentation in our training pairs. Since the primary focus of this paper is to address the resolution limitations of using pre-trained base models for VSR, extending our approach to handle real-world noise remains an interesting direction for future work. Secondly, our approach inherits an iterative inference scheme from the base model, which introduces efficiency issues for practical applications. For VSR, the low-resolution input video provides significant opportunities for algorithm optimization to reduce the number of inference steps, which will be another area of future research.

5. Conclusion

In this paper, we present the first exploration of utilizing a pre-trained T2V base model for patch-level video super-resolution (VSR). To accomplish this, we propose an effective dual-branch adapter consisting of a patch condition branch and a global context branch. The global context branch is crucial for bridging the gap between full-video generative capabilities and patch-level detail synthesis. Additionally, our proposed multi-patch joint modulation scheme achieves consistent results across patches. Thanks to the flexibility of our PatchVSR approach, it outperforms existing VSR methods in generating faithful details while maintaining high computational efficiency. We hope this work will inspire further research in this area.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC) under Grant No.62406167, U24B6012.

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning (ICML)*, 2023. 4

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

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

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 2

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022.

[6] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 2, 5, 6

[7] 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, 4

[8] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2, 4

[9] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 2

[10] Shian Du, Xiaotian Cheng, Qi Qian, Henglu Wei, Yi Xu, and Xiangyang Ji. Efficient personalized text-to-image generation by leveraging textual subspace. *arXiv preprint arXiv:2407.00608*, 2024. 2

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. 3

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 5

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[14] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17441–17451, 2022. 2

[15] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, Ying Shan, and Bihan Wen. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 4

[16] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024. 1, 2, 4, 5, 6

[17] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. 4

[18] 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. 2

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 4

[20] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 5

[21] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent

structure-detail network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 645–660. Springer, 2020. 2

[22] Mehran Jeelani, Noshaba Cheema, Klaus Illgner-Fehns, Philipp Slusallek, Sunil Jaiswal, et al. Expanding synthetic real-world degradations for blind video super resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1199–1208, 2023. 2

[23] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 5

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 4

[25] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5687–5696, 2022. 2

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5

[27] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 5

[28] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 2

[29] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[30] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 2

[31] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[32] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[33] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 5

[34] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 2

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

[36] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 4

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 5

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[40] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *Advances in Neural Information Processing Systems*, 35:36081–36093, 2022. 2

[41] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[42] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 2024. 4

[43] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[44] 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. 2, 5, 6

[45] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023. 5

[46] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2

[47] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying

Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 2

[48] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4781–4790, 2021. 2

[49] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 4

[50] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024. 5

[51] 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. *International Journal of Computer Vision*, pages 1–15, 2024. 2

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[54] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[55] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2545, 2024. 1, 2, 4, 5, 6