Cyprus University of Technology

杭州電子科技大学

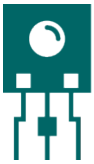
HANGZHOU DIANZI UNIVERSITY

Master's thesis

Research On Video Super-Resolution Technology Based On Diffusion Model

Chengzhang Wang

Limassol, February 2024



MSc in Electronics Science and Technology

CYPRUS UNIVERSITY OF TECHNOLOGY

Faculty of Engineering and Technology

Department of Electrical Engineering, Computer Engineering, and Informatics

Master's thesis

Research On Video Super-Resolution Technology Based On Diffusion Model

Chengzhang Wang

Supervisor

Prof. Sotirios Chatzis

Limassol, February 2024

Approval Form

Master's thesis

Research On Video Super-Resolution Technology Based On Diffusion Model

Presented by

Chengzhang Wang

Supervisor: Sotirios Chatzis

Member of the committee: Panagiotis Ilia

Member of the committee: Andreas Diavastos

Cyprus University of Technology
Limassol, February 2024

Copyrights

Copyright © 2024 Student Name

All rights reserved.

The approval of the dissertation by the Department of Electrical Engineering, Computer Engineering, and Informaticsdoes not necessarily imply the approval by the Department of the views of the writer.

Acknowledgements

I would like to express their sincere gratitude to Professor Sotirios Chatzis and the research team at the Cyprus university of technology for their generous support, insightful guidance, and continuous encouragement throughout this project. Special thanks are also extended to our senior PhD mentor, whose invaluable advice and assistance significantly contributed to the progress and success of this work.

We are also deeply grateful to Professor Zhiwei He of HangZhou Dianzi university for his constant care, support, and constructive suggestions during the research process. Finally, we would like to thank HangZhou Dianzi university for providing this valuable opportunity for international academic exchange, which has greatly broadened our horizons and inspired further collaboration.

ABSTRACT

With the rapid development of digital video technology, people's demand for video quality is getting higher and higher. Super resolution video task aims to improve the visual clarity of video by improving the spatial resolution of video frames, which is of great significance for video surveillance, video conferencing, film and television post-production and other fields. video super-resolution (VSR) is the process of reconstructing a low-resolution video sequence into a high-resolution video sequence. In the study of video super resolution, diffusion model can not only process the image clarity of single video frame, but also generate clearer and more coherent video frame. Such models often involve complex mathematical algorithms and deep learning techniques that automatically learn and simulate complex relationships between video frames to achieve transitions from low resolution to high resolution. In short, video super resolution technology is an important direction of the development of digital video technology, it is of great significance for improving video quality, enhancing user experience and promoting the development of related industries. As technology continues to advance, we can expect to see more efficient, high-quality video super resolution solutions emerge in the future.

Keywords: Deep Learning; Artificial Intelligence; Video Super-resolution; Diffusion Model;

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

VSR	Video Super Resolution
VIT	Vision Transformer
LR	Low Resolution
HR	High Resolution

1 Introduction

1.1 Aims and Objectives

Video Super-Resolution (VSR) technology aims to reconstruct high-resolution (HR) video sequences from their low-resolution (LR) counterparts. As a critical task in video enhancement and restoration, VSR has wide-ranging applications in areas such as medical imaging, video surveillance, digital archiving, and entertainment media, where legacy or low-quality video content often needs to be upscaled for modern high-resolution displays.

With the rapid proliferation of high-definition (HD) and ultra-high-definition (UHD) devices, user demand for high-quality video content continues to grow, making VSR an increasingly essential research direction in both academia and industry. However, the VSR task is considerably more challenging than its image counterpart due to the complex spatiotemporal dynamics of video, such as frame-wise motion, temporal consistency, and varying degradation patterns across time. Achieving both high spatial fidelity and temporal coherence in reconstructed videos remains a fundamental difficulty.

Recently, diffusion models have emerged as a powerful generative paradigm, demonstrating state-of-the-art performance in image generation, editing, and super-resolution tasks. Compared to traditional methods such as convolutional neural networks (CNNs) or generative adversarial networks (GANs), diffusion-based models offer superior training stability, fidelity, and diversity of output. Their capacity to model complex data distributions through iterative denoising has positioned them as a leading approach in modern image synthesis research.

Despite their success in static image domains, applying diffusion models to video super-resolution remains a highly non-trivial extension. This is primarily due to the higher dimensional complexity of video data. In particular, whereas images are typically represented as 4D tensors [B,C,H,W], video inputs introduce an additional temporal dimension, resulting in 5D tensors [B,C,T,H,W], where T denotes the number of frames. This increases both the computational load and the memory footprint significantly, making model training and inference more resource-intensive.

Furthermore, the inclusion of the temporal dimension introduces additional challenges: the model must not only reconstruct spatial details within each frame but also learn inter-frame dependencies to maintain temporal consistency and avoid artifacts such as flickering or ghosting. In practice, this requires more sophisticated architectures and training strategies, especially when dealing with real-world degradation factors like compression noise, motion blur, and dynamic lighting.

Another practical limitation arises from the difficulty in dataset construction. High-quality paired datasets for VSR are limited, and existing ones—such as REDS or UDM10—often involve large-scale video collections that demand long training times and high-performance hardware to process. This data bottleneck further complicates the deployment of large generative models like diffusion models in video restoration tasks.

To address these challenges, recent studies have explored two primary directions: (1) modifying the UNet-based denoising architecture of diffusion models to handle 3D spatiotemporal inputs, often by

integrating 3D convolutions or frame-wise attention modules; and (2) designing fully 3D diffusion models that treat the entire video clip as a unified 3D data volume. While both strategies have shown promising results in video generation and restoration, they often involve significant architectural changes and high computational costs.

In summary, while diffusion models offer a promising framework for tackling video super-resolution tasks, scalability, efficiency, and adaptability remain key research bottlenecks. There is a growing need for solutions that can leverage the power of pre-trained diffusion backbones without requiring task-specific redesign. Addressing these limitations could pave the way for more generalizable, lightweight, and plug-and-play VSR solutions suitable for real-world deployment. As technology advances, it is expected that VSR powered by generative diffusion models will become more capable and accessible, enabling a broad spectrum of applications across industries.

1.2 Research Questions

The primary objective of this study is to explore how to effectively enhance the spatial resolution of low-resolution video sequences while maintaining their temporal consistency across frames. Video Super-Resolution (VSR) is not merely a frame-wise upsampling problem—it fundamentally requires the model to capture and reconstruct high-frequency spatial details while also ensuring coherent motion and visual smoothness over time. This leads to a complex interplay between spatial fidelity and temporal stability, both of which are essential to generating perceptually pleasing and realistic video outputs.

To this end, the research is guided by the following central questions:

-How can we design a conditional generative framework that enables accurate reconstruction of high-resolution frames from low-resolution video inputs without sacrificing temporal consistency?

-In the context of diffusion models, which have shown strong performance in image generation, how can we leverage their generative capabilities to serve the needs of video super-resolution tasks, especially in real-world settings with unknown degradation?

-Can we reuse existing text-to-video diffusion models in a zero-shot or minimally supervised manner by adapting their input conditioning mechanism to accept visual (rather than textual) cues, and still achieve high-quality super-resolution?

-What kind of conditioning representation best guides the denoising process in diffusion-based VSR: Is it possible to “disguise” visual tokens as language embeddings and inject them into cross-attention layers to activate pre-trained generative capabilities without retraining the entire model?

-To what extent does the choice of UNet backbone (e.g., task-specific versus generic text-to-video) impact the final VSR quality, and what is the role of the adapter in balancing detail recovery and motion consistency?

-What architectural or algorithmic components are most critical in preserving high-frequency texture details while avoiding common temporal artifacts such as flickering, ghosting, or misalignment across frames?

-How can we systematically evaluate the performance of a VSR system not only in terms of traditional

metrics like PSNR and SSIM, but also with respect to perceptual quality and temporal stability across diverse datasets?

This study aims to address these questions through the design and empirical validation of an adapter-based VSR framework. By reformatting low-resolution video features into pseudo-language tokens compatible with pre-trained text-to-video diffusion models, we explore a novel plug-and-play strategy for high-quality video restoration. The overarching hypothesis is that an effective cross-modal adapter can serve as a minimal yet powerful intervention to unlock the generative potential of large-scale diffusion models in video super-resolution applications.

1.3 Contribution

To address these limitations, we propose a novel adapter-based conditional modeling framework for video super-resolution (VSR). The central idea is to repurpose a pre-trained text-to-video diffusion model for super-resolution tasks without modifying its original architecture. Specifically, we introduce a lightweight adapter module that reformulates low-resolution video frames into pseudo-textual embeddings. These embeddings are structurally aligned with the language inputs that the diffusion model was originally trained on, allowing seamless integration into the model's cross-attention layers as encoder hidden states. This enables the original diffusion backbone—comprising the UNet and VAE—to be reused in a zero-modification and zero-finetuning setting.

Our contributions are summarized as follows:

- **Architecture-preserving design:** The proposed method operates without altering the backbone of the pre-trained diffusion model. This eliminates the need for retraining large-scale generative components such as the UNet, significantly reducing the computational cost of adaptation.
- **Modularity and plug-and-play deployment:** The adapter is self-contained and easily attachable to any existing text-to-video diffusion framework. It supports flexible integration with a wide range of generative backbones, facilitating fast adaptation to downstream tasks like VSR.
- **Competitive performance without task-specific supervision:** Despite its simplicity, the adapter-based conditioning mechanism achieves performance that is comparable to or better than existing VSR-specific models across multiple benchmarks (REDS, UDM10, VID4). Our model delivers strong results not only in pixel-wise reconstruction metrics such as PSNR and SSIM, but also in perceptual quality indicators like LPIPS and temporal consistency scores.
- **Enhanced high-frequency detail recovery:** Qualitative and quantitative results reveal that the adapter module plays a pivotal role in restoring fine-grained textures and structural details, which are often lost in baseline generative models. This suggests that the adapter effectively bridges the modality gap between visual and language representations, providing more informative conditioning signals to guide the denoising process.
- **Toward scalable, unified generative modeling:** By decoupling the conditioning modality from the model architecture, our method paves the way for modality-adaptive transfer learning in video generation. It demonstrates a practical path for reusing powerful pre-trained generative models in novel tasks without architectural redesign or full fine-tuning.

1.4 Structure of the Thesis

This study primarily focuses on the structural innovation and optimization of neural network architectures within the diffusion model paradigm. The goal is to adapt and repurpose pre-trained text-to-video diffusion models for video super-resolution tasks through lightweight and modular conditional design.

Specifically, Section 2 provides a theoretical and literature-based foundation for our work. Section 2.1 presents a concise derivation of the denoising diffusion probabilistic model and introduces the architecture of the latent diffusion model (LDM), which serves as the backbone for most modern generative models. Section 2.2 reviews recent progress in applying diffusion models to image super-resolution, highlighting the differences between pixel-space and latent-space denoising strategies. Section 2.3 extends the discussion to video generation and video super-resolution, summarizing how temporal consistency, conditioning, and cross-modal adaptation have been addressed in prior works.

Section 3 introduces the proposed methodology. It details our adapter-based conditional modeling framework, which enables the reuse of existing diffusion backbones without retraining. The section covers the design of the visual-to-language adapter, the integration strategy with pre-trained U-Net architectures, and the novel conditioning pipeline built on ViT-based visual tokens.

Section 4 presents comprehensive experimental results and analysis. This section is structured into several key evaluations. We first provide training details and baseline settings. Then, we systematically compare the impact of various components, including post-color correction modules, initialization strategies for denoising, inference timestep configurations, and different choices of U-Net and VAE backbones. A long-term temporal consistency evaluation is conducted to assess video-level stability. We benchmark our approach against state-of-the-art VSR models across multiple datasets and perform ablation studies to analyze the influence of input frame number, adapter layer design, and conditioning strategies. Additionally, we present a detailed comparison of computational complexity, including inference time and parameter count, to highlight the efficiency of our method.

Finally, Section 5 summarizes the conclusions of this study, reflecting on the effectiveness of the proposed framework and its implications for future research in scalable, plug-and-play diffusion-based generative modeling for video restoration tasks.

1.5 Summary

Video super-resolution (VSR) technology based on diffusion models is emerging as a highly promising research direction, driven by the rapid advancements in deep learning and generative modeling techniques. Diffusion models, initially successful in image generation and restoration tasks, have demonstrated remarkable capability in modeling complex data distributions and generating high-fidelity visual content. Their application in the domain of video super-resolution opens up new possibilities for recovering fine-grained spatial details while simultaneously preserving temporal coherence across frames.

As deep learning algorithms continue to evolve, particularly in terms of network architectures, training stability, and data representation, the integration of diffusion models into video processing workflows has shown substantial potential for enhancing video quality. Unlike traditional VSR methods that often rely on deterministic upsampling or optical flow-based alignment, diffusion-based VSR models offer a

generative framework that can better handle diverse degradation types and complex motion patterns in video sequences.

However, despite these promising prospects, the field still faces several significant technical challenges. One key issue lies in controlling the inherent randomness introduced during the diffusion and reverse sampling processes, which can lead to temporal inconsistencies or flickering artifacts in the reconstructed video. Another critical challenge is maintaining the structural consistency and fine detail alignment between consecutive frames, especially in scenarios involving fast motion or occlusion. Achieving high detail fidelity while avoiding artifacts such as ghosting or unnatural textures remains an open problem.

Furthermore, controlling the temporal continuity and ensuring semantic coherence across generated frames are essential for producing realistic and visually pleasing videos. These challenges necessitate the development of more sophisticated model architectures, improved temporal attention mechanisms, and tailored loss functions that balance spatial detail restoration with temporal smoothness.

With ongoing breakthroughs in generative modeling and computational resources, it is expected that video super-resolution solutions based on diffusion models will become increasingly mature and efficient. Such advancements will not only improve user experiences in applications such as video streaming and mobile entertainment but also significantly enhance video analysis capabilities in fields like intelligent surveillance, remote conferencing, medical imaging, and film and television post-production. As the demand for high-definition and ultra-high-definition video content continues to rise, the development of robust, high-performance VSR systems will play a crucial role in shaping the future of digital media technology.

2 Literature Review

With the continuous progress in the field of deep learning, the diffusion model has become the mainstream model for tasks such as image generation and image super-resolution due to its excellent generation performance and powerful generation capabilities. Next, this article will introduce papers in related fields from the following three aspects: First, introduce the derivation process of the denoising diffusion model and the basic structure of the current mainstream diffusion model LDM (latent difffusion model). Second, introduce the relevant literature on the diffusion model in the field of image super-resolution. Third, introduce the relevant literature on the diffusion model in the field of video super-resolution and generation.

2.1 Diffusion Model

2.1.1 DDPM Model

The Denoising Diffusion Probabilistic Model (DDPM), first proposed by Jonathan Ho et al. [1] in 2020, marked a pivotal advancement in the field of generative modeling. Unlike earlier methods such as Generative Adversarial Networks (GANs) [2], which often suffer from unstable training dynamics, mode collapse, and incomplete data distribution coverage, DDPM offers a theoretically grounded and empirically stable alternative. Its design draws inspiration from non-equilibrium thermodynamics and probabilistic inference, framing the image generation process as a learned reversal of a predefined stochastic degradation process.

At its core, the DDPM framework is based on the idea of learning to reverse a Markovian diffusion process. This process consists of two main stages: the forward diffusion process and the reverse denoising process. As illustrated in Figure 2.1, the forward process incrementally corrupts an image $\mathbf{x}_0$ by adding Gaussian noise over T discrete time steps. With each step, the image becomes increasingly noisy until it approaches an isotropic Gaussian distribution. Mathematically, the forward process is modeled as a Gaussian Markov chain:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.1}$$

where β_t is a predefined variance schedule that controls the rate of noise addition at each step t.

The reverse process, in contrast, is learned. It aims to reverse the diffusion trajectory by progressively denoising a sample drawn from the prior noise distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The model learns to predict the mean of the reverse distribution at each timestep:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_t), \tag{2.2}$$

where μ_θ is parameterized by a neural network, typically a U-Net architecture, and Σ_t is either fixed or learned.

In practice, the training objective is derived by minimizing the variational bound on the negative log-likelihood of the data. This is often simplified into a mean squared error (MSE) loss between the true noise added in the forward process and the model's predicted noise. Formally, the training loss becomes:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right], \tag{2.3}$$

where ϵ_θ denotes the noise predicted by the denoising network at timestep t.

Once training is complete, the model can be used to generate images by sampling from a standard normal distribution and iteratively applying the reverse denoising steps. Each reverse step refines the noisy latent toward a coherent sample from the learned data distribution. Despite its simplicity, this framework produces high-quality, diverse samples and avoids many of the pitfalls associated with adversarial training methods.

DDPM's design reflects a physical analogy to molecular diffusion, wherein particles diffuse from high concentration to low concentration states over time. By simulating this process in the latent space of images, DDPM provides a stable, tractable, and interpretable route to image synthesis. Furthermore, it supports extensions such as conditional generation, classifier guidance, and improved sampling techniques (e.g., DDIM [3]), which expand its applicability across domains including super-resolution, inpainting, video synthesis, and more.

In conclusion, DDPM offers a principled and scalable approach to generative modeling. Its theoretical elegance, training stability, and strong empirical performance have laid the foundation for a wide range of subsequent advancements in diffusion-based image generation.

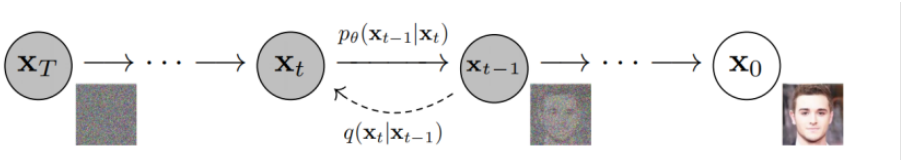


Figure 2.1: Denoising Diffusion Model Process

Figure 2.2 is the formula for training and verification. The training part on the left means that after adding noise step by step by inputting the original image X0 , the predicted noise composed of the noisy image and time steps calculated by the model is compared with the real noise in the denoising process to achieve the function of fitting the distribution of the noisy process. The right side is the verification part. Once each noise distribution is fitted, we can use the formula to subtract the noise distribution predicted by the neural network from the original noise image, so that the denoised image can be obtained.

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Figure 2.2: Noise addition and denoising formula

2.1.2 LDM Model

Since the introduction of the Denoising Diffusion Probabilistic Model (DDPM) in 2020, diffusion-based generative models have received widespread attention due to their impressive ability to synthesize high-fidelity images with diverse semantics and stable training dynamics. However, DDPM suffers from two significant limitations: (1) it operates directly in the high-dimensional pixel space, leading to extremely high computational cost during training and inference; (2) it lacks conditioning mechanisms, making it difficult to guide generation toward user-specified outputs.

To address these limitations, Rombach et al. [4] proposed the Latent Diffusion Model (LDM) in 2022, which improves upon the DDPM framework by introducing two critical modules: a Variational Autoencoder (VAE) and a Conditioning Network. As illustrated in Figure **??**, the LDM architecture enhances the original DDPM pipeline (depicted in green) by integrating the VAE module (red) and the condition module (white), enabling the model to operate more efficiently and flexibly on real-world image generation tasks.

The VAE module in LDM is composed of an encoder and a decoder. Its primary function is to project high-resolution images from the pixel space into a lower-dimensional latent space before the diffusion process begins. This step dramatically reduces computational complexity by allowing the denoising operations to be performed in the latent domain. In practice, the encoder downsamples the input image by a factor of 4 or 8, resulting in a latent space that is 16 or 64 times smaller in terms of spatial resolution. After the diffusion model performs the generative process in this compact space, the decoder upsamples the latent representation back to the original image resolution, yielding a high-fidelity reconstruction. This design effectively addresses the computational inefficiency of the original DDPM, making it feasible to process high-resolution images using moderate hardware.

In addition to the VAE, LDM introduces a conditioning module to incorporate external information (e.g., class labels, text descriptions, or reference images) into the generative process. Unlike the unconditional DDPM framework—where the same image is used for both input and output, making controlled generation impossible—LDM enables guided sampling. The conditioning network processes external inputs and reshapes them into feature embeddings compatible with the U-Net-based denoising network. These embeddings are then injected into the cross-attention blocks of the diffusion model, guiding the generation process toward producing content that aligns with the given conditions.

During training, the model learns to reconstruct images from noisy latent codes under the influence of conditioning inputs. This process ensures that the learned latent space remains semantically meaningful

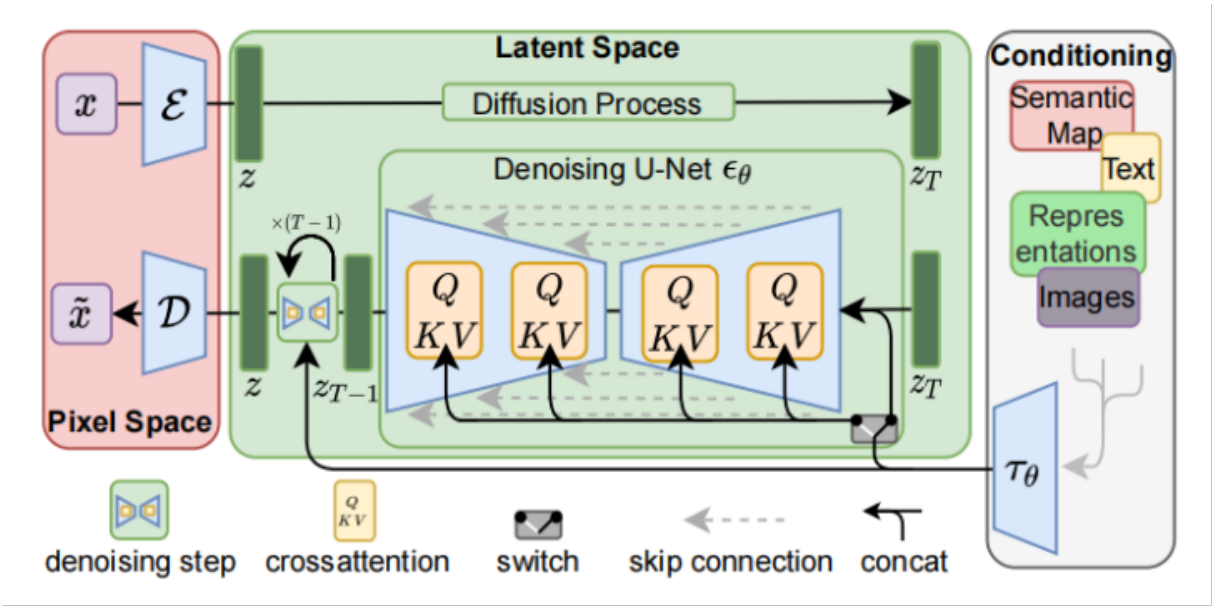


Figure 2.3: LDM model structure

and responsive to conditional guidance. During inference, pure Gaussian noise along with the desired condition (e.g., a prompt or reference image) is fed into the model. The denoising process then produces an image that not only resembles real data but also adheres to the specified condition.

The integration of these two components—VAE for latent compression and the conditioning module for controllability—makes LDM a powerful and efficient generative model. It addresses the inefficiency and lack of flexibility in standard DDPMs while maintaining high generative quality. As a result, LDM has become the de facto architecture for modern image synthesis, powering a wide range of tasks including text-to-image generation (e.g., Stable Diffusion), image inpainting, style transfer, and super-resolution.

In conclusion, the Latent Diffusion Model represents a significant step forward in diffusion-based generative modeling. Through architectural innovation and principled design, it combines the computational benefits of latent-space modeling with the controllability of conditional generation. This not only opens the door to high-resolution, high-quality image synthesis but also establishes a versatile framework for future research in multi-modal generation, cross-domain translation, and scalable generative learning.

2.1.3 DDIM Model

With the introduction of the Denoising Diffusion Probabilistic Model (DDPM), significant breakthroughs have been achieved in the field of generative modeling, particularly in high-quality image generation, image super-resolution, and semantic image editing. DDPM formulates the data generation process as a gradual denoising procedure that starts from pure Gaussian noise and progressively reconstructs a clean image through a sequence of learned denoising steps. Each step corresponds to a conditional probability distribution, grounded in a well-defined variational inference framework. This method provides a theoretically sound alternative to GANs, offering stable training and high sample diversity.

However, one of the major drawbacks of DDPM lies in its computational inefficiency. The sampling process typically requires hundreds or even thousands of iterative denoising steps to generate high-fidelity

outputs, as it is designed to reverse a finely discretized diffusion trajectory. At each step, a forward pass through a neural network is required, leading to substantial computational overhead during inference. This makes DDPM less suitable for real-time applications and motivates the search for faster alternatives.

To address this issue, Song et al. [3] proposed the Denoising Diffusion Implicit Model (DDIM), which extends the DDPM framework with several important innovations. First, DDIM reformulates the reverse diffusion process to be deterministic by removing the stochastic sampling of the variance term in the reverse step. Instead of drawing random noise at each timestep, DDIM uses the predicted noise directly to compute the denoised latent variable. This results in a deterministic sampling trajectory that preserves sample quality while eliminating the inherent variance of stochastic generation.

Second, DDIM relaxes the strict Markov assumption used in DDPM. While DDPM treats noise addition and removal as a Markov chain—requiring step-by-step conditional sampling—DDIM re-derives the process without this constraint. This decoupling from Markovian assumptions allows DDIM to define a non-Markovian path in the latent space, enabling high-quality generation in significantly fewer steps.

This deterministic reparameterization shares the same training objective as DDPM, yet defines a more efficient sampling path. The implications are profound for downstream tasks requiring high throughput and low latency, including real-time image synthesis, mobile inference, and video frame generation.

DDIM has also inspired further research into acceleration techniques for diffusion models. Its deterministic formulation supports adaptive-step samplers, hybrid stochastic-deterministic schedulers, and plug-and-play conditioning mechanisms. These advances expand the applicability of diffusion models to a broader set of real-world tasks where both sample quality and efficiency are paramount.

In summary, DDIM provides a principled and effective enhancement to the original DDPM framework. By eliminating stochasticity in the denoising process and relaxing the reliance on Markovian assumptions, it enables faster and more efficient generation without compromising quality. These improvements are especially valuable in compute-intensive scenarios such as image and video super-resolution, where balancing speed and fidelity is crucial.

$$X_{t-1} = \sqrt{\partial_{t-1}}\left(\frac{x_t - \sqrt{1-\partial_t}\varepsilon_\theta(x_t, t)}{\sqrt{\partial_t}}\right) + \sqrt{1-\partial_{t-1}}\varepsilon_\theta(x_t, t)$$

Figure 2.4: DDIM Format

2.2 Image Super-Resolution Task of Diffusion Model

Image super-resolution is a classic image restoration problem that involves reconstructing a high-resolution image from a low-resolution image [5–8]. With the development of digital image processing technology, the demand for high-resolution images is growing day by day, because they can provide more detailed information, thus bringing significant improvements in visual quality and application effects. Traditional image super-resolution methods, such as nearest neighbor interpolation, bilinear interpolation and bicubic interpolation, etc., although simple in calculation, often cannot effectively restore the detailed information of the image, especially in high-frequency areas. With the rise of deep learning technology, learning-based methods, especially methods based on convolutional neural networks, have made

breakthrough progress in the field of image super-resolution. These methods are able to learn complex mapping relationships from low-resolution to high-resolution images, thereby generating more visually satisfying results. However, despite the success of deep learning methods in image super-resolution, they usually require a large amount of high-resolution training data and may perform poorly in the face of unknown degradation models or complex scenes. In addition, these methods still face challenges in generating high-frequency details and maintaining image naturalness. To address these issues, we introduce diffusion models, a powerful generative model that generates data by simulating a gradual process from a data distribution to a noise distribution and then reversing the process. Diffusion models have shown their powerful capabilities in multiple fields such as image synthesis, image inpainting, and image super-resolution. The research on super-resolution images can be divided into the following research directions: the first one is to use the DDPM model to learn the mapping relationship from low-resolution images to high-resolution images; the second one is to use the LDM model as the main structure, and use the low-resolution images as the guide of the denoising process to guide the generation of high-resolution images; the third one is to combine the diffusion model and the GAN model to learn the interpolation between the generated image of the diffusion model network and the real image to achieve the process of generating high resolution. The first one uses the DDPM model to learn the direct mapping relationship between high and low. This idea is consistent with the idea of convolutional neural networks and GAN neural networks. During the training phase, the model receives low-resolution and high-resolution image pairs as input and output. In this way, the network is trained to predict the conversion from low-resolution images to high-resolution images. This process involves not only upsampling of the image, but also extraction and reconstruction of features to ensure the generation of high-resolution images .

In the field of deep learning, one of the early applications of convolutional neural networks (CNNs) is image processing tasks, especially in the field of image super-resolution [9–12] . The core goal of this task is to use deep learning models to convert low-resolution images into high-resolution images while retaining or enhancing the details and quality of the images. This type of neural network uses modules such as convolution and residual networks. During training, low-resolution image pairs are introduced as the input and output of the model, and the network learns the mapping relationship between the two.

GAN-related networks [13–16] aim to use the GAN network to input LR images and output a fake high-resolution image, and then compare it with the real image for discriminator discrimination. When the discriminator cannot distinguish between the real and the fake, it means that the model has completed training. In practical applications, network models such as ESRGAN, BSRGAN, and SRGAN optimize certain modules, modify the operation logic of the discriminator, and add complex degradation logic, so that the model can cope with different degradation environments when processing this task and enhance the ability to reconstruct and capture image details. The above advantages of GAN and convolutional neural networks are clear ideas and relatively simple network design. However, due to the simple design of neural networks, they cannot process large data, and the GAN network generation is unstable and the parameter adjustment is complex, resulting in unsatisfactory super-resolution effects. REFUSSION [17]is based on DDPM and uses SDE to derive the formula. The resulting network transforms the low-resolution image into a high-resolution one from the perspective of probability distribution, as shown in Figure 2.5. Compared with the GAN network, its generation process is closer to the denoising process of the diffusion model, and the final result is better than the other two. In summary, the advantage of learning

the mapping relationship from low-resolution images to high-resolution images is that the idea is clear and the network design is relatively simple. However, since a linear mapping relationship is established, it is inevitable that it is difficult to establish complex degradation problems. Since it is difficult to establish complex degradation relationships, most of these networks can only be used for already set data sets, and the results are often not satisfactory for the complex degradation causes in the real world.

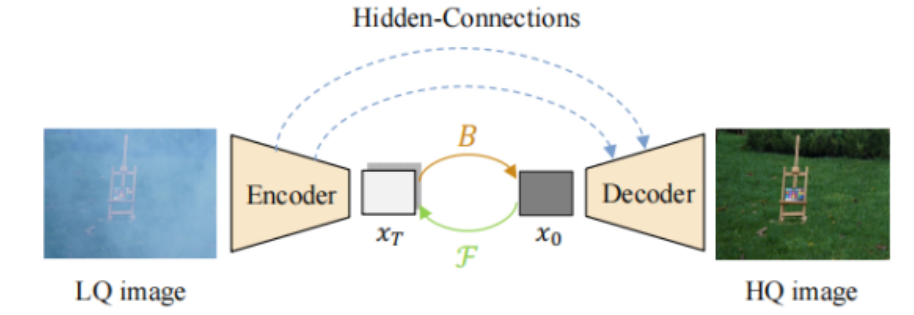


Figure 2.5: Refusion infrastructure

The second type uses the LDM model as the main structure to generate high-resolution images. This type of research focuses more on using low-resolution images as a guide condition and optimizing or replacing the denoising network to achieve better denoising generation. The papers based on the DDPM structure focus on improving the unet network of the denoising process, by replacing the number of network layers and the network itself to achieve better results.The SR3 [18]paper achieves super-resolution by using different numbers of layers for different depths and improving the residual block network. It is also the earliest paper to use the diffusion model for super-resolution tasks. The paper uses the high-resolution image and the noisy image as a whole after splicing them together, and adjusts some residual modules of the denoising network to better extract features.

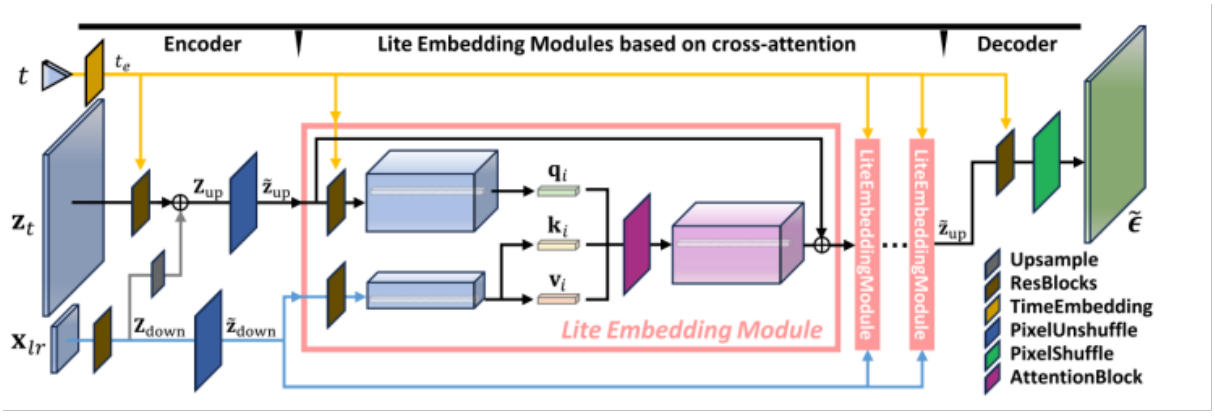


Figure 2.6: LWTDM basic structure

As shown in Figure 2.6, LWTDM [19] uses the attention structure to replace the structure of the unet network to implement the diffusion model. This article also provides guidance for using the diffusion model as a denoising architecture in the future. The subsequent DIT, VIT [20, 21] and other structures all use the attention structure to complete image-related tasks. Experiments have shown that when the data set is large and the amount of information is rich, the experimental results of attention are better than

those of the unet network, so many studies are trying to use attention for image-related tasks. The paper Denoising Diffusion Probabilistic Model for Retinal Image Generation and Segmentation [22] uses the DDPM model to complete retinal segmentation and super-resolution tasks . It learns the texture and retinal parts separately. After combining them, they can be better distinguished from each other and the edge texture of the image can be more obvious. This article also uses the cosine noise scheduler and re-reading training technology. The cosine noise scheduler controls the addition of noise through the cosine function, making the change of noise intensity smoother and more gradual, thereby affecting the training speed and generation effect of the model; the re-reading training technology (RTT) uses the same data to retrain the model. When denoising, the final result will be different due to the influence of noise, so the same input has different structures. Therefore, the same data can be used for training multiple times during training. In this way, the loss function of a training data will be smaller, and the model will be better able to maintain performance unchanged for changes and perturbations in the input data. By adding these two technologies, the paper successfully created a dataset related to retinal segmentation. DDNM [23] uses the zero-value domain decomposition method for mathematical derivation and concludes that for some processes from high-resolution images to low-resolution images due to position degradation, the value domain of the original image remains unchanged after degradation, and the zero domain disappears. Therefore, when we perform super-resolution tasks, we are essentially predicting the data of the disappeared zero domain. Therefore, based on the DDPM structure, this paper adds part of the original image to the predicted image at each step to generate a fused image, and uses this image as Xt-1 for the next prediction operation, as shown in the figure.

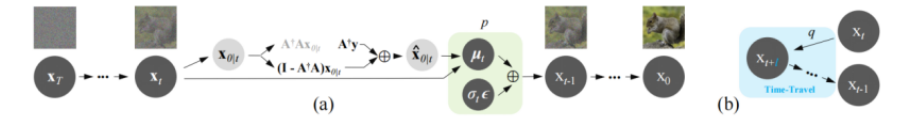


Figure 2.7: DDNM structure

The IDM [7] network solves the two problems of fixed magnification and artifacts. The architecture mainly fuses the LR and HR noise images, with the goal of better fusion and better integration of features into the image. As shown in the figure, the yellow part is used to extract features, and the blue part on the right adds position information. The final LR+HR noise image + position information is sent to the denoising network. First analyze the left side: X is the LR image, S is the scaling factor, which is used to select the scaling ratio. EDSR is a residual block network that establishes the initial features and then performs bicubic magnification. The final α is composed of three types of information: the first top branch converts the scaling factor S into a map and adds it to the image. The second branch U concats the LR and HR noise images, then downsamples and extracts features. Which one is more or less is determined by the ratio S. The final input to the downsampling is the LR magnification + LR magnification and HR noise fusion. Then, the coordinate information is added during the upsampling process to obtain the final result.

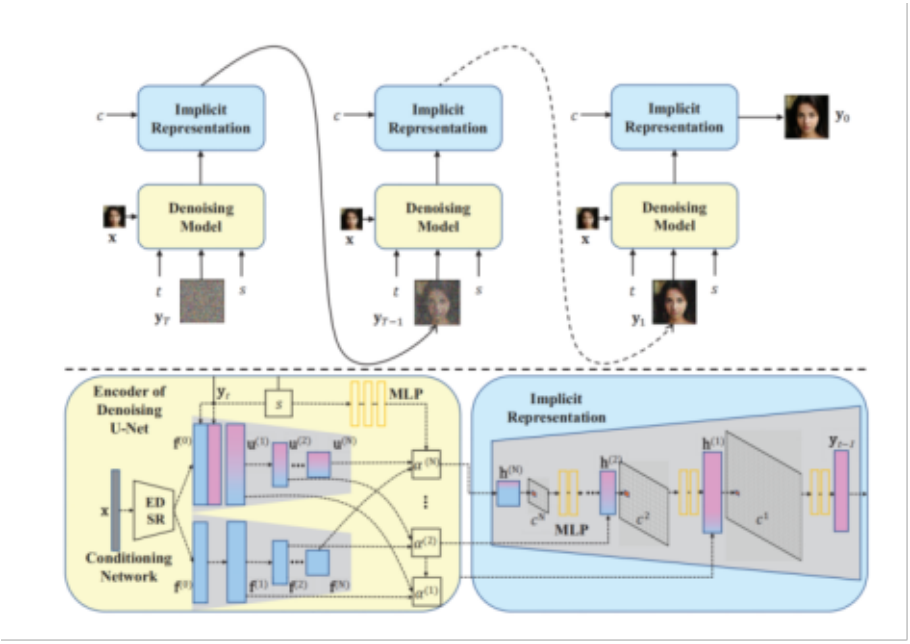


Figure 2.8: IDM structure

The LDDPM [24] network adopts the structure of the LDM network, but after the conditional guided input, a conditional feature extraction network is run in parallel to better fuse all the information of the two images. By fusing the information of the original denoising network and the feature extraction network, a better generated image and edge detail result is obtained. The pyramid-shaped PDDPM [25] network structure in the paper performs two times upsampling after each denoising, and after combining multiple upsampling, the upsampling size I need is obtained. In addition, this paper innovatively uses position encoding as input together with the image, that is, when inputting the low-resolution image, the extracted position information is encoded and the image is spliced as input, which contains more position information and the output image will not be deformed or distorted. This article fully demonstrates the importance of position encoding. More research focuses on the tasks of the LDM structure. Since the LDM structure itself does not learn the linear relationship of the image, the use of the LDM structure can represent more nonlinear relationships and the output is more flexible, which is why researchers favor this model. The biggest difference between the LDM architecture and other structures is that it uses low-resolution images as the conditions to guide the denoising process of the denoising model. Many papers and experiments modify the structure by modifying the conditional input structure, denoising network selection, high-resolution image input method, etc. to obtain better experimental results. SRDiff [26] did not use the attention method input when modifying the input structure, but used the interpolation of high-resolution and low-resolution images as input. The entire model predicts interpolation for noise prediction. This experiment shows that interpolation can extract more features, and further proves that the more conditional information is extracted and the more it is integrated into the denoising process, the better the generated image will be. The LDDPM article also proves the above conclusions well. In the main unet structure, the most important replacement method is to modify it into a unet structure. Among them, the most advanced structure is the SD3 model, which has achieved good output results on large data sets. The denoising network is designed to successfully predict noise and fuse LR images. Under the LDM framework, it can be roughly divided into the following categories: 1. Neural networks based

on unet, which are also the most common networks. They are usually modified by changing the residual blocks in unet, using different normalization layers, or combining attention with unet, and using attention at the bottom layer. 2. Neural networks with attention as the main structure. This structure is based on the VIT- DIT - SD3 [27] networks as the main idea. The reason for using attention is that the inductive bias of U-Net is not critical to the performance of the diffusion model. They can be easily designed according to the standard, and attention has been shown to have better training results when the amount of data is large. Finally, let's talk about the pre-training of HR images. In Latent Diffusion Models (LDMs), the encoder-decoder structure is used to compress image data into a low-dimensional latent space and then decode it back to the original data form. The encoder is usually a deep neural network, which is responsible for encoding high-dimensional image data into a low-dimensional latent space representation. This latent space representation captures the main features of the image and can process and generate images more efficiently due to its low dimensionality. The decoder is the inverse process of the encoder, which restores the representation of the latent space to the original image data. In LDM, the decoder is usually a network symmetrical to the encoder structure, which gradually restores the details of the image through a series of upsampling and convolution operations. The reason for adding encoders and decoders to LDM is to enable images to enter the latent domain for better calculation. There are relatively few studies in this area. The article that improves on this is REUSSION, which uses the jump connection of unet to make the input and output more matched.

Many articles have demonstrated that when LR is used as a conditional input, the more and richer the input features are, the better the effect and quality are. [13, 28, 29] In addition, during the denoising process, the higher the proportion of LR images, the more the image meets expectations, which is also what high-resolution tasks want, that is, to improve the resolution while retaining the original features (the parameter guidance scale represents the proportion of the condition. The higher the value, the more qualified images are generated; the lower the value, the greater the degree of freedom of the image obtained. In short, for super-resolution tasks, more conditions need to be considered) . Some methods to solve this problem include: multi-feature extraction with the goal of obtaining as high frequency information as possible (IDM, SD3); directly enlarging LR images and training with differences (SRdiff); direct and noise addition (LWTDM), and the synthetic caption method of DALL-E 3 [30] are all effective. DALL-E 3 is OpenAI's latest progress in the field of text-to-image generation. Its core innovation is to improve the model's prompt following ability by improving the image description. In previous studies, text-to-image models often have difficulties in processing detailed image descriptions, and often ignore or confuse the meaning of prompt words. This is mainly because the image descriptions in the training dataset are noisy and inaccurate. DALL-E 3 solves this problem by training a specialized image description generator and using it to regenerate the descriptions of the training dataset . The training process of DALL-E 3 consists of two stages: first, a small dataset that only describes the main body of the image is constructed for fine-tuning to generate short synthetic captions (SSC); then a large dataset of detailed description captions is created to generate descriptive synthetic captions (DSC). These captions not only describe the subject of the image, but also describe the environment, background, text, style, color and other details in the image . In the experiment, DALL-E 3 was compared with models such as DALL-E 2 and Stable Diffusion XL. The results showed that DALL-E 3 outperformed other models in terms of prompt following, style and consistency, and was favored by human evaluators. The article fully proves that using synthetic captions to get more semantic information for training will get better results. This point is also used in various

papers. Since most of the input conditional information is vague and lacking, the use of synthetic captions can obtain more refined conditional guidance generation. The results show that conditional captions may lead to unsatisfactory output results due to errors and insufficient descriptions. Therefore, richer descriptive descriptions are synthesized before input to make the conditional guidance more sufficient. This idea is also cited in the SD3 paper, which uses 50 persents of synthetic captions as conditions. I think it is possible to design a pre-trained network that allows LR images to combine position information and time information, expand some features, and enrich input features.

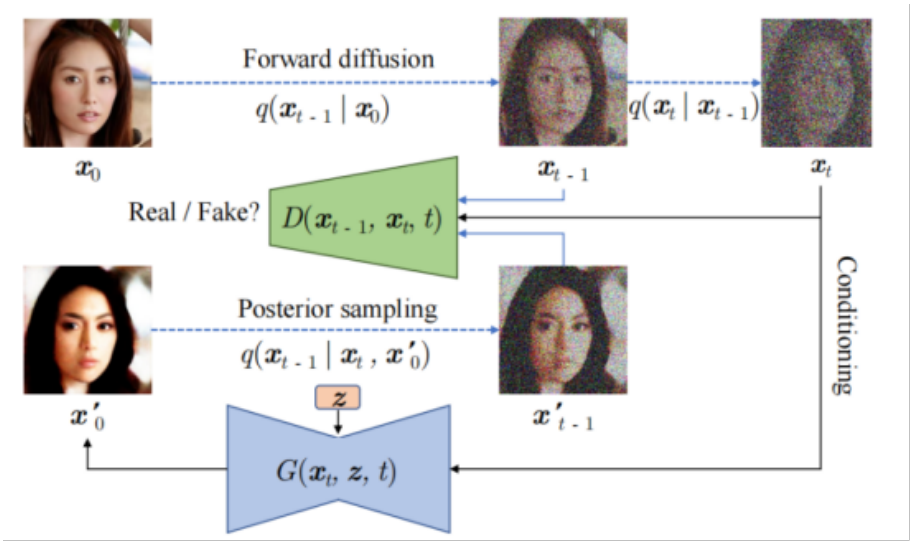


Figure 2.9: DDGAN network

The third idea is to use the DDGAN network for image generation [30–33] . This model was first proposed by Zhisheng Xiao in 2002. This network combines the GAN network and the diffusion model, which reduces the running time of the model based on the diffusion model. In traditional generative models, such as GAN, VAE, and Normalizing Flows, it is usually necessary to make a trade-off between the three requirements of quality, running speed, and generation diversity. For example, GAN can generate high-quality samples and has a fast sampling speed, but often lacks diversity. Although the denoising diffusion model can generate high-quality and diverse samples, the sampling process is slow and the computational cost is high. To address this problem, the authors of DDGAN [33] proposed a new model that uses conditional GAN to align the denoising distribution at each step, rather than assuming that the denoising distribution is a Gaussian distribution. This approach allows the use of larger step sizes for denoising, thereby reducing the number of sampling steps required and speeding up sampling. Experiments on the CIFAR-10 dataset show that this new approach is more competitive in terms of sample quality and diversity than the original diffusion model. In addition, the model demonstrates better mode coverage and sample diversity when compared to traditional GANs. To the best of the authors' knowledge, this is the first model that reduces the sampling cost of the diffusion model to a level that can be economically applied to practical applications. The paper also discusses the limitations of the Gaussian assumption of the denoising distribution under different conditions and proposes a multimodal distribution approach to better model the denoising process, especially when the marginal distribution of the data is not Gaussian. A key advantage of this approach is that it allows the model to be more stable during training and is able to generate more diverse samples. Overall, this paper proposes an innovative approach to solve the

trilemma in generative learning and experimentally verifies its effectiveness. This work provides valuable insights and directions for the future development of generative models. SRDDGAN, SRDIGAN [31, 32] are all modified through this related model to adapt to super-resolution tasks. Experimental results show that SRDDGAN is superior to existing diffusion model-based methods in terms of PSNR and perceptual quality indicators, while achieving It achieves a fast sampling speed, approximately 11 times that of the existing diffusion model SR3, making it more suitable for real-world applications. In addition, SRDDGAN also introduces a low-resolution (LR) encoder module to more effectively utilize the information in low-resolution images and use it as conditional input to improve the model's fidelity and detail recovery. In summary, SRDDGAN proposes a new single image super-resolution (SISR) method by combining the advantages of denoising diffusion models and GANs, which can significantly increase the sampling speed while maintaining image quality, and improve Generate sample diversity. SRDIG [34] is a paper that explores how to combine the denoising diffusion model to simultaneously handle the motion deblurring and super-resolution problems of images. This work targets the common motion blur problem in low-resolution (LR) images and attempts to enhance the image resolution while restoring its details. The proposed method achieves competitive sample quality and diversity compared to diffusion model-based image super-resolution models when handling dynamic blur and super-resolution tasks. In particular, compared to diffusion models such as SR3, the sampling process only requires four steps and is approximately 11 times faster. Compared with traditional GANs, the proposed model achieves significant improvements in training stability and sample diversity while maintaining competitiveness in sample fidelity. Work using the DDGAN network has done sufficient experiments on remote sensing images, medical images, etc., showing the extremely high sampling speed of this model. In summary, countless neural network-related studies have proven the effectiveness of using DDPM and LDM structures, so that when using neural networks, conditional constraints are used as guidance instead of directly establishing linear mappings to effectively face the complex degradation of the real world. . A series of image super-resolution task models are focusing on how to better integrate low-resolution information and location information. These studies show that better ways to extract information and better integrate this information into the denoising process will significantly improve the quality of the generation, that is, better prediction of the image content in the zero domain part. Therefore, we should make innovations in the extraction and fusion of condition information, so that we can better generate images.

2.3 Video Super-Resolution Task of Diffusion Model

Video Super Resolution (VSR) technology is a method of improving video resolution. It can enhance low-resolution video to high-resolution video, thereby providing more details and clarity. With the development of display technology, people have higher and higher requirements for video quality, and high-definition and even ultra-high-definition videos have become mainstream. However, due to limitations of shooting equipment, storage space, and transmission bandwidth, much video content still exists at lower resolutions. VSR technology effectively improves the viewing experience of these low-resolution videos through hardware or software methods. The development background of video super-resolution technology is closely related to image super-resolution technology. With the vigorous development of artificial intelligence, image super-resolution technology has achieved remarkable results. Video super-resolution technology is further developed on this basis. It not only needs to deal with the resolution improvement of

single-frame images, but also needs to consider the temporal continuity and motion compensation issues between frames in the video sequence. The application of diffusion models in the field of video super-resolution has become a research hotspot in recent years [35–43]. These models can establish mapping relationships between different resolutions by simulating the diffusion process of data, thereby achieving conversion from low-resolution images to high-resolution images. The ECDP [44] model aims to reduce time consumption and improve the efficiency of super-resolution image generation through the conditional diffusion process and probabilistic flow sampling technology. Experimental results show that ECDP significantly reduces model inference time while maintaining image quality. The DistriFusion [45] model aims at the computational cost of high-resolution image generation. DistriFusion takes advantage of the parallelism between multiple GPUs and divides the model input into multiple patches for processing, which reduces the computational cost and makes the model more efficient and faster. SinSR [46] proposed a single-step super-resolution generation method that accelerates the diffusion-based SR process by deriving a deterministic sampling process from state-of-the-art methods. This method not only improves the inference speed, but also reaches a level comparable to multi-step methods in performance. The ResShift [47] article improves the conversion efficiency by constructing a Markov chain to move the residuals in the conversion process of high-resolution images to low-resolution images. ResShift achieves performance comparable to or better than current state-of-the-art methods with only 15 sampling steps. Among the current studies on video diffusion models, the most important are Sora and Google's Image Video network, which have their own advantages. The core of Sora is a diffusion-based Transformer model, which achieves video generation through the following steps: Video compression network: compress the original video into a latent spatiotemporal representation. Spatiotemporal latent patches: extract a series of latent spatiotemporal patches from the compressed video. Its core is a diffusion Transformer model, which starts from a frame full of visual noise and gradually denoises it to generate a video. As an innovative technology, Sora demonstrates the great potential of AI in the field of video generation. It can not only simulate motion and interaction in the physical world, but also handle complex scenes in the digital world. Imagen Video [48] is a text-to-video generation model developed by Google, which converts text prompts into high-definition videos through a cascaded diffusion model. This technology can generate high-definition video clips up to 5.3 seconds long with a resolution of 1280×768 and a frame rate of 24fps. The core of Imagen Video is a cascaded diffusion model composed of multiple sub-models, including a text encoder, a basic video diffusion model, a spatial super-resolution model, and a temporal super-resolution model, with a total of 11.6 billion parameters. The generation process includes multiple steps. First, an initial video is generated based on the text prompt, and then the resolution and frame rate of the video are improved through a series of spatial and temporal super-resolution models. Although Imagen Video has made significant progress in experimental results, it still faces some challenges, including the use and creation of datasets. Future development directions may include improving the quality and duration of video generation, enhancing the generalization ability of the model, and exploring new application scenarios. Imagen Video demonstrates the potential of AI in understanding and generating complex video content, and research in this area still has broad prospects. The above two networks are structurally modified in the image super-resolution network LDM to meet the requirements of video tasks, which also shows that this task is closely related to the image task. In summary, video-related tasks are still an emerging field. Relevant research shows that we need to focus on several issues in video processing: First, we need to let the machine learn the temporal correlation from the same video; second, we

need the model to generate coherent, temporally connected images to form a video. The study improved some layers of the model to 3D layers, making the model perceptual; and introduced optical flow to assist the model in understanding the changes in action postures brought about by time. In general, the diffusion model is gaining momentum in the field of super-resolution. It has not only become a hot topic in current research, but also has great potential in model optimization and innovation. With the continuous advancement of technology, the diffusion model has achieved remarkable results in quality improvement and detail enhancement. Scholars are working to further improve the diffusion model and explore new algorithms and strategies in order to achieve more accurate and efficient processing results in the field of image super-resolution. The research prospects in this field are broad, indicating that there will be more breakthroughs in image processing technology in the future.

3 Research Methodology

3.1 Overview

To reduce the complexity of both model design and training while fully leveraging the expressive capabilities of large-scale pre-trained diffusion models, we adopt a text-to-video latent diffusion model as the backbone of our video super-resolution (VSR) framework. This approach allows us to reuse the powerful generative priors of diffusion models trained on large-scale datasets, without retraining the backbone network. Instead of modifying the original architecture or introducing complex task-specific modules, we design a lightweight conditional adapter that connects the task-specific input to the model's native cross-attention interface in a plug-and-play manner.

As shown in Figure 3.1, the framework operates over a sequence of noisy latent variables $\mathbf{z}_t$ across t diffusion steps. At each step, the UNet performs noise estimation guided by external conditional signals. Unlike conventional diffusion models that rely on textual embeddings, our framework uses low-resolution video frames $\mathbf{x}_{\text{LR}}$ as the only condition source. Each frame is independently encoded using a frozen Vision Transformer (ViT-base-patch16-224), which extracts patch-level semantic features and maintains consistent representations across frames.

These visual features are then passed through a learnable adapter module that transforms them into pseudo-text tokens, structurally aligned with the language embeddings used during original pre-training. The resulting token sequences are injected into the cross-attention blocks of the frozen UNet as encoder hidden states, enabling the model to interpret visual input as if it were textual conditioning. This approach maintains full compatibility with the original model architecture while introducing only minimal additional parameters.

To validate the generalization ability of the proposed conditional framework, we compare two representative UNet backbones:

- **UNet (upscale)** [49]: A task-specific architecture integrating a pre-trained image-level Latent Diffusion Model (LDM) with a self-trained 3D convolutional block, optimized for VSR.
- **UNet (ali)** [50]: A general-purpose diffusion backbone from the Ali team, with no VSR-specific architectural enhancements, used to evaluate the transferability of the adapter module.

We also adopt the pre-trained Variational Autoencoder (VAE) from the Upscale-A-Video framework as the latent encoder-decoder. This VAE employs a 3D architecture optimized for temporal information and incorporates a Spatial Feature Transform (SFT) layer to inject structural information from the low-resolution input during decoding, improving spatial fidelity and reconstruction accuracy.

During training, we freeze both the UNet and VAE, and optimize only the adapter. This ensures that the original generative capabilities of the backbone remain intact. The training follows the standard diffusion paradigm by minimizing the noise prediction loss across denoising steps. Through this setup, the adapter learns to condition the model effectively, aligning visual semantics with the learned language-based representation space, and enabling high-quality video super-resolution with minimal computational overhead.

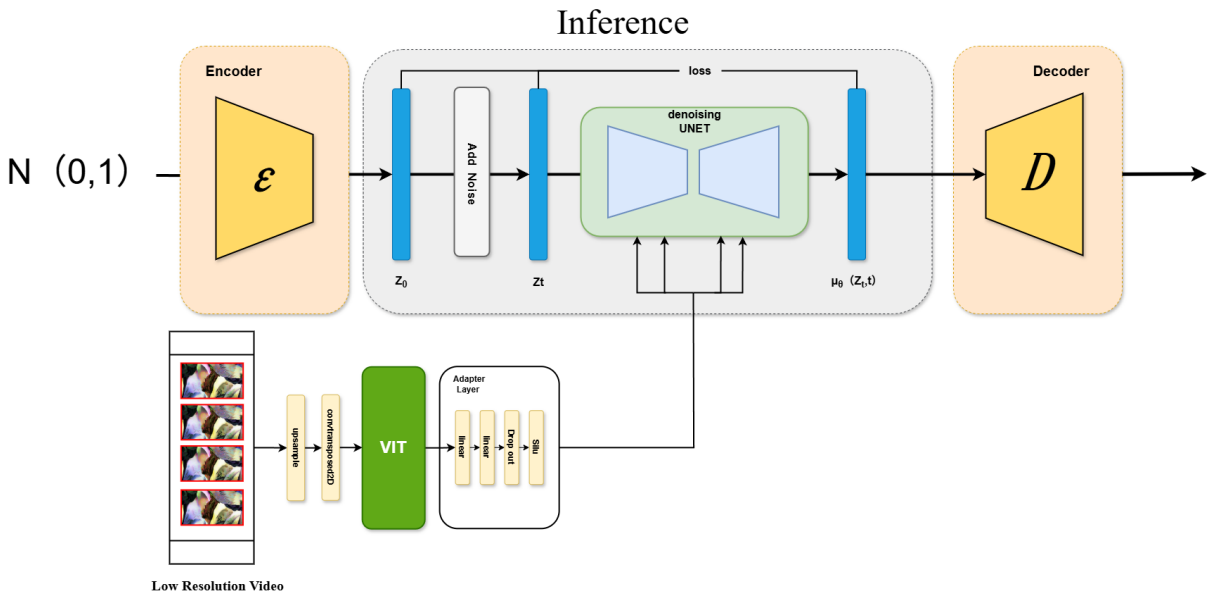


Figure 3.1: Overall model structure

3.2 Adapter-Based Conditioning Design

To enable the effective adaptation of a pre-trained text-to-video latent diffusion model for the video super-resolution (VSR) task, we propose a conditional modeling framework centered around an adapter-driven design. The core motivation behind this approach is to repurpose the powerful generative capabilities of large-scale diffusion models—originally trained with natural language conditioning—by introducing an intermediate representation that reformats visual input into a language-compatible embedding space. This allows the model to be applied to image-to-video or video-to-video tasks in a zero-shot or few-shot setting without modifying its original architecture or retraining its backbone components.

As illustrated in Figure 3.2, the adaptation pipeline begins with the encoding of each frame in the low-resolution input video using a frozen Vision Transformer (ViT-base-patch16-224). This module operates in a frame-wise manner, transforming each 80×80 frame into a sequence of patch-level feature tokens. These features preserve high-level semantic content and structural information, ensuring consistent representation across temporal dimensions while avoiding entanglement with low-level pixel noise. The use of a frozen ViT also guarantees that the visual features remain stable and comparable across frames and across videos, contributing to temporal coherence.

To reconcile the discrepancy between visual embeddings and the language-based conditioning interface expected by the pre-trained diffusion model, we introduce a lightweight, learnable adapter module positioned after the ViT encoder. This adapter is responsible for transforming the ViT-generated visual token sequences into pseudo-text tokens—feature embeddings that simulate the statistical and positional patterns of textual representations learned during the original model's training. Through this transformation, the adapter effectively "masks" the visual origin of the data, allowing the cross-attention layers in the UNet backbone to interpret them as if they were language tokens.

Importantly, this transformation is entirely non-invasive to the original architecture: neither the ViT encoder nor the UNet decoder nor the VAE are structurally modified. The adapter operates as a modality

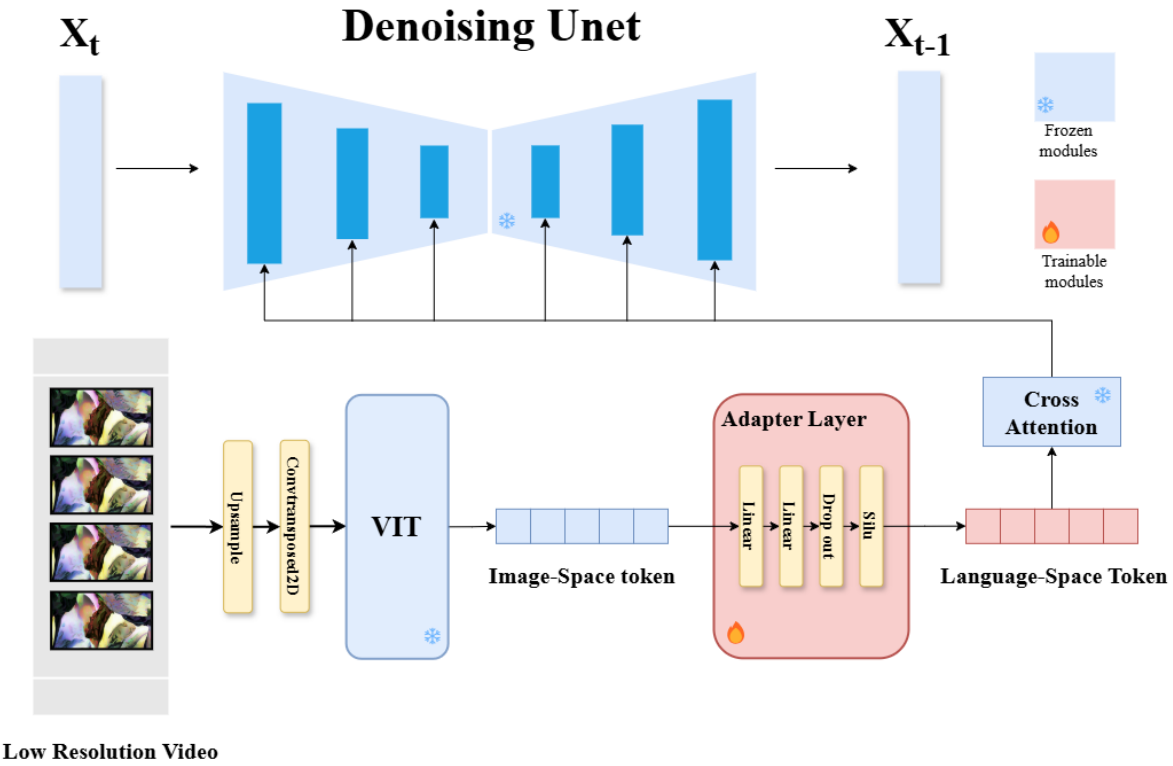


Figure 3.2: Conditional model architecture.Frame-wise image tokens extracted by ViT are fused by the adapter and mapped into the language space, serving as encoder hidden states for U-Net cross-attention.

translator, bridging vision and language at the representation level rather than through structural integration. The resulting pseudo-text tokens are injected directly into the cross-attention blocks of the frozen UNet at each denoising step of the diffusion process. In doing so, they provide effective guidance for frame reconstruction and upsampling, conditioned on the semantic content of the input video, rather than natural language prompts.

This plug-and-play conditioning mechanism enables the diffusion model to function as a generic conditional generator, extending its application from text-to-video generation to visual super-resolution without extensive retraining. It also provides an elegant decoupling between modality and architecture, which is a key factor in enabling cross-task generalization. Because the conditioning pathway is separated from the core generative pathway, different types of tasks (e.g., video inpainting, frame interpolation, stylization) can be supported simply by modifying or re-training the adapter, without affecting the backbone model.

Overall, our approach demonstrates how the conditioning modality can be flexibly reinterpreted without architectural redesign, unlocking the versatility of pre-trained generative models. The adapter not only enables compatibility but also enhances the scalability, modularity, and efficiency of video super-resolution systems built upon powerful diffusion backbones.

Our training pipeline enables pseudo-language conditioning by optimizing an adapter module that transforms visual inputs into language-aligned representations. Specifically, we simulate the forward diffusion

process by adding Gaussian noise to high-resolution video frames $\mathbf{x}_{\text{HR}}$ at timestep t:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_{\text{HR}} + \sqrt{1 - \bar{\alpha}_t} \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{3.1}$$

Meanwhile, the low-resolution input $\mathbf{x}_{\text{LR}}$ is processed by a pre-trained and frozen Vision Transformer and Adapter to obtain a pseudo-text condition embedding:

$$\tau_\theta(\mathbf{x}_{\text{LR}}) = \text{Adapter}\left(\text{ViT}(\mathbf{x}_{\text{LR}})\right). \tag{3.2}$$

The denoising U-Net then predicts the added noise using the noisy latent and the pseudo-text condition:

$$\hat{\boldsymbol{\epsilon}} = \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{x}_{\text{LR}})). \tag{3.3}$$

The training objective is defined as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathbf{x}_{\text{HR}}, \mathbf{x}_{\text{LR}}, \boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I}), t} \left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{x}_{\text{LR}}))\|^2 \right]. \tag{3.4}$$

During inference, the model starts from Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ and progressively denoises through the learned diffusion steps conditioned on $\tau_\theta(\mathbf{x}_{\text{LR}})$ to generate the high-resolution output.

3.3 Evaluation Metrics

To comprehensively evaluate the performance of the proposed method in the video super-resolution (VSR) task, we adopt four widely recognized quantitative evaluation metrics. These metrics are carefully chosen to cover multiple aspects of video quality, including pixel-level fidelity, structural preservation, perceptual realism, and temporal consistency. By leveraging this multi-faceted evaluation scheme, we aim to provide a thorough and objective assessment of the effectiveness of our approach.

PSNR (Peak Signal-to-Noise Ratio): PSNR is a traditional yet important metric for evaluating image and video restoration quality. It measures the average logarithmic ratio between the maximum possible pixel value and the mean squared error (MSE) between the generated frame and its corresponding high-resolution ground-truth frame. Higher PSNR values indicate lower distortion and better pixel-wise accuracy. While PSNR is sensitive to subtle pixel-level changes, it does not always correlate well with human perception, especially for perceptual or semantic distortions.

SSIM (Structural Similarity Index): SSIM is a perceptual metric that evaluates the similarity between two images based on structural information rather than absolute pixel differences. It compares local patterns of pixel intensities after normalizing for luminance, contrast, and structure, thereby providing a more perceptually aligned assessment of visual quality. SSIM is particularly effective at capturing local distortions and structural degradations, making it a valuable complement to PSNR in assessing the quality of reconstructed frames.

LPIPS (Learned Perceptual Image Patch Similarity): LPIPS is a deep learning-based perceptual similarity metric that compares image patches using feature activations from pre-trained convolutional neural networks, such as AlexNet or VGG. Unlike PSNR and SSIM, which rely on pixel-level statistics, LPIPS

captures high-level semantic differences that are more consistent with human visual perception. Lower LPIPS values indicate greater perceptual similarity to the reference images, making it a crucial metric for evaluating the realism and naturalness of generated content.

Ewarp (Temporal Consistency Metric): Temporal consistency is a critical aspect of video generation tasks. Ewarp evaluates the stability and coherence of motion and structure across consecutive video frames. It computes a temporal warping error by extracting frame-wise features using a VGG16 backbone and estimating how well temporal correspondences align between the generated video and the reference sequence. Lower Ewarp scores suggest better frame-to-frame consistency, reduced flickering artifacts, and more temporally stable outputs.

Together, these four metrics offer a comprehensive evaluation framework: PSNR and SSIM assess the spatial reconstruction accuracy and structural fidelity of individual frames; LPIPS measures the perceptual alignment with human visual judgment; and Ewarp evaluates the temporal dynamics and inter-frame consistency of generated video sequences. By jointly analyzing these metrics, we can obtain a holistic understanding of our method's performance in enhancing both the spatial and temporal aspects of video quality, thereby validating its robustness and practical effectiveness in real-world video super-resolution scenarios.

3.4 Summary

In this work, we propose a simple yet effective conditioning framework tailored for the video super-resolution (VSR) task, built upon a powerful pre-trained text-to-video latent diffusion model. Rather than redesigning the entire network architecture or introducing substantial structural modifications, we focus on maximizing the utility of the existing backbone through a minimal yet expressive conditional mechanism. Our core idea is to seamlessly adapt visual inputs — specifically low-resolution multi-frame sequences — to the conditioning interface of the latent diffusion model, originally designed for text-based guidance.

To achieve this, we introduce a novel adapter-based conditioning module that efficiently encodes spatiotemporal information across video frames. In our framework, each frame in the input video sequence is first processed individually using a frozen Vision Transformer (ViT), which extracts rich semantic representations from each frame without any fine-tuning. These per-frame features are then aggregated and fused through a lightweight adapter network designed to capture both spatial and temporal dependencies, while maintaining computational efficiency. The output of the adapter is a set of transformed embeddings that align with the language embedding space expected by the pre-trained diffusion model.

These pseudo-language embeddings are then injected into the UNet-based denoising network through its existing cross-attention layers. Importantly, this design does not require any modification to the architecture of the backbone components — namely, the UNet and the Variational Autoencoder (VAE) used in the latent diffusion model. As a result, our framework preserves the integrity and stability of the original model while enabling it to perform video super-resolution tasks guided by visual context, rather than textual input.

The key strength of our approach lies in its modularity, parameter-efficiency, and simplicity. By leveraging fixed, high-quality pre-trained components and introducing only a lightweight adapter for task-

specific conditioning, our framework minimizes the overhead typically associated with fine-tuning or structural expansion. It also avoids the need for elaborate loss function design or handcrafted supervision signals. Despite its simplicity, our method demonstrates strong performance in generating high-resolution video sequences with both perceptual quality and temporal consistency.

This design paradigm not only enhances the practicality and scalability of diffusion-based VSR but also highlights the potential of repurposing large-scale generative models across diverse vision tasks through adaptable and interpretable interfaces.

4 Experimental Results and Discussion

4.1 Experimental Results

4.1.1 Training Details

Considering the computational limitations associated with high-resolution video data and the resource-intensive nature of diffusion model inference, we adopt a center cropping strategy to preprocess the raw video inputs. This cropping ensures uniform spatial dimensions across all samples while retaining the most informative regions of each frame. Specifically, the input low-resolution (LR) video clips are resized to dimensions of $80 \times 80 \times 8$, where 80×80 denotes the spatial resolution and 8 indicates the number of frames in each temporal clip. Correspondingly, the high-resolution (HR) targets are cropped and resized to $320 \times 320 \times 8$, establishing a consistent $4\times$ spatial upscaling ratio.

To train our model, we utilize paired LR-HR video sequences from a high-quality video super-resolution dataset. The objective is to learn a robust mapping from the degraded LR space to the HR domain while preserving both spatial detail and temporal consistency. In alignment with our architectural philosophy of minimal intervention, we freeze the entire UNet backbone and train only the conditional input module — comprising the frozen Vision Transformer (ViT) for feature extraction and a lightweight adapter network for conditional embedding generation.

The training follows the standard inference pipeline of latent diffusion models. Each training iteration simulates the reverse diffusion process by performing 30 denoising steps, where at each step, the UNet predicts the noise component to be removed. The conditional module provides guidance through the injection of visual conditioning information into the cross-attention modules of the UNet, enabling the model to adapt the generic generative process to the specific VSR task context.

Optimization is performed using the Adam optimizer with an initial learning rate of 1×10^{-7}. A batch size of 1 is used due to GPU memory constraints, especially considering the multi-frame nature of the input. To enhance convergence efficiency and improve training stability, we incorporate a cosine learning rate scheduler, which gradually reduces the learning rate in a smooth and non-linear fashion. Additionally, we adopt an early stopping mechanism to prevent overfitting and unnecessary computation. The early stopping criterion is based on validation loss, with a patience of 2 epochs. If the validation loss fails to improve significantly over two consecutive evaluation periods, the training process is automatically terminated.

As illustrated in Figure 4.1 and Figure 4.2, both the training and validation loss curves exhibit a steady downward trend throughout the training process. This consistent decrease indicates that the model is effectively learning the underlying spatial and temporal correlations from the training data. Furthermore, the close alignment between training and validation losses suggests good generalization capability and the absence of severe overfitting. These observations collectively demonstrate that the proposed conditional module, despite its lightweight design, is capable of effectively guiding the diffusion process to reconstruct high-quality video sequences under the constraints of limited training and computational budgets.

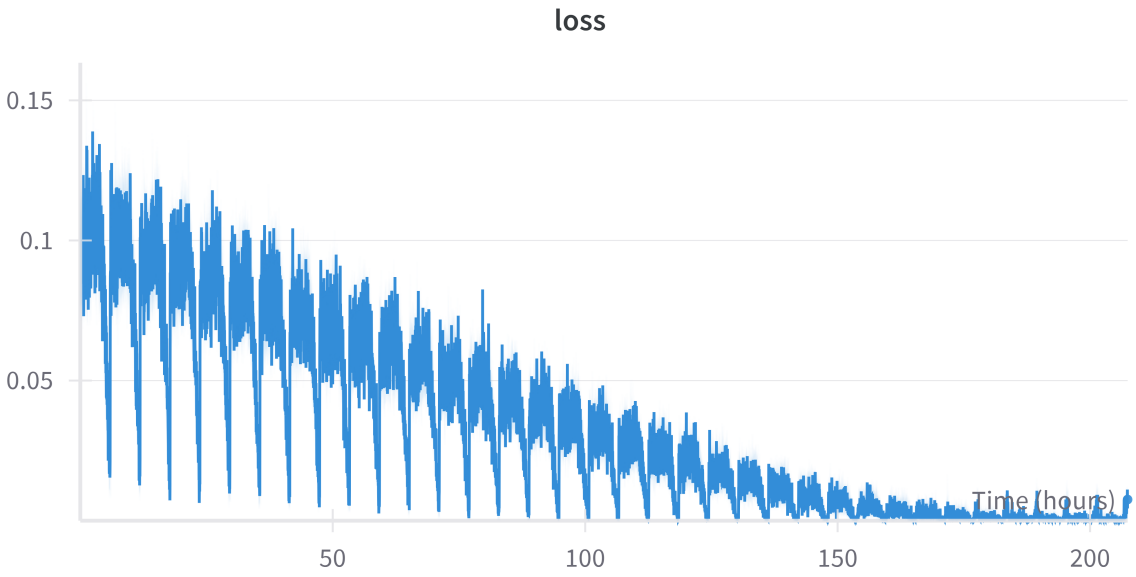


Figure 4.1: loss

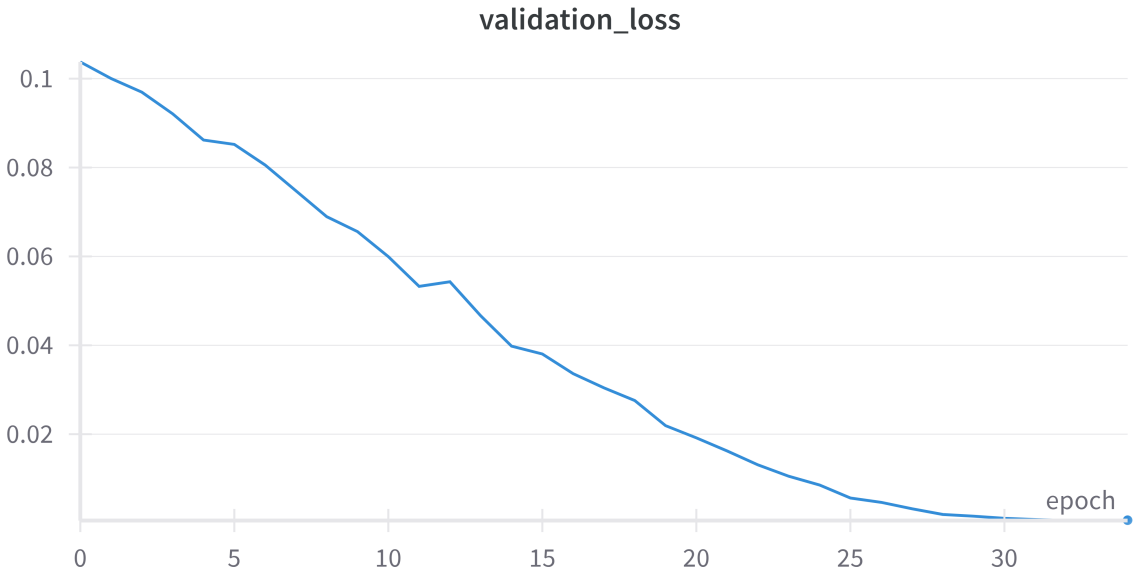


Figure 4.2: validation loss

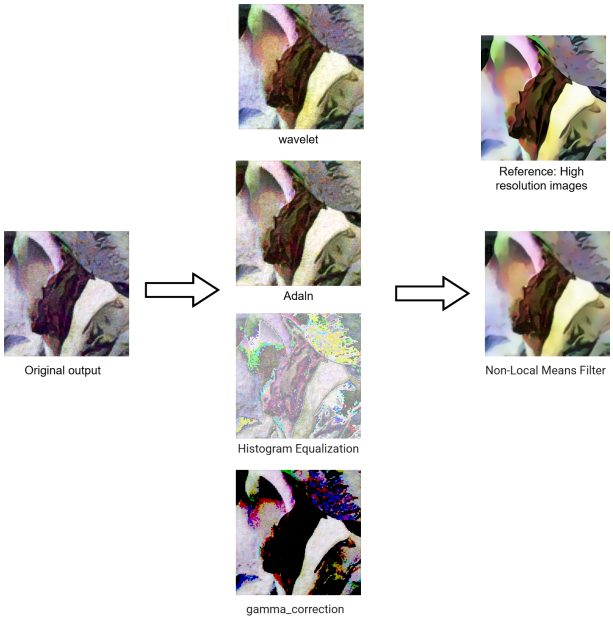


Figure 4.3: Post-Processing Strategies for Enhancing Raw VAE Output

4.1.2 Comparison With Different Post-color correction methods

The initial raw outputs generated by the pre-trained Variational Autoencoder (VAE), as used in our diffusion-based video super-resolution (VSR) framework, frequently exhibit suboptimal visual quality. In particular, we observe two consistent issues: (1) inconsistent color representation across frames, leading to temporal flickering and unrealistic tone variations; and (2) noticeable noise artifacts, especially in low-texture or dark regions, which degrade the perceptual clarity of reconstructed videos. These artifacts undermine the visual fidelity of the final output and limit the practical applicability of the system, especially in scenarios requiring high perceptual quality, such as media production or broadcast enhancement.

To address these issues, we initially investigated a model-level solution by introducing a trainable residual layer appended to the end of the pre-trained VAE decoder. The goal was to allow a limited degree of fine-tuning, correcting residual errors in generated frames without retraining the entire VAE. However, empirical results demonstrated that this approach was largely ineffective. Most layers of the pre-trained VAE had already converged to near-optimal weights during large-scale pretraining. Consequently, gradients from the residual layer had negligible influence on the output quality. Moreover, the training loss curve during fine-tuning failed to show meaningful or consistent downward trends, indicating poor optimization dynamics and limited capacity for improvement through shallow architectural modification.

Given these limitations, and considering the substantial computational cost of retraining or partially reinitializing the VAE backbone, we shifted our focus to post-processing techniques as a practical and low-cost alternative for improving output quality. This decision was further motivated by the fact that post-processing methods are modular, task-agnostic, and can be applied during inference without affecting model structure or training time.

As shown in Figure 4.3, we systematically evaluated a set of post-processing strategies targeting two main aspects: (1) correction of inter-frame color inconsistency, and (2) reduction of high-frequency noise. We explored both traditional image enhancement techniques and modern denoising filters. In particular, we

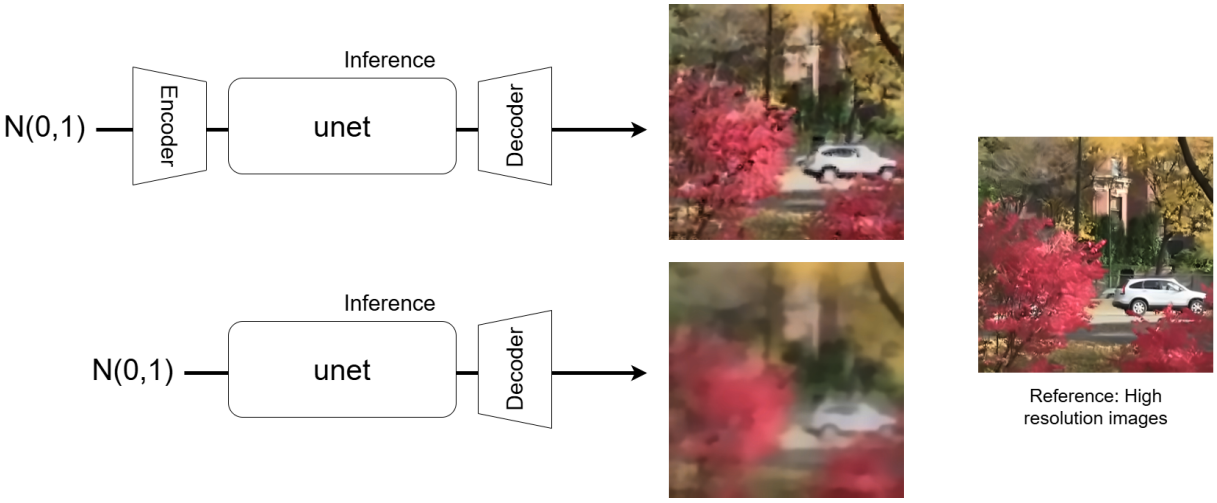


Figure 4.4: Effect of Denoising Initialization Strategies on Video Generation Quality

tested adaptive histogram equalization, color transfer based on frame statistics, and a range of denoising approaches including bilateral filtering, Gaussian blur, non-local means (NLM), and wavelet-based smoothing.

All tests were conducted under identical conditions to ensure fair comparison: the same video dataset, a fixed number of diffusion steps ($t = 30$), and identical pre-trained UNet and VAE components. This controlled setting allowed us to isolate the impact of each post-processing technique.

Among all evaluated configurations, we found that a hybrid strategy combining wavelet-based color correction and non-local means filtering yielded the best results. This combination effectively suppresses noise while preserving fine details and edge sharpness. Moreover, it significantly improves visual coherence between consecutive frames by stabilizing chromatic distributions over time. Qualitative results indicate a clear enhancement in realism and perceptual smoothness, especially in challenging regions such as skin tones, sky gradients, and motion boundaries.

In summary, while simple fine-tuning strategies failed to resolve quality issues arising from pre-trained VAE outputs, post-processing techniques provide a lightweight yet effective alternative. Without modifying or retraining the backbone, such methods can substantially improve the perceptual quality of diffusion-based VSR outputs, offering practical value in real-world deployment scenarios.

4.1.3 Comparison With Different Denoising Initialization Strategies

To investigate the influence of denoising initialization strategies on the quality of generated videos, we conducted a comparative experiment using two different approaches to define the starting point of the denoising process within the diffusion pipeline.

See Figure 4.4,In the first approach, we passed pure Gaussian noise through the VAE encoder to obtain an encoded latent representation, which was then used as the initial noisy input for the U-Net-based denoising process. In the second approach—the standard setting—we directly sampled random Gaussian noise as the input for U-Net without encoding.

The visual outcomes from these two strategies were notably different. Using the encoder-processed noise

Figure 4.5: Effect of Diffusion Steps on Visual Quality and Efficiency

as the starting point led to clearer and sharper reconstructions, preserving more fine-grained details in the generated frames. However, this approach also introduced more visible noise artifacts, especially in homogeneous regions of the image. On the other hand, the standard approach, which initializes denoising directly from unstructured noise, produced videos that were overall smoother and more visually coherent, but at the cost of reduced spatial sharpness and more blurring of fine details.

These results highlight a trade-off between perceptual clarity and visual smoothness depending on the choice of initialization. While encoder-based initialization enhances structural fidelity, it also amplifies noise, whereas random noise initialization results in more stable and artifact-free outputs at the cost of detail degradation. This observation provides useful insight into the role of initialization in diffusion-based video generation, and suggests potential for hybrid or adaptive strategies depending on specific task requirements.

4.1.4 Comparison With Different Inference Timestep

As shown in Figure 4.5, we conduct a series of controlled experiments to investigate how the number of diffusion steps affects both the visual quality of generated frames and the corresponding inference time. Specifically, we evaluate multiple settings by varying the number of denoising steps t while keeping all other model parameters fixed, including the dataset, the pre-trained VAE and U-Net backbones, and the inference resolution.

From the visual comparisons, we observe that increasing the diffusion steps leads to a noticeable improvement in reconstruction sharpness and the preservation of fine image structures. In particular, higher values of t enable the model to better refine high-frequency details and suppress artifacts such as blurring or texture inconsistency, which are often present in lower-step outputs. This improvement, however, is accompanied by a significant increase in inference time, as each step requires a full forward pass through the denoising network, and the diffusion process is inherently iterative.

To quantitatively analyze this trade-off, we measure the average per-frame inference time under different step settings (e.g., $t = 10$, $t = 30$, $t = 100$, $t = 200$). Our findings reveal that while configurations with $t = 100$ or $t = 200$ indeed achieve marginally better visual fidelity, the perceptual improvement becomes increasingly less significant beyond a certain threshold (typically after $t = 30$). Meanwhile, the computational cost grows linearly with the number of steps, making large-step configurations impractical for real-time or resource-constrained applications.

Taking both visual quality and efficiency into account, we select $t = 30$ as the default configuration for our final model. This choice provides a favorable balance: the generated outputs are visually coherent

Figure 4.6: Qualitative comparison of visual results generated using different combinations of VAE and U-Net architectures. The figure demonstrates the impact of architectural choices on reconstruction clarity, detail preservation, and artifact suppression.

and rich in detail, while the inference time remains within an acceptable range for practical deployment. Our analysis underscores the importance of step count tuning in diffusion-based video generation and highlights $t = 30$ as a sweet spot that combines stability, clarity, and performance.

4.1.5 Comparison With Different Unet And VAE

To gain deeper insight into the role of individual components within the backbone architecture for video super-resolution (VSR), we conducted a comprehensive set of comparative experiments across different combinations of U-Net and VAE designs. All experiments were conducted under consistent conditions, with the number of diffusion inference steps fixed to 30. This setup ensures a fair evaluation by isolating architectural effects from other variables such as sampling depth or optimization parameters.

We examined two distinct U-Net architectures:

- **Upscale-U-Net**: A specialized U-Net variant developed for video super-resolution tasks, incorporating multi-scale upsampling and temporal feature aggregation mechanisms.
- **Ali-U-Net**: A general-purpose U-Net architecture derived from a pre-trained text-to-video diffusion model developed by the Ali team. It was originally not tailored for super-resolution, making it a suitable candidate for evaluating the generalizability of our conditioning framework.

In parallel, we assessed two different VAE configurations:

- **Standard 3D VAE**: A baseline latent encoder-decoder that reconstructs video frames directly from latent variables using a 3D convolutional structure.
- **SFT-guided VAE**: An enhanced variant that incorporates low-resolution inputs through a Spatial Feature Transform (SFT) module, allowing additional structural guidance to be injected during the decoding process.

As illustrated in Figure 4.6, all combinations of U-Net and VAE were evaluated under the same conditional setup, where the visual input was processed using a frozen ViT and a lightweight adapter module. This ensured that variations in performance could be attributed solely to differences in the U-Net and VAE architectures.

Experimental results reveal that the SFT-guided VAE consistently outperforms the standard VAE across all U-Net settings. It produces sharper image details, smoother and more continuous textures, fewer

visual artifacts, and superior temporal coherence. These findings validate the effectiveness of integrating structural priors via SFT and highlight the SFT-VAE's role in improving reconstruction quality, especially in low-frequency and global structure restoration.

Interestingly, we observe that when paired with our unified conditional model, the performance disparity between different U-Net architectures becomes marginal. Despite its lack of VSR-specific design, Ali-U-Net achieves perceptual quality and structural fidelity comparable to the more specialized Upscale-U-Net. This outcome underscores the strength and versatility of our adapter-based conditional model, which effectively enables general-purpose backbones to perform well on highly structured visual tasks like super-resolution.

Overall, these experiments highlight a complementary relationship between the VAE and the conditional adapter. The VAE—particularly the SFT-augmented variant—plays a crucial role in reconstructing global structure and maintaining low-frequency consistency across frames. Meanwhile, the conditional model excels at enhancing high-frequency details such as fine textures, edge sharpness, and semantic alignment through cross-modal guidance. Although the quantitative improvements from the adapter are relatively modest, its consistent cross-architecture compatibility and plug-and-play design confirm its value as a generic modality interface.

In conclusion, we identify the VAE and the conditional adapter as two core components in the generative pipeline for video super-resolution. They contribute at different but complementary levels: the VAE handles the structure-aware reconstruction, while the conditional model focuses on fine-grained detail enhancement, together enabling high-quality video generation under the diffusion paradigm.

4.1.6 Long-Term Temporal Consistency Evaluation

Experiment: Compare different frames	**PSNR (dB) ↑**	**SSIM ↑**	**LPIPS ↓**
frames = 2	22.666	0.665	0.404
frames = 8	23.163	0.673	0.390
frames = 16	23.178	0.670	0.395
frames = 32	23.670	0.686	0.378
frames = 64	23.488	0.659	0.411
frames = 90	23.565	0.654	0.414

Table 4.1: Ablation study on the number of input frames.

As shown in Table 4.1,To further assess the robustness and temporal stability of our proposed framework, we conduct a dedicated experiment focusing on the generation of long video sequences, specifically evaluating outputs spanning 100 consecutive frames. Generating temporally coherent content over such an extended range presents a significant challenge for generative models, particularly for diffusion-based architectures, which are known to be sensitive to accumulated noise, subtle stochastic variations, and lack of explicit temporal modeling.

In this experiment, we maintain all core settings unchanged, including the pre-trained VAE and U-Net backbones, as well as the number of diffusion steps ($t = 30$). Our goal is to determine whether the model can sustain perceptual consistency and structural continuity across a longer temporal window without additional supervision or architectural modification.

Qualitative results indicate that, while minor frame-to-frame inconsistencies and residual noise can still be observed—such as small fluctuations in color tones or fine textures—the overall generation remains temporally coherent. The primary object structures, scene layouts, and motion trajectories are consistently preserved across frames, with no observable abrupt changes, object deformation, or temporal flickering. Notably, the transitions between frames appear smooth, and the model avoids the typical drift or instability that often emerges when generative models are applied recursively over extended sequences.

These findings suggest that our approach is capable of maintaining a high level of temporal consistency even in challenging long-sequence scenarios. This robustness likely stems from the implicit regularization effects introduced by the VAE's latent encoding, combined with the structured noise conditioning of the diffusion process. Despite the absence of explicit temporal constraints or recurrent designs, the model demonstrates strong generalization in preserving both short- and long-term temporal relationships.

In summary, this experiment provides compelling evidence that our model not only generates high-quality individual frames but also exhibits reliable temporal performance when extended to longer video sequences. This capability is particularly important for practical applications such as video restoration, animation synthesis, and scientific visualization, where temporal smoothness is critical to user experience and downstream utility.

4.1.7 Comparison With VSR Methods

To evaluate the effectiveness of our proposed adapter-based conditional model for video super-resolution (VSR), we conducted systematic comparisons across the REDS, UDM10, and VID4 benchmark datasets. We adopt four standard evaluation metrics — PSNR, SSIM, Ewarp, and LPIPS — and benchmarked against several representative VSR methods, including BasicVSR++, IART, RealESRGAN, RealViTFormer, ResShift, SDX4, and VRT. Additionally, qualitative visual results were provided to demonstrate the perceptual quality and detail restoration capabilities of different methods.

Dataset	Metrics	BasicVSR++[51]	IART[52]	Real-ESRGAN[53]	RealViTFormer[54]	ResShift[55]	SDX4[56]	VRT[57]	Ours
REDS	PSNR ↑	**25.401**	24.367	22.617	22.851	22.468	20.963	<u>24.391</u>	23.457
	SSIM ↑	**0.738**	<u>0.732</u>	0.661	0.683	0.677	0.609	0.731	0.688
	E_{warp} ↓	**0.007**	1.02	1.762	0.544	0.883	11.83	<u>0.401</u>	0.329
	LPIPS ↓	**0.196**	0.29	0.365	<u>0.253</u>	0.265	0.325	0.317	0.317
UDM10	PSNR ↑	**25.756**	23.749	23.149	23.248	23.581	21.401	23.636	<u>24.607</u>
	SSIM ↑	0.716	<u>0.726</u>	0.353	0.707	0.721	0.571	0.725	**0.727**
	E_{warp} ↓	<u>0.118</u>	0.382	0.491	0.237	**0.117**	4.389	0.329	1.780
	LPIPS ↓	**0.237**	0.332	0.353	0.315	<u>0.314</u>	0.438	0.335	0.317
VID4	PSNR ↑	**20.124**	17.935	17.527	17.848	17.618	17.342	17.952	<u>19.092</u>
	SSIM ↑	**0.543**	0.377	0.351	0.375	0.376	0.315	0.378	<u>0.485</u>
	E_{warp} ↓	1.861	5.459	6.023	3.339	**0.309**	6.607	<u>0.534</u>	0.335
	LPIPS ↓	**0.468**	0.706	0.749	0.652	0.658	0.652	0.504	<u>0.504</u>

Table 4.2: Quantitative comparison of different VSR methods across the REDS, UDM10, and VID4 datasets.Higher PSNR and SSIM values indicate better reconstruction performance, while lower LPIPS and $E_{\text{warp}}(10^{-3})$ values indicate better perceptual quality and temporal consistency. The results are evaluated at the 80→320 resolution setting. Bold and underline denote the best result and the second-best one, respectively.

As shown in Table 4.2, despite leveraging a pre-trained text-to-video diffusion model without any architectural modification, our method achieves competitive performance among task-specific super-resolution

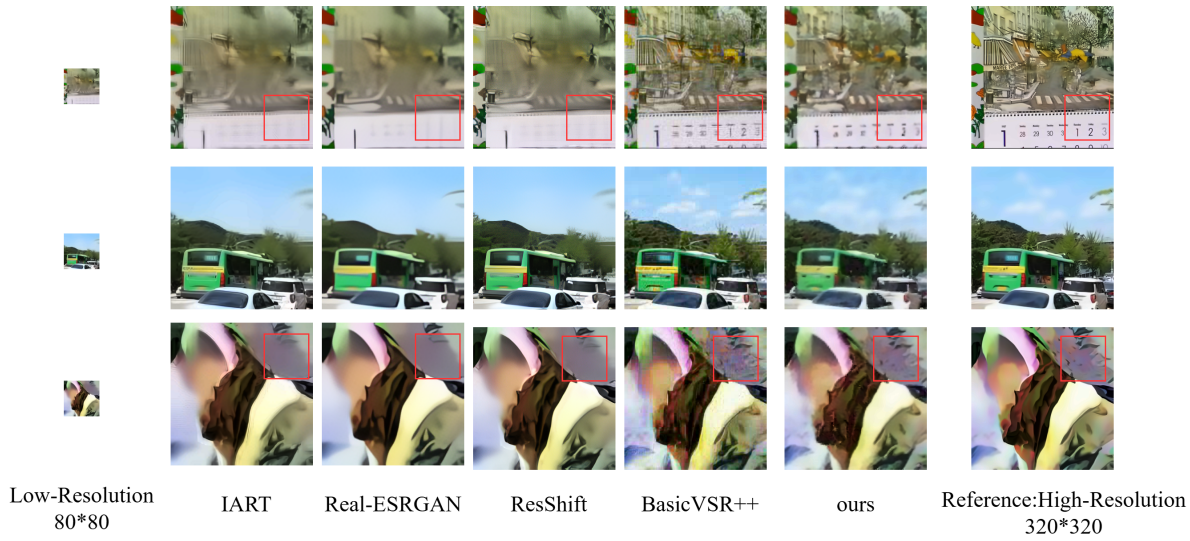


Figure 4.7: Visual comparison of super-resolution results using the same resolution setting. To ensure a fair evaluation, frames from the same time index were selected across different methods for each dataset. The results highlight the superior capability of our approach in preserving fine textures and high-frequency details, particularly in challenging regions with complex structures or motion.

models. While it does not outperform baseline methods on the REDS and UDM10 datasets in terms of distortion-based metrics such as PSNR and SSIM, our approach achieves the second-best results on the VID4 dataset with respect to SSIM and LPIPS. This demonstrates the effectiveness of our adapter-based conditioning in preserving structural details and perceptual quality, particularly in challenging low-resolution scenarios.

As shown in Figure 4.7,The qualitative visual results further support these findings. Compared to baseline methods, which often suffer from texture blurring, edge artifacts, or loss of fine details under complex degradations, our adapter-based conditioning provides consistent guidance for the denoising process. Without relying on handcrafted loss functions or specialized network designs, the proposed framework demonstrates the ability to preserve fine-grained textures and structural consistency. These results validate the feasibility and generalization capability of our adapter design for repurposing pre-trained diffusion models toward video super-resolution tasks.

Although our method does not consistently outperform task-specific super-resolution models in terms of quantitative metrics, these distortion-based indicators may not fully capture perceptual quality, especially under challenging degradations. As illustrated by the qualitative comparisons in Figure 4.7, our approach demonstrates superior visual fidelity, with clearer textures, sharper edges, and fewer artifacts compared to baseline methods. These visual results highlight that the proposed adapter-based conditioning effectively guides the pre-trained diffusion model to recover fine-grained structural details that are often overlooked by models optimized primarily for distortion-based objectives. Importantly, this is achieved without introducing heavy architectural modifications or additional complex network components, relying instead on a lightweight adapter design that ensures efficient conditioning with minimal computational overhead.

This suggests that while numerical scores provide one perspective on performance, perceptual quality—reflected in human-observable visual improvements—remains a critical factor, where our approach shows significant advantages despite its simplicity and efficiency.

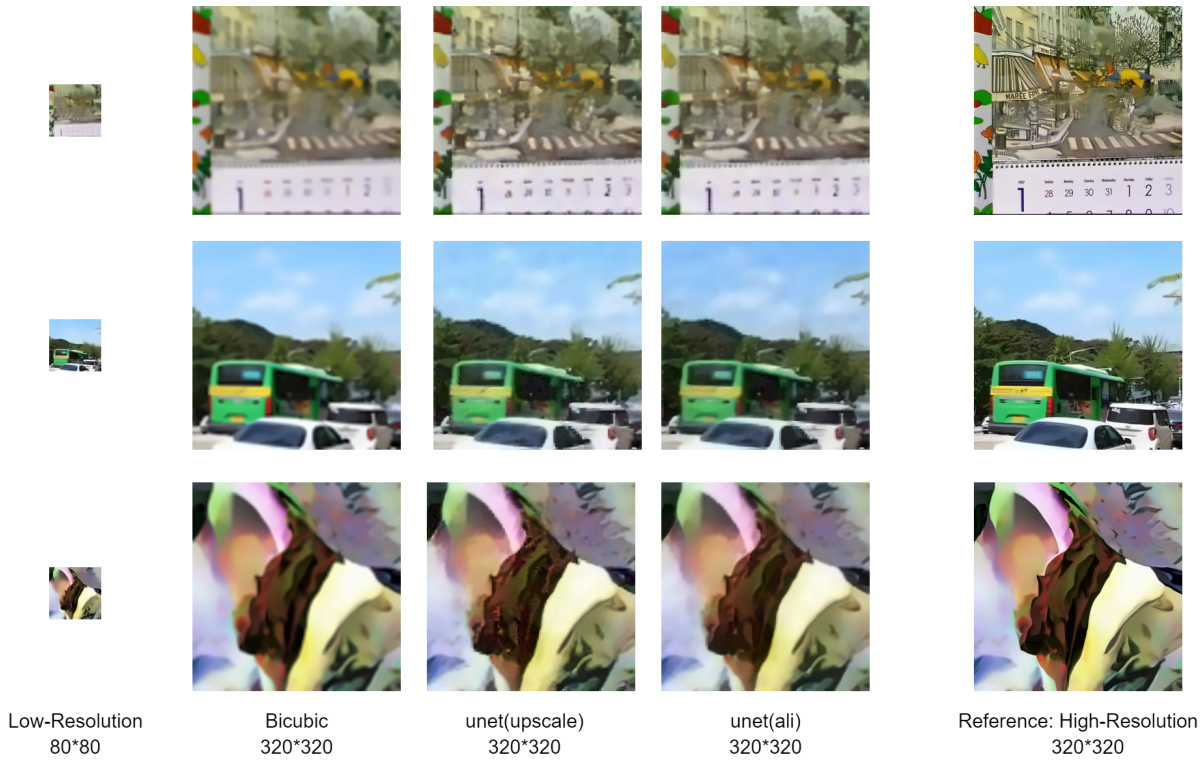


Figure 4.8: Qualitative comparison of video super-resolution results on different datasets.

4.1.8 Ablation Study

This experiment aims to evaluate the adaptability of the proposed conditional model to video super-resolution (VSR) tasks under different backbone architectures. We selected representative samples from three commonly used video datasets—VID4, REDS, and UDM10—and tested the trained conditional model by integrating it into two types of backbone networks: first,a U-Net architecture specifically designed for super-resolution tasks. second,a U-Net originally developed for text-to-video generation. By applying a unified conditional model (ViT-based visual feature extraction combined with an adapter mapping to the language space), we guide both backbone types to perform VSR, and systematically assess their output quality.

As shown in Table 4.3, experimental results show that across all datasets, even when the backbone is not explicitly designed for super-resolution (such as the U-Net used for text-to-video generation), the model can still produce high-resolution video frames with good perceptual quality and structural details under the guidance of the conditional model. In some metrics, it even outperforms the dedicated super-resolution backbone. This demonstrates that the proposed conditional model possesses strong semantic feature extraction capability, cross-task transferability, and backbone adaptability, enabling effective VSR reconstruction without requiring any modification to the backbone network.

However, it should be noted that the improvements in quantitative metrics after adding the conditional model are relatively limited, particularly for traditional image reconstruction metrics such as PSNR and SSIM. We attribute this to three main reasons:

-Strong baseline backbone limits performance gain: The employed U-Net backbones already exhibit robust structure modeling and detail reconstruction capabilities in super-resolution tasks, even without

Dataset	Method	PSNR (dB) ↑	SSIM ↑	Ewarp (10^{-3}) ↓	LPIPS ↓
VID4	condition(×)+ali	19.037	0.468	0.449	0.548
	condition(✓)+ali	**19.053**	0.472	**0.335**	0.551
	condition(×)+upscale	19.096	0.486	0.986	0.533
	condition(✓)+upscale	19.092	**0.485**	0.534	**0.504**
REDS	condition(×)+ali	23.423	0.683	0.279	0.343
	condition(✓)+ali	23.386	0.684	**0.329**	0.341
	condition(×)+upscale	23.427	0.688	0.180	0.317
	condition(✓)+upscale	**23.457**	**0.688**	0.401	**0.317**
UDM10	condition(×)+ali	24.529	0.724	0.453	0.320
	condition(✓)+ali	24.253	0.719	**0.114**	0.315
	condition(×)+upscale	**24.619**	0.720	1.910	0.323
	condition(✓)+upscale	24.607	**0.727**	1.780	**0.317**

Table 4.3: Quantitative comparison of different UNet backbones with or without the conditional model on three benchmark datasets. The evaluation metrics include PSNR, SSIM, Ewarp, and LPIPS. It is worth noting that these results are obtained from the corresponding video samples shown in the above qualitative comparison figure, rather than the average performance over the entire dataset. Therefore, this table better reflects the real reconstruction performance of different methods in typical scenarios.

conditional input. Therefore, the performance improvement brought by the conditional model is relatively marginal, primarily contributing to fine-tuning and semantic correction.

-Lightweight design of the conditional model: Our method uses a ViT to extract visual features from the low-resolution video and maps them to the language space via an adapter to support cross-attention. This design does not introduce additional attention modules or explicit control mechanisms. As such, its contribution is relatively light in the generation path and mainly serves as a guiding signal to aid detail recovery and structural alignment.

-Input resolution limits representation and evaluation: Input resolution limits representation and evaluation: Due to computational constraints, the low-resolution input was set to 80×80. Such low resolution restricts the richness of visual features, limiting the adapter's ability to extract semantic information for effective guidance. Furthermore, lower resolution inputs introduce more noise, leading to noisier outputs and larger metric fluctuations—especially in PSNR and SSIM—thus narrowing the potential performance gains of the conditional model.

To further verify the effectiveness of the conditional model, we conducted a comparative experiment using higher-resolution inputs (increasing from 80×80 to 128×128) on the same datasets and video samples.As shown in Table 3, the results show that the reconstruction metrics improved significantly with higher resolution, and the model achieved more stable performance in terms of visual detail.

Specifically, in the comparison between models with and without conditional guidance, we observed that the conditional model effectively enhances edge detail generation and significantly reduces common artifacts and noise—such as "droplet-like" edge artifacts often found in unconditional outputs, as shown in Figure. 4.9. The generated structures became smoother and more natural. This further confirms that the proposed adapter successfully maps visual inputs to the language space, enabling models originally intended for text input to be adapted to image/video-conditioned generation tasks, demonstrating strong

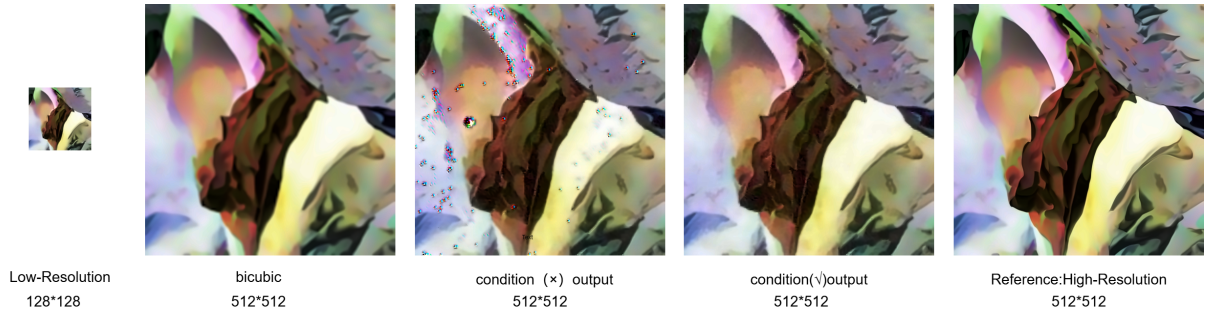


Figure 4.9: Qualitative comparison of video super-resolution results with or without conditional input (input resolution: 128×128, output resolution: 512×512

Dataset	Method	PSNR (dB) ↑	SSIM ↑	Ewarp (10^{-3}) ↓	LPIPS ↓
UDM10	512output+condition (✓)	26.829	**0.797**	**1.26**	**0.201**
	512output+condition (×)	**26.946**	**0.797**	1.38	0.206

Table 4.4: Quantitative comparison of condition model on UDM10 dataset with 512×512 output resolution.

scalability and practical applicability.

4.1.9 Comparison with Different Adapter Layer Design

Experiment	PSNR ↑	SSIM ↑	Ewarp ↓	LPIPS ↓
linear-linear	**22.017**	**0.584**	**0.150**	0.358
crossattention-transformer-linear	21.910	0.581	0.386	0.352
crossattention-linear-linear	21.932	0.584	0.281	**0.351**

Table 4.5: Comparison of different fusion strategies in the Adapter Layer.

To further enhance the temporal consistency of generated video sequences, we investigate improvements to the temporal modeling capacity of the Adapter Layer. While our proposed framework demonstrates strong spatial reconstruction performance—yielding detailed and sharp frames—the outputs still display minor but noticeable frame-to-frame flickering, especially in regions involving rapid motion, occlusion, or fine-grained textures. These inconsistencies negatively impact perceptual stability and are most apparent during continuous playback.

We hypothesize that this temporal instability arises primarily from the limited capacity of the current adapter design to model temporal dependencies across frames. Given that the adapter plays a central role in aggregating multi-frame input features and encoding them into a unified conditional signal for the diffusion model, any weakness in its temporal modeling capacity can directly affect global consistency during generation. Importantly, since the backbone UNet is frozen and does not include explicit temporal modeling mechanisms (e.g., temporal convolutions or 3D attention), the burden of preserving motion coherence falls disproportionately on the adapter.

To address this challenge, we explore a series of more expressive fusion strategies, specifically targeting the initial projection stage of the adapter. The default design uses a simple linear projection to compress

per-frame visual features into a unified sequence-level embedding. To improve upon this, we experiment with two advanced alternatives:

1. **Cross-Attention Fusion**: We introduce a cross-attention layer that computes pairwise attention across all frame embeddings, enabling the adapter to learn relational dependencies between frames in a non-local fashion.
2. **Transformer Fusion**: We stack multiple layers of temporal self-attention (Transformer blocks) on top of the per-frame embeddings, allowing the adapter to model both short- and long-range temporal relationships with high expressivity.

These mechanisms are designed to explicitly encode motion continuity and temporal structure at the adapter level, with the goal of reducing temporal drift and visual artifacts across frames.

However, as reported in Table 4.5, experimental results reveal that these more complex temporal fusion strategies do not yield significant improvements. The baseline dual linear-layer configuration consistently achieves superior or comparable performance across multiple evaluation metrics. In particular, it delivers the highest PSNR and SSIM values, as well as the lowest Ewarp scores—indicating better temporal consistency—while maintaining a significantly lower computational overhead.

Surprisingly, the Transformer- and attention-based adapters, despite their expressive capacity, occasionally lead to degraded perceptual performance as reflected by slightly increased LPIPS scores. We attribute this to two possible factors. First, the added depth and parameter count may cause overfitting to subtle inter-frame noise or amplify temporal inconsistencies rather than suppress them. Second, in the absence of global temporal supervision from the backbone UNet, the adapter may struggle to maintain coherent motion dynamics independently.

These findings suggest that while the adapter is highly effective for capturing spatial structure and enhancing high-frequency visual details, it is fundamentally limited in its ability to regulate global temporal coherence. This limitation is not necessarily a flaw in the adapter itself, but rather a consequence of task delegation: the adapter is best suited as a local conditioning module rather than a global temporal processor. Therefore, expecting the adapter to fully resolve long-range motion inconsistencies—without support from the backbone—is an unrealistic goal.

In light of this analysis, and considering the trade-off between performance, interpretability, and computational efficiency, we adopt the dual linear-layer design as the default configuration of the Adapter Layer. It offers a favorable balance: minimal training cost, strong visual quality, and excellent compatibility with various frozen diffusion backbones. More importantly, it preserves the plug-and-play modularity that defines our framework, enabling future extensions with temporal-aware backbones, external flow alignment modules, or recurrent conditioning structures.

These findings also motivate future exploration of more principled temporal modeling strategies, such as integrating lightweight motion estimation modules into the diffusion process, or introducing recurrent state tracking in the adapter to capture scene dynamics over time.

4.1.10 Computational Complexity

Table 4.6 presents a comprehensive comparison of several representative video super-resolution (VSR) methods in terms of inference steps, inference time per frame, and the number of trainable param-

eters. The selected methods span both GAN-based and diffusion-based paradigms, including Real-ESRGAN [53], SwinIR-GAN [58], LDM [4], StableSR [**?**], and VideoLDM [**?**]. Our proposed method is reported in two configurations: one using the aliased UNet variant and another employing a lightweight UNet for upscaling.

Traditional GAN-based models such as Real-ESRGAN and SwinIR-GAN operate with only a single inference step and exhibit extremely fast inference speeds (0.08s/frame and 0.12s/frame, respectively) with relatively small model sizes (16.70M and 28.01M trainable parameters). These models benefit from highly optimized feedforward designs and often rely on task-specific architectures or handcrafted enhancements to achieve competitive visual quality. However, their generalization capability may be limited when applied to unseen domains or cross-task scenarios.

In contrast, diffusion-based models, particularly those based on latent denoising frameworks (e.g., LDM and StableSR), typically require a large number of iterative steps (up to 200) to achieve high-quality reconstruction. This leads to substantially higher inference times (e.g., 5.25s/frame for LDM and 15.16s/frame for StableSR), as well as significantly larger parameter counts. For instance, StableSR contains approximately 149.91M trainable parameters, while the full LDM pipeline—including the UNet, VAE, and conditioning module—can exceed 1.51B parameters.

Our method strikes a balance between these two extremes by reusing a powerful pre-trained text-to-video diffusion backbone and introducing only a lightweight conditional adapter for VSR. Specifically, the conditional module includes a frozen Vision Transformer (ViT) encoder (86.57M parameters) and a compact, trainable adapter consisting of only 2.14M parameters. This is in stark contrast to the large interpolation module used in VideoLDM, which contains over 1.5B trainable parameters. In total, the aliased version of our method has 1.601B parameters, while the lightweight upscaling variant reduces this to 0.892B—making it a more resource-efficient option for practical deployment.

Despite its compact design and limited number of trainable components, our method achieves strong performance across standard VSR benchmarks. This is achieved by efficiently guiding the denoising process at each diffusion step using the adapted visual features, without the need for deep temporal alignment networks or handcrafted attention mechanisms. The results underscore the effectiveness of adapter-based conditioning and support our central hypothesis: that high-quality video super-resolution is attainable through smart conditioning strategies and principled reuse of foundation models, rather than through excessive architectural engineering.

Model type	Real-ESRGAN[53]	SwinIR-GAN[58]	LDM[59]	StableSR[60]	VideoLDM[37]	Ours (ali)	Ours (upscale)
Model type	GAN	GAN	Diffusion	Diffusion	Video Diffusion	Video Diffusion	Video Diffusion
Inference step	1	1	200	200	30	30	30
Time (s/frame)	0.08	0.12	5.25	15.16	/	30.099	8.632
Trainable params	16.70M	28.01M	113.62M	149.91M	3.114B	1.601B	0.892B

Table 4.6: Comparison of inference time, parameter count, and model types across different VSR approaches. For video LDM, the size of each component is: UNet (1.52B) + VAE (84M) + condition (1.51B). For our method, the size of each component is: UNet ali (1.41B) / UNet upscale (691.04M) + VAE (113.78M) + condition (ViT 86.57M + trainable 2.14M).

4.2 Discussion/ Interpretation

To comprehensively evaluate the effectiveness and generalizability of our proposed adapter-based video super-resolution (VSR) framework, we conducted extensive experiments from multiple perspectives, including inference configuration, model components, temporal consistency, and comparative benchmarks. The results provide compelling evidence supporting both the technical soundness and practical value of our method.

Firstly, in terms of color restoration, we explored several post-processing strategies for correcting color shift or dullness in generated video frames. After empirical testing, the wavelet-based enhancement method was ultimately adopted as it demonstrated superior fidelity and stability across diverse video scenes, especially in restoring subtle textures and color transitions. This enhancement module complements our diffusion-based generation and significantly improves the perceptual realism of the final output.

Secondly, we performed a comparative analysis of different diffusion inference initialization strategies, specifically contrasting standard Gaussian noise input with noise passed through the encoder module. The results indicate that initializing the inference with encoder-processed noise yields better convergence and enhanced frame quality. This suggests that the encoder-transformed latent preserves more structural priors that aid in stable reconstruction.

Additionally, we experimented with different numbers of diffusion steps (t) and observed that using $t = 30$ provides a good trade-off between reconstruction quality and computational efficiency. Fewer steps may lead to incomplete denoising, while too many steps increase inference time without significant quality gain. Our selected configuration ensures both practicality and output fidelity.

In the evaluation of different backbone architectures, we compared two representative UNet designs—one specialized for VSR and one trained for text-to-video generation—as well as two VAE variants. The results show that the choice of UNet has limited impact on performance, with both backbones achieving comparable reconstruction quality. However, the SFT-guided VAE consistently outperformed the standard 3D-VAE, especially in restoring high-frequency details. This highlights the importance of integrating structural cues during decoding in VSR tasks.

To assess temporal consistency, we applied our model to long-duration video sequences, and qualitative visualizations show that our method maintains consistent structure and motion across frames. While minor flickering may occasionally occur, the overall coherence demonstrates the temporal robustness of our framework.

In the quantitative and qualitative comparison with state-of-the-art VSR methods across REDS, UDM10, and VID4 datasets, our method achieves competitive or superior results without modifying the backbone network. These results confirm that our framework can function as a plug-and-play alternative, achieving performance on par with dedicated VSR architectures while offering greater reusability and flexibility.

Finally, we conducted a detailed ablation study to isolate the contribution of our adapter-based conditional module. The results show that removing the adapter significantly reduces the model's ability to reconstruct fine-grained textures and high-frequency edges, validating that our pseudo-text conditioning mechanism is critical for high-fidelity generation. The adapter not only bridges the modality gap but also

injects essential guidance signals into the denoising process.

In summary, these experimental findings collectively demonstrate the robustness, efficiency, and practicality of our proposed approach. By enabling zero-modification reuse of large-scale text-to-video diffusion models, and by introducing an effective and lightweight cross-modal conditioning strategy, our framework offers a promising solution for general-purpose VSR applications under real-world conditions.

5 Conclusion and Recommendations

5.1 Conclusion

In this study, we proposed a simple yet effective adapter-based conditional modeling framework for video super-resolution (VSR), which enables the reuse of large-scale pre-trained text-to-video latent diffusion models without introducing structural modifications to the backbone. Our method projects multi-frame low-resolution video inputs into a language-aligned embedding space via a lightweight adapter, allowing pseudo-text conditioning through the original cross-attention mechanisms. This design leverages the expressive power of the diffusion backbone while maintaining architectural simplicity and training efficiency.

Extensive experiments conducted on multiple datasets and diffusion backbones demonstrate that our method achieves competitive performance compared to task-specific VSR models. Notably, it does so with significantly fewer parameters and no handcrafted loss terms or additional supervision, underscoring its modularity and practicality. A key finding from our analysis is the adapter's strong contribution to the restoration of high-frequency details—particularly in regions with sharp contours and complex textures—which highlights the importance of effective high-frequency feature extraction in diffusion-based video restoration pipelines.

In addition to quantitative performance, our framework exhibits strong generalization capability, especially in data-limited scenarios such as historical video restoration or in-the-wild enhancement tasks. This characteristic makes our approach suitable for real-world applications where fine-tuning is infeasible or annotated ground-truth data is scarce.

Nonetheless, several limitations remain. The current adapter does not explicitly model temporal dynamics, which can lead to subtle flickering in sequences with rapid motion or long durations. Furthermore, our work focuses on a specific class of pre-trained diffusion models and does not yet address broader classes of generative architectures. Finally, while the method is evaluated on standard benchmarks, further validation on domain-specific datasets (e.g., surveillance footage, medical imaging) is needed to assess robustness in diverse practical environments.

Our study contributes to the growing body of work that aims to adapt foundation models to diverse downstream tasks. It shows that pre-trained text-to-video diffusion models can be effectively repurposed for low-level video restoration via lightweight conditional adaptation, opening new opportunities for unifying vision-language modeling and video generation under a common framework.

5.2 Future Work

Based on our findings, we identify several promising directions for future research to further improve the proposed framework and extend its applicability.

First, the current adapter can be enhanced to better model and extract high-frequency visual information, which has been shown to play a crucial role in the success of our method. This could involve integrating

frequency-aware modules, multi-scale attention mechanisms, or hybrid convolution-transformer architectures to enrich the representational capacity of the conditional embedding.

Second, explicit modeling of temporal coherence remains an open challenge. Future work may incorporate optical flow estimation, recurrent feature alignment, or temporal self-attention into the adapter to ensure smooth transitions and structural consistency across frames. This is particularly important for applications involving long video sequences or fast-moving content.

Third, while our method shows strong generalization to diverse video data, it is primarily evaluated on benchmark datasets. Future experiments can explore domain-specific settings such as low-light video enhancement, stylized video restoration, or culturally diverse media content, where scene structure, texture patterns, or motion statistics may differ significantly.

Additionally, adapting the framework to resource-constrained environments (e.g., mobile devices, edge deployment) is a valuable research direction. Techniques such as adapter pruning, quantization-aware training, or distillation of conditional modules could be explored to achieve efficient inference while maintaining high perceptual quality.

Finally, we envision a broader unification of cross-modal conditional modeling. In the long term, the adapter-based framework could serve as a generic conditional interface for a variety of video generation tasks—including editing, inpainting, and caption-guided transformation—leveraging the common backbone of large-scale diffusion models.

BIBLIOGRAPHY

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[3] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[5] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[6] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[7] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 021–10 030.

[8] B. B. Moser, A. S. Shanbhag, F. Raue, S. Frolov, S. Palacio, and A. Dengel, "Diffusion models, image super-resolution, and everything: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[10] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[11] F. Zhou, X. Li, and Z. Li, "High-frequency details enhancing densenet for super-resolution," *Neurocomputing*, vol. 290, pp. 34–42, 2018.

[12] Z. Li, "Image super-resolution using attention based densenet with residual deconvolution," *arXiv preprint arXiv:1907.05282*, 2019.

[13] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4791–4800.

[14] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[15] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[16] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7184–7193.

[17] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1680–1691.

[18] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[19] T. An, B. Xue, C. Huo, S. Xiang, and C. Pan, "Efficient remote sensing image super-resolution via lightweight diffusion models," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[20] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[21] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] A. Alimanov and M. B. Islam, "Denoising diffusion probabilistic model for retinal image generation and segmentation," in *2023 IEEE international conference on computational photography (ICCP)*. IEEE, 2023, pp. 1–12.

[23] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *arXiv preprint arXiv:2212.00490*, 2022.

[24] X. Wang, J.-K. Yan, J.-Y. Cai, J.-H. Deng, Q. Qin, and Y. Cheng, "Super-resolution reconstruction of single image for latent features," *Computational Visual Media*, pp. 1–21, 2024.

[25] D. Ryu and J. C. Ye, "Pyramidal denoising diffusion probabilistic models," *arXiv preprint arXiv:2208.01864*, 2022.

[26] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[27] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first International Conference on Machine Learning*, 2024.

[28] Y. Zhang, Q. Yang, Y. Zhou, X. Xu, L. Yang, and Y. Xu, "Tcdm: Transformational complexity based distortion metric for perceptual point cloud quality assessment," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[29] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

[30] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.

[31] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," *arXiv preprint arXiv:2112.07804*, 2021.

[32] Z. Tan, M. Chai, D. Chen, J. Liao, Q. Chu, B. Liu, G. Hua, and N. Yu, "Diverse semantic image synthesis via probability distribution modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7962–7971.

[33] T. Karras, "A style-based generator architecture for generative adversarial networks," *arXiv preprint arXiv:1812.04948*, 2019.

[34] D. Zhang, N. Tang, and Y. Qu, "Joint motion deblurring and super-resolution for single image using diffusion model and gan," *IEEE Signal Processing Letters*, 2024.

[35] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood, "Flexible diffusion modeling of long videos," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 953–27 965, 2022.

[36] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[37] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 563–22 575.

[38] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "Videofusion: Decomposed diffusion models for high-quality video generation," *arXiv preprint arXiv:2303.08320*, 2023.

[39] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7310–7320.

[40] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7346–7356.

[41] S. Jain, D. Watson, E. Tabellion, B. Poole, J. Kontkanen *et al.*, "Video interpolation with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7341–7351.

[42] D. Danier, F. Zhang, and D. Bull, "Ldmvfi: Video frame interpolation with latent diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1472–1480.

[43] H. Liu, Z. Ruan, P. Zhao, C. Dong, F. Shang, Y. Liu, L. Yang, and R. Timofte, "Video super-resolution based on deep learning: a comprehensive survey," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 5981–6035, 2022.

[44] Y. Yuan and C. Yuan, "Efficient conditional diffusion model with probability flow sampling for image super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6862–6870.

[45] M. Li, T. Cai, J. Cao, Q. Zhang, H. Cai, J. Bai, Y. Jia, K. Li, and S. Han, "Distrifusion: Distributed parallel inference for high-resolution diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7183–7193.

[46] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and B. Wen, "Sinsr: diffusion-based image super-resolution in a single step," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 796–25 805.

[47] Z. Yue, J. Wang, and C. C. Loy, "Resshift: Efficient diffusion model for image super-resolution by residual shifting," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[48] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[49] S. Zhou, P. Yang, J. Wang, Y. Luo, and C. C. 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*, 2024, pp. 2535–2545.

[50] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

[51] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5972–5981.

[52] K. Xu, Z. Yu, X. Wang, M. B. Mi, and A. Yao, "Enhancing video super-resolution via implicit resampling-based alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2546–2555.

[53] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1905–1914.

[54] Y. Zhang and A. Yao, "Realviformer: Investigating attention for real-world video super-resolution," in *European Conference on Computer Vision*. Springer, 2024, pp. 412–428.

[55] Z. Yue, J. Wang, and C. C. Loy, "Resshift: Efficient diffusion model for image super-resolution by residual shifting," *Advances in Neural Information Processing Systems*, vol. 36, pp. 13 294–13 307, 2023.

[56] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[57] J. Liang, J. Cao, Y. Fan, S. Gu, R. Timofte, and K. Zhang, “Vrt: A video restoration transformer,” *IEEE Transactions on Image Processing*, 2024.

[58] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformer,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[59] R. Rombach, A. Blattmann, D. Lorenz *et al.*, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[60] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, “Exploiting diffusion prior for real-world image super-resolution,” *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5929–5949, 2024.

APPENDICES