

STAR: Spatial-Temporal Augmentation with Text-to-Video Models for Real-World Video Super-Resolution

Rui Xie^{1*}, Yinhong Liu^{1*}, Penghao Zhou², Chen Zhao¹, Jun Zhou³
 Kai Zhang¹, Zhenyu Zhang¹, Jian Yang¹, Zhenheng Yang², Ying Tai^{1†}
¹Nanjing University, ²ByteDance, ³Southwest University

<https://nju-pcalab.github.io/projects/STAR>

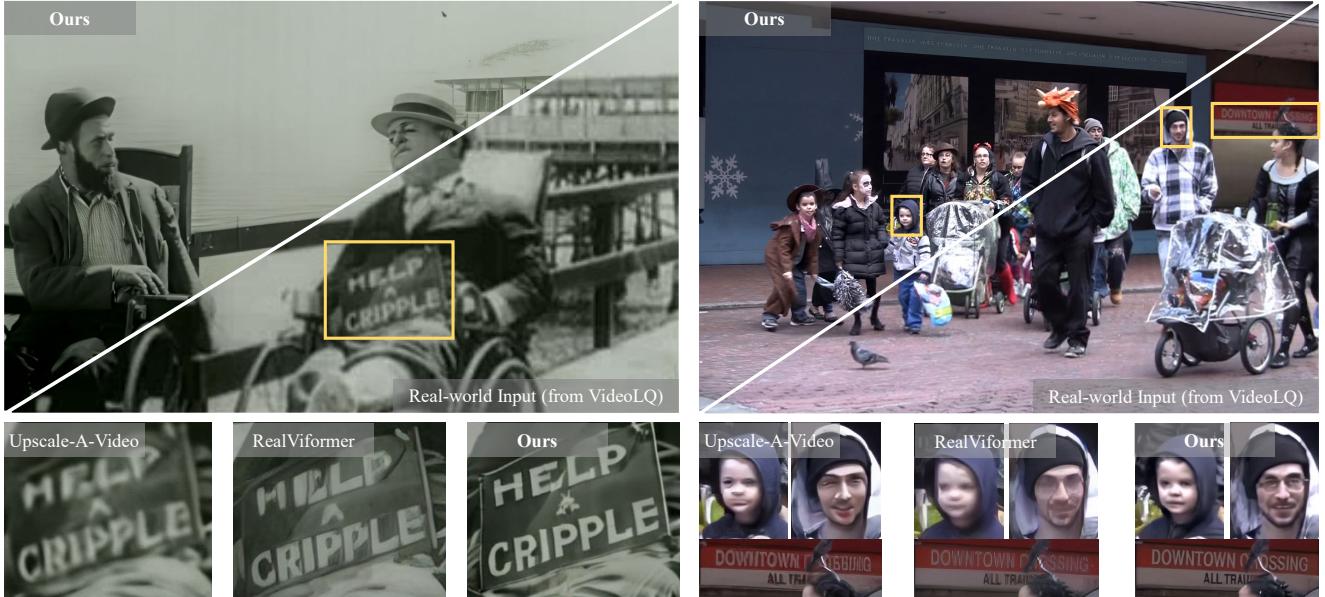


Figure 1. Visualization comparisons on both real-world and synthetic low-resolution videos. Compared to the state-of-the-art VSR models [73, 75], our results demonstrate more natural facial details and better structure of the text. (**Zoom-in for best view**)

Abstract

Image diffusion models have been adapted for real-world video super-resolution to tackle over-smoothing issues in GAN-based methods. However, these models struggle to maintain temporal consistency, as they are trained on static images, limiting their ability to capture temporal dynamics effectively. Integrating text-to-video (T2V) models into video super-resolution for improved temporal modeling is straightforward. However, two key challenges remain: artifacts introduced by complex degradations in real-world scenarios, and compromised fidelity due to the strong generative capacity of powerful T2V models (e.g., CogVideoX-5B). To enhance the spatio-temporal quality of restored videos, we introduce **STAR** (**Spatial-Temporal Augmentation with T2V models for Real-world video super-resolution**), a novel

approach that leverages T2V models for real-world video super-resolution, achieving realistic spatial details and robust temporal consistency. Specifically, we introduce a Local Information Enhancement Module (LIEM) before the global attention block to enrich local details and mitigate degradation artifacts. Moreover, we propose a Dynamic Frequency (DF) Loss to reinforce fidelity, guiding the model to focus on different frequency components across diffusion steps. Extensive experiments demonstrate **STAR** outperforms state-of-the-art methods on both synthetic and real-world datasets.

*Equal contributions. Work done during Rui Xie's ByteDance internship. † indicates corresponding author.

1. Introduction

Real-world video super-resolution (VSR) aims to generate high-resolution (HR) videos with clear details and strong temporal consistency from low-resolution (LR) inputs with unknown degradations. Most VSR methods [10, 22, 50, 60] only focus on simple, known degradations like downsampling [15, 21] or camera-related issues [62]. However, real-world scenarios often involve *unexpected degradations* such as noise, blur, and compression, making it difficult for models to capture both spatial and temporal information needed for high-quality, consistent restoration.

GAN-based methods [11, 51, 58, 62, 73] are widely used in real-world VSR for improving details through adversarial learning. By incorporating optical flow maps, they also improve temporal consistency, yielding smooth motion across frames. However, their limited generative capacity often results in oversmoothing, as illustrated in Figure 1. Recently, image diffusion models [43] have been applied to real-world VSR for realistic video generation. Methods like [14, 63, 67, 75] incorporate temporal blocks or optical flow maps to improve temporal information capture. However, since these models are primarily trained on image data rather than video data [13, 36, 49, 53], simply adding temporal layers often fails to ensure high temporal consistency (see Figure 8). VEnhancer [17] and LaVie-SR [52] incorporate T2V models for super-resolving AI-generated videos. However, two key challenges still remain: *artifacts introduced by complex degradations* in real-world settings, and *compromised fidelity* due to the strong generative capacity of powerful T2V models (*e.g.*, CogVideoX).

To fully leverage the T2V prior [64, 72] to enhance practical VSR, we introduce STAR, a novel Spatial-Temporal Augmentation approach for Real-world VSR that achieves realistic spatial details and robust temporal consistency. Specifically, 1) To address artifacts, we introduce a Local Information Enhancement Module (LIEM) before global self-attention to evaluate its impact on T2V models for real-world VSR. This approach stems from our observation that most T2V models rely solely on a global information extraction module (*i.e.*, global self-attention), whereas capturing local details is crucial for video restoration. 2) To improve fidelity, we propose a Dynamic Frequency (DF) Loss, guiding the model to prioritize low- or high-frequency information at different diffusion steps. This is based on our observation that during the reverse diffusion process, our model tends to first recover structure and then refine details. This approach decouples fidelity requirements, reduces learning difficulty, and enhances restoration fidelity.

In summary, our main contributions are as follows:

- We propose STAR, a Spatio-Temporal quality Augmentation framework for Real-world VSR. To our best knowledge, we are the first to integrate diverse, powerful text-to-video diffusion priors into real-world VSR, improv-

ing both spatial details and temporal consistency.

- We introduce LIEM to enhance local details and ease degradation removal, effectively mitigating artifacts. Moreover, we propose DF loss to guide the model in learning frequency-specific information across diffusion steps, decoupling fidelity requirements and ultimately improving overall fidelity.

- Our STAR achieves the highest clarity (DOVER scores) across all datasets compared to state-of-the-art methods, while maintaining robust temporal consistency.

2. Related Work

Video Super-Resolution. Traditional VSR methods can be roughly divided into two categories: recurrent-based [16, 20, 28, 44, 46] and sliding-window-based [8, 27, 29, 59, 65] methods. Recurrent-based methods process LR video frame by frame using recurrent neural networks [34]. In contrast, sliding-window-based methods divide a video sequence into segments, using each as input to super-resolve the video. However, both approaches suffer from degradation mismatch, leading to significant performance drops in real-world applications. Recently, there has been a growing focus on real-world VSR, targeting complex, unknown degradations. RealBasicVSR [11], an extension of BasicVSR [9], introduces a pre-cleaning module to mitigate artifacts. RealVifomer [73] discovers that channel attention is less sensitive to artifacts and uses squeeze-excite mechanisms and covariance-based rescaling to address these challenges further. While GAN-based and image diffusion models have made substantial progress, they still face issues such as *over-smoothing details and temporal inconsistency*.

Text-to-Video Diffusion Model. Large-scale pre-trained text-to-video (T2V) diffusion models have garnered significant attention, particularly with the impressive results from Sora [7, 37]. Numerous T2V models have since emerged, generally divided into: U-Net-based methods [4, 5, 19, 47] and DiT-based methods [3, 12, 40, 64]. I2VGen-XL [72], a U-Net-based method, employs a two-stage approach: first generating semantically and content-consistent LR videos, then using these as conditions to produce HR outputs. CogvideoX [64], built on DiT [39], introduces an adaptive LayerNorm to enhance text-video alignment and employs 3D attention to better integrate spatio-temporal information. Both models have large model capacities and are trained on large-scale datasets, enabling them to capture robust spatio-temporal priors. In this work, we propose STAR to *fully leverage T2V model prior for real-world VSR*.

Diffusion Prior for Super-Resolution. Several works [30, 48, 57, 61, 74] have leveraged generative diffusion priors for image and video super-resolution. StableSR [48] adds a time-aware encoder and feature warping module to the SD model. DiffBIR [30] integrates restoration and

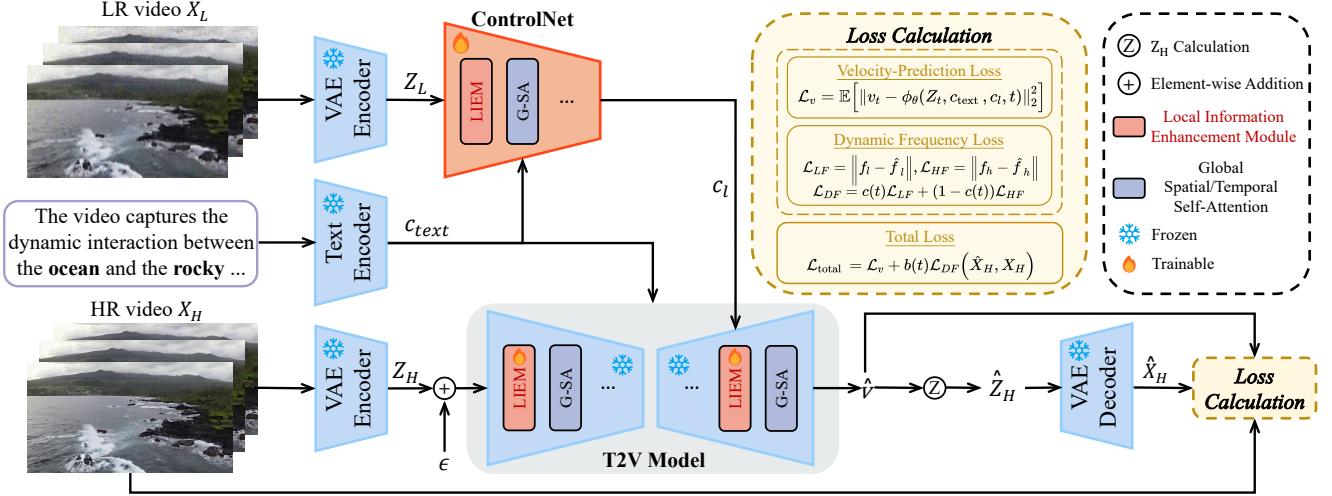


Figure 2. Overview of the proposed STAR.

generative modules via ControlNet, while PASD [61] and SeeSR [57] embed semantic information in U-Net to guide diffusion. These methods balance fidelity and perceptual quality, achieving high-resolution image details. Methods like Upscale-A-Video [75], MGLD-VSR [63], Inflating with Diffusion [67], and SATeCo [14] have adapted text-to-image diffusion priors [19, 43] for VSR by adding temporal layers. However, rooted in text-to-image models, they often struggle with temporal consistency. More recently, VEnhancer[17] and LaVie-SR[52] have incorporated T2V models to super-resolve AI-generated videos but struggle with complex degradations in practical environments. In contrast, we are the *first* to integrate powerful T2V diffusion priors for real-world VSR, introducing the LIEM module to address spatial artifacts and DF loss to enhance fidelity.

3. Methodology

3.1. Overview

Modules. The STAR primarily includes four modules: VAE [24], text encoder [41, 42], ControlNet [70] and T2V model [64, 72] with Local Information Enhancement Module (LIEM) to alleviate the artifacts (further analysis is provided in Sec. 3.2). As depicted in Figure 2, the VAE encoder takes HR videos X_H and LR videos X_L as input to generate latent tensors Z_H and Z_L , respectively. The text encoder is responsible for generating text embeddings c_{text} to provide high-level information. ControlNet takes Z_L and c_{text} as input to guide the T2V model output. Finally, the T2V model ϕ_θ with LIEM receives noisy input $Z_t = \alpha_t Z_H + \sigma_t \epsilon$ (t denotes diffusion step, α_t and σ_t are noise scheduler parameters), c_{text} and the control signal from ControlNet c_l to predict the velocity $v_t \equiv \alpha_t \epsilon - \sigma_t Z_H$ [45].

Losses. We utilize v-prediction objective in optimization:

$$\mathcal{L}_v = \mathbb{E}[\|v_t - \phi_\theta(Z_t, c_{text}, c_l, t)\|_2^2]. \quad (1)$$

Given the strong generalization ability of T2V models, relying solely on the v-prediction objective for optimization may lead to restored outputs with low fidelity, an essential factor in video super-resolution tasks. To address this, we introduce Dynamic Frequency (DF) Loss, which adaptively adjusts the constraint on high- and low-frequency components of the predicted \hat{X}_H across different diffusion steps. The overall optimization objective for STAR is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_v + b(t)\mathcal{L}_{DF}(\hat{X}_H, X_H), \quad (2)$$

where $b(t) = 1 - \frac{t}{t_{max}}$ is a weighting function (t_{max} is set to 999) to balance \mathcal{L}_v and \mathcal{L}_{DF} . With the proposed LIEM and DF loss, STAR achieves high spatio-temporal quality, reduced artifacts and enhanced fidelity.

3.2. Local Information Enhancement Module

Motivation. Most T2V models primarily use a global attention mechanism [31], which is well-suited to text-to-video tasks by capturing global information to generate complete videos from scratch. However, this approach may be suboptimal for real-world video super-resolution, where complex degradations occur and local details are crucial [25]. Relying solely on global attention mechanisms presents two drawbacks for real-world video super-resolution: 1) It *complicates degradation removal*, as it processes the entire degraded video at once (the first and second columns in Figure 3 (right)). 2) It *lacks local details*, resulting in blurry outputs (the third column in Figure 3 (right)).

Details of LIEM. To address the above issues, we propose a simple but effective approach: adding a *Local Information Enhancement Module* (LIEM) before the global

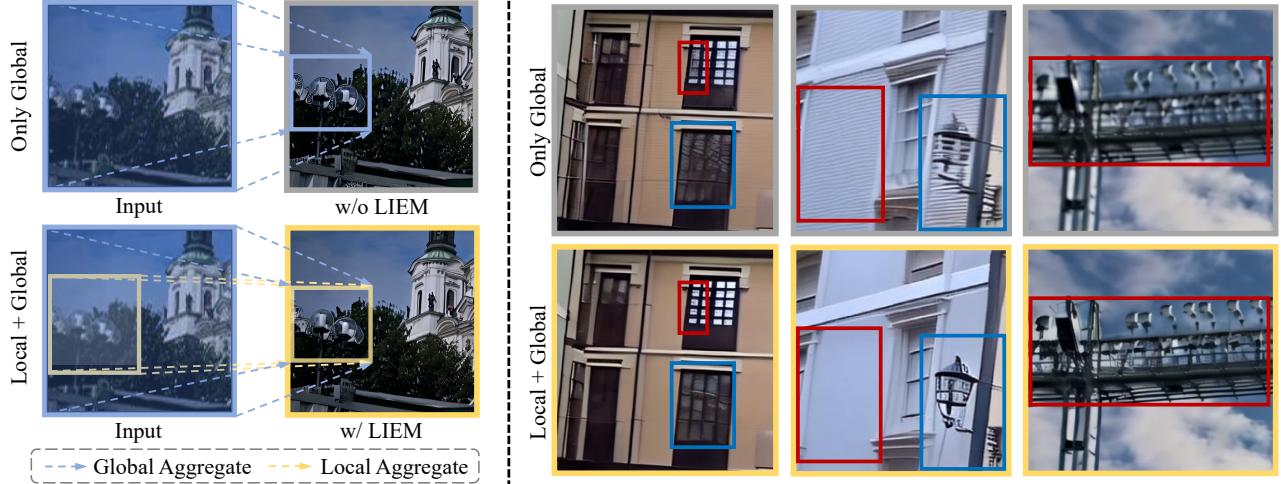


Figure 3. **Motivation of LIEM.** **Left:** schematic diagram illustrating the impact of using only global structure versus a combination of local and global structures. **Right:** visual comparison on real-world and synthetic videos. (**Zoom-in for best view**)

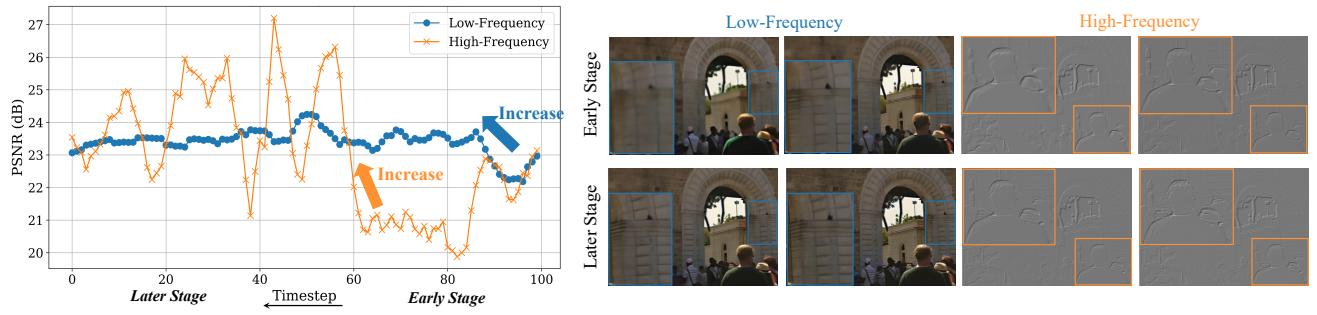


Figure 4. **Motivation of DF Loss.** **Left:** PSNR curves of low- and high-frequency components relative to ground truth across diffusion steps. The low-frequency PSNR increases during the early diffusion steps, while the high-frequency PSNR rises in the later diffusion steps. **Right:** visual results of low- and high-frequency components at different diffusion stage. (**Zoom-in for best view**)

attention block to make T2V model pay more attention to local information. It can be expressed by:

$$L(F_I) = \text{Sigmoid}(\text{Conv}_{3 \times 3}(\text{Concat}(AP(F_I), MP(F_I)))), \quad (3)$$

$$F_O = G(L(F_I) \cdot F_I) + F_I, \quad (4)$$

where $AP(\cdot)$ and $MP(\cdot)$ denote average pooling and max pooling, respectively. F_I and F_O represent the input and output features, while $G(\cdot)$ and $L(\cdot)$ refer to the global attention block and LIEM. We adopt the local attention block in CBAM [55] as LIEM for simplicity. Additional analysis on the impact of adding LIEM is provided in Sec. 4.3. Intuitively, as shown in the second row of Figure 3 (left), incorporating LIEM enables the T2V model to address local region degradation first and then aggregate global features. This approach reduces the complexity of degradation removal and mitigates artifacts. Furthermore, the T2V model with LIEM produces clearer, more detailed results due to the enriched local information.

3.3. Dynamic Frequency Loss

Motivation. The powerful generative capacity of diffusion models may compromise the fidelity in restored result [57, 66]. In Figure 4 (Right), an interesting pattern emerges when examining restored results at each diffusion step during inference. In the early stages, the model primarily reconstructs structure with low frequency, whereas in later stages, after the structure is largely complete, focus shifts to refining details with high frequency. To further illustrate this phenomenon, Figure 4 (Left) presents PSNR curves of low- and high-frequency components against the ground truth across diffusion steps. The low-frequency PSNR rises in the early stages, while the high-frequency PSNR increases later, aligning with the visual results.

Fidelity can be divided into two types: 1) Low-frequency fidelity, encompassing large structures and instances. 2) High-frequency fidelity, including edges and textures, aligning with the characteristics of the denoising process. This raises a question: Can we design a loss func-

Table 1. Quantitative evaluations on diverse VSR benchmarks from synthetic (UDM10, REDS30, OpenVid30) and real-world (VideoLQ) sources. The best performance is highlighted in **bold**, and the second-best in underlined. E_{warp}^* refers to E_{warp} ($\times 10^{-3}$).

Datasets	Metrics	Real-ESRGAN ICCVW 2021	DBVSR ICCV 2021	RealBasicVSR CVPR 2022	RealVifomer ECCV 2024	ResShift NeurIPS 2023	StableSR IJCV 2024	Upscale-A-Video CVPR 2024	MGLDVSR ECCV 2024	Ours
UDM10	PSNR↑	22.41	19.65	23.64	24.00	22.90	23.50	21.29	23.74	23.91
	SSIM↑	0.6476	0.4747	0.6842	<u>0.6896</u>	0.5451	0.6599	0.5967	0.6826	0.7164
	LPIPS↓	0.2769	0.4566	0.2514	0.2325	0.4036	0.2785	0.3006	<u>0.2195</u>	0.1885
	DOVER↑	0.4831	0.0959	0.5039	0.5055	0.3252	0.3490	<u>0.5309</u>	0.4896	0.5422
REDS30	$E_{warp}^* \downarrow$	11.17	12.56	5.14	3.57	12.69	8.89	<u>2.83</u>	6.03	2.68
	PSNR↑	19.56	14.85	<u>20.85</u>	20.86	19.93	20.32	19.71	20.57	20.29
	SSIM↑	0.4862	0.2941	0.5469	0.5377	0.4261	0.5043	0.4315	0.5113	<u>0.5411</u>
	LPIPS↓	0.3376	0.5915	0.2899	<u>0.2597</u>	0.4422	0.3857	0.3443	0.2240	0.2804
OpenVid30	DOVER↑	0.3182	0.0600	0.3483	0.3400	0.2221	0.2519	0.2857	<u>0.3857</u>	0.4017
	$E_{warp}^* \downarrow$	19.1	18.00	8.32	6.06	17.40	22.14	15.65	12.28	<u>7.30</u>
	PSNR↑	24.62	21.14	24.63	26.21	24.29	24.91	24.41	24.73	<u>25.30</u>
	SSIM↑	0.7778	0.5887	0.7759	<u>0.8080</u>	0.6070	0.7633	0.7167	0.7686	0.8371
VideoLQ	LPIPS↓	0.1994	0.4207	0.2297	<u>0.1881</u>	0.3902	0.2102	0.2479	0.2074	0.1011
	DOVER↑	0.6992	0.1819	<u>0.7345</u>	0.7275	0.5435	0.6368	0.7201	0.7191	0.7393
	$E_{warp}^* \downarrow$	8.46	12.11	4.12	<u>2.52</u>	9.78	8.87	4.72	4.82	1.82
	ILNIQE↓	27.95	27.19	26.29	26.11	25.92	29.97	<u>24.49</u>	23.94	25.35
VideoLQ	DOVER↑	0.4967	0.3392	<u>0.5285</u>	0.4804	0.4113	0.4775	0.4833	0.5319	0.5431
	$E_{warp}^* \downarrow$	8.00	7.75	6.52	5.10	8.33	9.26	10.89	7.82	<u>6.38</u>

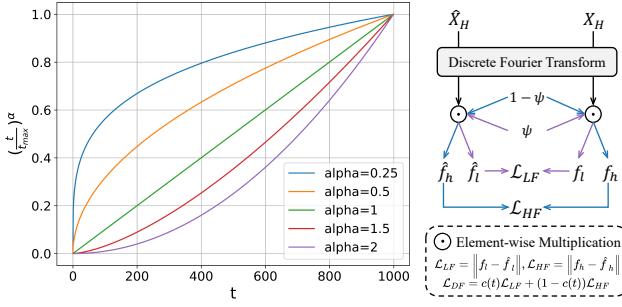


Figure 5. **Dynamic Frequency Loss.** **Left:** curves of weighting function $c(t)$ for different α . **Right:** details of DF loss.

tion that exploits this characteristic to decouple fidelity and simplify optimization? Specifically, we aim to guide the model to *prioritize low-frequency components in the early stages*, shifting focus to *high-frequency components later*.

Details of DF Loss. Here, we propose Dynamic Frequency Loss. Specifically, in each diffusion step t , we use the following equation to obtain the estimated \hat{Z}_H :

$$\hat{Z}_H = \sigma_t^{-1}(\alpha_t \epsilon - \phi_\theta(Z_t, c_{text}, c_l, t)). \quad (5)$$

Then, we use the decoder to convert the latent \hat{Z}_H back to the pixel space, resulting in \hat{X}_H . After that, we apply Discrete Fourier Transform (DFT) to transform \hat{X}_H into the frequency domain as shown in Figure 5. We predefine a low-frequency pass filter ψ to obtain the low- and high-frequency:

$$\hat{f}_l = \mathcal{F}(\hat{X}_H) \odot \psi, \quad \hat{f}_h = \mathcal{F}(\hat{X}_H) \odot (1 - \psi), \quad (6)$$

where $\mathcal{F}(\cdot)$ is DFT, \odot is element-wise multiplication. \hat{f}_l and \hat{f}_h denote the low and high frequency of \hat{X}_H . The pro-

Table 2. Training dataset comparison.

Method	Dataset	Size	#Frames	Resolution
UAV[75]	WebVid [2] + YouHQ [75]	335K+37K	-	336×596, 1080×1920
RealVifomer[73]	REDS [35]	300K	100	720×1280
Ours	OpenVid [36]	200K	32	720×1280

posed DF loss can be written as:

$$\mathcal{L}_{LF} = \|f_l - \hat{f}_l\|, \quad \mathcal{L}_{HF} = \|f_h - \hat{f}_h\|, \quad (7)$$

$$\mathcal{L}_{DF} = c(t)\mathcal{L}_{LF} + (1 - c(t))\mathcal{L}_{HF}, \quad (8)$$

where f_l / f_h stand for low- / high-frequency of X_H , respectively. $c(t) = (t/t_{max})^\alpha$ is the weighting function.

4. Experiments

4.1. Datasets and Implementation

Training Datasets. We train STAR using the subset of OpenVid-1M [36], containing $\sim 200K$ text-video pairs. The OpenVid-1M dataset is a high-quality video dataset consisting of over 1 million in-the-wild video clips with detailed captions, where the minimum resolution is 512×512 and the average length is 7.2 seconds. Utilizing this large-scale high-quality data for training further improves our model's restoration capacity for real-world VSR. More training dataset comparisons can be found in Table 2. We generate the LR-HR video pairs following the degradation strategy in Real-ESRGAN [51], combined with video compression operations, resulting in severe degradation similar to the approach used in RealBasicVSR [11].

Testing Datasets. We evaluate our method on both synthetic and real-world datasets. As for synthetic testing datasets, we follow the same degradation pipeline in training to generate LR videos from HR ones to construct three

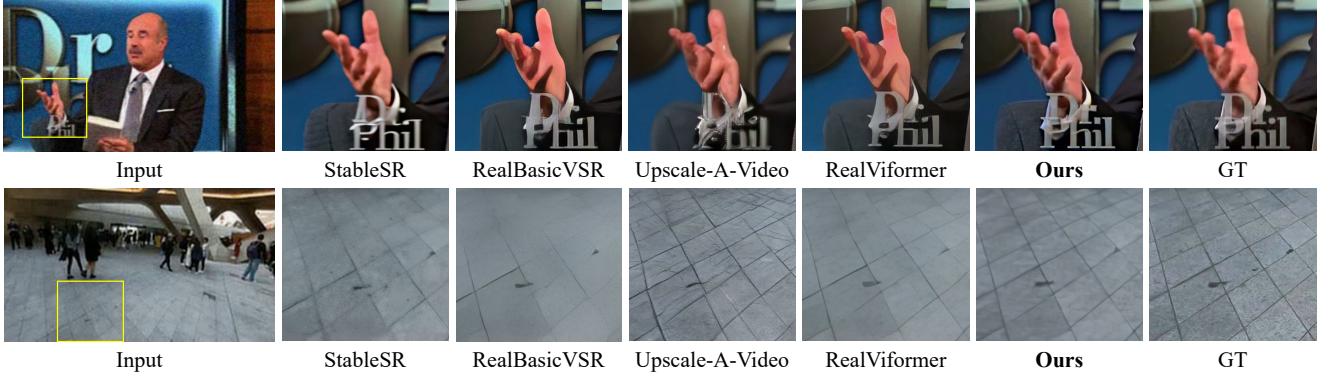


Figure 6. Qualitative comparisons on synthetic LR videos from OpenVid30 and REDS30[35]. (Zoom-in for best view)

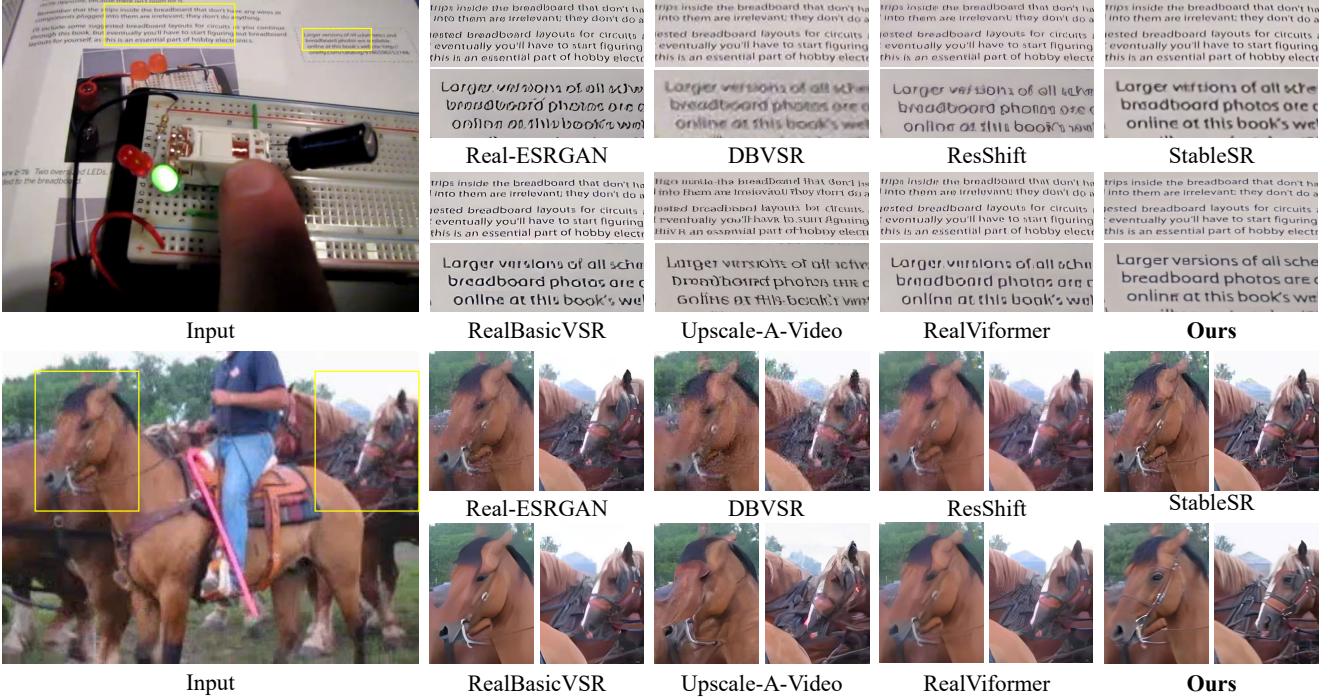


Figure 7. Qualitative comparisons on real-world test videos in VideoLQ [11] dataset. (Zoom-in for best view)

synthetic datasets (*i.e.*, UDM10 [65], REDS30 [35], and OpenVid30). The OpenVid30 is split from OpenVid-1M [36] ensuring no overlap with the training dataset and comprises the first approximately 100 frames of 30 videos. For the real-world dataset, we choose VideoLQ [11] which contains 50 videos, each with 100 frames.

Training Details. By default, we adopt I2VGen-XL [72] as our T2V backbone. For fast convergence, we initialize the model using the weights from VEnhancer [17]. We then train the ControlNet and inserted LIEM to adapt the T2V model for the real-world VSR task. Specifically, we train STAR on 8 NVIDIA A100-80G GPUs with 15K iterations and a batch size of 8. The training data is 720×1280

with 32 frames. We use AdamW [33] as the optimizer with a learning rate of 5e-5.

Evaluation Metrics. We adopt six metrics to evaluate the VSR outputs from several different perspectives: image fidelity (PSNR), perceptual similarity (SSIM [54], LPIPS [71]), quality (ILNIQE [69]), video clarity (DOVER [56]) and temporal consistency (E_{warp}^* [26, 32]). For synthetic datasets, we calculate PSNR, SSIM and LPIPS between the output and ground-truth frames, along with DOVER and flow warping error (*i.e.*, E_{warp}^*) of output videos. For real-world dataset, because of no ground-truth videos, we use three non-reference metrics: ILNIQE, DOVER, and E_{warp}^* .

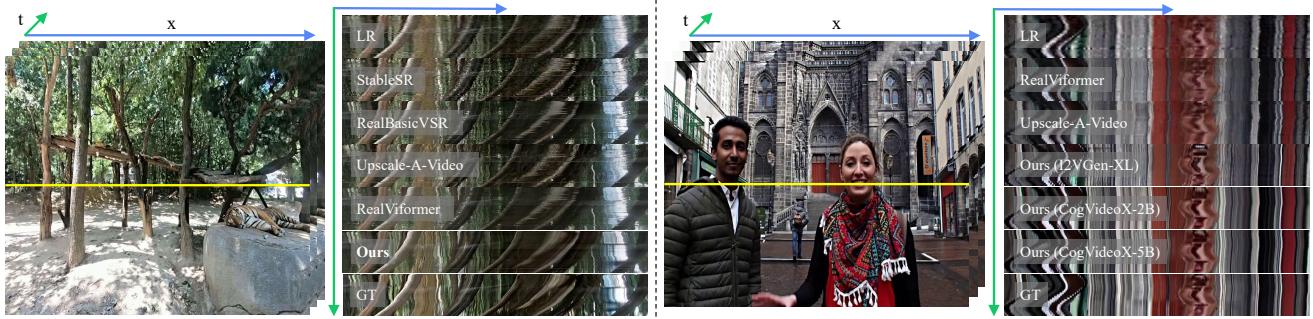


Figure 8. Qualitative comparisons on temporal consistency in REDS30 [35] and OpenVid dataset. (**Zoom-in for best view**)

4.2. Comparisons

To verify the effectiveness of our approach, we compare STAR with several state-of-the-art methods, including Real-ESRGAN [51], DBVSR [38], RealBasicVSR [11], RealVifomer [73], ResShift [68], StableSR [48], and Upscale-A-Video [75].

Quantitative Evaluation. As shown in Table 1, we calculate five metrics on each synthetic benchmark. Our STAR achieves the best scores in four out of these five metrics (SSIM, LPIPS, DOVER, and E_{warp}^*) on both UDM10 and OpenVid30 datasets, along with the second-best PSNR scores. This indicates that STAR can generate realistic details with good fidelity and robust temporal consistency. Moreover, we evaluate three non-reference metrics on a real-world dataset. On this dataset, STAR achieves the best score in DOVER and the second-best scores in IL-NIQE and E_{warp}^* . These results demonstrate that STAR can effectively restore real-world videos with high spatial and temporal quality. Additionally, our visual results on both real-world and synthetic datasets are preferred by human evaluators, as detailed in the User Study section (see Appendix).

Qualitative Evaluation. To intuitively demonstrate the effectiveness of the proposed STAR, we present visual results on both synthetic and real-world datasets in Figure 6 and 7, respectively. As shown, our STAR generates the most realistic spatial details and exhibits the best degradation removal capability. Specifically, the first example in Figure 7 illustrates that STAR reconstructs the text structure most effectively, thanks to the T2V prior efficiently capturing temporal information, and the DF loss that improves the fidelity. Furthermore, the T2V model has a strong spatial prior, which helps generate more realistic details and structures, such as the human hand in Figure 6 and the horse shape and fur in Figure 7.

We also compare the temporal consistency in Figure 8. As observed in the left of Figure 8, StableSR demonstrates the most temporal inconsistency, primarily because it is originally designed for image super-resolution. Although

Table 3. Ablation of LIEM position.

Position	Spa-Local	Temp-Local	PSNR↑	LPIPS↓	$E_{warp}^* \downarrow$
(i)	✓	✓	23.14	0.2015	2.83
			23.61	0.2013	2.82
			23.65	0.1945	2.92
			<u>23.69</u>	0.1943	<u>2.74</u>
(ii)	✓	✓	23.27	0.2363	3.57
(iii)			24.51	0.2094	1.99

RealBasicVSR, Upscale-A-Video, and RealVifomer incorporate optical flow maps to enhance temporal consistency, they still face challenges in generating consistent results under complex degraded video conditions, as the optical flow maps may not always be accurate. In contrast, our proposed STAR achieves the best temporal consistency, thanks to the powerful temporal prior inherent in the T2V model, which effectively helps reconstruct temporal information *even without the use of optical flow maps*.

4.3. Ablation Study

Local Information Enhancement Module. We primarily investigate the impact of introducing LIEM in different ways. First, we find that adding LIEM on both spatial and temporal blocks achieves the best results as shown in Table 3. Second, we consider three connection types as shown in Figure 9 (Left). From visual results in Figure 9 (Right) and quantitative results in Table 3, we find that position (i) achieves the best results. This phenomenon can be attributed to the fact that, with most weights frozen to preserve the prior, the newly added blocks can influence the model’s mapping process. However, the impact at positions (ii) and (iii) is too large, making it difficult for the model to fine-tune and adapt to this change, resulting in poor performance.

Dynamic Frequency Loss. First, we investigate the impact of different variants of frequency loss. As shown in Table 4, “Separate” indicates whether the frequency components are separated into high and low frequency, constraining them individually. “Type” refers to the specific

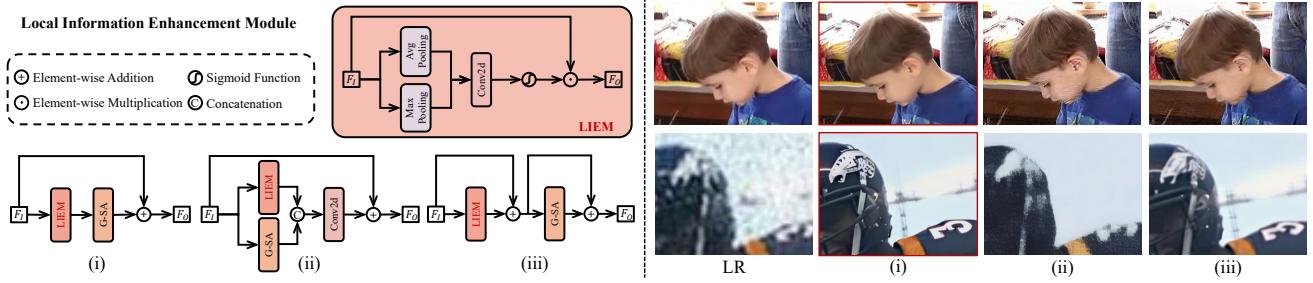


Figure 9. Ablation study about LIEM. **Left:** illustration of different insertion positions of LIEM and the structure of LIEM. **Right:** visual comparison on real-world and synthetic videos with different LIEM positions.

Table 4. Ablation of different variants of DF loss.

Separate	Type	PSNR↑	LPIPS↓	$E^*_{warp} \downarrow$
w/o Frequency Loss		23.69	0.1943	2.74
-	-	23.72	0.1941	2.71
✓	Inverse	23.67	0.1945	2.83
✓	Direct	23.85	0.1903	2.69

Table 5. Ablation of $b(t)$ and α in $c(t)$.

$b(t)$	α	PSNR↑	LPIPS↓	$E^*_{warp} \downarrow$
Linear	0.25	23.76	0.2030	2.72
	0.5	23.71	0.2010	2.75
	1	<u>23.85</u>	<u>0.1903</u>	<u>2.69</u>
	1.5	23.53	0.1928	2.81
Exponential	2	23.91	0.1885	2.61
		23.68	0.1990	2.78

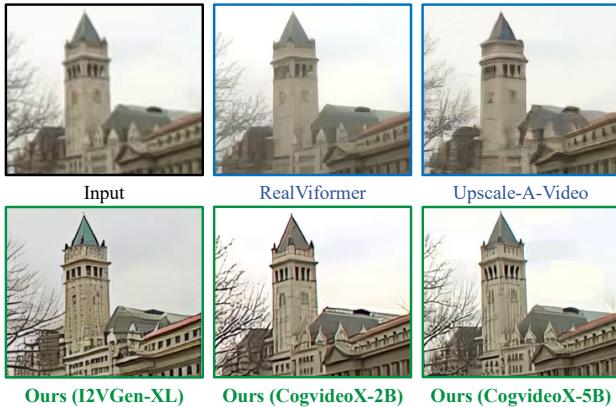


Figure 10. Illustration on scaling up with larger t2v models on a real-world low-quality video. (**Zoom-in for best view**)

definition of the DF loss: if set to “inverse,” a higher weight is given to high frequencies in the early stages and a lower weight to low frequencies; if set to “direct”, a higher weight is given to low frequencies initially and a lower weight to high frequencies, which is matching the analysis in Sec. 3.3.

Table 6. Effectiveness of T2V diffusion prior for real-world VSR.

Metrics	UAV	RealVifomer	Ours		
			I2VGen-XL	CogX-2B	CogX-5B
PSNR↑	22.46	22.90	21.46	<u>23.18</u>	23.60
SSIM↑	0.6552	0.6944	0.6715	<u>0.7112</u>	0.7400
LPIPS↓	0.2035	0.1823	0.1779	<u>0.1571</u>	0.1314
DOVER↑	0.6609	0.4286	<u>0.7267</u>	0.6955	0.7350
$E^*_{warp} \downarrow$	5.424	4.75	5.529	3.68	<u>4.56</u>

As observed, separating the frequency components and prioritizing low-frequency reconstruction early on yield the best perceptual quality while maintaining high fidelity. Second, we explore the optimal settings for $b(t)$ and α in $c(t)$. As shown in Table 5, using a linear form for $b(t)$ with $\alpha = 2$ for $c(t)$ yields the best results. Therefore, we adopt this DF loss configuration for training our model and comparing it with other state-of-the-art methods.

Scaling up with Larger T2V Models. To further validate the effectiveness of T2V diffusion priors for real-world VSR, we replace I2VGen-XL with larger DiT-based [39] T2V models (*i.e.*, CogVideoX [1, 64]), and evaluate results both quantitatively and qualitatively. Since CogVideoX only supports inputs at 480×720 resolution, we created a new test set by cropping 10 videos from OpenVid-1M [36] to this size. As shown in Table 6, the powerful CogVideoX models yield consistent improvements across all metrics. Notably, SSIM improves from 0.6944 to 0.7400, and DOVER increases from 0.6609 to 0.7350, marking a substantial enhancement in visual quality. The robust spatio-temporal priors in CogVideoX enable realistic details and clear building structures (Figure 10), while maintaining high temporal consistency (Figure 8 Right). Inspired by scaling law [18, 23] and our findings, we believe larger, more powerful T2V models will further advance VSR tasks.

5. Conclusion

In this paper, we present STAR, a real-world VSR framework that leverages T2V diffusion prior to restore videos with fewer artifacts, higher spatial fidelity, and stronger

temporal consistency. Specifically, we introduce a Local Information Enhancement Module into the original T2V backbone to improve its ability to handle degradations and reconstruct fine details. Additionally, we propose a Dynamic Frequency Loss that guides the model to focus on restoring different frequency components at each diffusion step, thereby enhancing fidelity. Furthermore, we demonstrate that a powerful T2V model can effectively generate high-quality results in both spatial and temporal dimensions. Extensive experiments show that STAR achieves superior performance in both spatial and temporal quality. We hope our work lays a solid foundation for applying T2V models in real-world VSR and inspires future advancements in the field.

References

- [1] Cogvideox-5b, 2024. <https://huggingface.co/THUDM/CogVideoX-5b>. 8
- [2] Max Bain, Arsha Nagrani, Gü̈l Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 5
- [3] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024. 2
- [4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2
- [5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023. 2
- [6] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 12
- [7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2
- [8] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, pages 4778–4787, 2017. 2
- [9] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*, pages 4947–4956, 2021. 2
- [10] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, pages 5972–5981, 2022. 2
- [11] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *CVPR*, pages 5962–5971, 2022. 2, 5, 6, 7, 12
- [12] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6441–6451, 2024. 2
- [13] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pages 13320–13331, 2024. 2
- [14] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In *CVPR*, pages 9232–9241, 2024. 2, 3
- [15] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 2
- [16] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, pages 3897–3906, 2019. 2
- [17] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024. 2, 3, 6
- [18] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. 8
- [19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3
- [20] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE TPAMI*, 40(4):1015–1028, 2017. 2
- [21] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *ECCV*, pages 645–660. Springer, 2020. 2
- [22] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, pages 3224–3232, 2018. 2
- [23] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec

- Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 8
- [24] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3
- [25] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *CVPR*, pages 766–776, 2022. 3
- [26] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 6
- [27] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*, pages 335–351. Springer, 2020. 2
- [28] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *NeurIPS*, 35:378–393, 2022. 2
- [29] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE TIP*, 2024. 2
- [30] Xinqi Lin, Jingwen He, Ziyuan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2
- [31] Yichao Liu, Zongru Shao, and Nico Hoffmann. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv preprint arXiv:2112.05561*, 2021. 3
- [32] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, pages 22139–22149, 2024. 6
- [33] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [34] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, pages 1045–1048. Makuhari, 2010. 2
- [35] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, pages 0–0, 2019. 5, 6, 7, 12
- [36] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 2, 5, 6, 8
- [37] OpenAI. Sora, 2024. <https://openai.com/index/sora>. 2
- [38] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *ICCV*, pages 4811–4820, 2021. 7
- [39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2, 8
- [40] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3
- [44] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, pages 6626–6634, 2018. 2
- [45] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3
- [46] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *NeurIPS*, 35:36081–36093, 2022. 2
- [47] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2
- [48] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, pages 1–21, 2024. 2, 7
- [49] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *arXiv preprint arXiv:2403.06098*, 2024. 2
- [50] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, pages 0–0, 2019. 2
- [51] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 2, 5, 7
- [52] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2, 3

- [53] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 2
- [54] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6
- [55] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 4
- [56] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 6
- [57] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pages 25456–25467, 2024. 2, 3, 4
- [58] Yanze Wu, Xintao Wang, Gen Li, and Ying Shan. Animesr: Learning real-world super-resolution models for animation videos. *NeurIPS*, 35:11241–11252, 2022. 2
- [59] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *CVPR*, pages 6388–6397, 2021. 2
- [60] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 127:1106–1125, 2019. 2
- [61] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 2, 3
- [62] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *ICCV*, pages 4781–4790, 2021. 2
- [63] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. 2024. 2, 3, 12
- [64] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3, 8
- [65] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, pages 3106–3115, 2019. 2, 6
- [66] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, pages 25669–25680, 2024. 4
- [67] Xin Yuan, Jinoo Baek, Keyang Xu, Omer Tov, and Hongliang Fei. Inflation with diffusion: Efficient temporal adaptation for text-to-video super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 489–496, 2024. 2, 3
- [68] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 36, 2024. 7
- [69] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE TIP*, 24(8):2579–2591, 2015. 6
- [70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3
- [71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6
- [72] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2, 3, 6
- [73] Yuehan Zhang and Angela Yao. Realviforner: Investigating attention for real-world video super-resolution. *ECCV*, 2024. 1, 2, 5, 7, 12
- [74] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8281–8291, 2024. 2
- [75] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *CVPR*, pages 2535–2545, 2024. 1, 2, 3, 5, 7, 12

A. Perception-Distortion Trade-Off

The trade-off between perception and distortion [6] is a widely recognized challenge in the super-resolution domain. Thanks to our *DF Loss*, our method can easily control the model to favor either fidelity or perceptual quality in the generated results. We can adjust the hyper-parameter β in the $b(t)$ to achieve this goal. The total loss in our STAR is:

$$\mathcal{L}_{total} = \mathcal{L}_v + b(t)\mathcal{L}_{DF}, \quad (9)$$

The $b(t)$ can be written as follows:

$$b(t) = \beta \cdot \left(1 - \frac{t}{t_{max}}\right), \quad (10)$$

Where t is the timestep and β is the hyper-parameter that adjusts the weight between \mathcal{L}_v and \mathcal{L}_{DF} , which we set to 1 by default. From equations (1) and (2), we can observe that a larger β increases the weight of the DF loss at each timestep, thereby further enhancing the fidelity of the results. In contrast, a smaller β reduces the influence of the DF loss at each timestep, allowing the v-prediction loss to have a greater impact and produce more perceptual results. The $b(t)$ - t curves under different β are shown in Figure 11.

We conduct experiments under these settings to demonstrate the ability to achieve the perception-distortion trade-off. The quantitative results are shown in Table 7. From Table 7, we can observe that increasing β improves the PSNR and E_{warp}^* , leading to better fidelity. Conversely, decreasing β reduces the LPIPS score, indicating better perceptual quality.

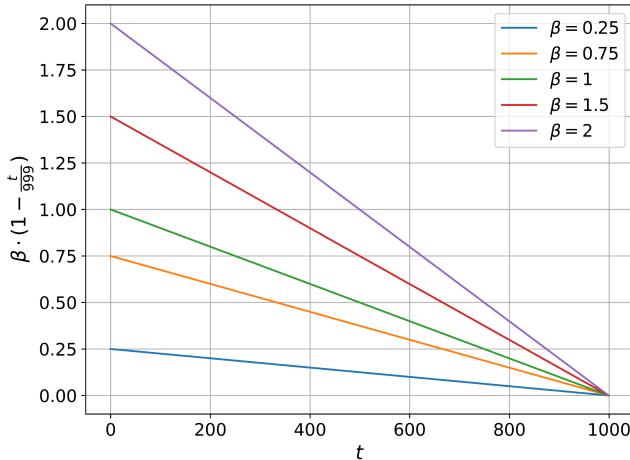


Figure 11. Ablation on $b(t)$. Higher hyper-parameter β produces results with greater fidelity, while lower β emphasizes more perceptual quality.

Table 7. Qualitative comparison under different β of $b(t)$.

β	PSNR↑	LPIPS↓	$E_{warp}^* \downarrow$
0.25	23.55	0.1825	2.88
0.75	23.76	0.1842	2.74
1.0	23.91	0.1885	2.68
1.5	24.08	0.2272	2.53
2.0	24.41	0.3339	2.21

B. More Results

B.1. User Study

To find the human-preferred results between our STAR and other state-of-the-art methods, we conduct a user study that evaluate the results on both real-world and synthetic datasets. Specifically, we use the real-world dataset VideoLQ [11] and the synthetic dataset REDS30 [35]. We select two image-diffusion-model-based methods, Upscale-A-Video [75] and MGLD-VSR [63]; and one GAN-based method, RealVformer [73] for comparison. We invite 12 evaluators to participate in the user study. For each evaluator, we randomly select 10 videos from each dataset and present four results: one from our STAR and three from the compared methods. The evaluators were asked to choose which result had the best visual quality and temporal consistency. The results of the user study are depicted in Figure 12, indicating that our STAR is preferred by most human evaluators for both visual quality and temporal consistency.

B.2. Qualitative Comparisons

We provide more visual comparisons on synthetic and real-world datasets in Figure 13 and Figure 14 to further highlight our advantages in spatial quality. These results clearly demonstrate that our method preserves richer details and achieves greater realism. To demonstrate the impact of scaling up with larger text-to-video (T2V) models, we present additional results in Figure 15. It is evident that scaling up the T2V model further improves the restoration effect, indicating that a large and robust T2V model can serve as a strong base model for video super-resolution.

B.3. Video Demo

We provide a demo video [STAR-demo.mp4] in the supplementary material, showcasing the temporal and spatial advantages of our proposed STAR more intuitively. This video includes additional results and comparisons on synthetic, real-world, and AIGC videos.

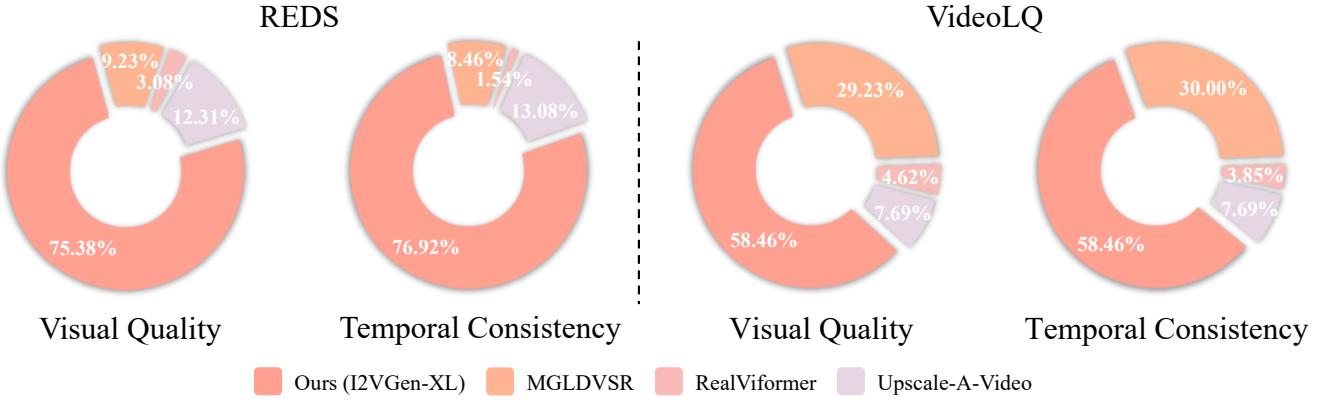


Figure 12. User study results. Our STAR is preferred by human evaluators for both visual quality and temporal consistency.

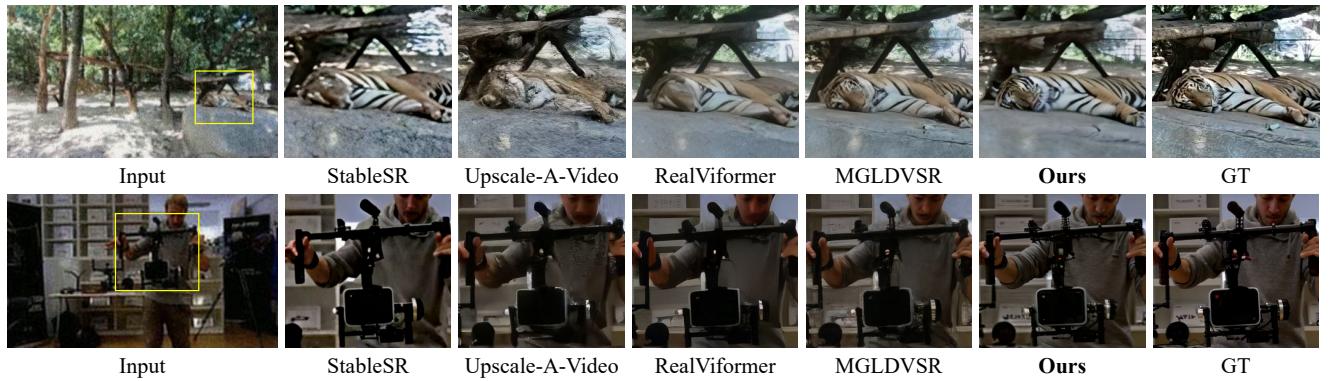


Figure 13. Qualitative comparisons on synthetic datasets. Our STAR generates more detailed and realistic results. (**Zoom-in for best view**)



Figure 14. Qualitative comparisons on real-world datasets. Our STAR produces the clearest facial details and the most accurate text structure. (**Zoom-in for best view**)



Figure 15. Qualitative comparisons on synthetic and real-world datasets with larger T2V models. Scaling up the T2V model enhances detail and realism in video super-resolution results. (**Zoom-in for best view**)