Enhancing Perceptual Quality in Video Super-Resolution through Temporally-Consistent Detail Synthesis using Diffusion Models - Supplementary Material -

Claudio Rota[1], Marco Buzzelli[1], and Joost van de Weijer[2]

[1] University of Milano-Bicocca, Milan, Italy
{claudio.rota, marco.buzzelli}@unimib.it
[2] Universitat Autònoma de Barcelona, Barcelona, Spain
joost@cvc.uab.es

This supplementary file provides additional details that were not included in the main paper due to page limitations. Demo videos are available on the project page[3].

1 Additional methodology details

1.1 Description of the pre-trained LDM for SISR

The proposed StableVSR is built upon a pre-trained Latent Diffusion Model (LDM) for single image super-resolution (SISR). We use Stable Diffusion $\times 4$ Upscaler (SD$\times$4Upscaler)[4]. It follows the LDM framework [14], which performs the iterative refinement process into a latent space and uses the VAE decoder $\mathcal{D}$ [7] to decode latents into RGB images. Starting from a low-resolution RGB image LR (conditioning image) and an initial noisy latent x_T, the denoising UNet ϵ_θ is used to generate the high-resolution counterpart via an iterative refinement process. In this process, noise is progressively removed from x_t guided by LR. After a defined number of sampling steps, the obtained latent x_0 is decoded using the VAE decoder $\mathcal{D}$ [7] into a high-resolution RGB image $\overline{\text{HR}}$. The obtained image $\overline{\text{HR}}$ has a $\times 4$ higher resolution than the low-resolution image LR, as $\mathcal{D}$ performs $\times 4$ upscaling. In practice, the low-resolution RGB image LR and the initial noisy latent x_T are concatenated along the channel dimension and inputted to the denoising UNet.

1.2 Bidirectional information propagation in the Frame-wise Bidirectional Sampling strategy

We show in Figure 1 a graphical representation of the proposed Frame-wise Bidirectional Sampling strategy to better show the bidirectional information propagation. We take a sampling step t in video time $i = 1, ..., N$ before moving to the

[3] https://github.com/claudiom4sir/StableVSR
[4] https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler

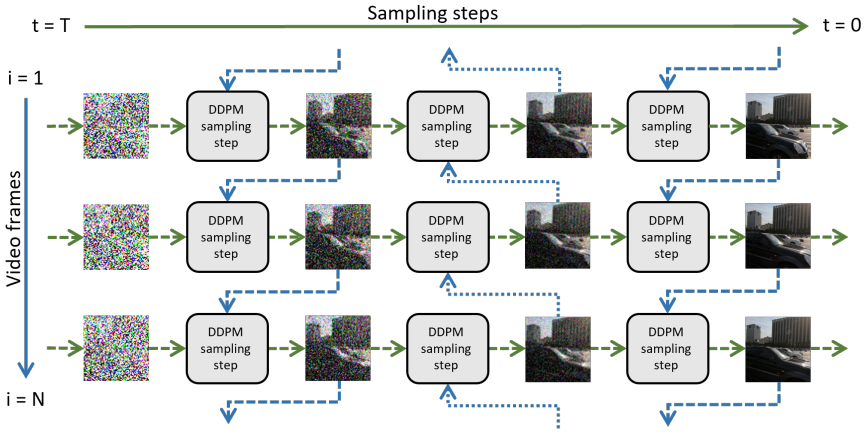


Fig. 1: Graphical representation of the proposed Frame-wise Bidirectional Sampling strategy. The green flow propagates information forward in sampling time while the blue flow alternately propagates it forward and backward in video time. Forward propagation is shown with dashed lines, while backward propagation with dotted lines.

next sampling step $t-1$. At every sampling step, we invert the video time order for processing: from $i = 1, ..., N$ to $i = N, ..., 1$. For the generation of x^i_{t-1}, we start from $\tilde{x}^{i-1}_0$ and x^i_t. Since $\tilde{x}^{i-1}_0$ is related to the previous frame, it provides information from the past. In addition, since x^i_t is generated starting from $\tilde{x}^{i+1}_0$ and x^i_{t+1}, it contains information from future frames, which is implicitly propagated to the current sampling step. As a consequence, x^i_{t-1} benefits from past information from $\tilde{x}^{i-1}_0$ due to the forward direction of the current sampling step, and future information from x^i_t due to the backward direction of the previous sampling step.

2 Additional experiments

2.1 Architecture details

We report the StableVSR architecture details in Table 1. We can identify three main components: denoising UNet, Temporal Conditioning Module (TCM), and VAE decoder [7]. Following ControlNet [21], we freeze the weights of the denoising UNet during training. We only train TCM for video adaptation. We apply spatial guidance on the low-resolution frame via concatenation, *i.e.* the noisy latent x^i_t (4 channels) is directly concatenated with the low-resolution frame LR^i (3 channels) along the channel dimension. The temporal guidance is instead provided via TCM, which receives Temporal Texture Guidance $\widetilde{\text{HR}}^{i-1\rightarrow i}$ as input (3 channels). Once the iterative refinement process is complete, the VAE decoder $\mathcal{D}$ [7] receives the final latent of a frame i as input, *i.e.* x^i_0, and converts it into an RGB frame. This latent-to-RGB conversion applies $\times 4$ upscaling, hence

Table 1: Architecture details of StableVSR.

	Denoising UNet	Temporal Conditioning Module	VAE decoder
Downscaling	×8	×8	-
Upscaling	×8	-	×4
Input channels	7	3	4
Output channels	4	-	3
Trainable	No	Yes	No
Parameters	473 M	207 M	32 M

Table 2: Additional quantitative comparison with state-of-art methods for VSR using no-reference perceptual metrics. Best results in bold text. Almost all the metrics highlight the proposed StableVSR achieves better perceptual quality.

Method	Vimeo-90K-T			REDS4		
	MUSIQ↑	CLIP-IQA↑	NIQE↓	MUSIQ↑	CLIP-IQA↑	NIQE↓
Bicubic	23.27	0.358	8.44	26.89	0.304	6.85
ToFlow	40.79	0.364	8.05	-	-	-
EDVR	-	-	-	65.44	0.367	4.15
TDAN	46.54	0.386	7.34	-	-	-
MuCAN	49.84	0.379	7.22	64.85	0.362	4.30
BasicVSR	48.97	0.376	7.27	65.74	0.371	4.06
BasicVSR++	50.11	0.383	7.12	67.00	0.381	3.87
RVRT	50.45	0.387	7.12	67.44	0.392	3.78
RealBasicVSR	-	-	-	67.03	0.374	**2.53**
StableVSR (ours)	**50.97**	**0.414**	**5.99**	**67.54**	**0.417**	2.73

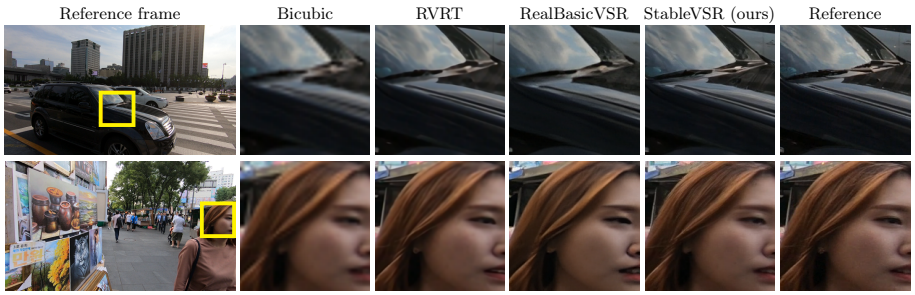
the output of the decoder represents the upscaled frame. The overall number of parameters in StableVSR (including the VAE decoder [7]) is about 712 million.

2.2 Additional comparison with state-of-the-art methods

As in the main paper, we compare the proposed StableVSR with ToFlow [20], EDVR [17], TDAN [15], MuCAN [10], BasicVSR [2], BasicVSR++ [3], RVRT [11], and RealBasicVSR [4].

Frame quality results. We report additional results using no-reference perceptual quality metrics, including MUSIQ [9], CLIP-IQA [16] and NIQE [12]. The results are reported in Table 2. All the metrics highlight the proposed StableVSR achieves superior perceptual quality. The only exception is NIQE [12] on REDS4 [13], which indicates StableVSR achieves the second-best results. We show in Figure 2 an additional qualitative comparison with BasicVSR++ [3] and RVRT [11] on Vimeo-90K-T [20] (Figure 2a) and with RVRT [11] and RealBasicVSR++ [4] on REDS4 [13] (Figure 2b). We can observe the proposed StableVSR is the only method that correctly upscales complex textures while the other methods fail, producing blurred results.

(a) Results on Vimeo-90K-T.



(b) Results on REDS4.

Fig. 2: Additional qualitative comparison with state-of-the-art methods for VSR. Only the proposed StableVSR correctly upscales complex textures.

Table 3: Comparison with the DM video baseline. Perceptual metrics are marked with ⋆, reconstruction metrics with ⋄, and temporal consistency metrics with •. Best results in bold text. The proposed StableVSR achieves better results in terms of frame quality and temporal consistency. Results computed on center crops of 512×512 resolution of REDS4.

Method	tLP•↓	tOF•↓	LPIPS⋆↓	DISTS⋆↓	PSNR⋄↑	SSIM⋄↑
Video baseline	13.08	2.92	0.113	0.075	26.27	0.771
StableVSR (ours)	**6.16**	**2.84**	**0.095**	**0.067**	**27.14**	**0.799**

Temporal consistency results. We can qualitatively assess the temporal consistency aspect of the proposed StableVSR in the demo videos. We compare StableVSR with SD×4Upscaler, which represents the baseline model used by StableVSR, and RealBasicVSR [4], which represents the second-best method on REDS4 [13] in terms of temporal consistency.

2.3 Comparison with the DM video baseline

We compare the proposed StableVSR with a DM video baseline containing 3D convolutions and temporal attention. Starting from the same pre-trained DM for SISR we use in StableVSR, i.e. SD×4Upscaler, we implement the video baseline by introducing a temporal layer (3D convolutions + temporal attention) after each pre-trained spatial layer, as done in previous video generation methods [1, 8, 19]. For training, we set the temporal window size to 5 consecutive frames and use the same training settings as in StableVSR. The only difference is the batch size, which is set to 8 instead of 32 due to memory constraints. We freeze the spatial layers and only train the temporal layers. Table 3 reports the results, where we can see the proposed StableVSR achieves better performance in both frame quality and temporal consistency. We attribute the lower performance of the DM video baseline to the limited temporal view, the inability to capture fine-detail image information, and the lack of proper frame alignment. StableVSR does not suffer from these problems, achieving better results.

2.4 Additional ablation study

Temporal Texture Guidance. In Figure 3, we provide additional results related to the ablation study on the Temporal Texture Guidance. We can observe that only the proposed design for the Temporal Texture Guidance ensures temporal consistency at the fine-detail level over time.

2.5 Impact of sampling steps

We study how the performance changes as the number of sampling steps varies. Figure 4 shows the results obtained by increasing the number of sampling steps

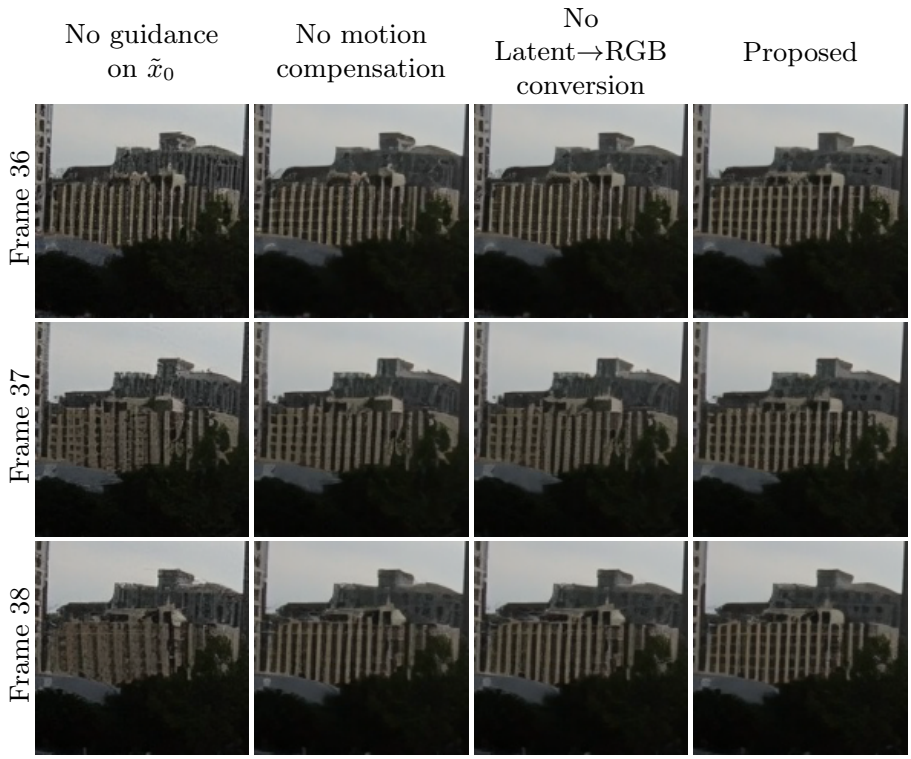


Fig. 3: Additional ablation experiments for the Temporal Texture Guidance. We show the results obtained on three consecutive frames. Only the proposed solution ensures temporal consistency at the fine-detail level over time. Results on sequence 015 of REDS4.

from 10 to 100. Reconstruction quality metrics, i.e. PSNR and SSIM [18], deteriorate with more sampling steps. Conversely, perceptual quality metrics, i.e. LPIPS [22], DISTS [6], MUSIQ [9], CLIP-IQA [16], NIQE [12], improve. We can attribute this behavior to the iterative refinement process of DMs, which progressively refines realistic image details that may not be perfectly aligned with the reference. We can observe the temporal consistency metric tLP [5] reaches the best value using 30 steps, while tOF [5] values are better as the number of sampling steps increases. According to these results, 50 sampling steps represent a good balance between perceptual quality and temporal consistency.

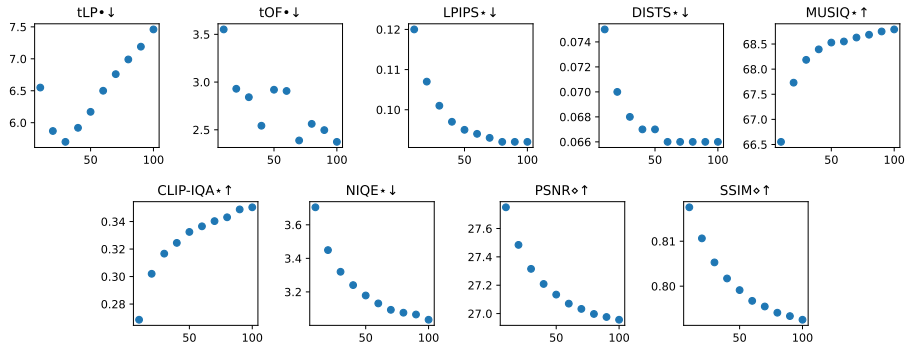


Fig. 4: Performance changes as the number of sampling steps varies. The x axis represents sampling steps, while the y axis metric values. Perceptual metrics are marked with $\star$, reconstruction metrics with $\diamond$, and temporal consistency metrics with $\bullet$. Increasing the sampling steps improves perceptual quality while deteriorating reconstruction quality. Results computed on center crops of 512×512 resolution of REDS4.

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
2. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4947–4956 (2021)
3. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5972–5981 (2022)
4. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5962–5971 (2022)
5. Chu, M., Xie, Y., Mayer, J., Leal-Taixé, L., Thuerey, N.: Learning temporal coherence via self-supervision for gan-based video generation. ACM Transactions on Graphics **39**(4), 75–1 (2020)
6. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2567–2581 (2020)
7. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
8. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022)

9. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
10. Li, W., Tao, X., Guo, T., Qi, L., Lu, J., Jia, J.: Mucan: Multi-correspondence aggregation network for video super-resolution. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 335–351. Springer (2020)
11. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems **35**, 378–393 (2022)
12. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
13. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1996–2005 (2019)
14. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
15. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3360–3369 (2020)
16. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
17. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops. pp. 1954–1963 (2019)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
19. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
20. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**, 1106–1125 (2019)
21. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
22. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)