

SeedVR: Seeding Infinity in Diffusion Transformer Towards Generic Video Restoration

Supplementary Material

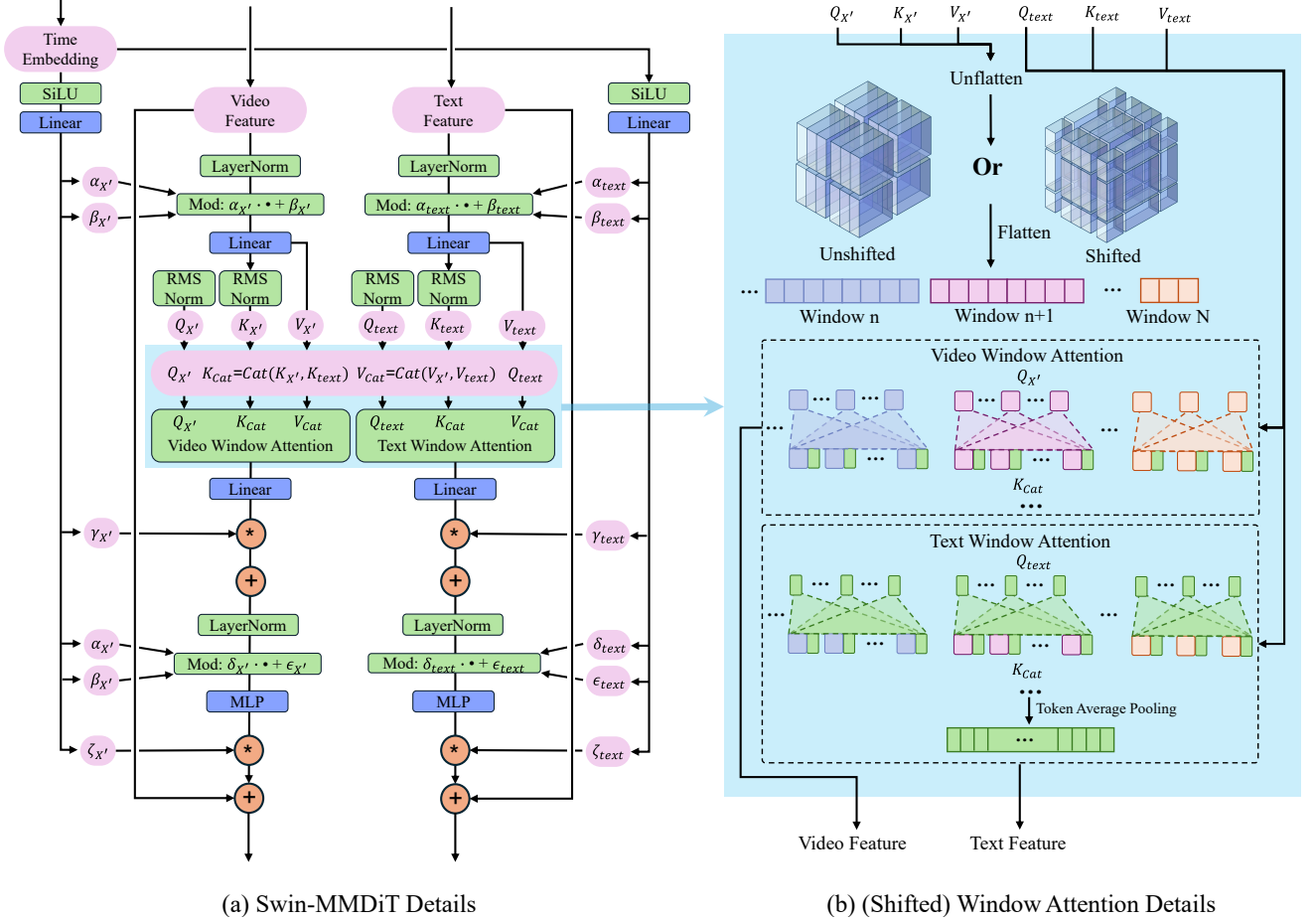


Figure 5. Detailed architecture of Swin-MMDiT and the illustration of shifted window mechanism inside Swin-MMDiT. By introducing the proposed shifted window mechanism into the transformer block, our approach is capable of overcoming the resolution constraint of vanilla attention. We further adopt large attention windows around the center and variable-sized windows near the boundary, enabling long-range dependency capturing given inputs of any length and size.

6. Video Demos

We have added some video demos in the supplementary material for visualization. For more video demos generated by our SeedVR, please refer to the project page: <https://iceclear.github.io/projects/seedvr/> for details.

In the video demo, we mainly compare SeedVR with two state-of-the-art VR approaches [20, 64] due to the space limit. Specifically, MGLD-VSR [64] achieves high metric performance compared with other existing approaches in Table 1. And VEnhancer [20] demonstrates superior enhance-

ment performance compared with other existing baselines on AIGC videos. We adopt its latest checkpoint, i.e., venhancer_v2.pth for comparison. As shown in the demo, our SeedVR outperforms these two methods by a large margin on real-world videos.

7. Detailed Architecture of Swin-MMDiT

As mentioned in Sec. 3.1, our Swin-MMDiT enhances the vanilla MMDiT [17] via introducing a shifted window mechanism. Here, we further illustrate the detailed architecture in Figure 4a as well as the window attention mechanism in Figure 4b.

Our design follows the principle of partitioning the vanilla full attention into window attention to overcome the resolution constraints. Hence, we apply two window attentions for video and text features, respectively.

For video features, we adopt shifted large window, *i.e.* $5 \times 64 \times 64$ to achieve long-range dependencies and introduce additional *key* and *value* from text features for text guidance. To keep the flexibility for arbitrary resolutions, these text features interact with each video window feature. And a 3D rotary position embedding [48] is further adopted as a relative positional embedding in each window. Such a flexible position embedding makes the model aware of varying-sized windows, bypassing the need to adjust the input resolution to be multiples of the window size.

For text features, directly making attention with the whole video features will bring the resolution constraint. Instead, we assume the windows from the same video share the same text. Thus, the text features are first repeated to interact with each window. Then, the repeated features are combined via average pooling to maintain the same sequence length.

With a careful design of the window mechanism inside the Swin-MMDiT, our SeedVR is capable of handling any video input effectively and efficiently.