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

Supplementary Material

1. Model Architecture

Base Model. The architecture of our base model is shown in Fig. 1. Our PatchVSR is built upon a pre-trained DiT-based video diffusion model, which comprises four main components: spatial self-attention (SSA), spatial cross-attention (SCA), temporal self-attention (TSA) and feed-forward network (FFN). Text prompts, time step and other micro conditions (aspect ratio, FPS, etc) are injected via the modulation mechanism [5]. The pre-trained model is trained on 77-frames 512×512 high-quality video data with diverse aspect ratio using NaViT [2]. During training, we additionally fine-tune the base model using a LoRA module for all four components. We have also introduced global cross-attention (GCA) modules in SCA blocks to fuse global context feature with the backbone feature.

Patch Condition Branch. The patch condition branch comprises several transformer blocks, which shares the same architecture as base model. After projected to the latent space by 3D VAE, the latent is diffused with time step 250 (where 1000 step denotes pure noises in our noise scheduling) before input to the adapter in each denoise time step during inference. The parameters of patch condition branch are initialized from base model and the layer indices are chosen at equal intervals following [4].

Global Context Branch. To provide initialization weight, we also follow the structure of the pre-trained model and sample the layer indices from it at equal intervals. To obtain high-level global semantic token, we design a transformer-based encoder architecture to compress the token number layer to one quarter layer by layer. The transformer blocks are evenly split for each token level. The token number is reduced for the output of the transformer blocks using the patchify operation [5]. The input to the global context branch is the latent of the resized full video encoded by the 3D VAE of the base model. We also diffuse the latent with time step 150 for noise augmentation. The choice of time step comes from a fact that the noise added to the full video will have a greater impact on local areas, and we set a smaller noise to maintain the noise intensity received by the corresponding local areas in patch condition branch and global context branch. Before input to the transformer blocks, the binary map is concatenated with the latent at the channel dimension, which indicates the target patch location within it.

2. Additional Results

2.1. Visual Comparisons on 2K Full Videos

AIGC videos. The comparisons of 2K AIGC videos are shown in Fig. 2. The results show that our method can effectively generalize to AI-generated videos, generating rich local details and textures while guaranteeing high fidelity.

Real-world Videos To adapt our model for real-world degraded videos, we further finetuned our model on a subset of our training data that involves degradation simulation pipeline like [1]. The qualitative comparison and quantitative evaluation are shown in Fig. 3 and Tab. 1 respectively.

2.2. Quantitative Comparisons on VideoScore

As evaluation by additional benchmark VideoScore [3], Tab. 2 tells the best performance of PatchVSR on *temporal consistency* and *dynamic degree* among all compared methods.

2.3. Visualization on 4K Full Videos

We have also performed 4K video super-resolution using 720P input video. Results can be seen in Fig. 4. It can be seen that our method successfully generate 4K-resolution video with high clarity and fine details.

2.4. Ablation Study

Qualitative Comparisons. The visual results of the ablation studies are shown in Fig. 5. While the use of local condition branch provides basic low-frequency information such as region structure and color, it lacks the ability to generate accurate high-frequency details. Besides, the lack of guidance from location embedding can make the global semantic tokens provided by global context branch less accurate, preventing the model from generating further high-definition details and textures. We also observe that removing LoRA makes the local high-frequency details of the generated results missing and inaccurate, consistent with our view that adding LoRA can facilitate the matching of pre-trained model distribution with patch video distribution.

Fidelity Evaluation. We provide the comparison in Tab. 3. It shows that, although introducing global branch brings slight fidelity degradation, using location embedding will partly mitigate it. Also, using global prompt only will lead to the mismatch between global context and video patch, which severely affect the performance.

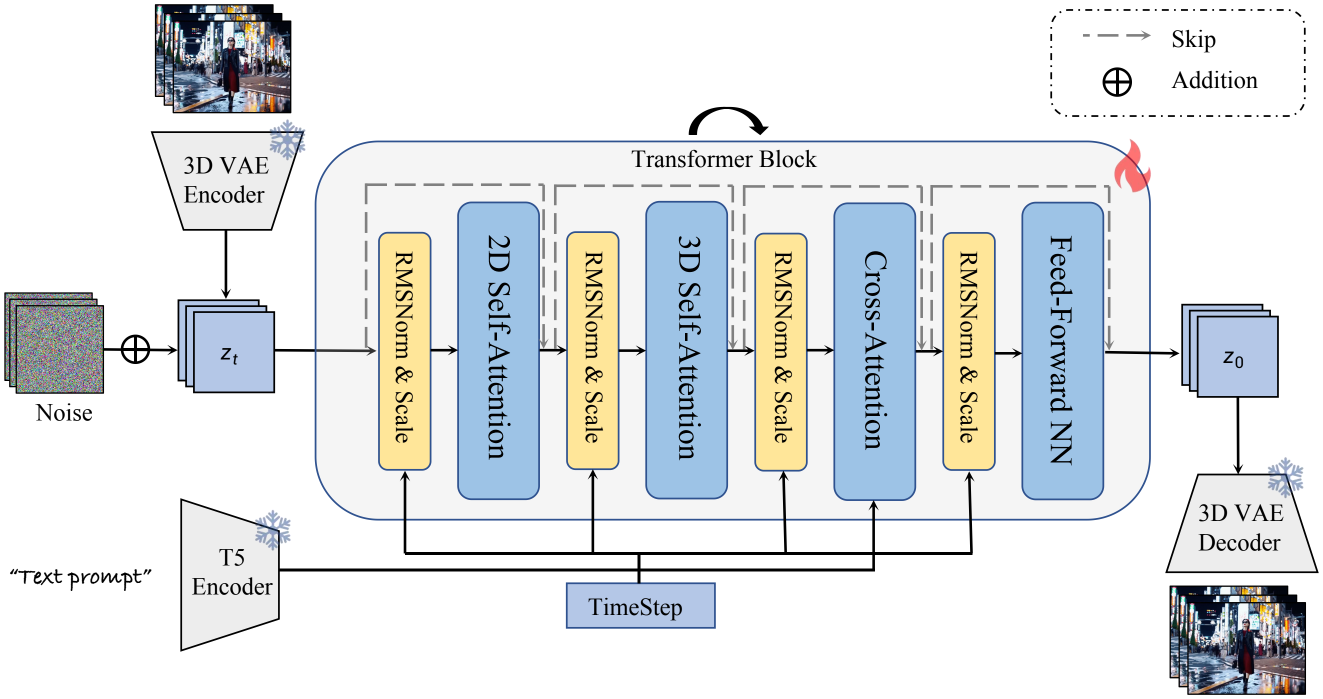

Figure 1. An overview of the architecture of our base model.

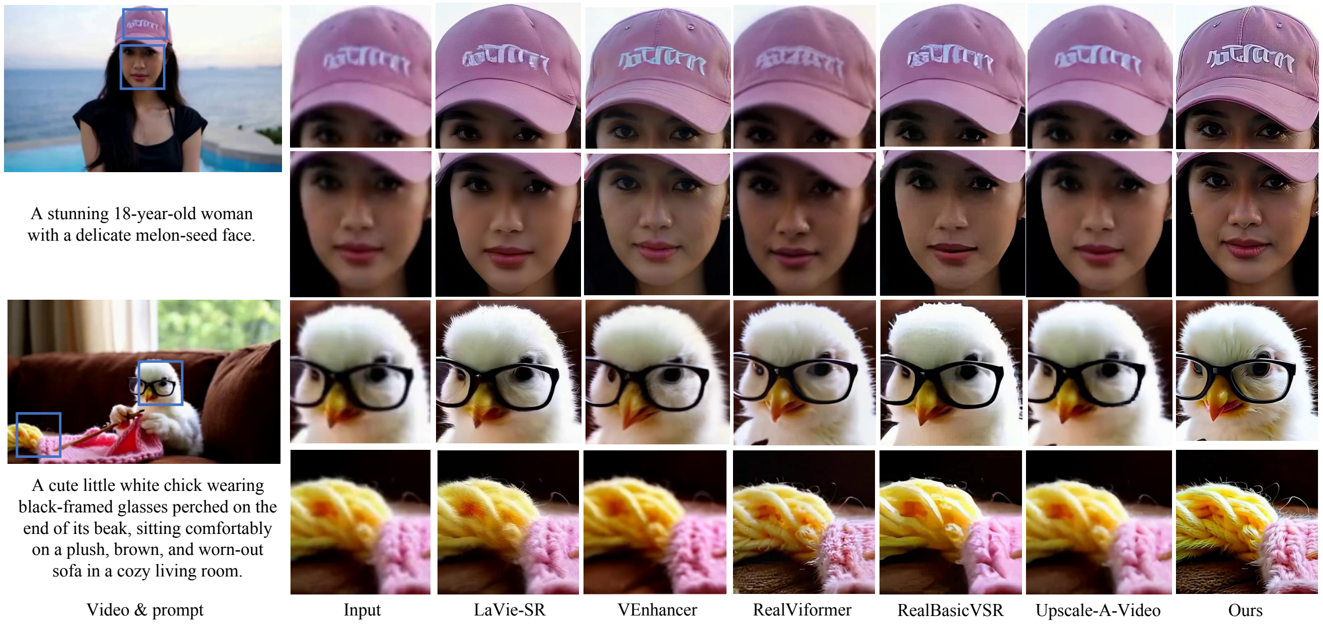
A stunning 18-year-old woman with a delicate melon-seed face.
A cute little white chick wearing black-framed glasses perched on the end of its beak, sitting comfortably on a plush, brown, and worn-out sofa in a cozy living room.
Video & prompt
Input
LaVie-SR
VEnhancer
RealViformer
RealBasicVSR
Upscale-A-Video
Ours

Figure 2. Comparisons of 2K full videos on VideoGen30 (test). **Zoom in for best view.**

2.5. Stitched Schemes

Although original multi-patch modulation successfully solves the seam problem of 2K full videos, it inevitably occurs black holes due to feature mismatch at the intersection of multiple mismatches. To tackle this problem, we further propose a weighted multi-patch modulation technique to obtain smooth feature. Specifically, for the four pixels adjacent to the black hole, their features in latent space all originate from three patches (including one primary patch and two auxiliary patches). When a uniform mixing strategy is applied to the features of the auxiliary patches, any

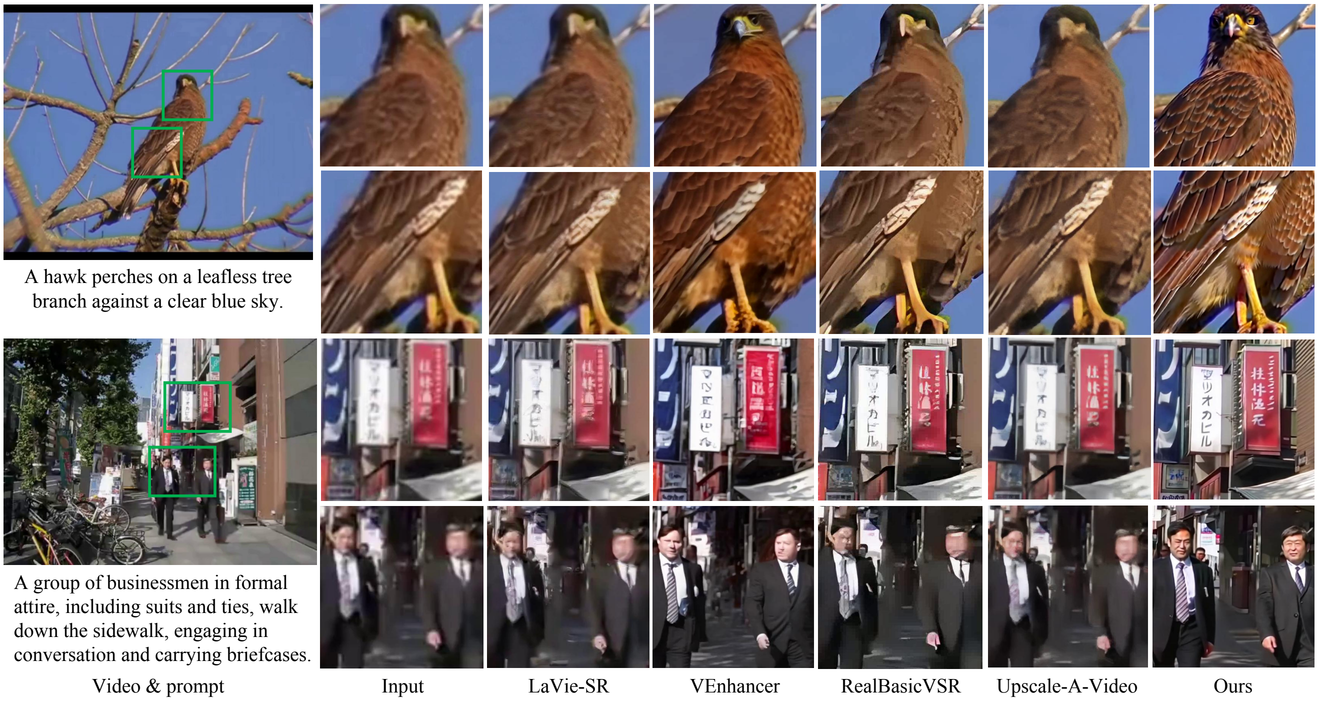
A hawk perches on a leafless tree branch against a clear blue sky.
A group of businessmen in formal attire, including suits and ties, walk down the sidewalk, engaging in conversation and carrying briefcases.
Video & prompt
Input
LaVie-SR
VEnhancer
RealBasicVSR
Upscale-A-Video
Ours

Figure 3. Qualitative comparisons on VideoLQ test set.

Metrics	RealBasicVSR	LaVie-SR	Upscale-A-Video	VEnhancer	Ours
DOVER ↑	0.497	0.371	0.350	0.411	**0.501**
MUSIQ ↑	**52.074**	36.204	24.599	35.157	37.342
Aesthetics ↑	0.531	0.521	0.526	0.522	**0.549**

Table 1. Quantitative results on VideoLQ test set.

two of these four points have two patches of different origins, leading to a severe black hole phenomenon. After applying the linear blending strategy, only one patch source is different while the other different patch source has a weight of 0 here, thus attenuating the black hole phenomenon in the stitched video. However, due to the feature mismatch of the primary patch, the black hole still occurs. To eliminate the black hole, we manually set larger weight for the feature of the shared patch (one of the auxiliary patches) to amplify its feature, which results in smooth feature transition, thus preventing from generating black holes. The comparisons of different stitch schemes are shown in Fig. 6. The results show that when the blending weight is set to 5.0, the seam position transitions smoothly and the black holes are effectively removed for all examples.

References

[1] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 1

[2] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[3] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 1

[4] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024. 1

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

Metrics	RealBasicVSR	LaVie-SR	Upscale-A-Video	VEnhancer	Ours
Visual Quality ↑	2.669	2.749	2.685	2.686	**2.811**
Temporal Consistency ↑	2.669	2.848	2.818	2.844	**2.897**
Dynamic Degree ↑	2.481	2.488	2.392	2.380	**2.619**
Text-to-video Alignment ↑	2.506	**2.598**	2.541	2.563	2.490
Factual Consistency ↑	2.613	2.703	2.673	2.686	**2.768**

Table 2. VideoScore evaluation on VideoGen30 test set.

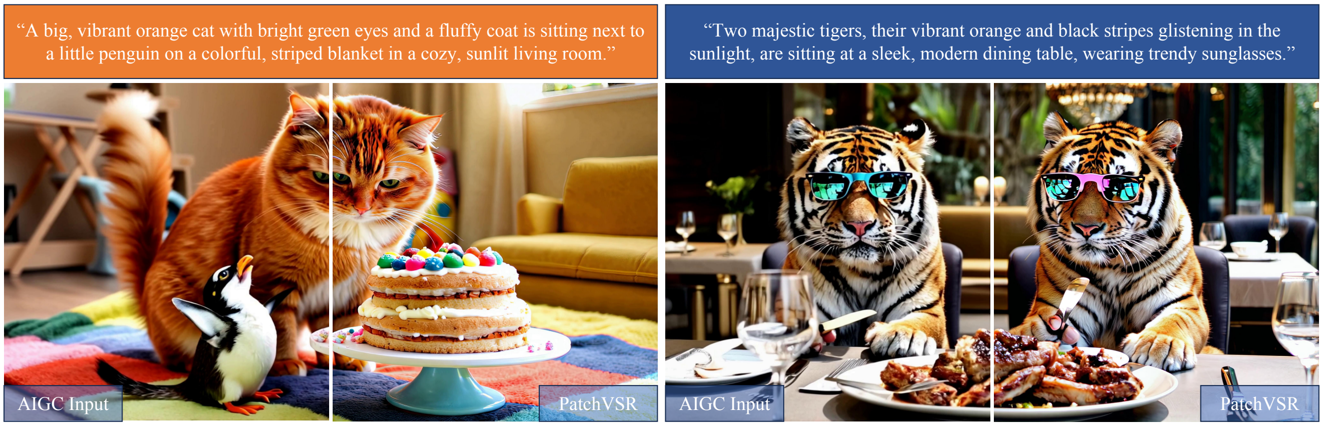
"A big, vibrant orange cat with bright green eyes and a fluffy coat is sitting next to a little penguin on a colorful, striped blanket in a cozy, sunlit living room."
"Two majestic tigers, their vibrant orange and black stripes glistening in the sunlight, are sitting at a sleek, modern dining table, wearing trendy sunglasses."
AIGC Input
PatchVSR
AIGC Input
PatchVSR

Figure 4. Visualization of 4K full videos based on 720P input. **Zoom in for best view.**

2023. 1

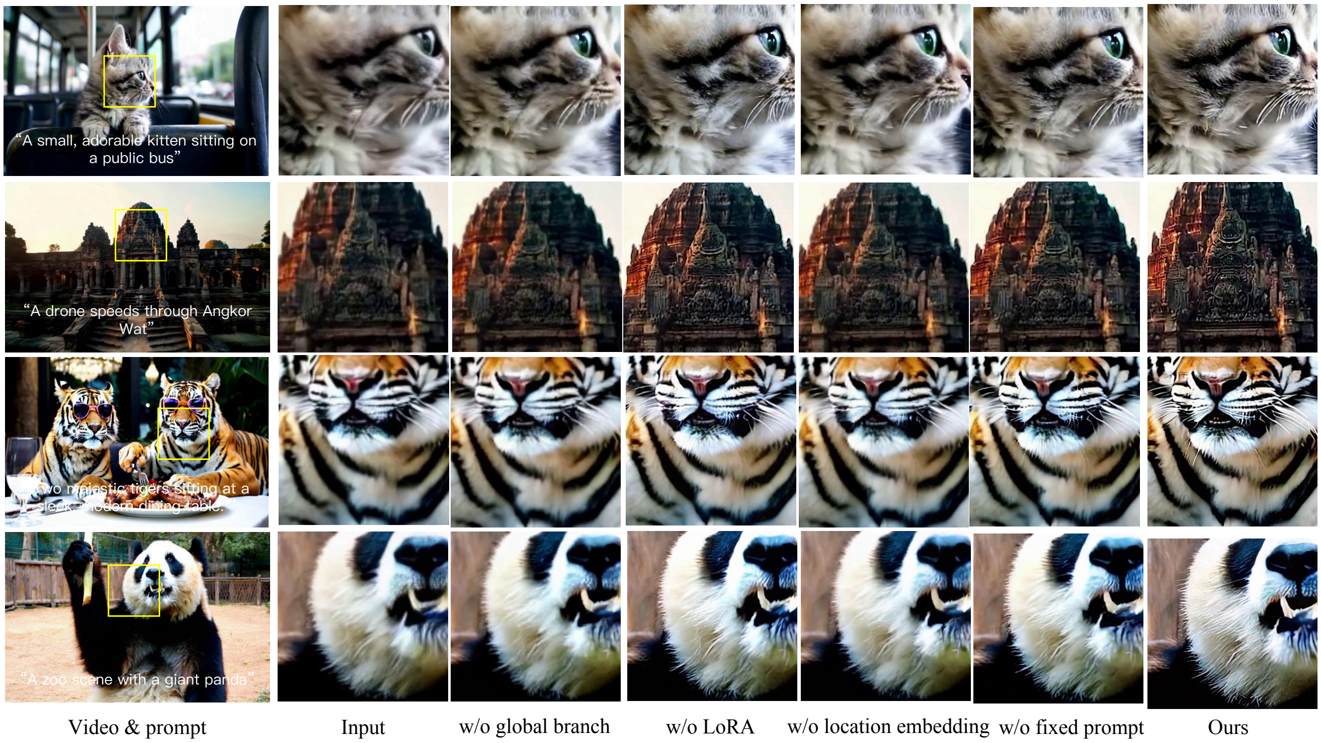
"A small, adorable kitten sitting on a public bus"
"A drone speeds through Angkor Wat"
"Two majestic tigers sitting at a sleek, modern dining table."
"A zoo scene with a giant panda"
Video & prompt
Input
w/o global branch
w/o LoRA
w/o location embedding
w/o fixed prompt
Ours

Figure 5. The qualitative comparisons of the ablation studies.

Metrics	w/o global branch	w/o LoRA	w/o location embedding	w/o fixed prompt	Ours
PSNR ↑	**31.245**	30.744	30.563	30.142	30.857
SSIM ↑	**0.744**	0.736	0.707	0.688	0.732
LPIPS ↓	**0.176**	0.185	0.195	0.201	0.183

Table 3. Fidelity evaluation on SynVideo30 test set.

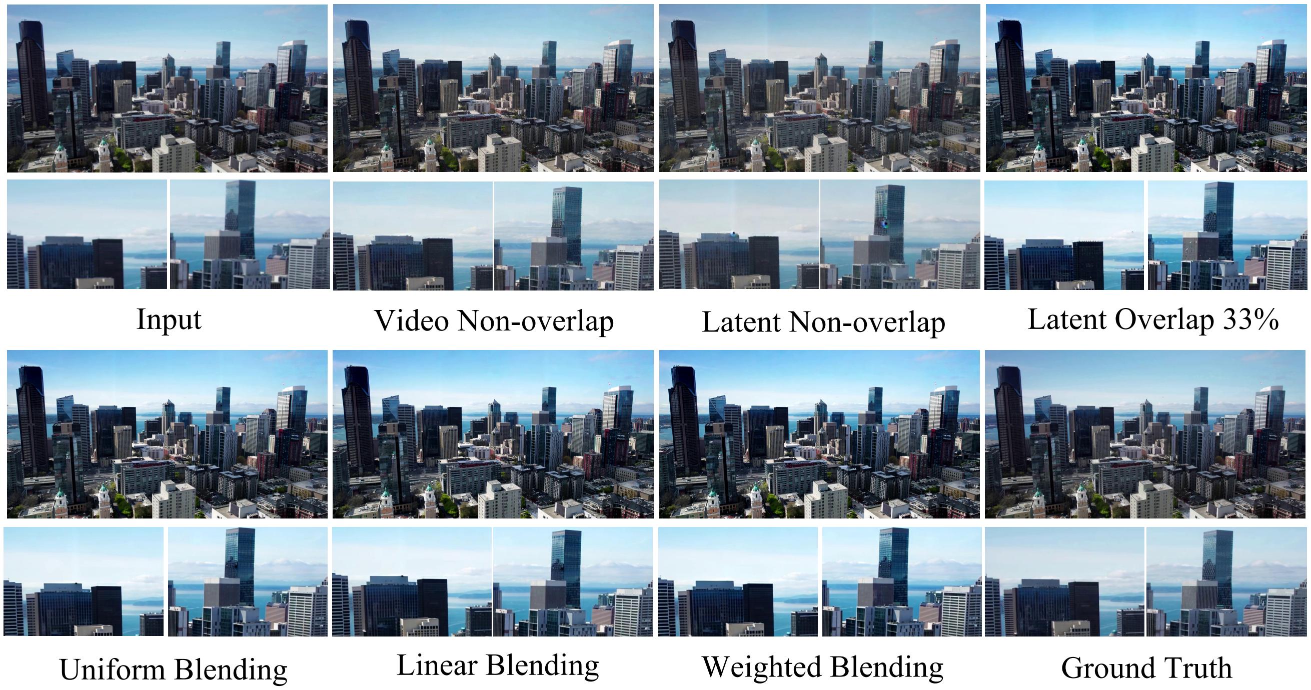

Figure 6. Comparisons of stitch schemes of 2K full video.