SimpleGVR

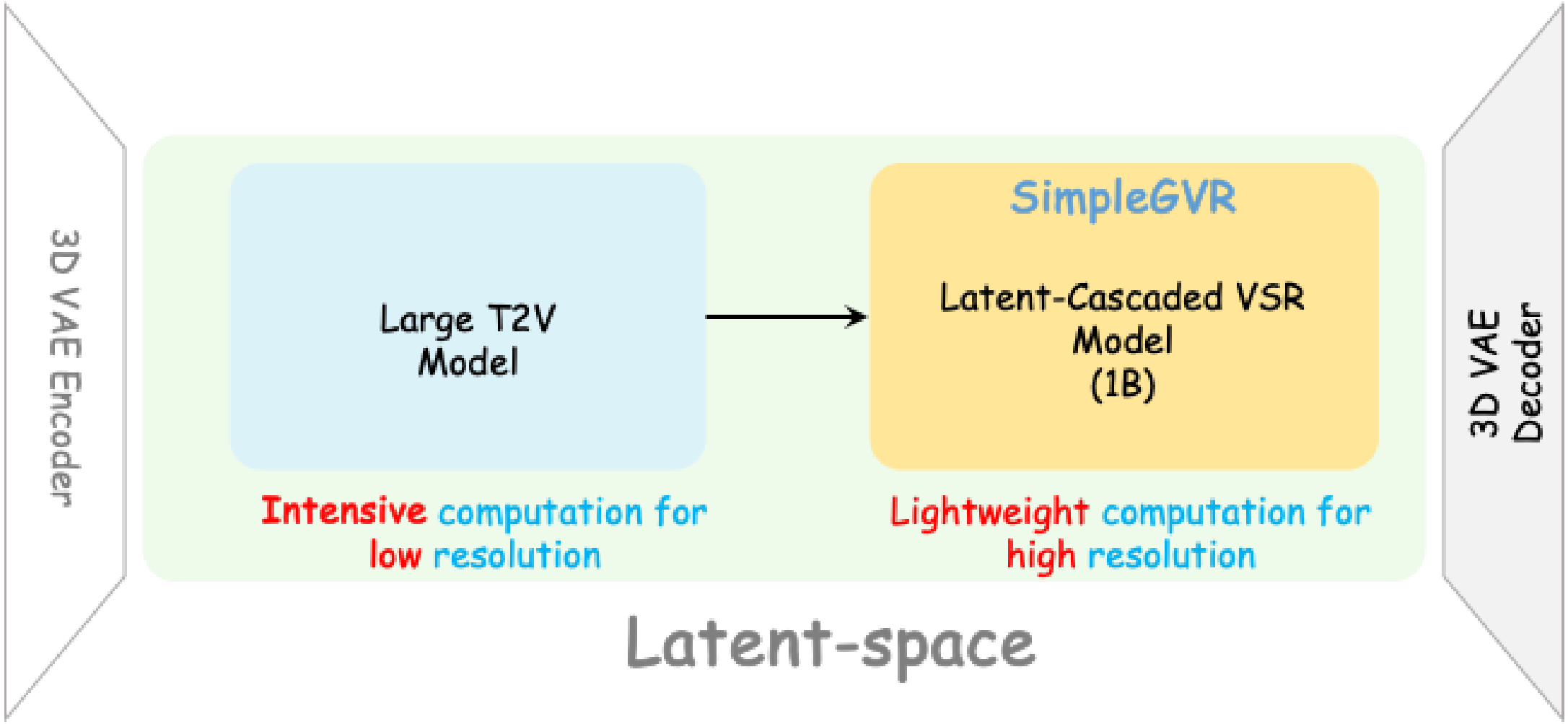
SimpleGVR: A Simple Baseline for Latent-Cascaded Video Super-Resolution

Liangbin Xie[1,2*] **Yu Li**[3] **Shian Du**[3] **Menghan Xia**[4†] **Xintao Wang**[4]
Fanghua Yu[2] **Ziyan Chen**[2] **Pengfei Wan**[4] **Jiantao Zhou**[1†] **Chao Dong**[2,5]
[1]State Key Laboratory of Internet of Things for Smart City, University of Macau
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3]Tsinghua University [4]Kuaishou Technology [5]Shenzhen University of Advanced Technology
https://simplegvr.github.io/

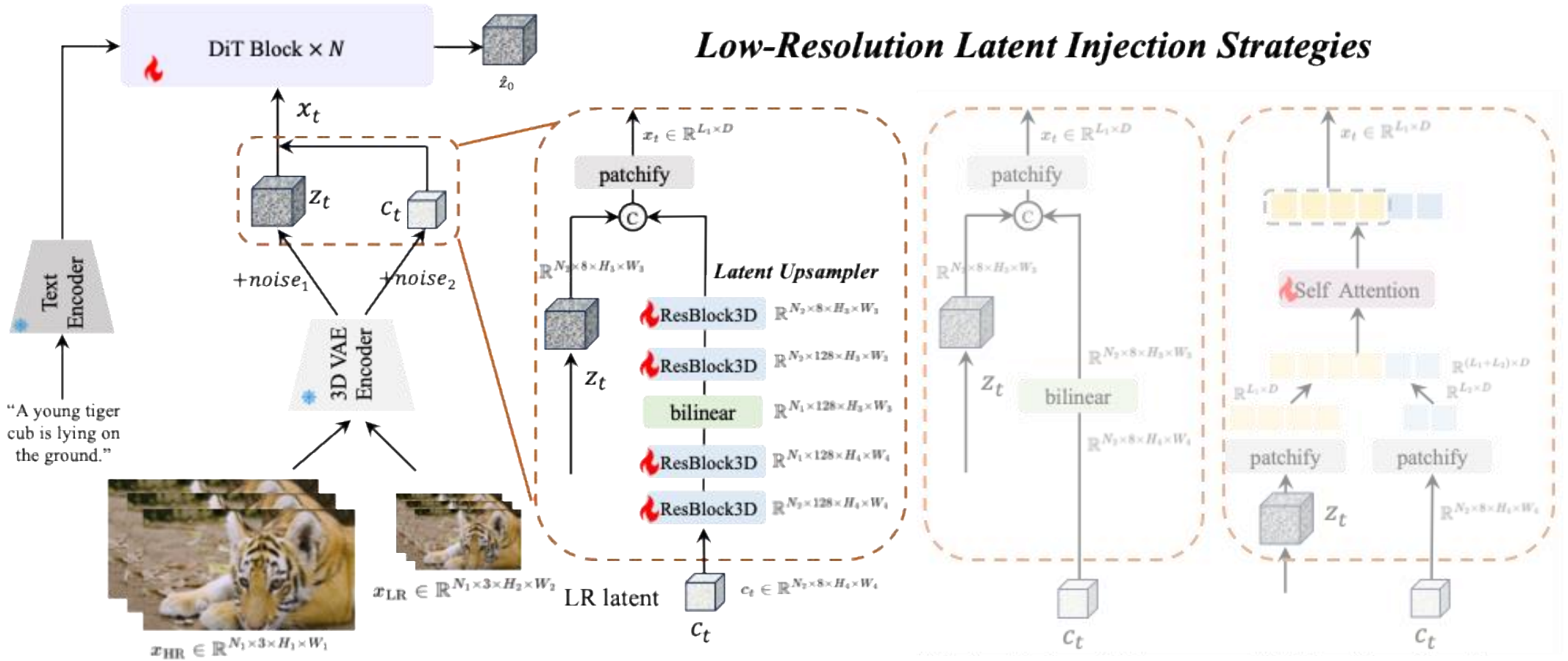
Figure 1: Built upon the low-resolution latent outputs (e.g., 384 × 672 resolution) from the first-stage Large T2V model, SimpleGVR generates high-quality results that even surpass the 1080p outputs of the Large T2V model. Compared to FlashVideo, which also adopts a cascaded architecture, SimpleGVR produces more realistic and finer details.

Motivation

Computation decomposition for high-resolution **text-to-video generation**.



Real latent-space upscaling



Overall Architecture

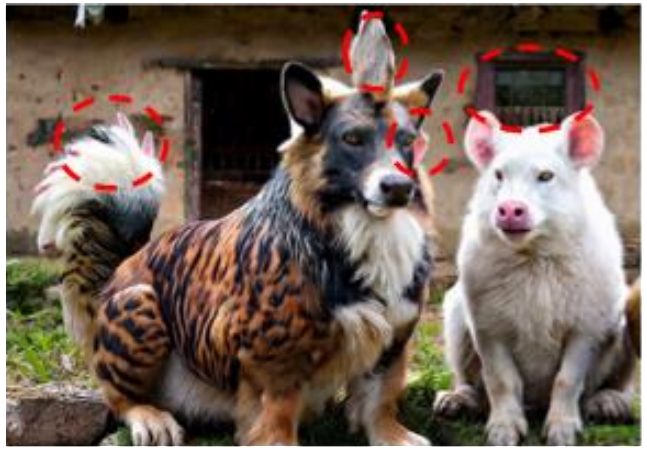
(a) Latent Upsampler + Channel Concatenation

(b) Latent Interpolation + Channel Concatenation

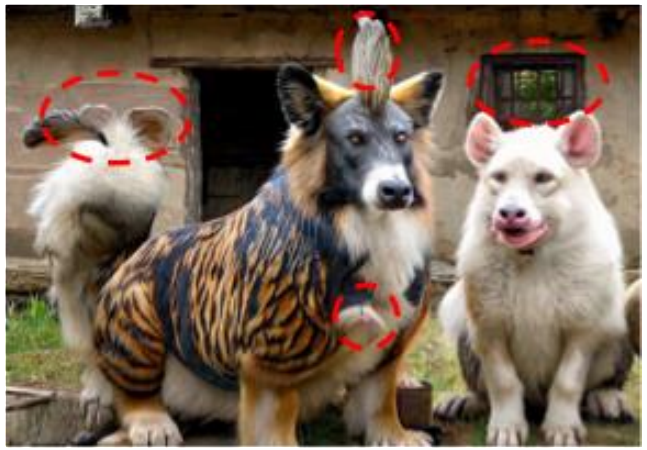
(c) Token Concatenation + Self Attention + Token Drop

Real latent-space upscaling

(a) Input



(b) Latent Interpolation + Channel Concatenation



(c) Token Concatenation

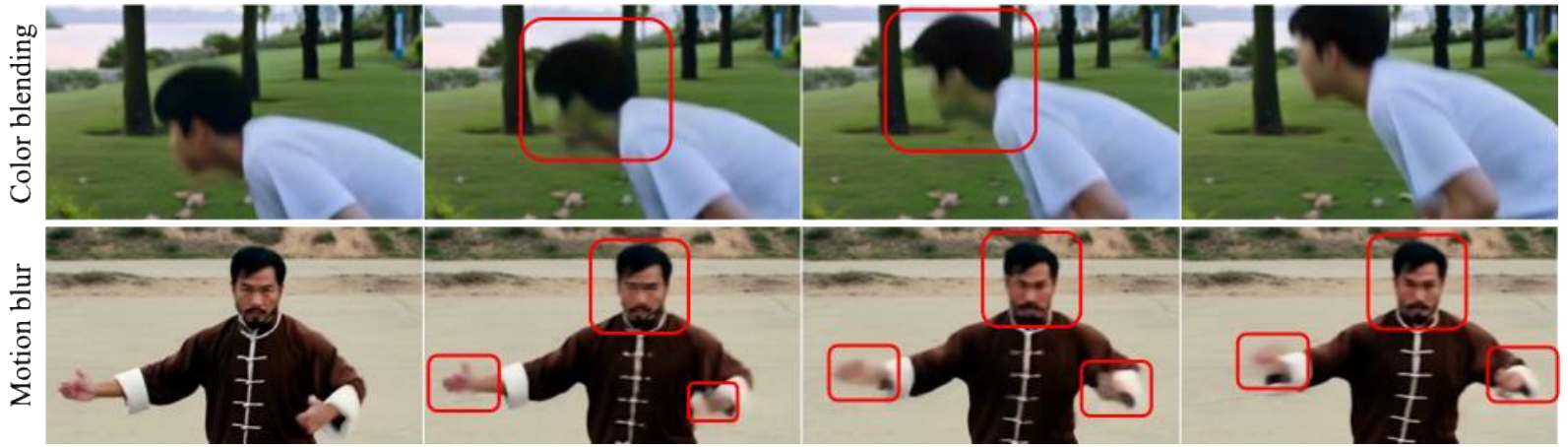
(d) Latent Upsampler + Channel Concatenation

(e) 3D ResBlocks + Latent Interpolation + Channel Concatenation

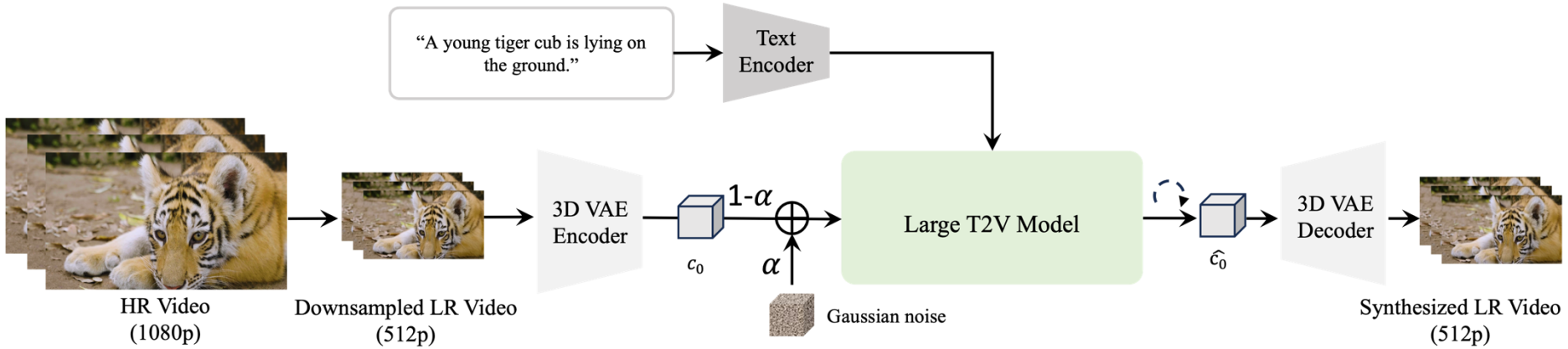
- Our proposed latent upsampler can better preserve the layout and semantic

Degradation Modeling

- **Flow-based** degradation, where optical flow guides motion-aware color blending and adaptive blurring.

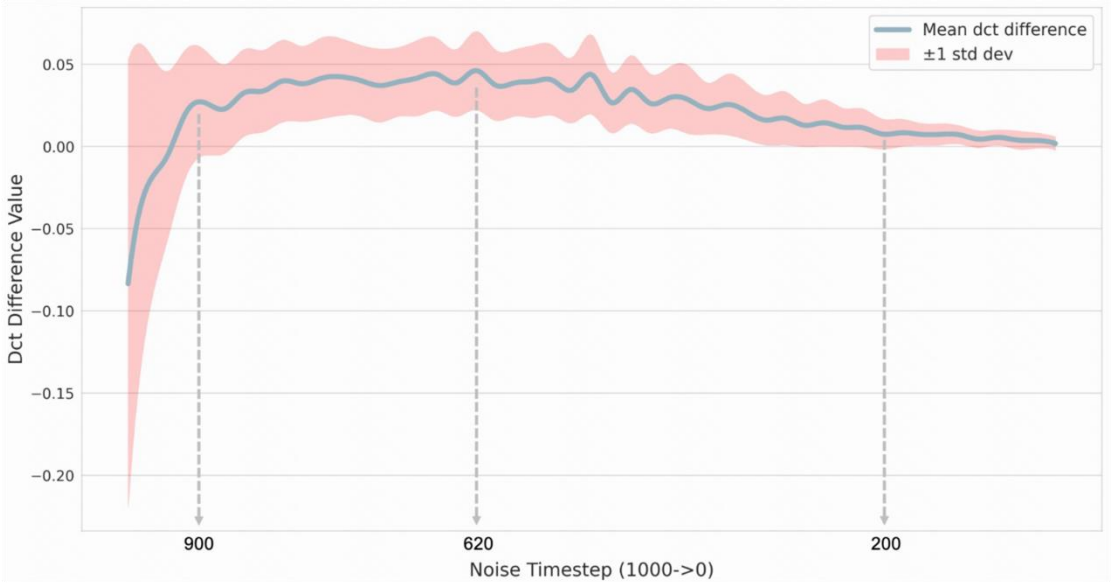


- **Model-guided** degradation, where noise is added to low-resolution video frames and partially denoised using the base T2V model.



Training Configurations

- **Detail-aware sampler**, where higher sampling probabilities are assigned to the timestep intervals that contribute more to detail enhancement.

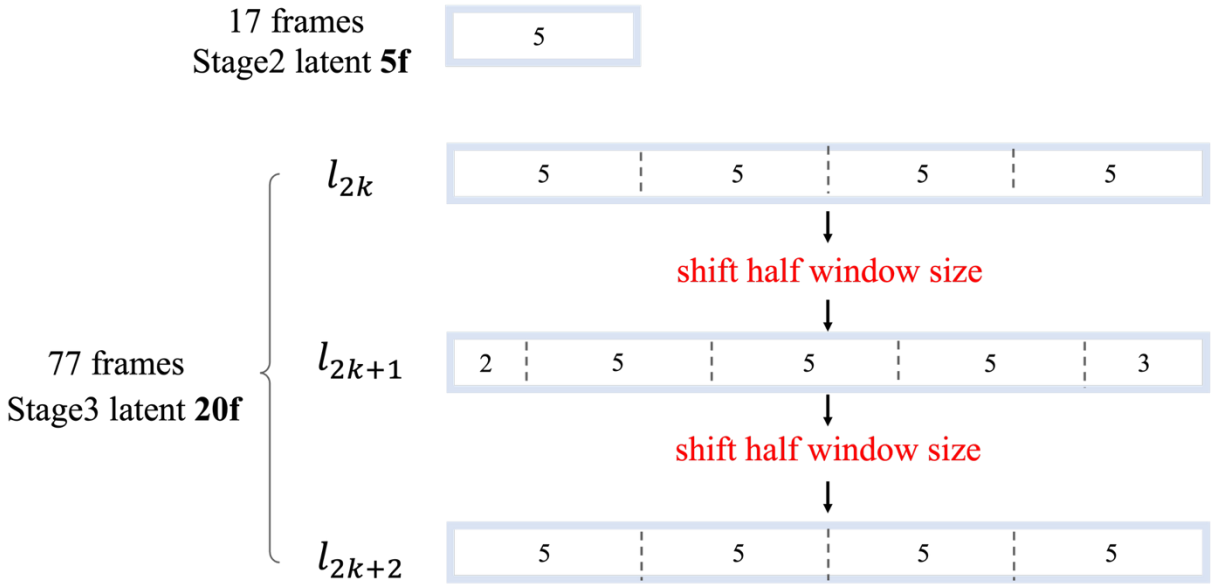


High-frequency variation curve over timesteps during inference.

- **Middle noise augmentation interval (0.3~0.6)**. When SimpleGVR is trained with noise in this interval, it is capable of enhancing high-frequency details while still being able to correct structural errors in the input frames.

Efficient Pseudo-global Computation

- **Interleaving temporal unit**, which expands SimpleGVR's capability to handle 77 frames in an effective way.



Qualitative Comparisons

SimpleGVR (Ours)

STAR

FlashVideo

RealBasicVSR

Upscale-A-Video

VEnhancer

Quantitative Comparisons

Table 1: Quantitative comparison on AIGC100 dataset. **Bold** and underline indicate the best and second best performance.

Method	MUSIQ	DOVER			Vbench Metrics					
		Technical	Aesthetic	Overall	Background Consistency	Subject Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness	Average Score
RealBasicVSR	57.55	12.27	98.66	61.84	93.73	93.98	61.63	72.76	98.70	84.16
VEnhancer	40.03	15.38	98.32	62.54	94.59	94.44	59.98	64.22	99.16	82.48
Upscale-A-Video	36.35	12.43	98.29	59.04	95.96	94.41	61.26	63.85	98.99	82.89
STAR	46.73	18.17	98.66	67.76	96.17	94.43	62.24	67.24	99.01	83.82
Flashvideo	53.65	15.97	98.61	65.38	95.49	94.75	60.76	69.11	98.45	83.71
Ours	**62.35**	**20.44**	**98.88**	**71.34**	95.35	94.32	62.84	71.91	98.74	**84.63**

T2V: Cascade vs. End-to-End

Table 3: Quantitative comparison between two different T2V paradigms on AIGC100 dataset.

Method	MUSIQ	DOVER			Vbench Metrics					
		Technical	Aesthetic	Overall	Background Consistency	Subject Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness	Average Score
End-to-End	56.77	18.82	97.27	62.32	96.04	95.16	63.45	67.69	98.89	84.25
Cascaded	**62.35**	**20.44**	**98.88**	**71.34**	95.35	94.32	62.84	**71.91**	98.74	**84.63**

T2V: Cascade vs. End-to-End

Large T2V (384x762) **+ SimpleGVR**

Large T2V (1080P)

T2V: Cascade vs. End-to-End

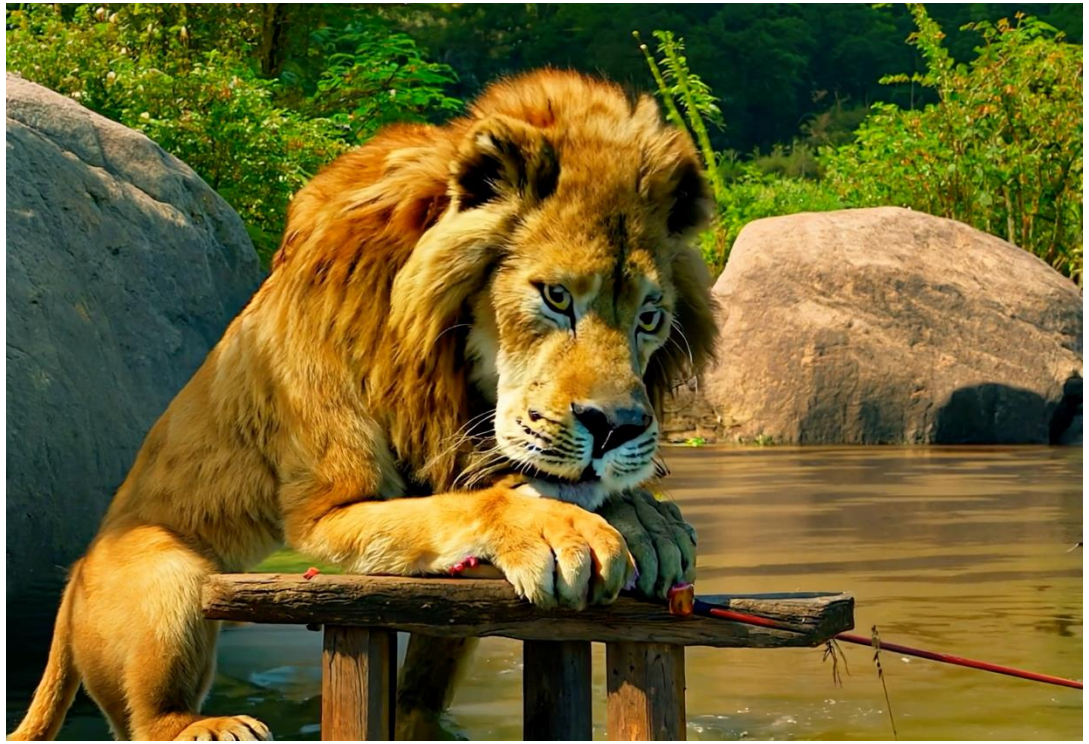
Large T2V (384x762) **+ SimpleGVR**

Large T2V (1080P)

Generalization: CogVideoX

CogVideoX (480 × 720)

CogVideoX + SimpleGVR



Generalization: Wan

Wan (480p)

Wan + SimpleGVR