

## A APPENDIX

### A.1 THE USE OF LLM

The use of LLMs was strictly limited to the final manuscript writing stage, specifically for the purpose of correcting grammatical errors and polishing the text.

### A.2 ADDITIONAL DETAILS OF DC-AE-V

#### A.2.1 MODEL ARCHITECTURE

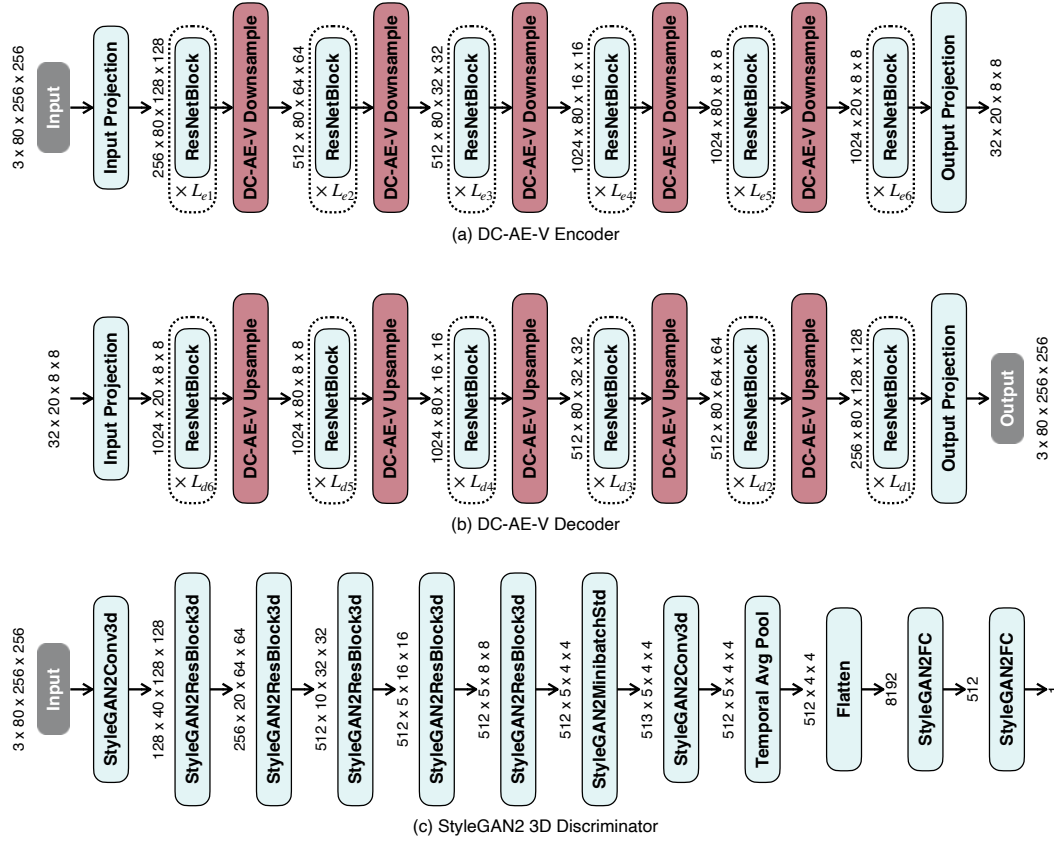


Figure 9: Model Architecture of DC-AE-V.

Figure 9 presents the detailed model architecture of an f32t4c32 DC-AE-V. Both the encoder and decoder are composed of six stages, each built from 3D ResNet blocks. The first five stages perform only spatial downsampling and upsampling, while the final stage handles temporal downsampling and upsampling. Following DC-AE (Chen et al., 2024a), we incorporate Residual Autoencoding to facilitate optimization during downsampling and upsampling. For adversarial training, we extend the StyleGAN2 discriminator (Karras et al., 2020) to process video inputs.

#### A.2.2 DATASET

Our DC-AE-V is trained on a mixture of video and image datasets. The video datasets include subsets of Panda70m (Chen et al., 2024c) and OpenVid1m (Nan et al., 2024). The image datasets include ImageNet21k (Ridnik et al., 2021), Mapillary Vistas (Neuhoud et al., 2017), DataComp (Gadre et al., 2023), WiderFace (Yang et al., 2016), WiderPerson (Zhang et al., 2019), TextCaps (Sidorov et al., 2020), and Unsplash (Unsplash, 2020).

Video Autoencoder	Config	Compress. Ratio	Panda70m				UCF101				ActivityNet				Kinetics600			
			PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$		PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$
VideoVAEPlus [53]	f8t4c16	48	36.88	0.968	0.009	35.79	0.959	0.016	2.11	35.68	0.955	0.016	0.96	36.73	0.960	0.014	0.93	
CogVideoX VAE [58]	f8t4c16	48	35.54	0.961	0.021	34.53	0.949	0.034	8.32	34.47	0.946	0.034	5.16	35.40	0.951	0.032	4.18	
HunyuanVideo VAE [23]	f8t4c16	48	35.46	0.960	0.015	34.40	0.950	0.024	3.80	34.41	0.943	0.024	3.16	35.40	0.950	0.022	2.59	
IV VAE [50]	f8t4c16	48	35.32	0.959	0.017	34.84	0.955	0.025	3.71	35.03	0.948	0.025	1.88	36.27	0.956	0.022	1.52	
Wan 2.1 VAE [45]	f8t4c16	48	34.15	0.952	0.017	33.81	0.943	0.024	3.71	33.82	0.938	0.025	1.76	35.04	0.946	0.022	1.49	
Wan 2.2 VAE [45]	f16t4c48	64	35.12	0.958	0.013	34.27	0.948	0.022	4.02	34.41	0.943	0.021	1.56	35.57	0.950	0.019	1.51	
StepVideo VAE [31]	f16t8c64	96	32.17	0.930	0.043	32.17	0.930	0.043	8.23	32.08	0.922	0.047	5.29	33.02	0.931	0.044	4.62	
Video DC-AE <sub>finetune &amp; blending</sub> [36]	f32t4c128	96	34.10	0.952	0.023	33.65	0.945	0.034	14.22	33.55	0.938	0.033	7.92	34.73	0.946	0.030	6.81	
Video DC-AE [36]	f32t4c128	96	31.73	0.915	0.040	31.52	0.914	0.047	26.30	31.34	0.901	0.049	17.52	32.39	0.915	0.044	15.43	
LTX Video VAE [18]	f32t8c128	192	32.41	0.928	0.039	31.12	0.910	0.059	70.92	31.29	0.900	0.058	45.51	32.26	0.911	0.056	42.30	
DC-AE-V	f32t4c256	48	39.54	0.979	0.008	37.13	0.967	0.018	2.03	37.28	0.965	0.016	0.94	38.11	0.969	0.016	0.88	
	f32t4c128	96	37.36	0.968	0.013	34.83	0.951	0.027	5.49	35.06	0.948	0.025	2.59	35.91	0.953	0.024	2.47	
	f32t4c64	192	35.03	0.953	0.020	32.72	0.931	0.036	13.04	33.02	0.927	0.035	6.22	33.87	0.935	0.033	6.24	
	f32t4c32	384	33.07	0.933	0.027	30.83	0.909	0.046	29.11	31.08	0.901	0.045	13.83	32.01	0.912	0.042	13.05	
	f64t4c128	384	32.83	0.932	0.030	30.65	0.907	0.049	32.16	30.90	0.899	0.049	15.63	31.77	0.910	0.047	15.50	

Table 5: Additional Video Reconstruction Results.

### A.2.3 EVALUATION

We evaluate all video autoencoders on  $80 \times 256 \times 256$  videos using PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and FVD. The evaluation set includes 1,000 unseen videos from Panda70m (Chen et al., 2024c), 3,783 test videos from UCF101 (Soomro et al., 2012), 5,044 test videos from ActivityNet 1.3 (Caba Heilbron et al., 2015), and the first 5,000 test videos from Kinetics600 (Carreira et al., 2018).

For VideoVAEPlus, we use the ‘16z’ version. For CogVideoX VAE (Yang et al., 2024), we use the model from THUDM/CogVideoX-2b. For HunyuanVideo VAE (Kong et al., 2024), we use the model from hunyuanvideo-community/HunyuanVideo. For IV VAE (Wu et al., 2025), we use the ‘ivvae\_z16\_dim96’ version. For Wan 2.1 VAE (Wan et al., 2025), we use the model from Wan-AI/Wan2.1-T2V-14B-Diffusers, and for Wan 2.2 VAE (Wan et al., 2025), we use the model from Wan-AI/Wan2.2-Ti2V-5B. For StepVideo VAE (Ma et al., 2025), we use the ‘vae\_v2’ from stepfun-ai/stepvideo-t2v. For Video DC-AE (Peng et al., 2025), we use the model from hpcal-tech/OpenSora-v2-Video-DC-AE. Finally, for LTX Video VAE (HaCohen et al., 2024), we use the model from Lightricks/LTX-Video-0.9.7-dev. When an autoencoder cannot process 80-frame videos, we pad extra frames at the end and exclude them from the reconstructions when computing evaluation metrics.

### A.2.4 ADDITIONAL RECONSTRUCTION RESULTS

Table 5 presents the full reconstruction results. Our DC-AE-V consistently achieves superior accuracy and generalizes effectively to longer videos across a range of benchmarks.

Figure 10 presents additional reconstruction examples. Our DC-AE-V demonstrates superior reconstruction accuracy and generalization ability to longer videos, especially for small texts and human faces.

## A.3 ABLATION STUDY ON VIDEO EMBEDDING SPACE ALIGNMENT

Figure 11 presents additional ablation studies on video embedding space alignment. Aligning both the patch embedder and output head yields the best results, with the patch embedder alignment playing the most critical role in overall performance.

## A.4 DETAILED EVALUATION RESULTS ON VBENCH

Table 6 reports detailed metrics on VBench. DC-VideoGen-Wan-2.1-T2V-1.3B outperforms the base Wan-2.1-T2V-1.3B on 11 of the 16 metrics.

## A.5 DETAILED EFFICIENCY BENCHMARK RESULTS

Table 7 presents detailed efficiency results measured on an NVIDIA H100 GPU.



Figure 10: **Additional Video Autoencoder Reconstruction Visualization.** Under deep compression settings, causal video autoencoders suffer from low reconstruction quality. In contrast, non-causal video autoencoders achieve better reconstruction quality but generalize poorly to longer videos.

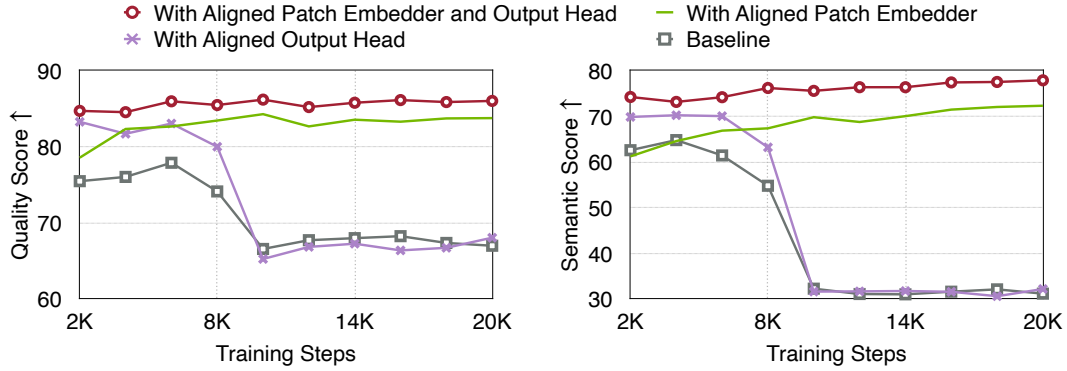


Figure 11: **Ablation Study on Video Embedding Space Alignment.**

Text-to-Video Generation Results on VBench 720×1280								
Models	Temporal Flickering	Subject Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Background Consistency	Overall Consistency
Wan-2.1-T2V-1.3B	99.15	94.97	<b>98.36</b>	67.78	70.20	68.44	97.99	<b>25.97</b>
DC-VideoGen-Wan-2.1-T2V-1.3B	<b>99.18</b>	<b>96.58</b>	98.34	<b>72.78</b>	<b>72.00</b>	<b>68.72</b>	<b>98.00</b>	25.41
Models	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style
Wan-2.1-T2V-1.3B	<b>89.11</b>	72.07	94.05	81.54	65.90	44.83	<b>21.61</b>	<b>23.22</b>
DC-VideoGen-Wan-2.1-T2V-1.3B	88.73	<b>75.98</b>	<b>94.64</b>	<b>89.16</b>	<b>78.20</b>	<b>44.86</b>	21.20	22.97

Table 6: **Detailed Results on VBench.**

## A.6 DETAILED TRAINING HYPERPARAMETERS OF AE-ADAPT-V

Table 8 lists the detailed hyperparameters for AE-Adapt-V.

(a) Latency (minutes per video)					(b) Latency (minutes per video)				
Models	Resolution				Models	Number of Frames			
	480×832	720×1280	1080×1920	2160×3840		80	160	320	640
Wan-2.1-1.3B [45]	1.49	5.76	25.46	375.12	Wan-2.1-1.3B [45]	5.76	20.18	75.77	296.30
DC-VideoGen-Wan-2.1-1.3B	0.24	0.70	2.27	25.41	DC-VideoGen-Wan-2.1-1.3B	0.70	1.99	6.03	20.86
Speedup	6.2×	8.2×	11.2×	14.8×	Speedup	8.2×	10.1×	12.6×	14.2×

Table 7: Detailed Efficiency Benchmark Results.

Training Stage	Hyperparameter	Value
Patch Embedder Alignment	learning rate	1e-4
	warmup steps	0
	batch size	4
	training steps	20k
	optimizer	AdamW, betas=[0.9, 0.999]
Output Head Alignment	learning rate	1e-4
	warmup steps	0
	batch size	32
	training steps	4k (Wan-2.1-1.3B) / 3k (Wan-2.1-14B)
	optimizer	AdamW, betas=[0.9, 0.999]
End-to-End Fine-Tuning	learning rate	5e-5
	warmup steps	1k
	training steps	20k (Wan-2.1-1.3B) / 6k (Wan-2.1-14B)
	batch size	32
	optimizer	AdamW, betas=[0.9, 0.999]
	weight decay	1e-3
	LoRA (rank, alpha)	(256, 512)
Resolution Increasing	480px→720px, training steps	1000
	720px→1080px, training steps	500
	1080px→2160px, training steps	200

Table 8: Training Hyperparameters of AE-Adapt-V.

## A.7 QUALITATIVE COMPARISON WITH THE PRE-TRAINED MODELS



Figure 12: Visual Comparison of DC-VideoGen-Wan2.1-I2V-14B and the Base Model Wan2.1-I2V-14B.



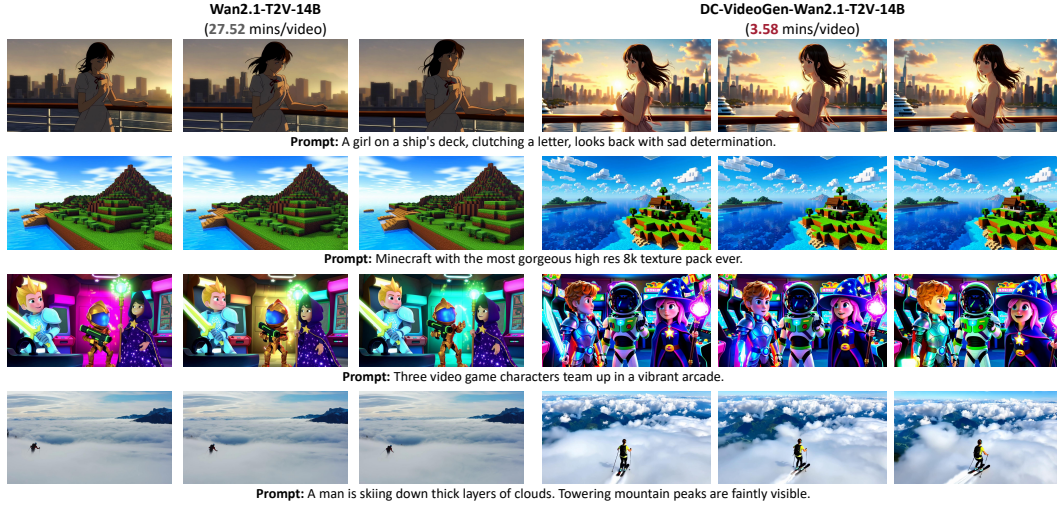


Figure 13: Visual Comparison of DC-VideoGen-Wan2.1-T2V-14B and the Base Model Wan2.1-T2V-14B.

Figure 12 and Figure 13 provide a qualitative comparison between DC-VideoGen and its base models. We observe that DC-VideoGen-Wan2.1-I2V-14B and DC-VideoGen-Wan2.1-T2V-14B retains the generation quality of Wan2.1-I2V-14B and Wan2.1-T2V-14B while reducing the latency by 87%.

#### A.8 LIMITATION AND FUTURE WORK

**Limitations.** As a post-training framework, DC-VideoGen accelerates video diffusion models through lightweight fine-tuning, effectively leveraging the rich knowledge encoded in the pre-trained model. Consequently, its performance is strongly dependent on the quality of the pre-trained model.

**Future Work.** DC-VideoGen substantially reduces the training and inference costs of video diffusion models, especially when scaling to higher resolutions. For the next step, we plan to extend our framework for long video generation, leveraging techniques from (Gu et al., 2025a; Yang et al., 2025a).