Supplementary Material of
REDUCIO! Generating 1K Video within 16 Seconds using Extremely Compressed Motion Latents

Rui Tian[1,2] Qi Dai[3*] Jianmin Bao[3] Kai Qiu[3] Yifan Yang[3]
Chong Luo[3] Zuxuan Wu[1,2*] Yu-Gang Jiang[1,2]
[1]Institute of Trustworthy Embodied AI, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing [3]Microsoft Research

A. Implementation details

***Reducio*-VAE.** We demonstrate the detailed architecture of *Reducio* in Fig. 1. Generally, we follow the training strategies of SD-VAE [12]. We employ a customized version of PatchGAN [8] based on 3D convolutions and optimize the model with L_1 loss, KL loss, perceptual loss [18], and GAN loss. While initializing the 2D encoder and the 3D VAE with SD-VAE pre-trained weights accelerates convergence, we find that freezing the 2D encoder leads to worse performance than training the full parameters. We follow LaMD [6] to feed the motion latent into a normalization layer to obtain the output of the VAE encoder. Therefore, in the stage of diffusion training, we use a scale factor of 1.0, which is multiplied by the input latent as input into DiT.

Table 1. Hyperparameters for *Reducio*-VAE

	Reducio-VAE	
z-shape[1]	$4 \times 8 \times 8 \times 16$	$8 \times 8 \times 8 \times 16$
Channels (3d)	128	
Channels (2d)	128	
Ch Multiplier (3d)	1,2,2,4,4,4	
Ch Multiplier (2d)	1,2,2,4,4	
Depth	2	
Batch size	32	24
Learning rate	4e-5	
Iterations	1,000,000	

During inference, We split videos with a resolution over 256×256 into overlapping spatial tiles, we fuse the encoded latent as well as the video output, in a similar manner with Movie Gen [11]. Note that since *Reducio*-VAE employs a spatial down-sampling factor of 32, *Reducio*-VAE can only process video inputs whose width and height are divisible by 32. Otherwise, videos should be padded to meet this requirement before being fed into the VAE.

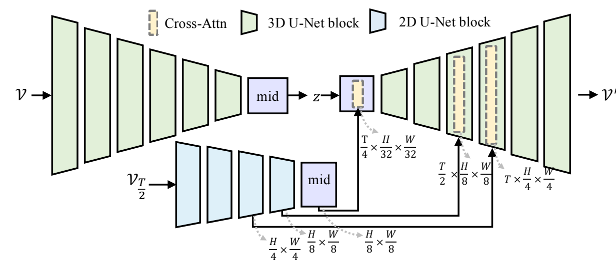


Figure 1. The detailed architecture of *Reducio*-VAE with $f_t = 4, f_s = 32$.

***Reducio*-DiT.** We elaborate on the details of content image conditions in Fig. 2. During inference, we use classifier-free guidance [5] for better generation quality and set the default scale to 2.5. During training, we randomly drop image conditions at a probability of 0.1, as well as drop 10% text conditions.

B. More ablations

Table 2. Ablation on the convolution types in *Reducio*-VAE.

f_s	**Conv**	**PSNR↑**	**SSIM↑**
64	2d	30.22	0.86
64	3d	35.51	0.94

Using 3d VAE in *Reducio*-VAE helps to reconstruct videos in a better quality. We keep f_t to 1 and f_s to 32 and implement VAE with 2d convolutions. During decoding, we duplicate the middle frame condition for T times to fuse with the latent of each frame respectively. As shown in Tab. 2, *Reducio*-VAE with 3d convolution outperforms its counterpart with 2d convolution. We believe that 3d convolution

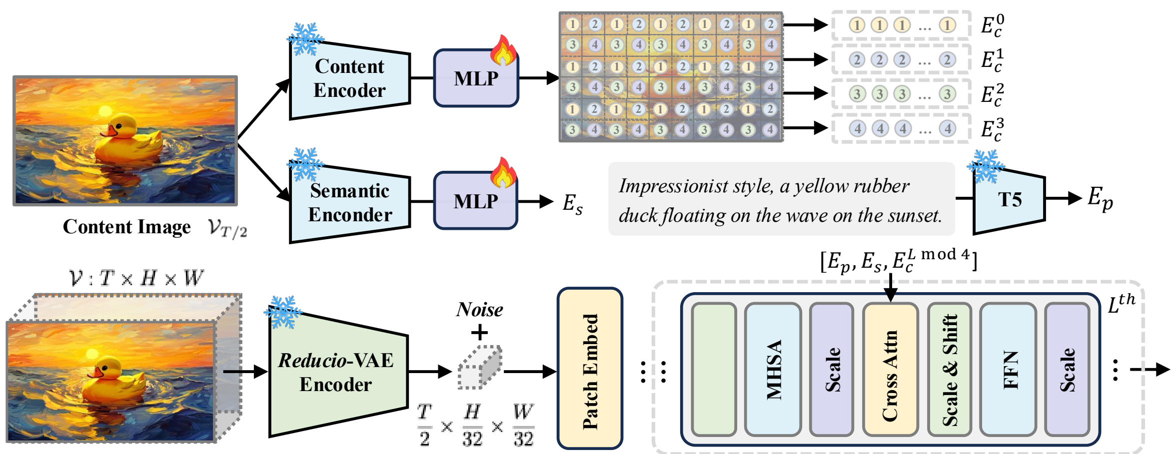


Figure 2. The overview of the efficient content condition solution for *Reducio*-DiT on high-resolution videos.

facilitates the VAE to model consistent motion and capture spatiotemporal differences.

Table 3. Ablation on the content frame choice in *Reducio*-VAE.

Content Frame	PSNR↑	SSIM↑
n/a	27.91	0.80
random	31.72	0.87
middle	35.88	0.94

Using middle frame in *Reducio*-VAE. The content frame in *Reducio*-VAE provides a strong content prior and hence leads to a promising reconstruction performance. On the other hand, relying on any given frame as the content image may not generalize perfectly in all scenarios, especially when certain entities appear only briefly or outside the chosen frame. As shown in Tab. 3, we choose the middle frame by default as it serves as a more stable and robust content guidance due to its temporal centrality. Meanwhile, *Reducio*-VAE without condition achieves significantly worse results in both PSNR (-7.97) and SSIM (-0.14). In consequence, the *Reducio*-DiT framework without condition-based 3D VAE leads to unsatisfactory results featured with blurry frames and obvious visual defects.

Table 4. Comparison with the state-of-the-art 2D Autoencoder with a significant spatial down-sampling factor.

Model	latent shape	$\lvert z\rvert$	PSNR↑	SSIM↑
DC-AE [3]	$16 \times 8 \times 8$	32	30.68	0.70
Reducio-VAE	$4 \times 8 \times 8$	16	35.56	0.97

Comparison between DC-AE and *Reducio*-VAE. We compare *Reducio*-VAE with DC-AE [3] on the Pexel test split with resolution of 512×512. As shown in the Table below, *Reducio*-VAE outperforms DC-AE on PSNR and SSIM by 4.88 and 0.27, respectively, highlighting the advantage of our framework in video domain.

Table 5. Ablation on the attention type in *Reducio*-DiT.

Attn	FVD↓	IS↑
2d + 1d	382.2	32.4
3d	337.6	34.1

Table 6. Comparison with more SOTA models on Vbench.

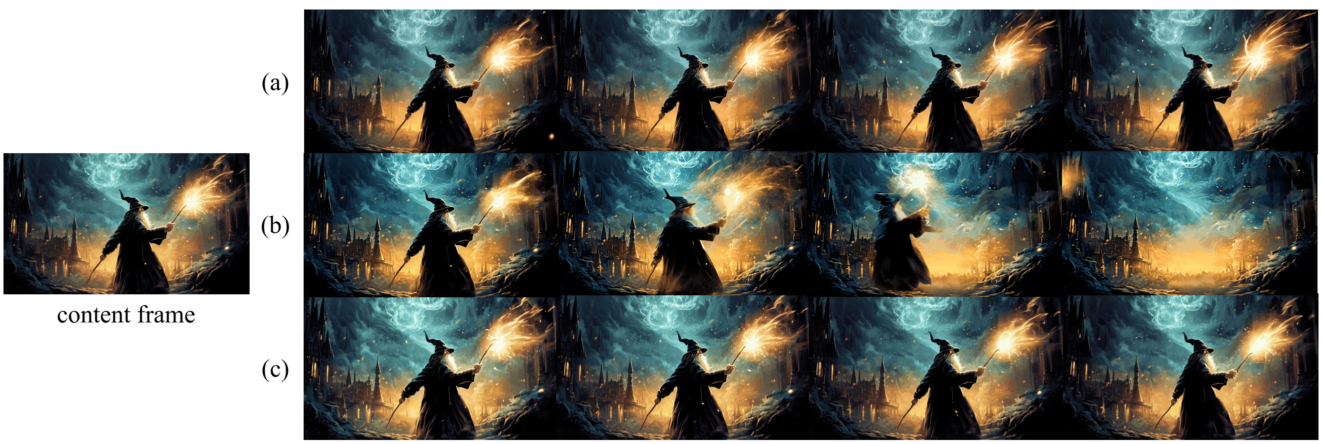
Model	Quality Score	Semantic Score	Total Score
Show-1 [17]	80.42	72.98	78.93
Lavie [14]	78.78	70.31	77.08
VideoCrafter [2]	81.59	72.22	79.72
OpenSora v1.2 [19]	81.35	73.39	79.76
Lavie-2 [14]	83.24	75.76	81.75
Pyramid Flow [9]	84.74	69.62	81.72
VideoCrafter-2 [4]	83.27	76.73	81.97
***Reducio*-DiT**	82.24	78.00	81.39
WAN [13]	84.92	80.10	83.96
STIV [10]	81.20	72.70	79.50
CausVid [16]	85.21	78.57	83.88

Using joint spatiotemporal 3D attention in *Reducio*-DiT outweighs using factorized spatial and temporal attention (*i.e.*, 2D + 1D attention) in generation quality. Interestingly, we observe that factorized attention leads to a faster convergence of training loss. However, with the same training step, as shown in Tab. 5, factorized attention lags behind its counterpart with joint 3D attention for 45 in FVD. We suppose the possible reason is that 2D + 1D scheme demands adding additional temporal layers and performs factorized

Table 7. Detailed quantitative comparison with state-of-the-art text-to-video generation models on VBench.

Model	subject consistency	background consistency	temporal flickering	motion smoothness	dynamic degree	aesthetic quality	imaging quality	object class	multiple objects	human action	color	spatial relationship	scene	appearance style	temporal style	overall consistency
Lavie [14]	91.41	97.47	98.30	96.38	49.72	54.94	61.90	91.82	33.32	96.80	86.39	34.09	52.69	23.56	25.93	26.41
Show-1 [17]	95.53	98.02	99.12	98.24	44.44	57.35	58.66	93.07	45.47	95.60	86.35	53.50	47.03	23.06	25.28	27.46
VideoCrafter [2]	95.10	98.04	98.93	95.67	55.00	62.67	65.46	78.18	45.66	91.60	93.32	58.86	43.75	24.41	25.54	26.76
OpenSora v1.2 [19]	96.75	97.61	99.53	98.50	42.39	56.85	63.34	82.22	51.83	91.20	90.08	68.56	42.44	23.95	24.54	26.85
Lavie-2 [14]	97.90	98.45	98.76	98.42	31.11	67.62	70.39	97.52	64.88	96.40	91.65	38.68	49.59	25.09	25.24	27.39
Pyramid Flow [9]	96.95	98.06	99.49	99.12	64.63	63.26	65.01	86.67	50.71	85.60	82.87	59.53	43.20	20.91	23.09	26.23
VideoCrafter-2 [4]	97.17	98.54	98.46	97.75	42.50	65.89	70.45	93.39	49.83	95.00	94.41	64.88	51.82	24.32	25.17	27.57
***Reducio*-DiT**	98.05	99.13	98.45	98.77	27.78	64.02	67.67	91.49	69.91	92.60	89.06	52.85	54.90	25.16	26.40	28.87

A dog wearing a Superhero outfit with red cape flying through the sky.



A dog wearing a Superhero outfit with red cape flying through the sky.

Figure 3. Comparison between frames generated given an identical frame and prompt, by (a) DynamicCrafter [15], (b) SVD-XT [1] and (c) *Reducio*-DiT, respectively. We resize the output frames from 1344×768 to 1024×576 to match with the former two baselines.

self-attention on a small set of tokens e ach, making it hard to model smooth open-set motion with the light computation. In contrast, 3D attention directly exploits the original parameters and collaborates all spatiotemporal tokens.

Quantitative results. We display the detailed performance comparison on VBench [7] in Tab. 7 and Tab. 6. Despite using only 3.2K A100 GPU hours and 5.4M training samples, *Reducio*-DiT achieves a promising semantic score of 78.00, beating a range of state-of-the-art LDMs. While the most recent models such as WAN [13] and CausVid [16] achieve higher overall scores than *Reducio*-DiT, we argue that our model uses a much smaller scale of training data and has a relatively small model scale, *i.e.*, 1.2B.

Visualizations. We present more examples of comparison between *Reducio*, SVD-XT [1] and DynamicCrafter [15] in Fig. 3. *Reducio*-DiT exhibits reasonable motion and preserves the details in the content frame well.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[2] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2, 3

[3] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *ICLR*, 2025. 2

[4] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024. 2, 3

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[6] Yaosi Hu, Zhenzhong Chen, and Chong Luo. Lamd: Latent motion diffusion for video generation. *IJCV*, 2025. 1

[7] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 3

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[9] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025. 2, 3

[10] Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, et al. Stiv: Scalable text and image conditioned video generation. *arXiv preprint arXiv:2412.07730*, 2024. 2

[11] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[13] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3

[14] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*, 2024. 2, 3

[15] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, 2024. 3

[16] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025. 2, 3

[17] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *IJCV*, 2024. 2, 3

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[19] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 2, 3