A. Appendix

A.1. Tokenizer Scalability

We briefly validate the scalability of our tokenizer in Fig. A1 by leveraging different hidden dimensions of 384, 512 and 768. As is shown, increasing the hidden dimension improves the performance at all stages of training, demonstrating the scalability of our approach. We ultimately select Tokenizer-L for our final tokenizer.

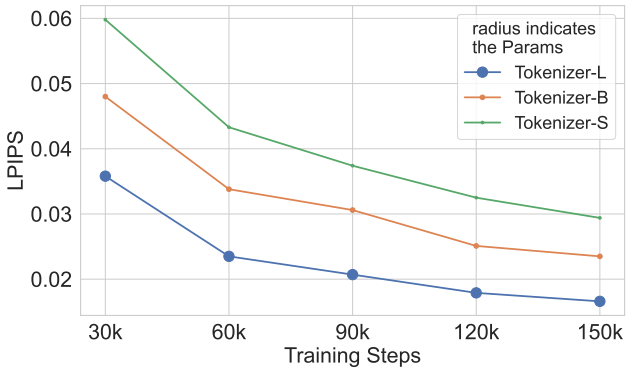

Figure A1. Scaling the tokenizer improves LPIPS at all stages of training. Tokenizer-S, Tokenizer-B, and Tokenizer-L have hidden dimensions of 384, 512, and 768, with parameter counts of 20.11M, 35.01M, and 77.35M, respectively.

A.2. Diffusion Denoising Steps

We evaluate the impact of the number of diffusion denoising steps on the interpolation performance of our model using DAVIS, with the visualization results in Fig. A2 and the quantitative results presented in Tab. A1. As observed, using random noise fails to capture accurate intermediate frame motion. Additionally, the model achieves satisfactory performance with just two denoising steps. Increasing the number of steps further significantly extends the denoising time while yielding limited improvement. Therefore, we ultimately select two denoising steps as the optimal choice.

Denoising Steps	DAVIS		
	LPIPS ↓	FloLPIPS ↓	RT (s)
0	0.7882	0.7907	0.06
1	0.0892	0.1216	0.09
2	0.0874	0.1201	0.12
5	0.0877	0.1201	0.28
20	0.0880	0.1204	1.17
50	0.0874	0.1200	2.81

Table A1. Performance comparison across different denoising steps on the DAVIS.

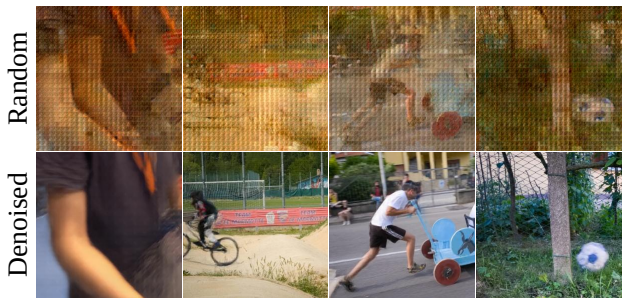

Figure A2. The results of EDEN generated from random noise and denoised latent.

A.3. 1D Tokenizer vs 2D VAE

The ablation results of Transformer-based tokenizer and CNN-based VAE are summarized in Tab. A2. As is shown, our tokenizer significantly outperforms 2D VAE of LBBDM in reconstructing intermediate frames.

	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓
LBBDM VAE	26.64	0.945	0.1471	0.2319
Our Tokenizer	**34.93**	**0.978**	**0.0428**	**0.0626**

Table A2. Performance comparison on the DAIN-HD544p dataset.

A.4. Increasing interval of input frames

We evaluate the performance of LBBDM and EDEN on DAIN-HD544p under increasing frame intervals, with the results summarized in Tab. A3. As is shown, EDEN's performance in fitting intermediate frames still consistently outperforms LBBDM.

	Method	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓
2x	LBBDM	26.64	0.945	0.1471	0.2319
	EDEN	**26.85**	**0.945**	**0.1321**	**0.2184**
4x	LBBDM	19.55	0.855	0.2665	0.3378
	EDEN	**20.91**	**0.869**	**0.2295**	**0.3064**
8x	LBBDM	16.09	0.764	0.4026	0.4376
	EDEN	**17.70**	**0.795**	**0.3328**	**0.3828**

Table A3. Comparison of LBBDM and EDEN under different temporally downsample rates on the DAIN-HD544p dataset.

A.5. Multi-fine-tuning Techniques

We conduct such ablation study on the DAIN-HD544p dataset. The results in Tab. A4 show a clear improvement in the tokenizer's performance after fine-tuning.

	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓
wo/fine-tuning	23.55	0.901	0.3314	0.3616
w/fine-tuning	**34.93**	**0.978**	**0.0428**	**0.0626**

Table A4. Performance comparison with and without fine-tuning.

A.6. Ablation studies of tokenizer dimension

An autoencoder's reconstruction quality directly constrains the achievable image quality in latent diffusion models. The contradiction in **??** of the main paper arises because we used only DiT-B in EDEN due to computational constraints. However, higher-dimensional tokenizers (24) require larger DiTs to fully capture distribution transitions.

To further clarify, we also provide ablation results of the tokenizer combined with DiT on DAIN-HD544p, training for 200k steps, as shown in Tab. A5. Clearly, the interpolation performance of the tokenizer aligns well with its reconstruction capability (as shown in the main paper) when integrated with DiT.

	PFFM	LPIPS↓	FloLPIPS↓
Tokenizer + DiT-B	×	0.1528	0.2531
Tokenizer + DiT-B	√	**0.1497**	**0.2503**

Table A5. Performance comparison of Tokenizer+DiT-B on the DAIN-HD544p dataset.

A.7. Different training datasets

We provide the results of LDMVFI and LBBDM trained on the same dataset, LAVIB, in Tab. A6. Clearly, EDEN still outperforms them when using the same training data.

	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓
LDMVFI	25.88	0.937	0.1501	0.2413
LBBDM	26.56	0.944	0.1477	0.2366
EDEN	**26.85**	**0.944**	**0.1321**	**0.2187**

Table A6. Performance comparison with the same training dataset on the DAIN-HD544p dataset.

A.8. Same number of diffusion steps

Tab. A7 shows the results on DAIN-HD544p using the same number of denoising steps. Clearly, EDEN achieves higher performance with faster speed compared to both LBBDM and LDMVFI.

	Denoising Steps	LPIPS ↓	FloLPIPS ↓	RT (s) ↓
LDMVFI	2	0.1501	0.2413	0.525
LBBDM		0.1477	0.2366	0.907
EDEN		**0.1321**	**0.2187**	**0.250**

Table A7. Runtime comparison of various methods (for interpolating a 544x1280 frame) at 2 denoising steps.

A.9. Limitations and Feature Work

Though our method demonstrates significant improvements in handling complex motions, it still has certain limitations. Specifically, it struggles with blurring when dealing with rapid changes in fine details (e.g., text). As illustrated in Fig. A3, while our method accurately captures the positions of moving car, the text appear blurred. A possible reason for this limitation is that our decoder directly applies pixel shuffle on the tokenizer decoder final layer's output to generate the image, which inherently introduces some degree of blurring. In future work, we plan to explore an effective pixel decoder network to transform the final output of the tokenizer decoder into sharper, more realistic images.

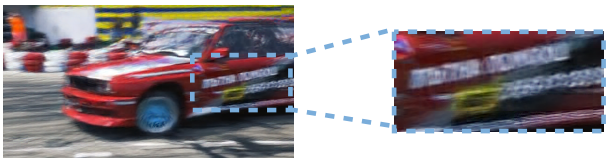

Figure A3. Visualization of results with blurred text.

A.10. More Visualizations

We provide additional visualization comparisons against previous state-of-the-art methods in Fig. A4. These results demonstrate that our method effectively handles complex or nonlinear motions in video frame interpolation. In comparison, prior methods struggle to accurately model such motions.

Figure A4. Visual comparison with different methods, examples selected from DAVIS. Ours outperforms previous methods in both capturing the motion of multiple objects and modeling fast, nonlinear motions.