S1. Additional Results.

We show additional results for both spatial reasoning and retrieval tasks for the maze dataset in Supp. Fig. S1 and S2. We additionally include results of our method generating long-horizon videos conditioned on short context in Supp. Fig. S3. These results are best viewed as videos, we encourage readers to refer to the website attached to the supplementary materials packet.

Additionally, we include FVD results of comparing our method against all relevant baselines under a long-horizon generation task of 240 context frames and 560 generated frames. As shown in Tab. S1, our method achieves the lowest FVD, even compared to a causal transformer trained on the entire context.

Method	FVD (240 + 560) ↓
Causal (192 Frame Context)	78.9
Causal (Full Context)	45.1
Mamba2	163
Mamba2 + Frame Local Attn	45.8
Ours	**38.9**

Table S1. **FVD scores for long-term generation.** Comparison of Fréchet Video Distance (FVD) scores over 560 generated frames given 240 context frames. Our method achieves the lowest FVD, outperforming even causal with full context.

Additional Results on RealEstate10K. We evaluate our method on datasets designed to exhibit long-range temporal dependencies, scene revisitation, and sufficient sequence lengths necessary for comprehensive evaluation of long-context video modeling architectures. Most existing real-world datasets do not meet these criteria. To complement our main experiments, we further assess our method on the RealEstate10K dataset, which provides photorealistic videos with structured camera motion.

Model	Reasoning (64 Frames)		
	SSIM ↑	LPIPS ↓	PSNR ↑
Ours	0.505	0.278	16.7
Causal (Full Context)	0.499	0.274	16.8

Table S2. **Results on RealEstate10K dataset.**

Following the setup described in the main paper, we compare our method against a causal diffusion transformer trained with full context on RealEstate10K. Each model is tasked with generating 64 future frames given 64 context frames, with the ground truth camera trajectory provided during training and inference. Both models have 350M parameters and are trained for 150K iterations. Qualitative and quantitative results show that our model generates photorealistic videos that closely follow the given camera trajectory, achieving metrics comparable to those of the causal baseline.

S2. Implementation Details.

Latent diffusion. Due to the dense information carried in videos, we trained our models on encoded versions of the data. For the Maze dataset, due to the large amount of frames, we use an internal VAE, which compresses token both spatially and temporally. For the minecraft dataset, due to the low number of raw frames, and fair comparisons, we use the same image VAE as DFot [60].

Models. We employ an architecture similar to CogVideo-X [84]. Each baseline model is built upon the same model architecture, opting to replace the attention blocks in each model block with the relevant mechanisms. Depending on the model, the parameter count per layer differs. For fairness, we kept parameter counts for all baselines and comparisons at 200M by adjusting the number of layers for each model.

Training. We train models for different number of iterations depending on task and dataset. For the maze reasoning task, we the first train our model on videos with 400 frames for 150K iterations, then fine-tune this model on 800 frame videos for another 250K steps. Similarly, for the maze retrieval task, we train the model for 100K steps on the 400 frame videos, and 50K extra steps on 800 frames videos. For the minecraft model, we trained on the full 300 frames for 100K steps. We employ our long-context training regime during all stages and for all models. Using a ratio of $p = 0.5$, we switch between two training modes: 1) sampling a random length prefix of the frame sequence to keep un-noised, and 2) sampling random noise levels for all frames. The length of the prefix must exceed half of the total length of the training sequence to further encourage long-context training. When we don't sample a prefix, we keep all tokens noised, in this case, training is the same as diffusion forcing. Note that diffusion forcing is a special case of long-context training, when prefix length is zero. For the memory maze retrieval task, we trained our models for a total of 150K iterations. This equates to approximately 4.3 H100 GPU days for our method, and 5.3 H100 GPU days for the causal architecture trained on the full context.

Frame local attention. We observed significant speedup when using our frame local attention, compared to a fully causal mask by utilizing FlexAttention [15]. For all of our experiments, we chose a frame window size of $k = 10$. For faster training and sampling speeds, we group frames into chunks of 5. In our implementation of frame local attention, frames in a chunk maintain bidirectionality, while also paying attention to frames in the previous chunk, making the effective frame window 10.

S3. Additional Ablations.

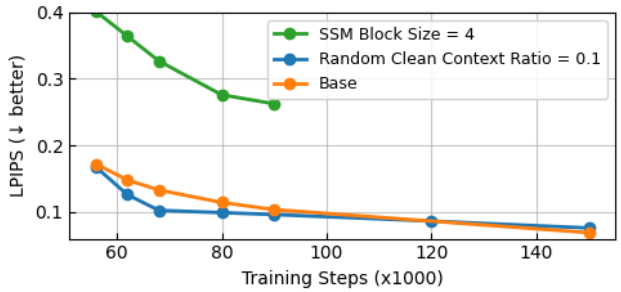
Ablation on frame local attention. We provide additional ablation results below for our model trained on the reasoning task without frame local attention. As expected, the local information provided is crucial for performance.

Model	Reasoning (200 Frames)		
	SSIM ↑	LPIPS ↓	PSNR ↑
Ours w/o Frame Local Attn.	0.813	0.150	25.7
Ours (Full)	**0.855**	**0.099**	**28.2**

Table S3. **Additional Ablation on Frame Local Attention.**

Effect of Hyperparameter Tuning. The additional components introduced in our architecture do expand the design space. Hyperparameters such as the SSM block size, the random clean context ratio, and the use of frame-local attention each influence the model's performance and convergence behavior. In the graph below, we highlight two illustrative modifications: reducing the random clean context ratio from 0.5 to 0.1, and setting all SSM block sizes to 4 across layers.

We observe that using a uniform block size of 4 slows down convergence, while lowering the clean context ratio accelerates convergence but results in slightly worse final performance.



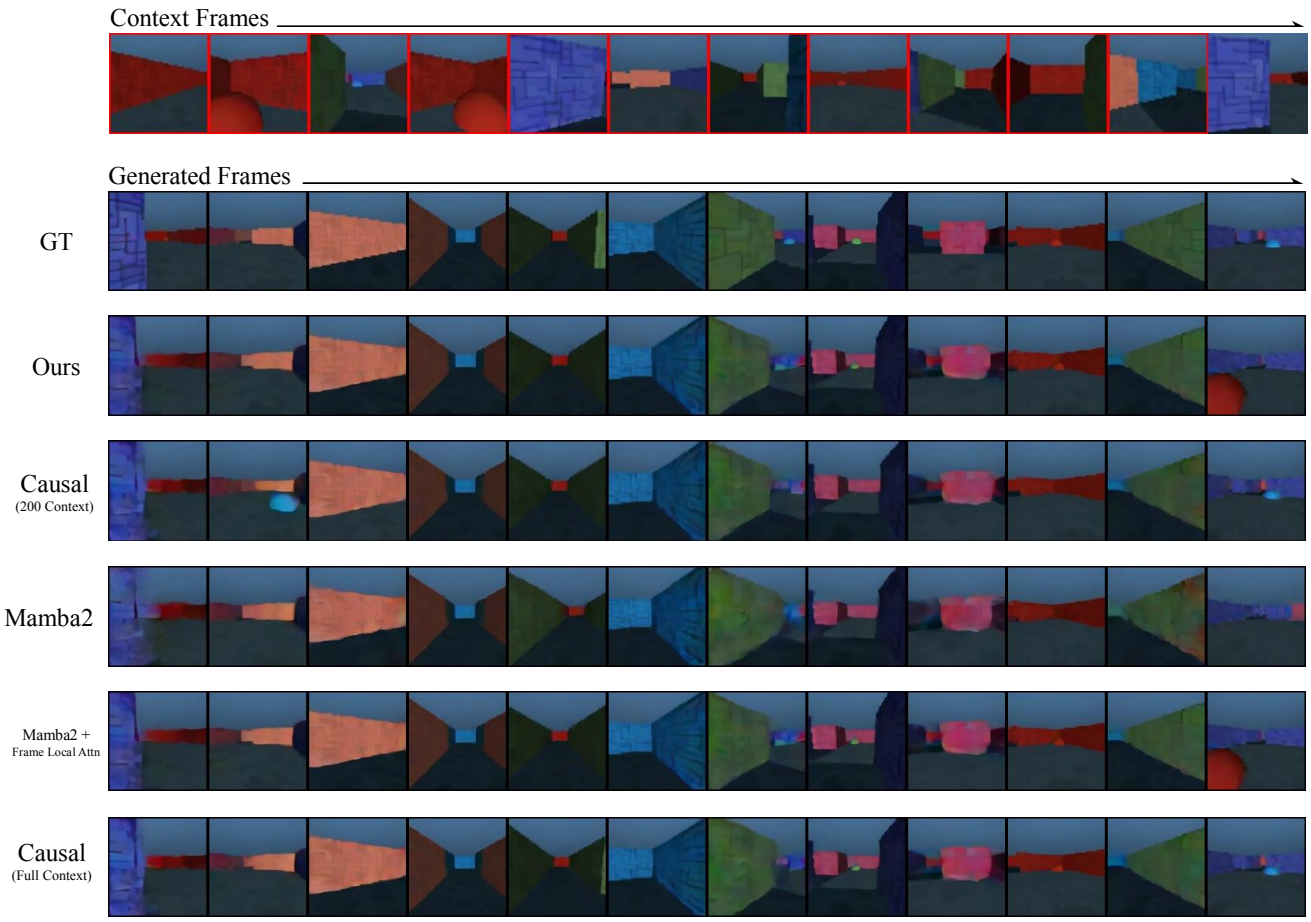


Figure S1. Additional results on reasoning task for the maze dataset.

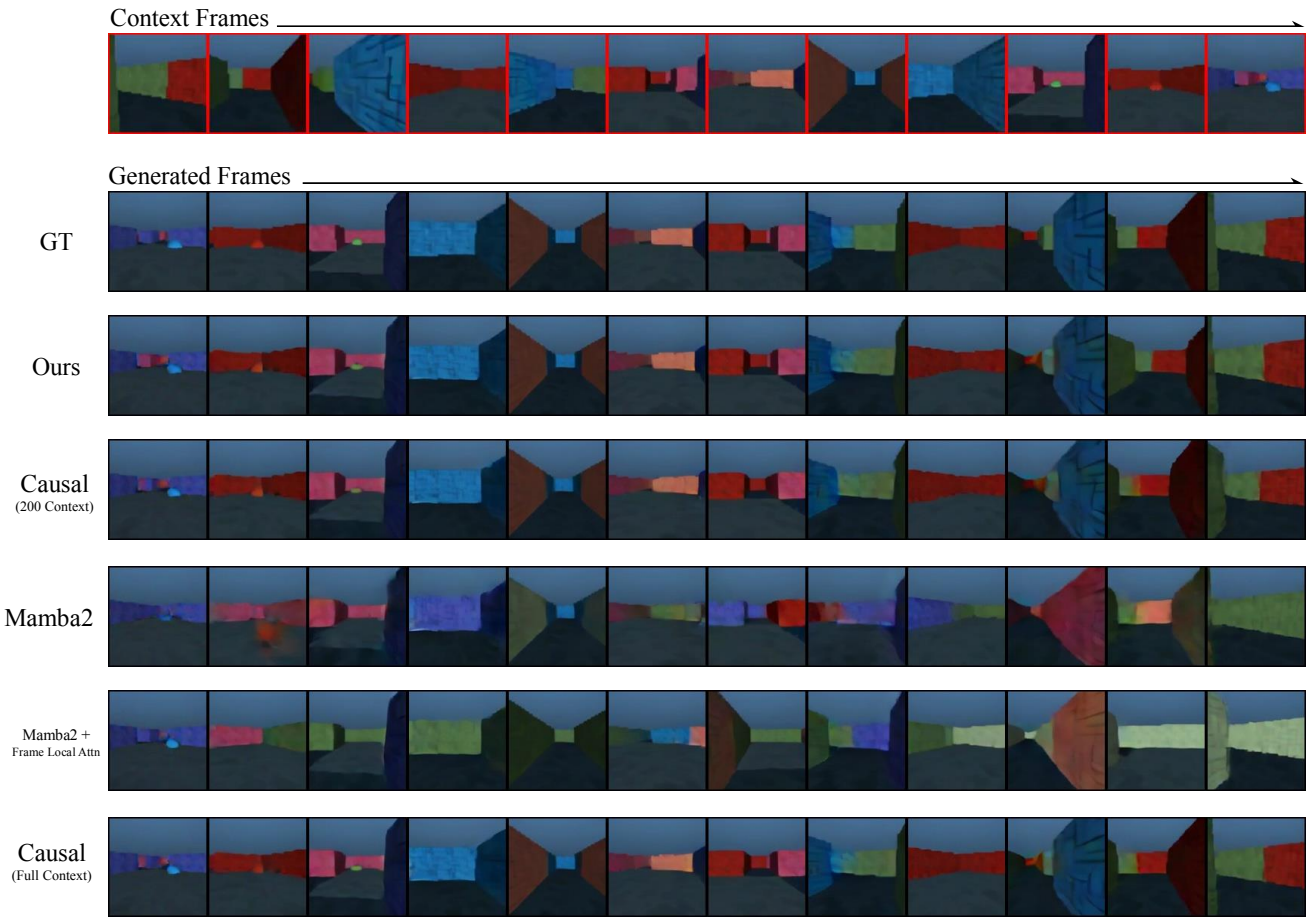


Figure S2. Additional results on retrieval task for the maze dataset.

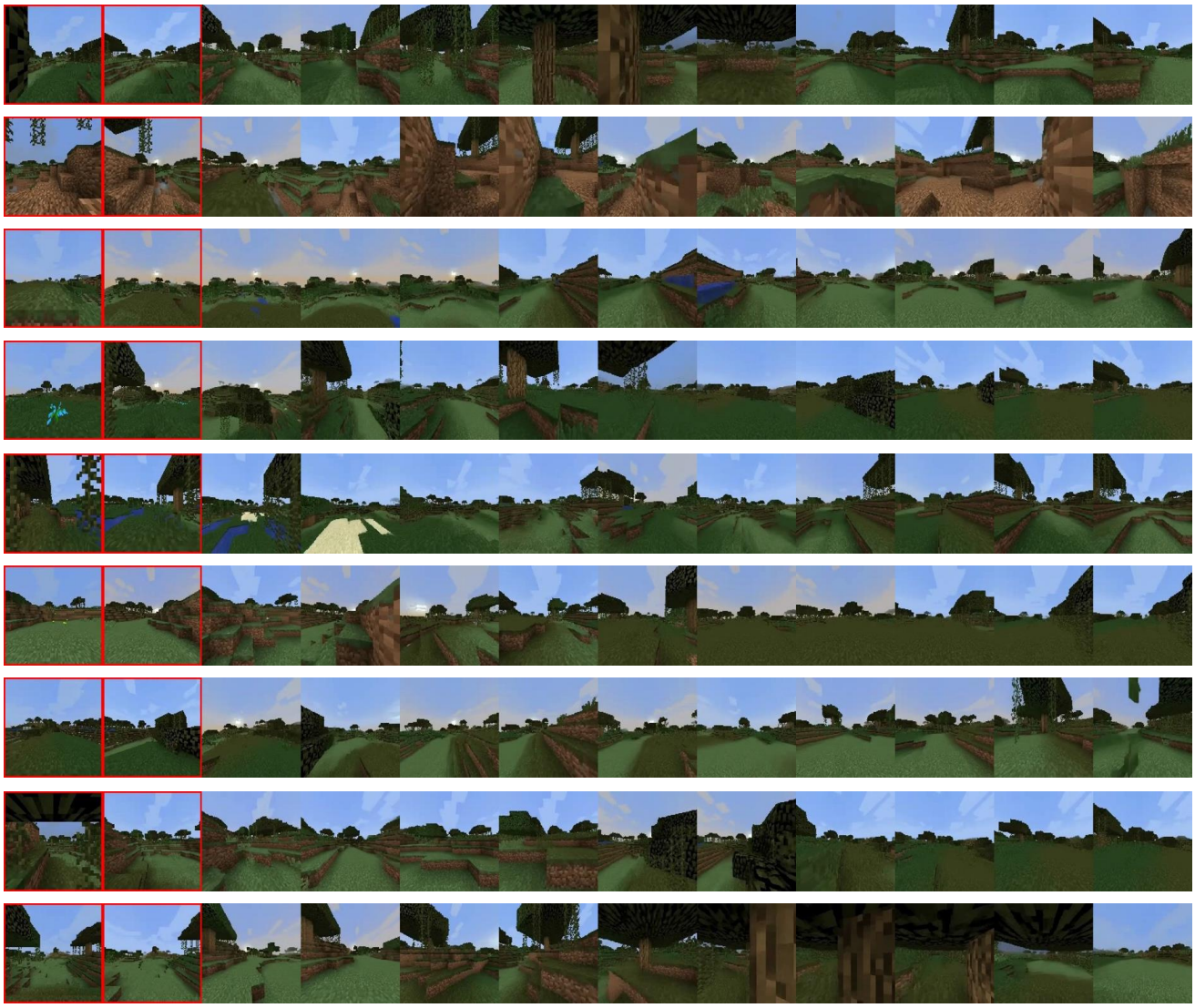


Figure S3. Additional results on long videos generated with our model trained on the minecraft dataset. Model if given 25 context frames and 125 random actions. Context frames are highlighted by a red border. Frames are sampled evenly from from all 150 output frames (including context).