

From Slow Bidirectional to Fast Autoregressive Video Diffusion Models

Supplementary Material

A. Discussion

Although our method is able to generate high-quality videos up to 30 seconds, we still observe quality degradation when generating videos that are extremely long. Developing more effective strategies to address error accumulation remains future work. Moreover, while the latency is significantly lower—by multiple orders of magnitude—compared to previous approaches, it remains constrained by the current VAE design, which necessitates the generation of five latent frames before producing any output pixels. Adopting a more efficient frame-wise VAE could reduce latency by an additional order of magnitude, significantly improving the model’s responsiveness.

Finally, while our method produces high-quality samples using the DMD objective, it comes with reduced output diversity. This limitation is characteristic of reverse KL-based distribution matching approaches. Future work could explore alternative objectives such as EM-Distillation [94] and Score Implicit Matching [56], which may better preserve the diversity of outputs.

While our current implementation is limited to generating videos at around 10 FPS, standard engineering optimizations (including model compilation, quantization, and parallelization) could potentially enable real-time performance. We believe our work marks a significant advancement in video generation and opens up new possibilities for applications in robotic learning [14, 91], game rendering [7, 81], streaming video editing [9], and other scenarios that require real-time and long-horizon video generation.

B. VBench-Long Leaderboard Results

We evaluate CausVid on the VBench-Long dataset using all 946 prompts across 16 standardized metrics. We refer readers to the VBench paper [27] for a detailed description of the metrics. As shown in Tab. 7, our method achieves state-of-the-art performance with the highest total score of 84.27. The radar plot in Fig. 5 visualizes our method’s comprehensive performance advantages. Our method is significantly ahead in several key metrics including dynamic degree, aesthetic quality, imaging quality, object class, multiple objects, and human action. More details can be found on the official benchmark website (https://huggingface.co/spaces/Vchitect/VBench_Leaderboard).

C. Extremely Long Video Generation

Our model demonstrates strong performance on videos exceeding 10 minutes in duration. As shown in Fig. 10, a

14-minute example video exhibits slight overexposure but retains overall high quality.

D. Qualitative Comparison with the Teacher

As demonstrated by VBench (Tab. 4) and human evaluations (Fig. 8), our distilled causal model obtains comparable overall quality to the bidirectional diffusion teacher model. In Fig. 7, we show qualitative comparisons between the two models. Additional qualitative results can be found on our project website (<https://causvid.github.io/>).

E. Limitations

E.1. Long-range Inconsistency

While our method demonstrates high-quality long video generation (see supplementary website), it faces a key limitation: the sliding window inference strategy discards context frames beyond a 10-second horizon. This temporal truncation can cause visual inconsistencies when previously-seen objects or environments reappear after extended periods. A potential solution could involve supervised fine-tuning on a curated set of long videos, inspired by practices in large language model training [11, 62].

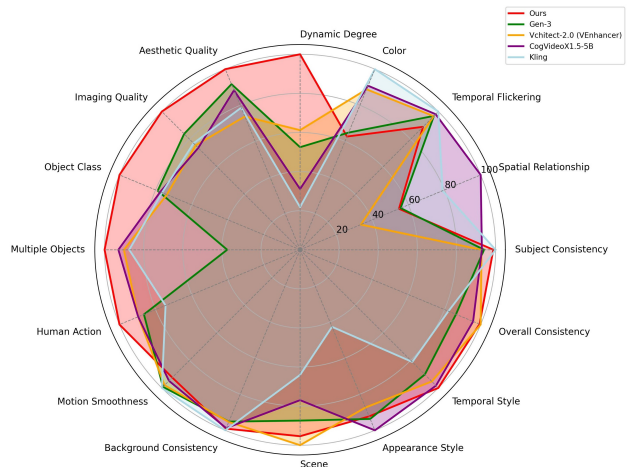


Figure 5. VBench metrics and comparison with previous state-of-the-art methods. Our method (red) performs strongly across different dimensions.

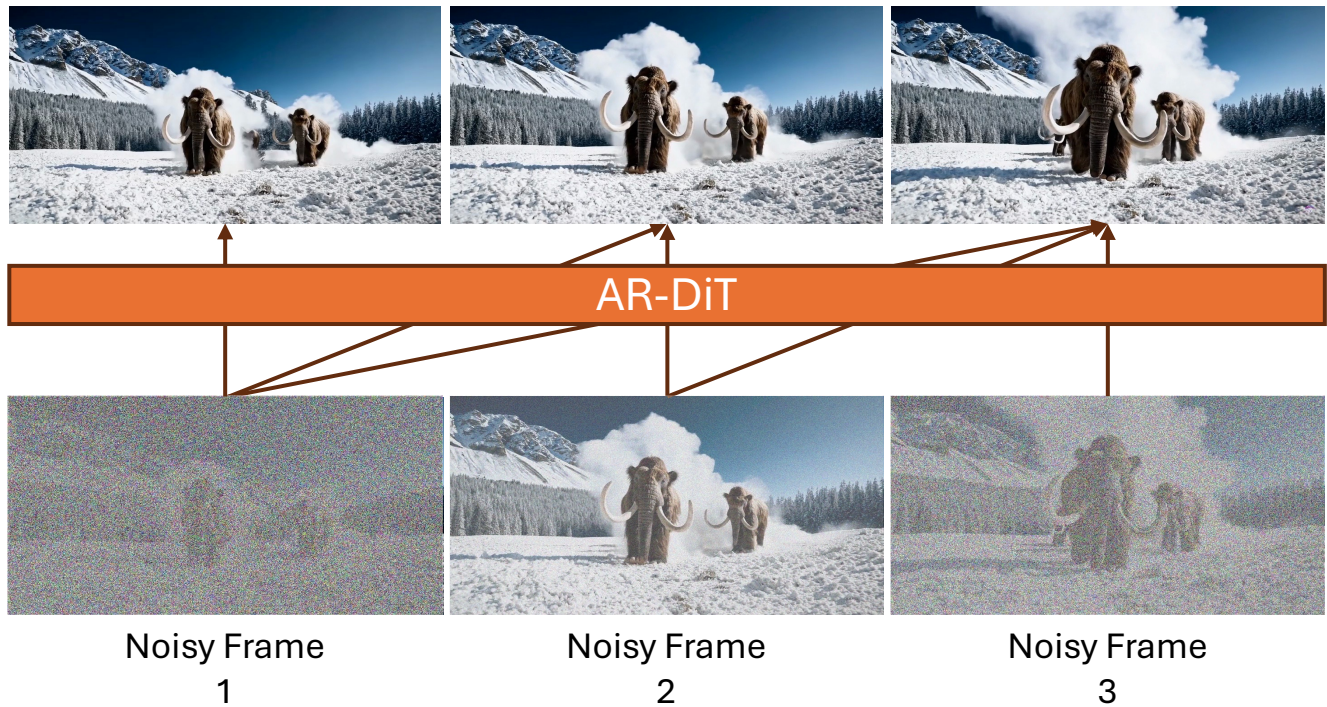


Figure 6. Overview of our autoregressive diffusion transformer architecture. Tokens in the current frame only attend to tokens from current and previous frames, but not the future.

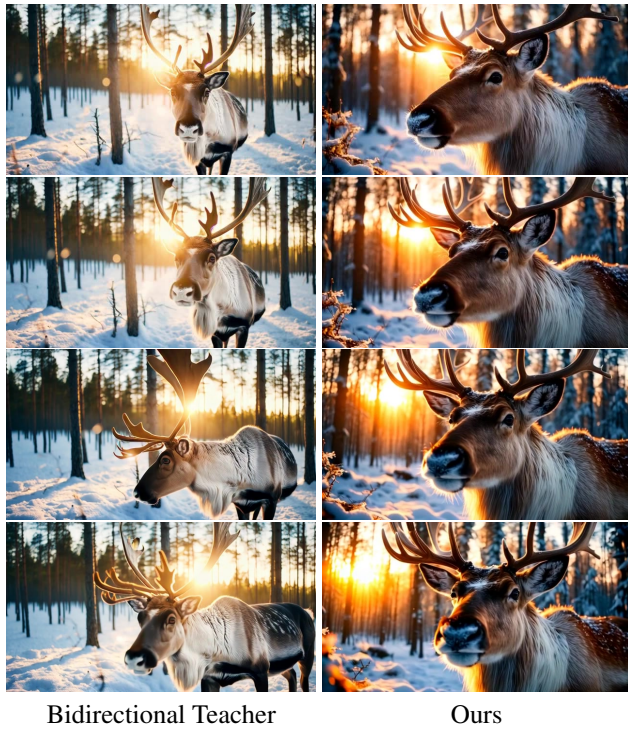


Figure 7. Qualitative comparison with the teacher.

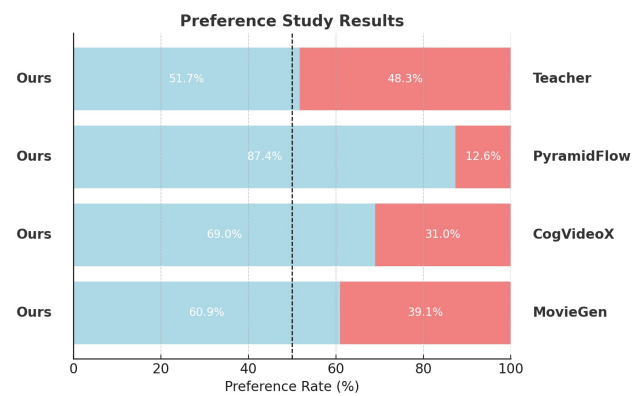


Figure 8. User study comparing our distilled causal video generator with its teacher model and existing video diffusion models. Our model demonstrates superior video quality (scores > 50%), while achieving a significant reduction in latency by multiple orders of magnitude.

Method	Total Score	Quality Score	Semantic Score	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
Vchitect	82.24	83.54	77.06	96.83	96.66	98.57	98.98	63.89	60.41	65.35	86.61	68.84	97.20	87.04	57.55	86.87	23.73	25.01	27.57
Jimeng	81.97	83.29	76.69	97.25	98.39	99.03	98.09	38.43	68.80	67.09	89.62	69.08	98.10	89.05	77.45	44.94	22.27	24.7	27.10
CogVideoX	81.61	82.75	77.04	96.23	96.32	98.66	96.92	70.97	61.08	62.90	85.23	62.11	99.40	82.81	66.35	53.20	24.91	25.38	27.59
Vidu	81.89	83.85	74.04	94.63	96.55	99.08	97.71	82.64	60.87	63.32	88.43	61.68	97.40	83.24	66.18	46.07	21.54	23.79	26.47
Kling	81.85	83.39	75.64	98.33	97.60	99.30	99.40	46.94	61.21	65.62	87.24	68.05	93.40	89.80	73.03	50.86	19.62	24.17	26.42
CogVideoX1.5-5B	82.17	82.78	79.76	96.87	97.35	98.88	98.31	50.93	62.79	65.02	87.47	69.65	97.20	87.55	80.25	52.91	24.89	25.19	27.30
Gen-3	82.32	84.11	75.17	97.10	96.62	98.61	99.23	60.14	63.34	68.82	87.81	53.64	96.40	80.90	65.09	54.57	24.31	24.71	26.69
CanaVid (Ours)	84.27	85.65	78.75	97.53	97.19	98.24	98.05	92.69	64.15	68.88	92.99	72.15	99.80	80.17	64.65	56.58	24.27	25.33	27.51

Table 7. Full comparison on VBench-Long using all 16 metrics. The best scores are in bold. Please zoom in for details.

Label


Which video is more aesthetic and faithfully follow the prompt shown above? (26 models remaining).

Label

Label


Prompt: Several giant wooly mammoths approach treading through a snowy meadow, their long wooly fur lightly blows in the wind as they walk, snow covered trees and dramatic snow capped mountains in the distance, mid afternoon light with wispy clouds and a sun high in the distance creates a warm glow, the low camera view is stunning capturing the large furry mammal with beautiful photography, depth of field.

Video



Vote me

Video



Vote me

Figure 9. An example interface for our user preference study, where videos generated by different methods are displayed in a randomized left/right arrangement.

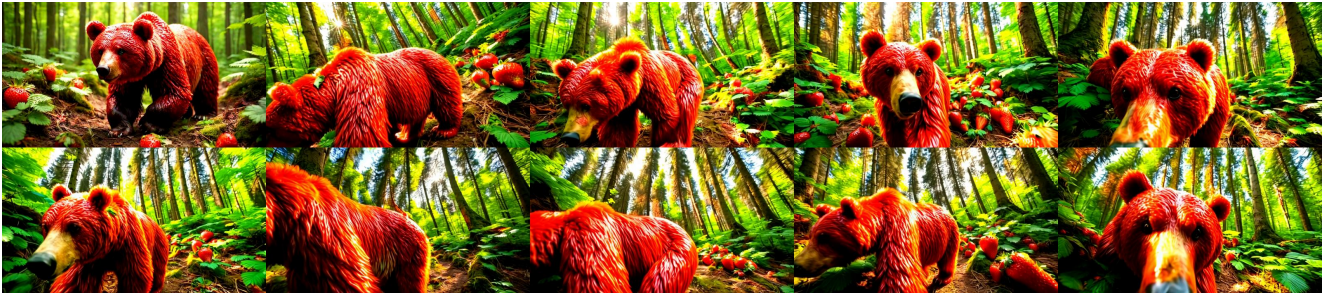


Figure 10. A 14-minute video example generated by our model. Frames are arranged temporally from left to right and top to bottom.



"Several giant woolly mammoths approach treading through a snowy meadow, their long woolly fur lightly blows in the wind as they walk [...]"



"A dynamic motion shot of a paper airplane morphing into a swan. The pointed nose becomes a graceful neck and head, wings unfolding and expanding [...]"



"The slow melting of a snowman, with water trickling down its sides and puddles forming around its base as the temperature warms."



"A breathtaking image of a meteor colliding with the surface of a planet, with bright flames and a massive explosion, illustrating the power and destruction of such an event."



"Cinematic closeup and detailed portrait of a reindeer in a snowy forest at sunset. The lighting is cinematic and gorgeous and soft and sun-kissed, with golden backlight and dreamy bokeh and lens flares [...]"



"in a beautifully rendered papercraft world, a steamboat travels across a vast ocean with wispy clouds in the sky. vast grassy hills lie in the distant background [...]"

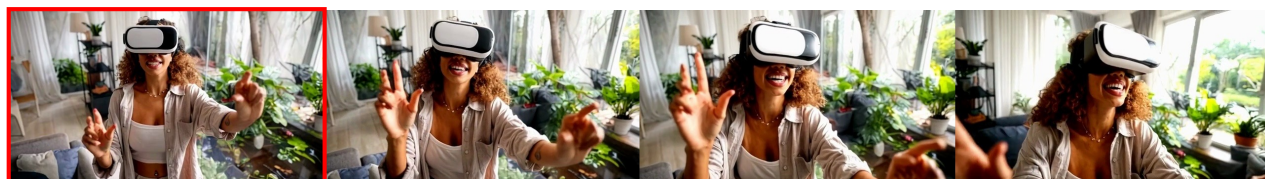


"a spooky haunted mansion, with friendly jack o lanterns and ghost characters welcoming trick or treaters to the entrance, tilt shift photography."

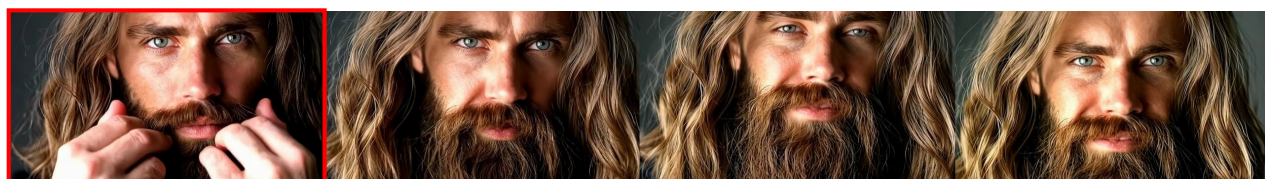
Figure 11. Our model, CausVid, demonstrates that autoregressive video diffusion can be effectively scaled up for general text-to-video tasks, achieving quality on par with bidirectional diffusion models. Moreover, when combined with distillation techniques, it delivers multiple orders of magnitude speedup. Please visit our website for more visualizations.



"Rocket blasting off from a laptop screen on an organized office table. The rocket leaves the screen and blast into space."



"Young woman watching virtual reality in VR glasses in her living room."



"close up portrait of young bearded guy with long beard."



"Hand holding a glowing digital brain, representing the concept of artificial intelligence and innovation in technology."



"The festive atmosphere highlights the celebration of the new year, showcasing bright lights and shimmering decorations for 2025."



"Illustration style of a lightbulb product shot in studio on a background of smaller lightbulbs representing ideas brainstorming."

Figure 12. Trained exclusively on text-to-video generation, our model, CausVid, can be applied zero-shot to image-to-video tasks thanks to its autoregressive design. In the examples shown, the first column represents the input image, while the subsequent frames are generated outputs. Please visit our website for more visualizations.