TurboDiffusion: Accelerating Video Diffusion Models by 100–200 Times

Jintao Zhang[123*], **Kaiwen Zheng**[1*], **Kai Jiang**[12*], **Haoxu Wang**[1*], **Ion Stoica**[3], **Joseph E. Gonzalez**[3], **Jianfei Chen**[1], **Jun Zhu**[12†]

[1]Tsinghua University, [2]Shengshu Technology, [3]UC Berkeley, [†]Corresponding author

https://github.com/thu-ml/TurboDiffusion

Abstract

We introduce `TurboDiffusion`, a video generation acceleration framework that can speed up end-to-end diffusion generation by 100–200× while maintaining video quality. `TurboDiffusion` mainly relies on several components for acceleration: (1) **Attention acceleration**: `TurboDiffusion` uses low-bit SageAttention and trainable Sparse-Linear Attention (SLA) to speed up attention computation. (2) **Step distillation**: `TurboDiffusion` adopts rCM for efficient step distillation. (3) **W8A8 quantization**: `TurboDiffusion` quantizes model parameters and activations to 8 bits to accelerate linear layers and compress the model.

We conduct experiments on the `Wan2.2-I2V-A14B-720P`, `Wan2.1-T2V-1.3B-480P`, `Wan2.1-T2V-14B-720P`, and `Wan2.1-T2V-14B-480P` models. Experimental results show that `TurboDiffusion` achieves 100–200× speedup for video generation on a single RTX 5090 GPU, while maintaining comparable video quality. The GitHub repository, which contains model checkpoints, training, and inference code, is available at https://github.com/thu-ml/TurboDiffusion.

Original *Latency: 184s*



TurboDiffusion *Latency:* **1.9s**



Figure 1 An example of a **5-second video** generated by `Wan2.1-T2V-1.3B-480P` **on a single RTX 5090**.

*Core contributors, co-first authorship.

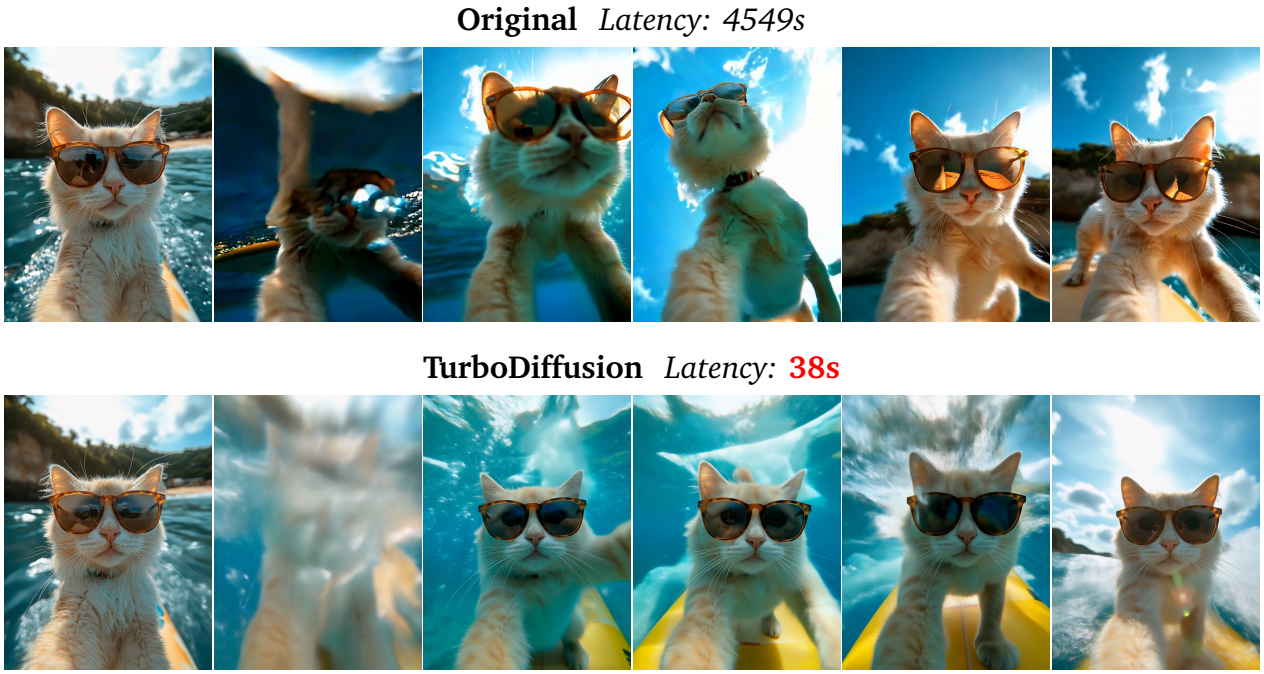


Figure 2 An example of a 5-second video generation on `Wan2.2-I2V-A14B-720P` **using a single RTX 5090**.

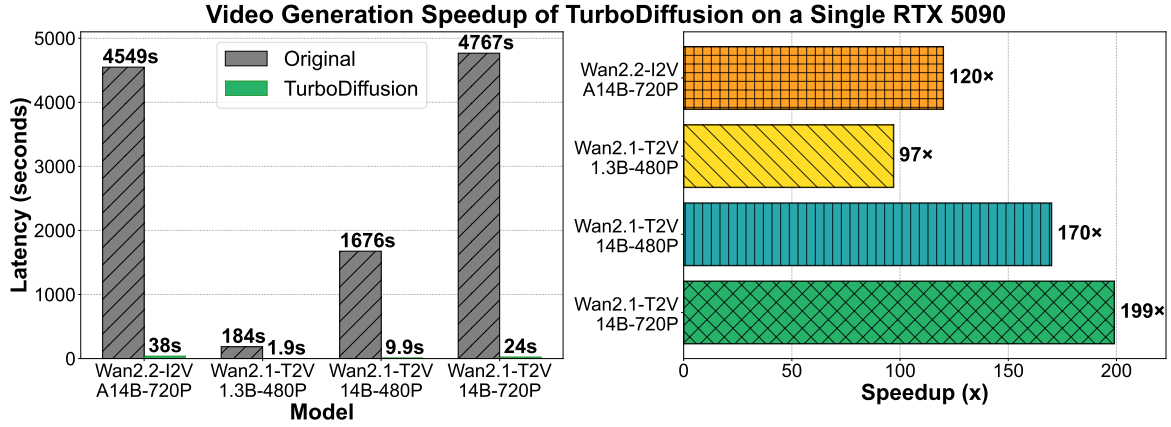


Figure 3 Speedup of `TurboDiffusion` on various video generation models **on a single RTX 5090**.
For Wan2.2-I2V-A14B-720P, the latency includes the switching overhead between the high-noise and low-noise models, resulting in a lower measured speedup compared to Wan2.1-T2V-14B-720P. In theory, the achievable speedup is identical.

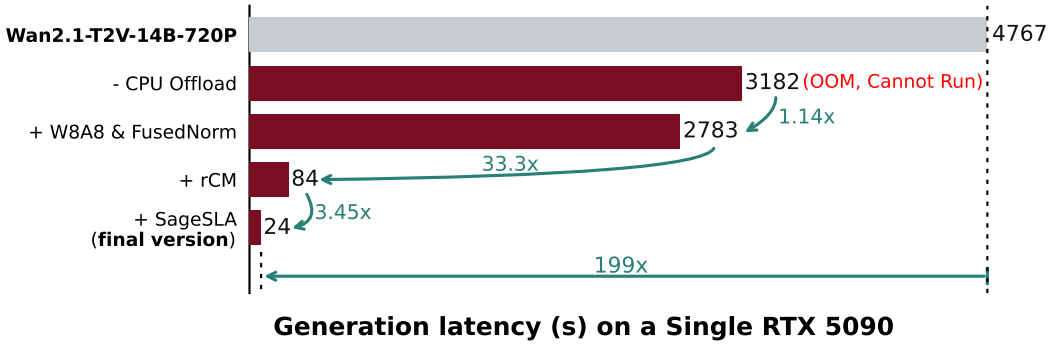


Figure 4 By algorithm and system co-optimization, `TurboDiffusion` reduces the diffusion inference latency of `Wan2.1-T2V-14B-720P` by **around 200× on a single RTX 5090**.

1 Method

We first describe the main techniques used in `TurboDiffusion` in Section 1.1. Then we introduce the training process and inference-time acceleration details of `TurboDiffusion` in Section 1.2 and Section 1.3.

1.1 Main Techniques

`TurboDiffusion` mainly uses four techniques to accelerate diffusion models. First, `TurboDiffusion` uses SageAttention [1–4] for low-bit quantized attention acceleration; specifically, `TurboDiffusion` uses the SageAttention2++ [3] variant. Second, `TurboDiffusion` uses Sparse-Linear Attention (SLA) [5] for sparse attention acceleration. Since sparse computation is orthogonal to low-bit Tensor Core acceleration, SLA can build on top of SageAttention to provide cumulative speedup. Third, `TurboDiffusion` uses rCM [6] to reduce the number of sampling steps, which is currently a state-of-the-art diffusion distillation method. Through model weights merging, rCM naturally inherits attention-level accelerations. Finally, `TurboDiffusion` uses W8A8 quantization for Linear layer acceleration. Specifically, the data type is INT8 and the quantization granularity is block-wise with a block size of 128×128.

1.2 Training

Given a pretrained video diffusion model, `TurboDiffusion` performs the following training process. First, we replace full attention with Sparse-Linear Attention (SLA) [5] and finetune the pretrained model to adapt to sparsity. In parallel, we use rCM [6] to distill the pretrained model into a student model with fewer sampling steps. Second, we merge the parameter updates from both the SLA finetuning and the rCM training into a single model. All training can utilize either real or synthetic data.

Please see our GitHub Code for more details.

1.3 Inference

Given a video generation model trained with Sparse-Linear Attention (SLA) [5] and rCM [6] as described in the previous section, we deploy inference-time acceleration as follows.

Attention acceleration. We replace SLA with SageSLA, which is a CUDA implementation of SLA built on top of SageAttention.

Step distillation. We reduce the number of sampling steps from 100 to a much smaller value, e.g., 4 or 3.

Linear layer quantization. First, we quantize the Linear layer parameters to INT8 with a block-wise granularity of 128×128. Second, during inference, we also quantize the activations in Linear layers to INT8 with the same block-wise granularity and use INT8 Tensor Cores to perform the Linear layer computation. In this way, we compress the model size by roughly half and achieve faster Linear layer computation.

Other optimizations. We reimplement several other operations, such as LayerNorm and RMSNorm, using Triton or CUDA for better efficiency.

Please see our GitHub Code for more details.

2 Evaluations

In this section, we evaluate the efficiency and video quality of `TurboDiffusion`.

2.1 Setup

Models and baseline. We evaluate `TurboDiffusion` on `Wan2.2-I2V-A14B-720P`, `Wan2.1-T2V-1.3B-480P`, `Wan2.1-T2V-14B-720P`, and `Wan2.1-T2V-14B-480P` video diffusion models. We use the official implementation of Wan [7] (denoted as `Original`) and `FastVideo` [8] as our main baselines.

Hyper-parameters. We set the Top-K ratio to 0.1, corresponding to 90% attention sparsity, and use 3 sampling steps. In practice, we recommend using a Top-K value in the range $[0.1, 0.15]$ and setting the number of steps

to 4 to consistently achieve the best video quality. For fastvideo, we use the default parameters in the official implementation (3 sampling steps and 0.8 sparsity in attention).

GPU. Our primary inference experiments are conducted on a single RTX 5090 GPU. In addition, although the speedup is not as large as on the RTX 5090, we also observe substantial acceleration on other GPUs, such as RTX 4090 and H100.

2.2 Efficiency and Quality

We compare the video generation quality and efficiency of the `Original`, `FastVideo`, and `TurboDiffusion`. For efficiency evaluation, we report the end-to-end diffusion generation latency, excluding the text encoding and VAE decoding stages. The following figures present visual comparisons on `Wan2.2-I2V-A14B-720P`, `Wan2.1-T2V-1.3B-480P`, `Wan2.1-T2V-14B-720P`, and `Wan2.1-T2V-14B-480P`, respectively. Since `FastVideo` does not provide an accelerated `Wan2.2-A14B-I2V-720P`, we only compare `TurboDiffusion` with `Original` on `Wan2.2-A14B-I2V-720P`.

From the figures below, we can see that `TurboDiffusion` not only achieves the highest efficiency but also maintains the video quality, demonstrating clear superiority to `FastVideo`.

2.2.1 Wan2.2-I2V-A14B-720P

Original *Latency: 4549s*



TurboDiffusion *Latency:* **38s**



Figure 5 5-second video generation on `Wan2.2-I2V-A14B-720P` **using a single RTX 5090**.
Image prompt is the first frame and the text prompt is "POV selfie video, ultra-messy and extremely fast. A white cat in sunglasses stands on a surfboard with a neutral look when the board suddenly whips sideways, throwing cat and camera into the water; the frame dives sharply downward, swallowed by violent bursts of bubbles, spinning turbulence, and smeared water streaks as the camera sinks. Shadows thicken, pressure ripples distort the edges, and loose bubbles rush upward past the lens, showing the camera is still sinking. Then the cat kicks upward with explosive speed, dragging the view through churning bubbles and rapidly brightening water as sunlight floods back in; the camera races upward, water streaming off the lens, and finally breaks the surface in a sudden blast of light and spray, snapping back into a crooked, frantic selfie as the cat resurfaces."

Original *Latency: 4549s*



TurboDiffusion *Latency:* **38s**



Figure 6 5-second video generation on `Wan2.2-I2V-A14B-720P` **using a single RTX 5090**.
Image prompt is the first frame and the text prompt is "A colorless, rugged, six-wheeled lunar rover—with exposed suspension arms, roll-cage framing, and broad low-gravity tires—glides into view from left to right, kicking up billowing plumes of moon dust that drift slowly in the vacuum. Astronauts in white spacesuits perform light, bouncing lunar strides as they hop aboard the rover's open chassis. In the far distance, a VTOL lander with a vertical, thruster-based descent profile touches down silently on the gray surface. Above it all, vast aurora-like plasma ribbons ripple across the star-filled sky, casting shimmering green, blue, and purple light over the barren lunar plains, giving the entire scene an otherworldly, magical glow."

Original *Latency: 4549s*



TurboDiffusion *Latency:* **38s**



Figure 7 5-second video generation on `Wan2.2-I2V-A14B-720P` **using a single RTX 5090**.
Image prompt is the first frame and the text prompt is "Uma Thurman's Beatrix Kiddo holds her razor-sharp katana blade steady in the cinematic lighting. Without warning, the entire metal piece loses rigidity at once, its material trembling like unstable liquid. The surface destabilizes completely—chunks sag off in slow folds, turning into streams of molten silver that ooze downward in drops. Within moments, the object becomes a collapsing, formless metallic mass, with no edges, and no structure remaining. Thick liquid metal spills from her grip, followed by sheets of shimmering fluid that tear away and fall to the floor. What she holds now is only a quivering blob of mercury-like liquid, constantly sagging and dripping. Her expression shifts from calm readiness to shock and confusion as the last remnants of solidity dissolve, tear apart, and pour through her fingers, leaving her defenseless and disoriented."

Original *Latency: 4549s*



TurboDiffusion *Latency:* **38s**



Figure 8 5-second video generation on `Wan2.2-I2V-A14B-720P` **using a single RTX 5090**.
Image prompt is the first frame and the text prompt is "Close-up on an elderly sailor in a weathered yellow raincoat, seated on the sun-lit deck of a gently rocking catamaran. With each small rise and dip of the hull, the shadows on his face shift subtly. He draws from his pipe, the ember brightening, and the exhale sends a thin ribbon of smoke that wavers and bends as the boat sways. His cat lies beside him, eyes half-closed, its body adjusting with soft, instinctive shifts whenever the deck tilts. Sunlight glints off the polished wood in flickering patterns as the surface of the water rolls beneath, causing brief flares of moving reflections. A pair of seabirds glide overhead, dipping slightly as the wind gusts. As the camera eases into a slow push-in, every motion becomes more pronounced—the smoke trembling, the cat's fur fluttering, the deck creaking with each gentle sway—turning the peaceful moment into a dynamically living scene afloat at sea."

Original *Latency: 4549s*



TurboDiffusion *Latency:* **38s**

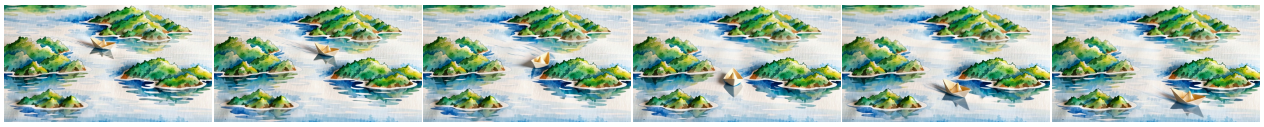


Figure 9 5-second video generation on `Wan2.2-I2V-A14B-720P` **using a single RTX 5090**.
Image prompt is the first frame and the text prompt is "Watercolor style. Wet suminagashi inks surge and spread rapidly across the paper, swirling outward as they form island-like shapes with actively shifting, bleeding edges. A tiny paper boat is pulled forward by a faster-moving stream of pigment, gliding swiftly toward the still-wet areas. The flow pushes it in small sudden bursts, creating sharper, overlapping ripples that distort its reflection. Ink currents twist around the boat, forming brief vortices and drifting streaks that keep redirecting its path. Soft natural side-light catches the moving sheen on the wet paper, enhancing the sense of continuous, fluid motion across the painted landscape."

Original *Latency: 4549s*



TurboDiffusion *Latency:* **38s**



Figure 10 5-second video generation on `Wan2.2-I2V-A14B-720P` **using a single RTX 5090**.
Image prompt is the first frame and the text prompt is "she looks up, and then looks back"

Original *Latency: 4549s*



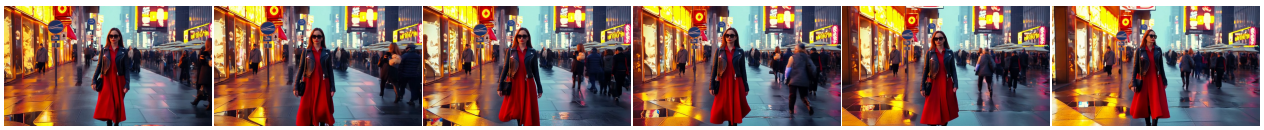
TurboDiffusion *Latency:* **38s**



Figure 11 5-second video generation on `Wan2.2-I2V-A14B-720P` **using a single RTX 5090**.
Image prompt is the first frame and the text prompt is "A man in a trench coat holding a black umbrella moves at a rapid, urgent pace through the streets of Tokyo on a rainy night, splashing hard through puddles. A handheld follow-cam tracks him from the side and slightly behind with quick, jittery motion, as if struggling to keep up. The focus stays locked on the man while neon signs streak into long, colorful bokeh trails from the fast movement. The scene has a cyberpunk, film-noir mood—mysterious, lonely, and restless. The slick pavement reflects vibrant neon light; raindrops cut sharply through the frame; a thin fog shifts as the man pushes forward with fast, determined steps."

2.2.2 Wan2.1-T2V-1.3B-480P

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**

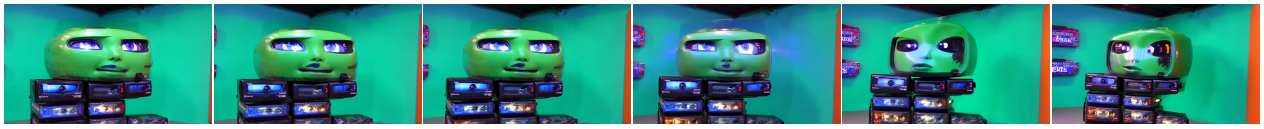


Figure 12 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*

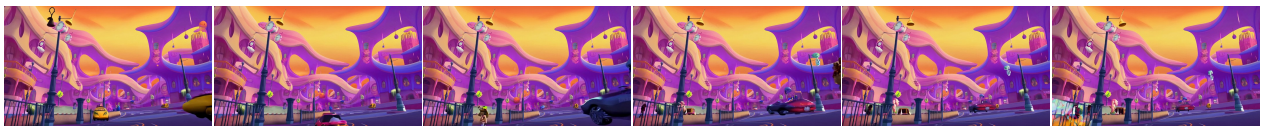


TurboDiffusion *Latency:* **1.9s**



Figure 13 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "The camera rotates around a large stack of vintage televisions all showing different programs — 1950s sci-fi movies, horror movies, news, static, a 1970s sitcom, etc, set inside a large New York museum gallery."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**

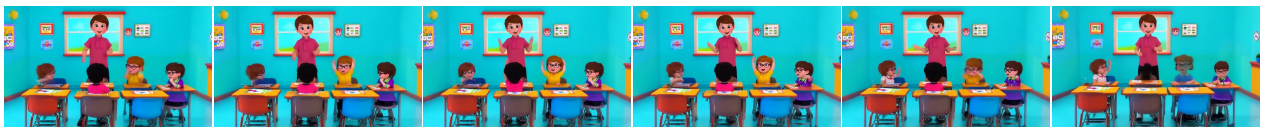


Figure 14 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "In the surreal world reminiscent of Salvador Dali's style, downtown Clowntown morphs into a mesmerizing dreamscape. The streets twist and bend in impossible ways, with melting clocks hanging from lampposts and floating objects suspended mid-air. Giant distorted buildings with soft, flowing edges loom over the scene, blending seamlessly into each other. The sky is a vivid blend of oranges, pinks, and purples, casting an ethereal glow over the bizarre landscape. Characters move slowly and dreamily through the streets, their expressions detached and contemplative. The scene is captured as a wide-angle view, emphasizing the surreal and dreamlike atmosphere."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**

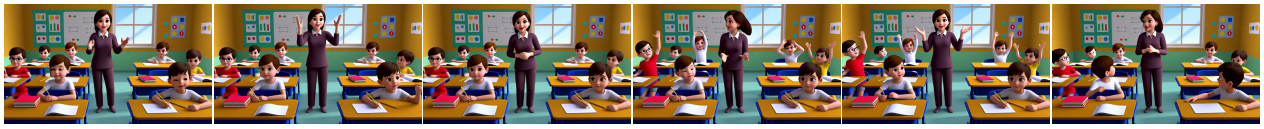


Figure 15 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "A 3D classroom scene filled with small children aged between 6 and 8 years old, all attentively sitting at their desks. A dedicated teacher stands in front of the class, engaged in lively interaction, using gestures and facial expressions to convey enthusiasm. The classroom is brightly lit, with colorful posters and charts adorning the walls. Each child has a focused expression, some nodding along, others raising their hands to ask questions. The desks are neatly arranged in rows, with books and pencils organized on each desk. The background showcases a window with sunlight streaming in, casting gentle shadows across the room. Wide shot, capturing the entire classroom from a frontal perspective."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*

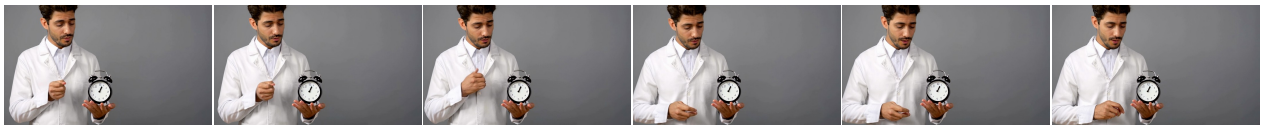


TurboDiffusion *Latency:* **1.9s**

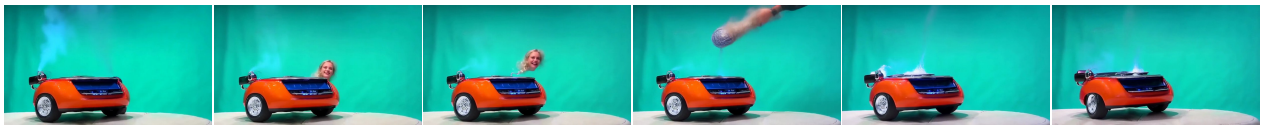


Figure 16 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "A Minecraft player character holding a torch enters a massive underground cave. The torchlight flickers against jagged stone walls, illuminating patches of iron and diamond ores embedded in the rock. Stalactites hang from the ceiling, lava flows in glowing streams nearby, and the faint sound of water dripping echoes through the cavern."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**

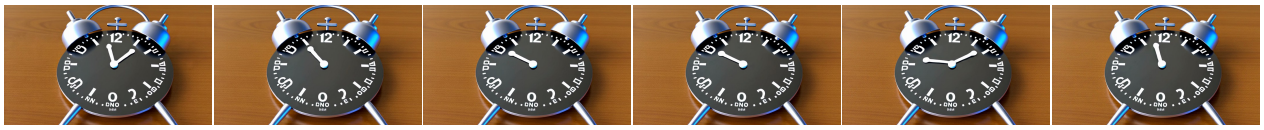


Figure 17 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "alarm clock".

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**



Figure 18 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "A close-up shot captures a steaming hot pot brimming with vegetables and dumplings, set on a rustic wooden table. The camera focuses on the bubbling broth as a woman, dressed in a light, patterned blouse, reaches in with chopsticks to lift a tender leaf of cabbage from the simmering mixture. Steam rises around her as she leans back slightly, her warm smile reflecting satisfaction and joy. Her movements are smooth and deliberate, showcasing her comfort and familiarity with the dining process. The background includes a small bowl of dipping sauce and a clay pot, adding to the cozy, communal dining atmosphere."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**



Figure 19 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = *"A dramatic and dynamic scene in the style of a disaster movie, depicting a powerful tsunami rushing through a narrow alley in Bulgaria. The water is turbulent and chaotic, with waves crashing violently against the walls and buildings on either side. The alley is lined with old, weathered houses, their facades partially submerged and splintered. The camera angle is low, capturing the full force of the tsunami as it surges forward, creating a sense of urgency and danger. People can be seen running frantically, adding to the chaos. The background features a distant horizon, hinting at the larger scale of the tsunami. A dynamic, sweeping shot from a low-angle perspective, emphasizing the movement and intensity of the event."*

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**



Figure 20 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "A futuristic robotic arm writing on a large glass screen. The robot has sleek metallic joints and a precise, articulated hand. It moves smoothly across the screen, leaving behind elegant, flowing text. The lighting highlights the reflective surfaces, creating dynamic shadows and highlights. The background is a high-tech lab with neon accents and advanced machinery. The scene focuses on the robotic arm as it writes, emphasizing the fluid motion and intricate design. Medium close-up shot, static camera."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**



Figure 21 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "A man and woman, walking hand in hand down a vibrant, winding street, their figures illuminated against a backdrop reminiscent of a Van Gogh painting. They are strolling leisurely, with the man wearing a casual, striped shirt and the woman in a floral dress, both with warm, content expressions. Behind them, a vintage bus passes slowly, adding to the lively atmosphere of the scene. The background features swirling brushstrokes of blues, yellows, and greens, typical of Van Gogh2019s style, with a charming village street filled with blooming flowers and quaint buildings. Medium shot, capturing the couple from mid-torso up, with the bus visible in the background."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**



Figure 22 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "A slow zoom into a bustling candy factory during operation, with smoke gently rising from multiple chimneys. The scene showcases workers actively engaged in various tasks such as mixing ingredients, operating machinery, and packaging candies. The interior is bright and colorful, filled with vibrant hues of the candies being produced. The exterior features a mix of old and modern industrial architecture, with large windows allowing glimpses of the production lines inside. The focus gradually narrows from an overview of the factory grounds to a detailed close-up of the chimneys emitting soft plumes of smoke. Wide shot transitioning to medium shot with a slow zoom effect."

Original *Latency: 184s*



FastVideo *Latency: 5.3s*



TurboDiffusion *Latency:* **1.9s**



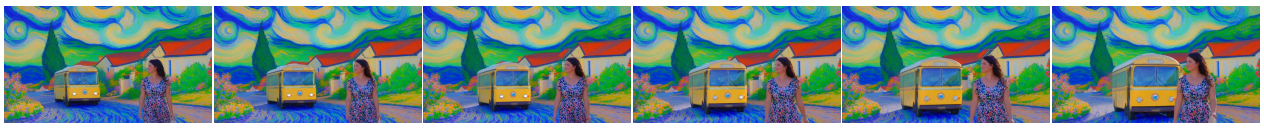
Figure 23 5-second video generation on `Wan2.1-T2V-1.3B-480P` **using a single RTX 5090**.
Prompt = "A dynamic and chaotic scene in a dense forest during a heavy rainstorm, capturing a real girl frantically running through the foliage. Her wild hair flows behind her as she sprints, her arms flailing and her face contorted in fear and desperation. Behind her, various animals—rabbits, deer, and birds—are also running, creating a frenzied atmosphere. The girl's clothes are soaked, clinging to her body, and she is screaming and shouting as she tries to escape. The background is a blur of greenery and rain-drenched trees, with occasional glimpses of the darkening sky. A wide-angle shot from a low angle, emphasizing the urgency and chaos of the moment."

2.2.3 Wan2.1-T2V-14B-720P

Original *Latency: 4767s*



FastVideo *Latency: 72.6s*



TurboDiffusion *Latency:* **24s**

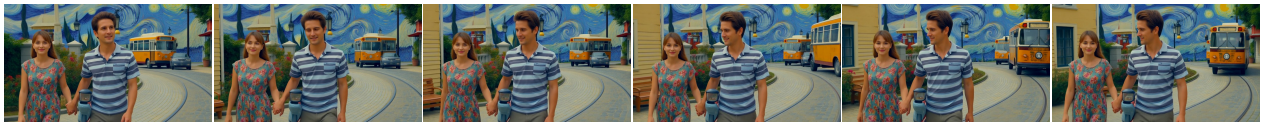
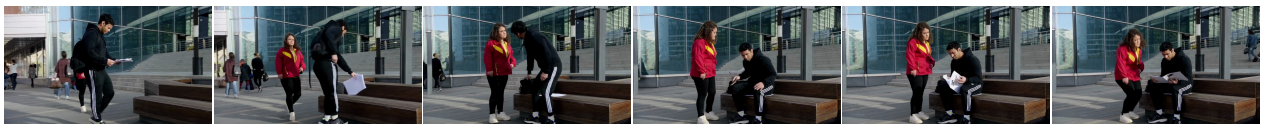


Figure 24 5-second video generation on `Wan2.1-T2V-14B-720P` **using a single RTX 5090**.
Prompt = "A man and woman, walking hand in hand down a vibrant, winding street, their figures illuminated against a backdrop reminiscent of a Van Gogh painting. They are strolling leisurely, with the man wearing a casual, striped shirt and the woman in a floral dress, both with warm, content expressions. Behind them, a vintage bus passes slowly, adding to the lively atmosphere of the scene. The background features swirling brushstrokes of blues, yellows, and greens, typical of Van Gogh̆2019s style, with a charming village street filled with blooming flowers and quaint buildings. Medium shot, capturing the couple from mid-torso up, with the bus visible in the background."

Original *Latency: 4767s*



FastVideo *Latency: 72.6s*



TurboDiffusion *Latency:* **24s**

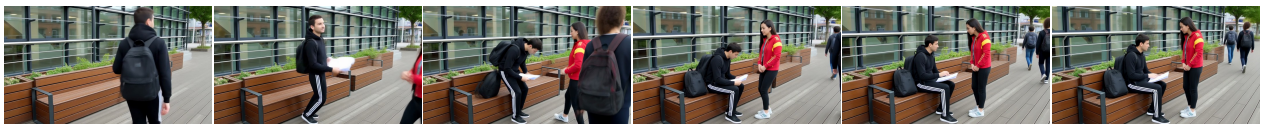


Figure 25 5-second video generation on `Wan2.1-T2V-14B-720P` **using a single RTX 5090**.
Prompt = "In an urban outdoor setting, a man dressed in a black hoodie and black track pants with white stripes walks toward a wooden bench situated near a modern building with large glass windows. He carries a black backpack slung over one shoulder and holds a stack of papers in his hand. As he approaches the bench, he bends down, places the papers on it, and then sits down. Shortly after, a woman wearing a red jacket with yellow accents and black pants joins him. She stands beside the bench, facing him, and appears to engage in a conversation. The man continues to review the papers while the woman listens attentively. In the background, other individuals can be seen walking by, some carrying bags, adding to the bustling yet casual atmosphere of the scene. The overall mood suggests a moment of focused discussion or preparation amidst a busy environment."

Original *Latency: 4767s*



FastVideo *Latency: 72.6s*



TurboDiffusion *Latency:* **24s**



Figure 26 5-second video generation on `Wan2.1-T2V-14B-720P` **using a single RTX 5090**.
Prompt = "A Minecraft player character holding a torch enters a massive underground cave. The torchlight flickers against jagged stone walls, illuminating patches of iron and diamond ores embedded in the rock. Stalactites hang from the ceiling, lava flows in glowing streams nearby, and the faint sound of water dripping echoes through the cavern."

Original *Latency: 4767s*



FastVideo *Latency: 72.6s*



TurboDiffusion *Latency:* **24s**



Figure 27 5-second video generation on `Wan2.1-T2V-14B-720P` **using a single RTX 5090**.
Prompt = "A documentary-style video focusing on Greenland's geopolitical landscape. The scene opens with aerial shots of Greenland's vast icy terrain, emphasizing its strategic location. Transition to close-ups of Greenlandic Inuit people discussing their aspirations for increased autonomy and recognition, conveying a sense of determination and hope. Interspersed with these are interviews with political figures and experts, highlighting the complexities and challenges posed by foreign interests. The background includes maps and satellite imagery illustrating Greenland's unique position. The overall tone is informative and balanced, presenting both opportunities and risks. Wide and medium shots throughout, with occasional zoom-ins on key points during interviews."

2.2.4 Wan2.1-T2V-14B-480P

Original *Latency: 1676s*



FastVideo *Latency: 26.3s*



TurboDiffusion *Latency:* **9.9s**

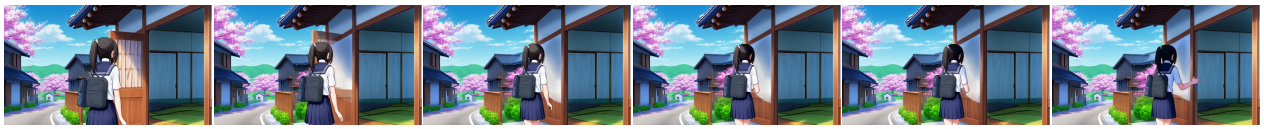


Figure 28 5-second video generation on `Wan2.1-T2V-14B-480P` **using a single RTX 5090**.
Prompt = "A Minecraft player character holding a torch enters a massive underground cave. The torchlight flickers against jagged stone walls, illuminating patches of iron and diamond ores embedded in the rock. Stalactites hang from the ceiling, lava flows in glowing streams nearby, and the faint sound of water dripping echoes through the cavern."

Original *Latency: 1676s*



FastVideo *Latency: 26.3s*



TurboDiffusion *Latency:* **9.9s**



Figure 29 5-second video generation on `Wan2.1-T2V-14B-480P` **using a single RTX 5090**.
Prompt = "Anime-style illustration, a young Japanese woman walking into a traditional house in a quaint Japanese town. She has shoulder-length black hair tied in a ponytail, wears a simple school uniform consisting of a white blouse and a navy blue pleated skirt, and carries a small backpack. The house features a typical Japanese entrance with a wooden door and a tatami mat inside visible through the partially opened door. Surrounding the house are traditional Japanese buildings and narrow streets lined with cherry blossom trees in full bloom. The sky is clear with soft pastel colors. Medium shot focusing on the woman as she steps into the house, showing her interaction with the traditional door."

3 Conclusion and Future Work

We present `TurboDiffusion`, a video generation acceleration framework that achieves 100–200× end-to-end diffusion speedup with negligible quality degradation. `TurboDiffusion` combines low-bit attention (SageAttention), Sparse-Linear Attention (SLA), step distillation via rCM, and W8A8 quantization, together with several additional engineering optimizations. Experiments on `Wan2.2-I2V-A14B-720P`, `Wan2.1-T2V-1.3B-480P`, `Wan2.1-T2V-14B-720P`, and `Wan2.1-T2V-14B-480P` demonstrate that `TurboDiffusion` reduces the generation time of a single video to <1 minute on a single RTX 5090 GPU, making high-quality video generation substantially more efficient and practical.

For future work, we plan to extend this framework to support more video generation paradigms, such as autoregressive video diffusion.

Acknowledgments

We thank Jia Wei, Pengle Zhang, and Kaichao You from Tsinghua University, and Yifan Qiao, Haocheng Xi, Shuo Yang, Roger Wang, and Simon Mo from Sky Lab, UC Berkeley, for their help.

References

[1] Jintao Zhang, Jia Wei, Pengle Zhang, Jun Zhu, and Jianfei Chen. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In International Conference on Learning Representations (ICLR), 2025.

[2] Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. In International Conference on Machine Learning (ICML), 2025.

[3] Jintao Zhang, Xiaoming Xu, Jia Wei, Haofeng Huang, Pengle Zhang, Chendong Xiang, Jun Zhu, and Jianfei Chen. Sageattention2++: A more efficient implementation of sageattention2. arXiv preprint arXiv:2505.21136, 2025.

[4] Jintao Zhang, Jia Wei, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Haoxu Wang, Kai Jiang, Jun Zhu, and Jianfei Chen. Sageattention3: Microscaling fp4 attention for inference and an exploration of 8-bit training. arXiv preprint arXiv:2505.11594, 2025.

[5] Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, Joseph E. Gonzalez, Jun Zhu, and Jianfei Chen. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention. arXiv preprint arXiv:2509.24006, 2025.

[6] Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. arXiv preprint arXiv:2510.08431, 2025.

[7] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

[8] The FastVideo Team. Fastvideo: A unified framework for accelerated video generation, April 2024.