JETDIFFUSION: PUSHING THE SPEED LIMITS OF VIDEO GENERATION WITH NVFP4 SPARSE ATTENTION

Anonymous authors
Paper under double-blind review

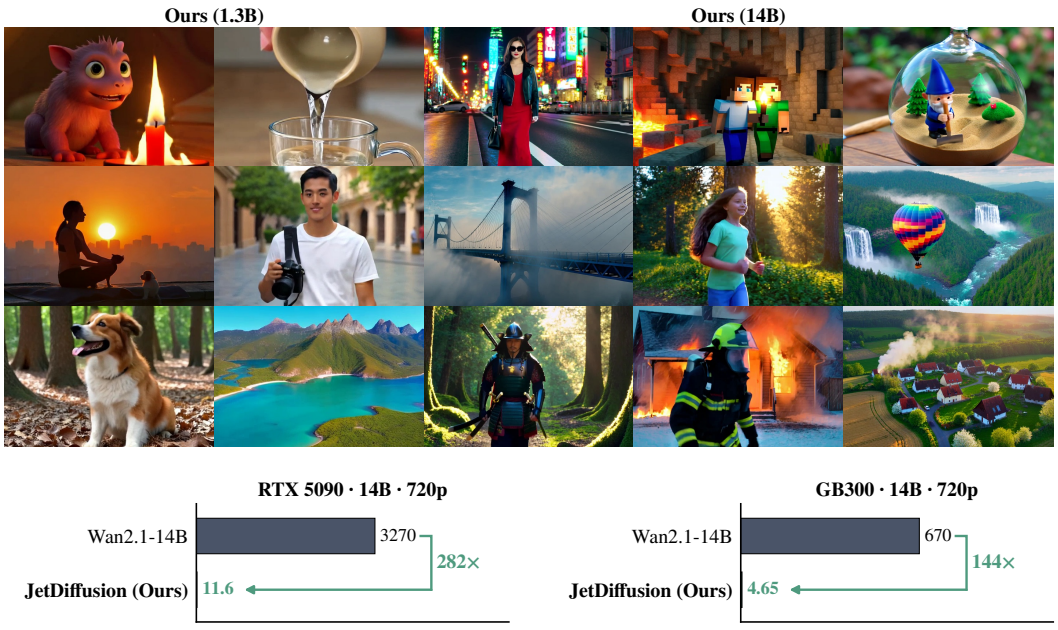


Figure 1: **Accelerating video generation with JetDiffusion.** Top: frames generated by JetDiffusion, from Wan2.1-1.3B in the left six panels and from Wan2.1-14B in the right nine. Bottom: end-to-end latency (s) of one Wan2.1-14B 720p generation on an RTX 5090 and a GB300, compared with the base model. Appendix D gives the full prompts and compares frames from each video with the base model.

ABSTRACT

Attention over long sequences is a major computational bottleneck in video generation, which can be alleviated through sparsity and low-bit quantization. We propose **JetDiffusion**, a video generation system built on trainable **V**ideo **S**parse **Q**uantized **A**ttention (**VSQA**), with custom inference kernels that combine native NVFP4 computation and kernel fusion. At 90% sparsity, VSQA accelerates attention by 17-29x compared with FlashAttention on different hardwares. To further accelerate video generation, we combine VSQA with NVFP4 linear quantization and step distillation. Although each technique can individually preserve generation quality, jointly adapting to all three remains challenging, particularly under aggressive NVFP4 quantization. Through empirical studies, we develop an effective staged training recipe that enables stable adaptation to quantization, sparsity, and few-step denoising. On Wan2.1-14B, the resulting model achieves 282x end-to-end video generation speedup on the RTX 5090 while maintaining comparable video quality.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Attention is a major computational bottleneck in video generation: higher resolution and longer duration increase the number of spatiotemporal tokens, and the cost of full attention grows quadratically with it. Sparse attention (Zhang et al., 2025d; Xi et al.; Zhang et al., 2026f; Li et al., 2026b; Zhang et al., 2026b) and low-bit attention (Zhang et al., 2025b; 2024; 2026e;g) reduce this cost in complementary ways: sparsity reduces the number of token interactions, while quantization lowers the cost of the interactions that remain. However, combining high sparsity with 4-bit quantization significantly perturb the pretrained model, causing compounding loss in generation quality.

Recent works apply post-training quantization to sparse attention without any training (Zhang et al.; Zhao et al., 2026b), or adapt the model to sparsity and quantization jointly through training (Liu et al., 2025a; Zhang et al., 2026d). These efforts mainly target 8-bit attention. Native NVFP4 support on NVIDIA Blackwell GPUs (Alvarez et al., 2025) makes 4-bit matrix multiplication substantially cheaper, offering an opportunity to further accelerate sparse attention. However, 4-bit precision leaves far less room for error than 8-bit. With NVFP4, quantization error grows sharply, and post-training calibration alone no longer preserves the quality of the pretrained model (Zhang et al., 2026g). Moreover, quantizing attention breaks the identities that standard attention kernels rely on in the backward pass, requiring dedicated kernels for trainable NVFP4 sparse attention.

We propose **V**ideo **S**parse **Q**uantized **A**ttention (VSQA), to our knowledge the first trainable attention method that combines high sparsity with NVFP4 quantization. VSQA keeps the coarse-to-fine structure of VSA (Zhang et al., 2026f), which pools tokens into cubes to select the cubes each query attends to, and computes both $QK^\top$ and PV of the selected token-level attention in NVFP4. The inference kernels use native NVFP4 matrix multiplication, traverse only the selected cube pairs, and fuse pooling with quantization so that non-matrix work stays small as sparsity increases. We implement both training and inference kernels for the RTX 5090 and the GB300.

To achieve acceleration in the end-to-end system, we extend VSQA into **JetDiffusion**, a complete system where we additionally quantize the linear layers to NVFP4 and distill the model to three denoising steps. Realizing these compression techniques without sacrificing generation quality requires adapting the model to the combined effects of low-precision computation, sparse attention, and few-step denoising. To maximize speedup, our quantization setup operates without smoothing tricks, making quantization error a central challenge for this adaptation.

Through empirical observations, we find that a dedicated reconstruction warm-up substantially stabilizes subsequent training by reducing quantization error. While conventional QAT/QAD objectives can reduce quantization error, our observations motivate a more targeted objective: layer-wise reconstruction, which directly matches quantized features to a frozen, unquantized reference (Li et al., 2021). Following this warm-up, we enable sparse attention and proceed with ODE initialization followed by DMD refinement, as in prior video distillation work (Yin et al., 2024b; 2025; Zhu et al., 2026), retaining reconstruction regularization during ODE initialization to limit distribution drift.

We apply this recipe to Wan2.1-1.3B at 480p and Wan2.1-14B at 720p (Wan et al., 2025). At 90% sparsity, VSQA accelerates attention on Wan2.1-14B by $17.2\times$ over dense FlashAttention-4 (Zadouri et al., 2026) on the GB300, and $28.5\times$ over FlashAttention-2 on the RTX5090, much faster than other implementations such as SageSLA (Zhang et al., 2025c). With NVFP4 linear layers and three-step generation, JetDiffusion achieves $282\times$ end-to-end speedup on Wan2.1-14B on the RTX 5090 and $144\times$ on the GB300, with VBench scores that match or exceed those of the dense BF16 base model.

Our main contributions are as follows:

- **Trainable NVFP4 sparse attention.** VSQA is, to our knowledge, the first trainable attention that combines high sparsity with NVFP4 quantization, with training kernels and inference kernels for the RTX 5090 and the GB300.
- **A stable training recipe for NVFP4 sparse few-step generation.** Guided by empirical findings on how quantization and sparsity are recovered, we separate quantization recovery from sparse few-step adaptation, using layer reconstruction and regularized ODE initialization before DMD refinement.

- **JetDiffusion: ultra-fast video generation.** On Wan2.1-14B at 720p, VSQA delivers a $17.2\times$ attention speedup over dense FlashAttention-4 on the GB300, and JetDiffusion delivers $282\times$ end-to-end speedup on the RTX 5090 and $144\times$ on the GB300 while preserving generation quality.

2 Related Work

Sparse and low-bit attention. Sparse attention reduces token interactions through structured patterns or adaptive token and block selection (Zhang et al., 2025d; 2026f; Xi et al.; Yang et al., 2026; Xia et al., 2025; Shen et al., 2025; Wu et al., 2026; Zhang et al., 2026b; Li et al., 2026b; Chen et al., 2026a; Sun et al., 2026; Tan et al., 2026; Zhan et al., 2025). Other methods compensate for omitted interactions through auxiliary approximations or residual reconstruction (Zhang et al., 2026c; Zhou et al., 2026; Taghavi et al., 2026). Low-bit attention accelerates computation by reducing numerical precision (Zhang et al., 2025b; Chen et al., 2024; Zhang et al., 2025a), using techniques such as smoothing to mitigate quantization error (Zhang et al., 2024; 2026e; Li et al., 2026c). Attn-QAT adapts models to 4-bit attention through quantization-aware training (Zhang et al., 2026g), while SageBwd studies trainable 8-bit attention (Zhang et al., 2026a). Combining sparsity and quantization, SpargeAttn and PAROAttention provide training-free approaches (Zhang et al.; Zhao et al., 2026b), whereas FPSAttention and SLA2 use training to preserve quality (Liu et al., 2025a; Zhang et al., 2026d). To the best of our knowledge, VSQA is the first trainable NVFP4 sparse attention method for video generation.

Few-step distillation. Few-step generation can be learned through consistency objectives (Song et al., 2023; Lu & Song, 2025; Zheng et al., 2026) or distribution matching methods such as Diff-Instruct and Distribution Matching Distillation (DMD), which align student and teacher output distributions (Luo et al., 2023; Yin et al., 2024b;a). For DMD in autoregressive video generation, CausVid initializes the student with teacher-generated ODE trajectories (Yin et al., 2025), and Causal Forcing examines their compatibility with the student's causal conditioning (Zhu et al., 2026). Related work reduces train–test conditioning gaps using student-generated histories (Huang et al., 2026a) and studies consistency-based initialization and its alignment with DMD refinement (Zhao et al., 2026a; Li et al., 2026a). Inspired by these works, we use ODE initialization to adapt the model to both few-step denoising and sparse attention before DMD refinement.

Combining compression and distillation. TurboDiffusion and LightWan2.2 combine step distillation, sparse attention, and quantization under FP8 (Zhang et al., 2025c; Lv et al., 2026). FastWan jointly trains sparse attention and few-step generation with DMD (FastVideo Team, 2025). FastWan-QAD applies quantization-aware fine-tuning and DMD with dense low-bit attention (FastVideo Team, 2026), while SQuad adapts sub-quadratic attention through flow-matching fine-tuning before DMD2 (Karnewar et al., 2026). LongLive-2.0 incorporates NVFP4 into autoregressive training and DMD (Chen et al., 2026b). VSQA targets ultrafast attention and end-to-end video generation by combining NVFP4 sparse attention, linear quantization, and few-step distillation.

3 VSQA: Video Sparse Quantized Attention with NVFP4

VSQA builds on the coarse-to-fine attention structure of VSA (Zhang et al., 2026f). We retain its cube-level routing and gated coarse branch, and apply NVFP4 quantization (Alvarez et al., 2025) to the token-level sparse attention branch.

Fixed top-k routing. Given query, key, and value matrices $Q, K, V \in \mathbb{R}^{N\times d}$ for N video tokens with head dimension d, we group spatially and temporally adjacent tokens into B cubes. Mean pooling within each cube gives the pooled queries, keys, and values $\bar{Q}, \bar{K}, \bar{V} \in \mathbb{R}^{B\times d}$. Attention over these pooled representations gives cube-level scores and outputs:

$$A = \frac{\bar{Q}\bar{K}^\top}{\sqrt{d}}, \qquad T = \mathrm{softmax}(A), \qquad O_c = T\bar{V}. \tag{1}$$

For each query cube, we select the k key cubes with the largest scores in the corresponding row of A. At target sparsity s, we use $k = \max(1, \lceil (1-s)B \rceil)$. All query tokens in a cube attend to the same selected key/value cubes, forming a block-sparse pattern without computing full token-level attention scores. The coarse output O_c also retains information from all cubes for the final output.

Native NVFP4 sparse attention. We use NVFP4-quantized operands for both $QK^\top$ and PV in the fine branch, while keeping softmax in high precision, following prior FP4 attention methods (Zhang et al., 2026e;g). We denote the NVFP4-quantized version of Q by $\widehat{Q}$, with block scales included, and use the same notation for K, V, and P. We encode these selections in a token-level mask M: $M_{ij} = 0$ when key token j belongs to a cube selected for query token i, and $M_{ij} = -\infty$ otherwise. The fine-branch computation is

$$S = \frac{\widehat{Q}\widehat{K}^\top}{\sqrt{d}} + M, \qquad P = \mathrm{softmax}(S), \qquad O_f = \widehat{P}\widehat{V}. \tag{2}$$

Here S contains the attention scores after masking, P contains the softmax probabilities over the selected tokens, and $\widehat{P}$ is their NVFP4 approximation used in the value product. We copy each cube's coarse output to its constituent tokens, and combine it with the fine output:

$$O = O_f + g \odot O_c, \tag{3}$$

where $g \in \mathbb{R}^{N\times d}$ is a learned elementwise gate. Each token therefore receives a quantized sparse-attention output and a gated contribution from the global cube-level attention.

Backward computation. We keep the top-k sparse selections invariant with forward pass during backward propagation, and train with a straight-through estimator (STE) (Bengio et al., 2013). Quantizing attention probabilities breaks the identity that allows FlashAttention (Dao et al., 2022) to compute the softmax gradient correction from the forward output. Following Attn-QAT (Zhang et al., 2026g), we apply fake quantization, recompute probabilities from the same quantized queries and keys used in forward, and save an auxiliary output $O_f^{\mathrm{aux}} = P\widehat{V}$ computed with unrounded probabilities for this correction. Appendix C compares videos generated with fake and real quantization, which show no visible difference.

Details of kernel implementation. Our inference kernels traverse only the selected cube pairs and use native low-precision matrix multiplication for the fine branch. On Blackwell, reducing non-matrix operations is important because their throughput scales more slowly than matrix multiplication (Zadouri et al., 2026). Pooling and input quantization still process all tokens, so their overhead becomes more prominent as sparsity increases. We fuse Q/K/V pooling and quantization around a shared input load: the loaded features supply both the unquantized routing inputs and the packed NVFP4 operands, avoiding separate reads for the two operations.

On the GB300, the kernel computes $QK^\top$ in NVFP4 and PV in FP8. As Attn-QAT reports (Zhang et al., 2026g), packing FP4 probabilities and transferring their scales add overhead, so an NVFP4 value product is not the faster choice on this GPU. Appendix C compares videos generated with the FP8 and NVFP4 value products and finds no visible difference, as for fake quantization above. For an NVFP4 value product, the GB300 kernel also reduces the overhead of quantizing the attention probabilities. A direct implementation exponentiates the scores, finds a scale for each group of probabilities, and rescales the probabilities before packing them. Because exponentiation preserves ordering, the kernel instead finds each group's maximum in the score domain and folds the rescaling into the exponentiation, and it accumulates the softmax denominator from the same unrounded values. This removes a separate reciprocal-and-rescaling pass.

4 Taming Compression for Stable Distillation

To push video generation efficiency to the extreme, building upon VSQA, we further extend NVFP4 quantization to the linear layers and reduce the number of sampling steps through step distillation. However, each form of compression perturbs the original model, and jointly adapting to these perturbations can make training less stable. Prior few-step distillation works have shown (Yin et al., 2025; Zhu et al., 2026; Zhao et al., 2026a; Li et al., 2026a; Karnewar et al., 2026; Cheng et al., 2025) that DMD typically relies on an initialization with basic few-step generation capability. Our empirical study indicates that a well-prepared initialization supports more stable DMD training in this compressed setting. In this compressed setting, a suitable initialization should absorb quantization error, establish basic few-step generation capability, and remain close to the original teacher distribution.

Recipe	Qual.	Sem.	Total
Base model	83.52	64.50	79.72
Direct DMD, FP8	85.25	70.68	82.33
Direct DMD, NVFP4	84.33	71.05	81.67
Ours, NVFP4	**85.28**	**72.47**	**82.72**

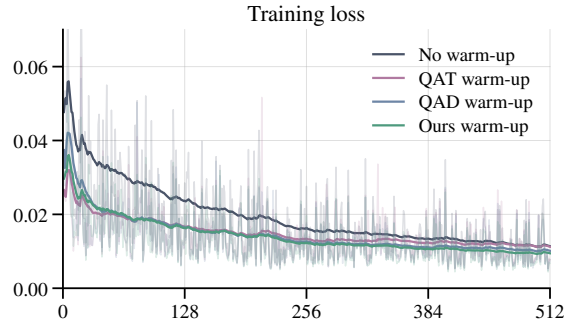


Figure 2: **The NVFP4 gap and its warmup.** Left: VBench scores on Wan2.1-1.3B under different training recipes. Right: subsequent training loss at 90% sparsity with different warmup objectives or no warmup.

4.1 Overcoming the NVFP4 Quantization Gap

Aggressive 4-bit quantization first introduces a pronounced low-bit initialization gap. Compared with milder 8-bit formats such as FP8, prior work has shown (Huang et al., 2026b; Li et al., 2026d; 2024; Wang et al., 2025; Liu et al., 2025b) that full 4-bit weights and activations are more sensitive to limited activation ranges, outliers, and timestep-dependent activation variations in diffusion models, resulting in larger quantization errors. Figure 2 (left) illustrates the challenge: direct DMD with NVFP4 attains lower Quality and Total scores than the FP8 run, motivating additional adaptation before distillation. To achieve extreme inference acceleration, we don't want to apply any additional compensation or techniques to the quantization process, which results in a larger initial perturbation and makes DMD training less stable. We therefore need to adapt to the dominant low-bit quantization errors before entering few-step distillation.

> Takeaway 1. NVFP4 introduces a large quantization perturbation that calls for a dedicated warmup before few-step distillation.

4.2 Why QAT and QAD Are Not the Best Choice

A natural choice for quantization error recovery is quantization-aware training (QAT) (FastVideo Team, 2026; Sui et al., 2024) or quantization-aware distillation (QAD) (Polino et al., 2018; Mishra & Marr, 2018; He et al., 2024), which can reduce quantization errors and recover standalone generation quality through finetuning or velocity distillation. For QAT, we consider fine-tuning the quantized student v_θ on real videos with the flow-matching objective of the base model (Wan et al., 2025):

$$\mathcal{L}_{\mathrm{FT}}(\theta) = \mathbb{E}_{x_0,\epsilon,t,c}\left[\|v_\theta(x_t,t,c) - (\epsilon - x_0)\|_2^2\right], \qquad x_t = (1-t)x_0 + t\epsilon, \tag{4}$$

where x_0 is a clean video latent, $\epsilon \sim \mathcal{N}(0, I)$, $t \in [0,1]$ is the noise level, and c is the condition. QAD instead matches the velocity prediction of the frozen unquantized teacher v_{T} on the same input:

$$\mathcal{L}_{\mathrm{VD}}(\theta) = \mathbb{E}_{x_t,t,c}\left[\|v_\theta(x_t,t,c) - v_{\mathrm{T}}(x_t,t,c)\|_2^2\right]. \tag{5}$$

Both objectives supervise only the final velocity prediction.

However, we find that the final-output objectives used by QAT or QAD are not necessarily well suited for DMD initialization. As shown in Figure 3, QAT and QAD recover the standalone generation quality of the quantized model, but the generated videos gradually deviate from the teacher in global layout and fine-grained content during warmup. Such deviations introduce additional teacher-student mismatch that may complicate subsequent DMD optimization.

> Takeaway 2. Recovering standalone quality with QAT or QAD does not ensure a suitable DMD initialization, because teacher alignment can deteriorate.

Inspired by reconstruction-based quantization methods (Nagel et al., 2020; Li et al., 2021; Hubara et al., 2021), which constrain quantization-induced deviations by reconstructing intermediate outputs at the layer or block level, we instead adopt fine-grained reconstruction supervision during

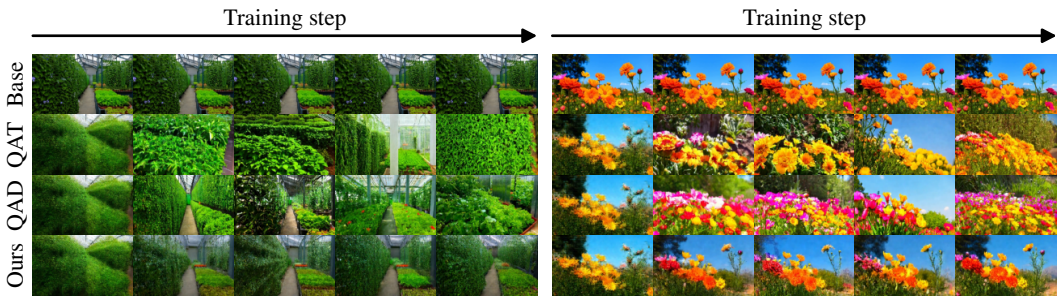


Figure 3: **Generations during quantization warmup.** Matched frames for two prompts, comparing the BF16 base model with QAT, QAD, and layer reconstruction. Training progresses from left to right.

quantization warmup. As illustrated in Figure 3, layer reconstruction substantially better preserves the layout and content of the original teacher generation throughout training.

This stronger teacher alignment comes at the cost of somewhat lower standalone generation quality than QAD. Since DMD serves as a refinement of an initialized few-step generator (Zhao et al., 2026a), we prioritize preserving teacher alignment while reducing disruptive quantization errors to provide a more stable starting point for subsequent quality refinement.

Following this principle, we introduce deeply supervised layer-wise reconstruction into our quantization warmup to reduce the additional deviation introduced during low-bit adaptation. With attention kept dense, the quantized and unquantized versions of each supervised layer receive the same teacher input $h_{\mathrm{T}}^{(\ell)}$, and the warmup minimizes the reconstruction error over the supervised layers $\mathcal{R}$:

$$\mathcal{L}_{\mathrm{rec}}(\theta) = \mathbb{E}_{x_t,t,c}\left[\sum_{\ell\in\mathcal{R}}\left\| f_\theta^{(\ell)}\big(\mathrm{sg}(h_{\mathrm{T}}^{(\ell)});t,c\big) - \mathrm{sg}\big(f_{\mathrm{T}}^{(\ell)}(h_{\mathrm{T}}^{(\ell)};t,c)\big)\right\|_F^2\right], \tag{6}$$

where $f_\theta^{(\ell)}$ and $f_{\mathrm{T}}^{(\ell)}$ denote the student and teacher layers, and sg stops gradients through the frozen teacher. Shared reference inputs isolate local quantization error from deviations accumulated in preceding layers.

We analyze teacher alignment during warmup using V-JEPA 2.1 (Mur-Labadia et al., 2026) features. Figure 4 shows that layer reconstruction settles into a stable offset from the teacher distribution while retaining stronger sample-wise correspondence. We interpret this residual shift as reflecting the standalone quality trade-off observed above, leaving room for subsequent DMD quality refinement. In contrast, QAD with velocity supervision exhibits fluctuating alignment across checkpoints, which may make subsequent DMD refinement less stable. This distinction favors a stable, teacher-aligned initialization with a remaining quality gap over a fluctuating warmup trajectory.

We further evaluate whether these warmup choices improve subsequent training. Figure 5 shows that layer reconstruction warmup yields the lowest RMSE throughout subsequent sparse adaptation and improves cosine similarity over training without warmup. QAT and QAD also improve cosine similarity and relative L1 error, but retain higher RMSE, indicating that these gains do not necessarily reduce absolute output deviations. Figure 2 (right) further shows lower training loss with reconstruction warmup than without warmup, particularly early in adaptation. Together, these results support reducing quantization error before introducing sparsity to ease subsequent optimization.

TAKEAWAY 3. Layer reconstruction preserves teacher alignment while reducing quantization error, supporting more stable subsequent distillation.

Meanwhile, following the ODE initialization scheme of (Yin et al., 2025), we construct trajectory pairs corresponding to the target few-step schedule from the teacher PF-ODE trajectory. The frozen teacher integrates

$$\frac{dx_t^{\mathrm{T}}}{dt} = v_{\mathrm{T}}(x_t^{\mathrm{T}},t,c), \qquad x_1^{\mathrm{T}} \sim \mathcal{N}(0,I), \tag{7}$$

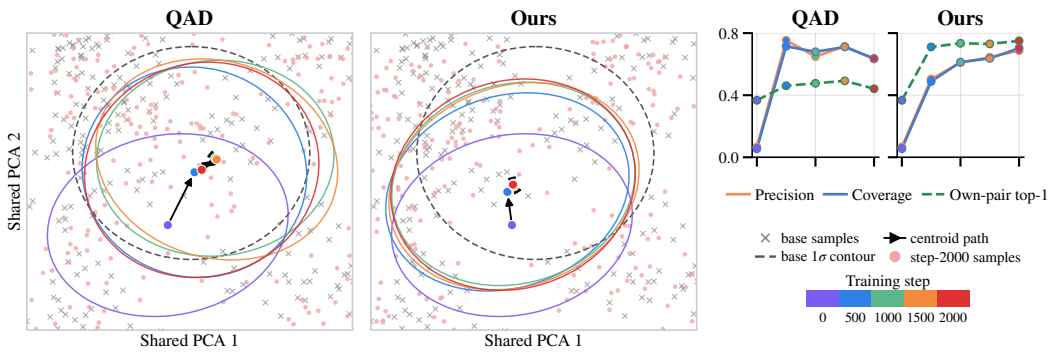


Figure 4: **V-JEPA feature alignment during quantization warmup.** Left: PCA projections relative to the base model. Right: set-level alignment (precision and coverage) and sample-wise correspondence (own-pair top-1).

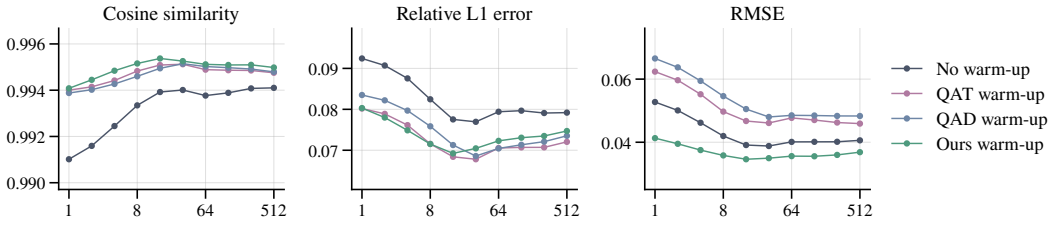


Figure 5: **Quantization error during subsequent sparse adaptation.** Cosine similarity, relative L1 error, and RMSE with different warmup objectives or no warmup, evaluated at the training sparsity of 90%.

from Gaussian noise at $t = 1$ to a clean latent x_0^{T} at $t = 0$. At the three student timesteps $t_1 > t_2 > t_3 > 0$, we retain the teacher states and their shared clean endpoint:

$$\mathcal{D}_{\mathrm{ODE}} = \left\{ (c, x_{t_k}^{\mathrm{T}}, t_k, x_0^{\mathrm{T}}) \right\}_{k=1}^{3}, \tag{8}$$

collected over teacher trajectories for different conditions and initial noise samples. Specifically, we extract the corresponding target states from the teacher PF-ODE trajectory and use these states for quantization warmup, instead of randomly sampling timesteps from the entire diffusion trajectory. This design better aligns quantization warmup with subsequent few-step training. In particular, quantization error recovery is concentrated on the trajectory states that the final generator actually visits during sampling, rather than being distributed across a large number of timesteps that are never directly encountered by the deployed few-step generator.

4.3 Anchoring Sparse Initialization to the Dense Teacher

After quantization warmup, we continue along the same teacher PF-ODE trajectory and perform standard few-step ODE initialization. Specifically, we train the student at the target sampling timesteps while enabling the sparse attention used for final deployment, allowing the model to acquire basic few-step generation capability while simultaneously adapting to sparse computation. Let $F_\theta(x_t, t, c)$ denote the student's clean-latent prediction. The teacher endpoints provide the supervision

$$\mathcal{L}_{\mathrm{ODE}}(\theta) = \mathbb{E}_{(c, x_t^{\mathrm{T}}, t, x_0^{\mathrm{T}}) \sim \mathcal{D}_{\mathrm{ODE}}} \left[\frac{1}{2} \left\| F_\theta(x_t^{\mathrm{T}}, t, c) - x_0^{\mathrm{T}} \right\|_2^2 \right]. \tag{9}$$

However, we further observe that directly applying standard ODE initialization to the sparse student introduces an additional sparsity-induced structural gap. Although the ODE objective enables the student to learn the target few-step mapping, sparse attention can cause its intermediate representations to deviate further from those of the dense teacher, resulting in additional teacher-student mismatch during the early stage of subsequent DMD training.

Table 1: **VBench results** (Huang et al., 2024) over 946 prompts and 16 dimensions at matched resolutions. Results use five seeds for 1.3B and one for 14B; all students use three-step VSQA with NVFP4 attention and linear layers. Bold/underline indicate the best/second-best students per model size. Per-dimension scores appear in Appendix B.2.

Method	Quality ↑	Semantic ↑	Total ↑
Wan2.1-1.3B base model, 480p	83.52	64.50	79.72
Direct DMD	84.33	71.05	81.67
Ours (JetDiffusion)	**85.28**	<u>72.47</u>	**82.72**
warmup replaced by QAD	<u>85.11</u>	70.54	82.19
w/o ODE initialization	84.73	**73.54**	<u>82.49</u>
w/o ODE initialization, QAD warmup	84.51	69.71	81.55
w/o reconstruction regularization	84.83	70.33	81.93
Wan2.1-14B base model, 720p	83.91	75.41	82.21
Ours (JetDiffusion)	**85.16**	**79.48**	**84.02**

TAKEAWAY 4. Sparse adaptation can leave a structural gap that endpoint supervision alone does not close, motivating layer reconstruction regularization.

To mitigate this gap, we extend the fine-grained teacher-alignment constraint introduced during quantization warmup and add a lightweight reconstruction regularizer on top of the standard ODE initialization objective. Specifically, we constrain the sparse attention outputs to match their corresponding dense-teacher outputs, allowing the student to acquire sparse few-step generation capability while maintaining better teacher alignment. We optimize

$$\mathcal{L}_{\text{init}}(\theta) = \frac{1}{1+\lambda}\mathcal{L}_{\text{ODE}}(\theta) + \frac{\lambda}{1+\lambda}\mathcal{L}_{\text{rec}}(\theta), \qquad \lambda \geq 0, \tag{10}$$

where $\mathcal{L}_{\text{rec}}$ now supervises the quantized sparse attention layers on the shared reference inputs defined in Equation 6, with $\mathcal{R}$ restricted to these layers. The weight λ controls the contribution of reconstruction relative to endpoint supervision. This regularization provides a better-aligned starting point for subsequent DMD training.

4.4 Distribution Matching Done Right

We use DMD as a refinement (Zhao et al., 2026a) of our initialized few-step student. Let $p_\theta(\cdot \mid c)$ and $p_{\text{T}}(\cdot \mid c)$ denote the student's and teacher's conditional generation distributions. DMD (Yin et al., 2024b;a) minimizes the reverse-KL objective

$$\mathcal{L}_{\text{DMD}}(\theta) = \mathbb{E}_c \left[D_{\text{KL}}(p_\theta(\cdot \mid c) \,\|\, p_{\text{T}}(\cdot \mid c)) \right], \tag{11}$$

with gradients approximated using teacher and student-distribution score estimates on noised samples. Our initialization supports more stable DMD refinement to improve generation quality, with quantization and sparsity enabled throughout training. Figure 2 (left) shows that the complete recipe achieves higher Quality, Semantic, and Total VBench scores than both the base model and direct NVFP4 DMD, demonstrating strong final generation quality despite aggressive compression. Table 1 further shows that, reconstruction warmup achieves higher scores on all three VBench metrics after DMD than QAD warmup, supporting its use as an initialization for subsequent refinement.

5 Experiments

5.1 Setup

Training. We train Wan2.1-1.3B at 480p ($61 \times 448 \times 896$) and Wan2.1-14B at 720p ($77 \times 768 \times 1280$) (Wan et al., 2025), with dimensions given as frames $\times$ height $\times$ width. All students use three denoising steps, 90% sparsity, and NVFP4 attention and linear layers. For native NVFP4 computation, we use $4 \times 4 \times 8$ and $4 \times 8 \times 8$ cubes at 480p and 720p, respectively. Training uses only text prompts from Wan-Syn (Zhang et al., 2026f) and VidProM (Wang & Yang, 2024).

Efficiency. We measure attention and total DiT latency on RTX 5090 at both resolutions and GB300 at 720p, using FP8 PV on GB300 for higher throughput. Dense baselines are FlashAttention-2 (Dao, 2024) on RTX 5090 and FlashAttention-4 (Zadouri et al., 2026) on GB300. RTX 5090 comparisons also include SageAttention2++ (Zhang et al., 2025a), SageAttention3 (Zhang et al., 2026e), and SageSLA (Zhang et al., 2025c). All sparse methods use 90% sparsity.

5.2 Generation Quality

Table 1 lists the VBench scores. JetDiffusion runs three denoising steps with 90% attention sparsity and NVFP4 attention and linear layers, and on both model sizes its scores stay at the level of the base model. Among the Wan2.1-1.3B students, the full recipe scores highest, and every part of it contributes. Removing ODE initialization or reconstruction regularization lowers the total score, and so does swapping our reconstruction warmup for a QAD warmup with the rest of the recipe unchanged, which falls behind mostly on the semantic score.

5.3 Inference Efficiency

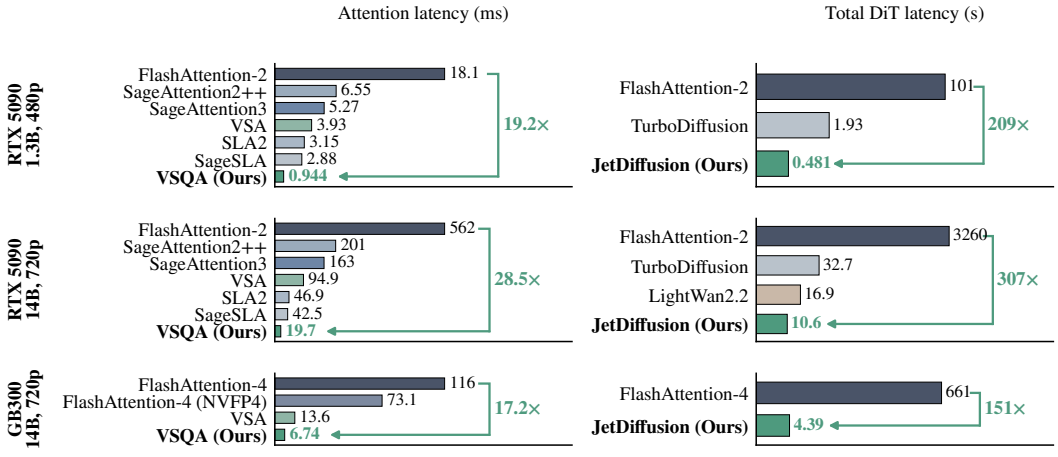


Figure 6: **Attention and total DiT latency on Blackwell GPUs.** Left: self-attention. Right: total DiT time per video at each method's default sampling steps (base: 50 steps with CFG). Annotations show speedups over the dense baseline on the same GPU.

Figure 6 shows that VSQA achieves 17.2–$28.5\times$ attention speedups over the respective dense FlashAttention baselines across the evaluated Blackwell configurations. On the RTX 5090, it also outperforms SageSLA by 2.15–$3.06\times$ at the same 90% sparsity.

Combined with linear quantization and few-step distillation, JetDiffusion reduces total DiT time per generation by 151–$307\times$ relative to the base model. On the RTX 5090, it is 3.08–$4.01\times$ faster than TurboDiffusion and $1.59\times$ faster than LightWan2.2 at 720p. Detailed timings are reported in Appendix B.3.

6 Conclusion

We presented VSQA, a trainable sparse attention method that combines high sparsity with native NVFP4 computation for video diffusion. VSQA achieves a $17.2\times$ attention speedup over dense FlashAttention-4 at 90% sparsity on NVIDIA GB300 while preserving generation quality. Building on VSQA, we further presented **JetDiffusion**, which combines sparse NVFP4 attention with NVFP4 linear layers and few-step distillation to accelerate the full video generation pipeline. Together, these techniques enable $282\times$ end-to-end acceleration on Wan2.1-14B on the RTX 5090, demonstrating the potential of jointly exploiting sparsity, low precision, and generation-step reduction for efficient video generation.

AI USE STATEMENT

We used generative AI tools to assist with polishing the manuscript. The authors take responsibility for the final content of this work, including its text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

This work improves video generation efficiency by combining sparse attention, low-bit quantization, and few-step distillation. Reducing inference cost may make video generation more accessible to users with limited computational resources. However, faster generation may also facilitate the creation of misleading or harmful synthetic content, and compression may alter the behavior of the base model, including its safety mechanisms. Deployment should therefore include safety evaluations of the compressed models, appropriate content safeguards, and clear disclosure of AI-generated videos.

REPRODUCIBILITY STATEMENT

We describe the sparse attention formulation and inference kernels in Section 3, and the training objectives and initialization procedure in Section 4. Section 5 specifies the base models, prompt sources, generation settings, hardware, and comparison methods used in our experiments. Appendix A provides additional implementation details, while Appendix B reports per-dimension quality scores and detailed inference timings.

REFERENCES

Eduardo Alvarez, Omri Almog, Eric Chung, Simon Layton, Dusan Stosic, Ronny Krashinsky, and Kyle Aubrey. Introducing NVFP4 for efficient and accurate low-precision inference. NVIDIA Technical Blog, June 2025. URL https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/. Online; accessed 2026-09-25.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Ollin Boer Bohan. Taehv: Tiny autoencoder for hunyuan video. https://github.com/madebyollin/taehv, 2025.

Pengtao Chen, Xianfang Zeng, Maosen Zhao, Mingzhu Shen, Wei Cheng, Gang Yu, and Tao Chen. Sparse-vdit: Unleashing the power of sparse attention to accelerate video diffusion transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2957–2965, 2026a.

Shimao Chen, Zirui Liu, Zhiying Wu, Ce Zheng, Peizhuang Cong, Zihan Jiang, Yuhan Wu, Lei Su, and Tong Yang. Int-flashattention: Enabling flash attention for int8 quantization. *arXiv preprint arXiv:2409.16997*, 2024.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, et al. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026b.

Hanbo Cheng, Peng Wang, Kaixiang Lei, Qi Li, Zhen Zou, Pengfei Hu, and Jun Du. From structure to detail: Hierarchical distillation for efficient diffusion model. *arXiv preprint arXiv:2511.08930*, 2025.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

FastVideo Team. FastWan: Generating a 5-Second Video in 5 Seconds via Sparse Distillation. FastVideo blog, August 2025. URL https://haoailab.com/blogs/fastvideo_post_training/. Technical blog; accessed 2026-09-25.

FastVideo Team. FastWan-QAD: FastVideo generates a 5-Second Video in 1.8 Seconds on a Single NVIDIA GeForce RTX 5090 via Quantization-Aware Distillation. FastVideo blog, June 2026. URL https://haoailab.com/blogs/fastwan-qad/. Technical blog; accessed 2026-09-25.

Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 15731–15750, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026a.

Yushi Huang, Ruihao Gong, Jing Liu, Yifu Ding, Chengtao Lv, Haotong Qin, and Jun Zhang. Qvgen: Pushing the limit of quantized video generative models. In *International Conference on Learning Representations*, volume 2026, pp. 155088–155121, 2026b.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Accurate post training quantization with small calibration sets. In *International conference on machine learning*, pp. 4466–4475. PMLR, 2021.

Animesh Karnewar, Denis Korzhenkov, Amirhossein Habibian, and Mohsen Ghafoorian. Squad: Sub-quadratic attention distillation for efficient video generation. *arXiv preprint arXiv:2608.16585*, 2026.

Jiaxing Li, Kai Zou, Cindy Zhou, Kaichen Huang, Junyao Gao, Zile Wang, Yang Liu, Bin Liu, Bo An, and Yangguang Li. Distillalign: Coordinating mode covering and mode seeking in autoregressive video distillation. *arXiv preprint arXiv:2607.26811*, 2026a.

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. *arXiv preprint arXiv:2411.05007*, 2024.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, et al. Radial attention: $\mathcal{O}(n \log n)$ sparse attention with energy decay for long video generation. *Advances in Neural Information Processing Systems*, 38:16822–16852, 2026b.

Xingyang Li, Dongyun Zou, Shining Zhang, Jiacheng Chen, Haocheng Xi, Lvmin Zhang, Jun-Yan Zhu, Song Han, Zhekai Zhang, Yujun Lin, et al. Vc-attention: Value smoothing and softmax casting for low-bit attention. *arXiv preprint arXiv:2609.15810*, 2026c.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

Zhiteng Li, Hanxuan Li, Junyi Wu, Kai Liu, Haotong Qin, Linghe Kong, Guihai Chen, Yulun Zhang, and Xiaokang Yang. Dvd-quant: data-free video diffusion transformers quantization. In *International Conference on Learning Representations*, volume 2026, pp. 35463–35475, 2026d.

Akide Liu, Zeyu Zhang, Zhexin Li, Xuehai Bai, Yizeng Han, Jiasheng Tang, Yuanjie Xing, Jichao Wu, Mingyang Yang, Weihua Chen, et al. Fpsattention: Training-aware fp8 and sparsity co-design for fast video diffusion. *arXiv preprint arXiv:2506.04648*, 2025a.

Xuewen Liu, Zhikai Li, Minghao Jiang, Mengjuan Chen, Jianquan Li, and Qingyi Gu. Dilatequant: Accurate and efficient quantization-aware training for diffusion models via weight dilation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8399–8408, 2025b.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023.

Chengtao Lv, Zihao Peng, and LightX2V Team. LightWan2.2-A14B: Extremely Fast Wan 2.2 14B Inference. LightX2V Blog, July 2026. URL `https://light-ai.top/LightX2V-BLOG/posts/LightWan22-A14B/`. Technical blog; accessed 2026-09-25.

Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *International Conference on Learning Representations*, 2018.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-jepa 2.1: Unlocking dense features in video self-supervised learning. In *European Conference on Computer Vision*, pp. 671–689. Springer, 2026.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, pp. 7197–7206. PMLR, 2020.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018.

Xuan Shen, Chenxia Han, Yufa Zhou, Yanyue Xie, Yifan Gong, Quanyi Wang, Yiwei Wang, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. Draftattention: Fast video diffusion via low-resolution attention guidance. *arXiv preprint arXiv:2505.14708*, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Yang Sui, Yanyu Li, Anil Kag, Yerlan Idelbayev, Junli Cao, Ju Hu, Dhritiman Sagar, Bo Yuan, Sergey Tulyakov, and Jian Ren. Bitsfusion: 1.99 bits weight quantization of diffusion model. *Advances in Neural Information Processing Systems*, 37:76775–76818, 2024.

Wenhao Sun, Rong-Cheng Tu, Yifu Ding, Jingyi Liao, Zhao Jin, Shunyu Liu, and Dacheng Tao. Vorta: Efficient video diffusion via routing sparse attention. *Advances in Neural Information Processing Systems*, 38:7837–7863, 2026.

Pardis Taghavi, Reza Langari, and Gaurav Pandey. Partition the support, reconstruct the residual: Training-free sparse attention for video generation and world models. *arXiv preprint arXiv:2608.18484*, 2026.

Xin Tan, Yuetao Chen, Yimin Jiang, Xing Chen, Kun Yan, Nan Duan, Yibo Zhu, Daxin Jiang, and Hong Xu. Dynamic sparsity in large-scale video dit training. In *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pp. 101–116, 2026.

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.

Haoxuan Wang, Yuzhang Shang, Zhihang Yuan, Junyi Wu, Junchi Yan, and Yan Yan. Quest: Low-bit diffusion model quantization via efficient selective finetuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15542–15551. IEEE, 2025.

Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

Jianzong Wu, Liang Hou, Haotian Yang, Ye Tian, Pengfei Wan, Di ZHANG, and Yunhai Tong. Vmoba: Mixture-of-block attention for video diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 132534–132548, 2026.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity, 2025. *URL https://arxiv. org/abs/2502.01776*, 9:10–29.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15982–15993. IEEE, 2025.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *Advances in Neural Information Processing Systems*, 38:96965–96991, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Ted Zadouri, Markus Hoehnerbach, Jay Shah, Vijay Thakkar, and Tri Dao. Flashattention-4: Algorithm and kernel pipelining co-design for asymmetric hardware scaling. *Proceedings of Machine Learning and Systems*, 8:912–926, 2026.

Chenlu Zhan, Wen Li, Chuyu Shen, Jun Zhang, Suhui Wu, and Hao Zhang. Bidirectional sparse attention for faster video diffusion training. *arXiv preprint arXiv:2509.01085*, 2025.

J Zhang, C Xiang, H Huang, J Wei, H Xi, J Zhu, and J Chen. Spargeattn: accurate sparse attention accelerating any model inference (2025). *URL https://arxiv. org/abs/2502.18137*.

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. *arXiv preprint arXiv:2411.10958*, 2024.

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*, 2025a.

Jintao Zhang, Pengle Zhang, Jun Zhu, Jianfei Chen, et al. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In *International Conference on Learning Representations*, volume 2025, pp. 71566–71585, 2025b.

Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Turbodiffusion: Accelerating video diffusion models by 100-200 times. *arXiv preprint arXiv:2512.16093*, 2025c.

Jintao Zhang, Marco Chen, Haoxu Wang, Kai Jiang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Sagebwd: A trainable low-bit attention. *arXiv preprint arXiv:2603.02170*, 2026a.

Jintao Zhang, Kai Jiang, Chendong Xiang, Weiqi Feng, Yuezhou Hu, Haocheng Xi, Jianfei Chen, and Jun Zhu. Spargeattention2: Trainable sparse attention via hybrid top-k+ top-p masking and distillation fine-tuning. *arXiv preprint arXiv:2602.13515*, 2026b.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, et al. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse–linear attention. In *International Conference on Learning Representations*, volume 2026, pp. 138288–138305, 2026c.

Jintao Zhang, Haoxu Wang, Kai Jiang, Kaiwen Zheng, Youhe Jiang, Ion Stoica, Jianfei Chen, Jun Zhu, and Joseph E Gonzalez. Sla2: Sparse-linear attention with learnable routing and qat. *arXiv preprint arXiv:2602.12675*, 2026d.

Jintao Zhang, Haoxu Wang, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Kai Jiang, Jianfei Chen, Jun Zhu, et al. Sageattention3: Microscaling fp4 attention for inference and an exploration of 8-bit training. *Advances in Neural Information Processing Systems*, 38:53901–53931, 2026e.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025d.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Faster video diffusion with trainable sparse attention. *Advances in Neural Information Processing Systems*, 38:152509–152534, 2026f.

Peiyuan Zhang, Matthew Noto, Wenxuan Tan, Chengquan Jiang, Will Lin, Wei Zhou, and Hao Zhang. Attn-qat: 4-bit attention with quantization-aware training. *arXiv preprint arXiv:2603.00040*, 2026g.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026a.

Tianchen Zhao, Ke Hong, Xinhao Yang, Xuefeng Xiao, Huixia Li, Feng Ling, Ruiqi Xie, Siqi Chen, Hongyu Zhu, Zhang Yichong, et al. Paroattention: Pattern-aware reordering for efficient sparse and quantized attention in visual generation models. *Advances in Neural Information Processing Systems*, 38:126484–126511, 2026b.

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. In *International Conference on Learning Representations*, volume 2026, pp. 2582–2603, 2026.

Xuanyi Zhou, Qiuyang Mang, Shuo Yang, Haocheng Xi, Jintao Zhang, Huanzhi Mao, Joseph E Gonzalez, Kurt Keutzer, Ion Stoica, and Alvin Cheung. Svg-ear: Parameter-free linear compensation for sparse video generation via error-aware routing. *arXiv preprint arXiv:2603.08982*, 2026.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A Implementation Details

NVFP4 represents values in E2M1 with E4M3 scales shared by groups of 16 values (Alvarez et al., 2025). For a group x_b of 16 values, our unit-global-scale notation is

$$a_b = \mathrm{round}_{\mathrm{E4M3}}\left(\frac{\max|x_b|}{6}\right), \qquad q_b = \mathrm{round}_{\mathrm{E2M1}}(x_b/a_b), \qquad \mathcal{Q}_4(x_b) = a_b q_b, \tag{12}$$

with zero output for zero-scale groups.

B Additional Experimental Results

B.1 Training Setup

Each student is trained in three stages from the base model. The quantization warmup trains the dense NVFP4 student with the layer reconstruction loss on cached teacher states. The sparse initialization then trains the sparse student at the three student timesteps on cached teacher PF-ODE trajectories (48 teacher steps, flow shift 8), with the ODE endpoint loss and the reconstruction regularizer weighted equally ($\lambda = 1$). DMD refines the result. Every warmup stage uses an effective batch size of 16 and a learning rate of 10^{-5}, and we take its checkpoint after 500 updates. Table 2 lists the DMD hyperparameters. All DMD runs in Table 1 had converged when training ended, and only then were their checkpoints evaluated.

Table 2: **DMD hyperparameters.**

Hyperparameters	Wan2.1-1.3B	Wan2.1-14B
Real score network	Wan2.1-T2V-1.3B	Wan2.1-T2V-14B
Real score CFG weight	3.5	6.0
Critic network initialization	Wan2.1-T2V-1.3B	Wan2.1-T2V-14B
Student timesteps	1000, 941, 800	1000, 941, 800
Batch size	16	16
Learning rate (G_θ)	2e-6	2e-6
Learning rate (f_ψ)	4e-7	4e-7
Generator/critic update ratio	1:5	1:1
EMA decay	none	none
Training iterations	1000	200
Prompt set	Wan-Syn	VidProM

B.2 Per-Dimension VBench Results

Tables 3 and 4 list every VBench dimension behind Table 1.

Table 3: **Per-dimension VBench scores on Wan2.1-1.3B at 480p.** Ours is JetDiffusion. Bold and underline mark the best and second-best students.

VBench dimension	Base model	Direct DMD	**Ours**	warmup replaced by QAD	w/o ODE init.	w/o ODE init., QAD warmup	w/o recon. reg.
Subject consistency	91.93	95.81	95.74	<u>95.86</u>	95.77	**96.75**	94.74
Background consistency	94.13	93.40	93.80	<u>94.37</u>	93.85	**95.73**	93.12
Temporal flickering	99.25	98.91	<u>99.25</u>	99.10	99.23	**99.32**	**99.32**
Motion smoothness	96.83	98.46	98.21	98.32	<u>98.65</u>	**98.80**	97.75
Aesthetic quality	60.45	60.30	61.23	61.55	<u>62.09</u>	**62.15**	60.79
Imaging quality	64.94	65.98	<u>66.74</u>	65.88	65.55	65.62	**67.43**
Dynamic degree	93.89	86.11	<u>94.17</u>	91.11	84.44	72.50	**94.72**
Quality	83.52	84.33	**85.28**	<u>85.11</u>	84.73	84.51	84.83
Object class	66.65	78.23	**81.41**	76.93	<u>80.89</u>	73.13	79.59
Multiple objects	41.48	58.87	<u>63.14</u>	59.25	**67.94**	61.69	58.70
Human action	71.40	70.60	<u>74.00</u>	**75.80**	71.80	72.80	70.80
Color	72.02	<u>82.95</u>	79.58	**85.40**	81.02	79.89	80.78
Spatial relationship	65.95	75.81	<u>77.87</u>	73.64	**81.21**	74.72	73.85
Scene	29.53	<u>38.60</u>	37.21	31.86	**39.77**	34.19	33.26
Appearance style	28.73	28.37	**28.69**	28.05	<u>28.52</u>	28.09	28.18
Temporal style	22.78	23.07	**23.97**	22.83	23.88	22.64	<u>23.94</u>
Overall consistency	23.26	23.04	<u>23.52</u>	23.32	**23.71**	22.94	23.43
Semantic	64.50	71.05	<u>72.47</u>	70.54	**73.54**	69.71	70.33
Total	79.72	81.67	**82.72**	82.19	<u>82.49</u>	81.55	81.93

Table 4: **Per-dimension VBench scores on Wan2.1-14B at 720p.** Ours is JetDiffusion.

VBench dimension	Base model	**Ours**
Subject consistency	95.78	95.74
Background consistency	97.24	96.03
Temporal flickering	99.53	99.26
Motion smoothness	98.17	97.76
Aesthetic quality	64.25	66.60
Imaging quality	65.97	69.30
Dynamic degree	61.11	73.61
Quality	83.91	85.16
Object class	85.13	86.00
Multiple objects	68.98	72.41
Human action	82.00	83.00
Color	72.92	83.89
Spatial relationship	76.22	78.03
Scene	44.19	56.98
Appearance style	31.08	30.13
Temporal style	23.11	24.17
Overall consistency	24.50	25.76
Semantic	75.41	79.48
Total	82.21	84.02

B.3 Inference Latencies

Tables 5 and 6 list the timings behind Figure 6, each number followed in parentheses by its speedup over the dense base model of the same GPU and resolution, that is, Wan2.1 with FlashAttention-2 on the RTX 5090 and with FlashAttention-4 on the GB300. Table 5 gives one self-attention call,

one DiT forward pass including the linear layers, and the total DiT time of one generation, which sums the forward passes at each method's default number of sampling steps, with 50 steps and CFG for the base model. Attention is compared across attention methods, and the DiT times across the base model with these attention methods applied directly and the few-step distilled systems TurboDiffusion (Zhang et al., 2025c) and LightWan2.2 (Lv et al., 2026). Table 6 compares one whole generation between the base model in its default pipeline and JetDiffusion. The non-DiT part covers everything else in producing one video, such as text encoding and VAE decoding. To push end-to-end latency further, JetDiffusion decodes with TAEHV (Boer Bohan, 2025), a tiny video autoencoder, while the base model keeps its default VAE. The other systems are left out of this table because their VAE choices differ, which would make end-to-end times incomparable. 480p is 61 frames at 448×896 and 720p is 77 frames at 768×1280. JetDiffusion uses VSQA with 128-token cubes at 480p and 256-token cubes at 720p, NVFP4 linear layers with dynamic scaling, and three denoising steps. VSA runs with its default 64-token cubes. SLA2 has no public implementation, so its DiT times are composed from its attention latency.

Table 5: **Attention and DiT timings.** One self-attention call (ms), one DiT forward pass (s), and the total DiT time of one generation (s), each followed in parentheses by its speedup over the dense base model of the same group. Sparse methods run at 90% sparsity.

Method	Steps	Attention (ms)	DiT / step (s)	DiT total (s)
RTX 5090, 1.3B, 480p				
FlashAttention-2	50	18.1 (1.00×)	1.00 (1.00×)	101 (1.00×)
SageAttention2++	50	6.55 (2.77×)	0.640 (1.57×)	64.0 (1.57×)
SageAttention3	50	5.27 (3.44×)	0.613 (1.64×)	61.3 (1.64×)
VSA	50	3.93 (4.61×)	0.647 (1.55×)	64.7 (1.56×)
SLA2	50	3.15 (5.76×)	0.558 (1.80×)	55.8 (1.80×)
TurboDiffusion (SageSLA)	4	2.88 (6.28×)	0.482 (2.08×)	1.93 (52.2×)
JetDiffusion (VSQA, Ours)	3	0.944 (19.2×)	0.160 (6.27×)	0.481 (209×)
RTX 5090, 14B, 720p				
FlashAttention-2	50	562 (1.00×)	32.6 (1.00×)	3260 (1.00×)
SageAttention2++	50	201 (2.79×)	17.7 (1.85×)	1770 (1.85×)
SageAttention3	50	163 (3.45×)	16.1 (2.03×)	1610 (2.03×)
VSA	50	94.9 (5.92×)	15.4 (2.12×)	1540 (2.12×)
SLA2	50	46.9 (12.0×)	12.0 (2.72×)	1200 (2.72×)
TurboDiffusion (SageSLA)	4	42.5 (13.2×)	8.17 (3.99×)	32.7 (99.8×)
LightWan2.2	4	32.5 (17.3×)	4.20 (7.77×)	16.9 (193×)
JetDiffusion (VSQA, Ours)	3	19.7 (28.5×)	3.57 (9.14×)	10.6 (307×)
GB300, 14B, 720p				
FlashAttention-4	50	116 (1.00×)	6.61 (1.00×)	661 (1.00×)
FlashAttention-4 (NVFP4)	50	73.1 (1.58×)	4.79 (1.38×)	479 (1.38×)
VSA	50	13.6 (8.49×)	2.89 (2.29×)	289 (2.29×)
JetDiffusion (VSQA, Ours)	3	6.74 (17.2×)	1.46 (4.52×)	4.39 (151×)

Table 6: **Generation timings (s).** Total DiT time of one generation, the non-DiT part of the pipeline, and end-to-end time, each followed in parentheses by its speedup over the dense base model of the same group. The base model runs its default pipeline and JetDiffusion decodes with TAEHV.

Method	Steps	DiT total (s)	Non-DiT (s)	End-to-end (s)
RTX 5090, 1.3B, 480p				
FlashAttention-2	50	101 (1.00×)	2.75 (1.00×)	103 (1.00×)
JetDiffusion (VSQA, Ours)	3	0.481 (209×)	0.39 (7.05×)	0.866 (119×)
RTX 5090, 14B, 720p				
FlashAttention-2	50	3260 (1.00×)	5.00 (1.00×)	3270 (1.00×)
JetDiffusion (VSQA, Ours)	3	10.6 (307×)	1.00 (5.00×)	11.6 (282×)
GB300, 14B, 720p				
FlashAttention-4	50	661 (1.00×)	8.75 (1.00×)	670 (1.00×)
JetDiffusion (VSQA, Ours)	3	4.39 (151×)	0.255 (34.4×)	4.65 (144×)

C Quantization in Training and Inference

Figure 7 compares one Wan2.1-14B video at 720p generated three ways from the same prompt: the training forward with fake quantization, inference with NVFP4 $QK^\top$ and PV on the RTX 5090, and inference with NVFP4 $QK^\top$ and FP8 PV on the GB300, where the kernel computes PV in FP8 for higher throughput. The first column against the other two compares fake and real quantization, and the last two columns compare NVFP4 and FP8 PV. The three videos show no visible difference. Training with fake quantization therefore carries over to both real inference kernels.



Figure 7: **Fake and real quantization, and NVFP4 versus FP8 PV.** Six evenly spaced frames of one Wan2.1-14B 720p video, from the first frame (top) to the last (bottom). Text prompt: *"A Chinese girl with a radiant smile, exuding beauty and confidence, stands in front of the camera under natural sunlight. She is dressed in a casual outfit with a harmonious blend of yellow and blue tones, complementing the vibrant hues of her surroundings. Her posture is relaxed and natural, capturing a positive and joyful perspective. The background showcases a picturesque natural landscape, featuring lush greenery and serene blue skies. The scene is captured in a medium close-up half-body shot, highlighting her engaging facial expressions and the cheerful ambiance."*

D Qualitative Results

Figures in this section compare videos from the base model and from JetDiffusion for the same text prompt. Each row shows six frames evenly spaced from the first frame to the last.

D.1 Wan2.1-T2V-1.3B

Wan2.1-T2V-1.3B



JetDiffusion (Ours)



Figure 8: Video generation with Wan2.1-T2V-1.3B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"Animated scene features a close-up of a short fluffy monster kneeling beside a melting red candle. The art style is 3D and realistic, with a focus on lighting and texture. The mood of the painting is one of wonder and curiosity, as the monster gazes at the flame with wide eyes and open mouth. Its pose and expression convey a sense of innocence and playfulness, as if it is exploring the world around it for the first time. The use of warm colors and dramatic lighting further enhances the cozy atmosphere of the image."*

Wan2.1-T2V-1.3B



JetDiffusion (Ours)



Figure 9: Video generation with Wan2.1-T2V-1.3B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A close-up shot of a ceramic teacup slowly pouring water into a glass mug. The water flows smoothly from the spout of the teacup into the mug, creating gentle ripples as it fills up. Both cups have detailed textures, with the teacup having a matte finish and the glass mug showcasing clear transparency. The background is a blurred kitchen countertop, adding context without distracting from the central action. The pouring motion is fluid and natural, emphasizing the interaction between the two cups."*

Wan2.1-T2V-1.3B



JetDiffusion (Ours)



Figure 10: Video generation with Wan2.1-T2V-1.3B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"In this serene video, a person sits cross-legged on a ledge in meditation, silhouetted against a vibrant orange sunset. They are captured in a moment of tranquility, with their eyes closed and hands resting on their knees in a mudra, as a small dog lies peacefully beside them. The background reveals a sprawling urban skyline, with buildings outlined against the glowing sun, which hangs low in the sky. The contrasting shadows beautifully highlight the figures, creating an atmosphere of peaceful coexistence between nature and city life. This captivating scene evokes a feeling of calm and reflection, perfectly encapsulated by the warm and harmonious colors of dusk."*

Wan2.1-T2V-1.3B



JetDiffusion (Ours)

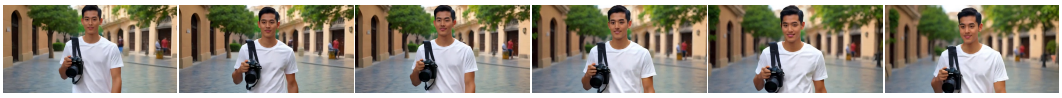


Figure 11: Video generation with Wan2.1-T2V-1.3B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A young man walks confidently through a picturesque plaza, clutching a professional camera over his shoulder. Dressed in a simple white t-shirt, his casual attire contrasts with the sophisticated camera equipment he carries, hinting at his expertise. The background is filled with charming historical architecture and lush greenery, suggesting a vibrant, bustling environment. The focus on the man's purposeful stride and the detailed lens of the camera creates an atmosphere of dynamic exploration and artistic intention."*

Wan2.1-T2V-1.3B



JetDiffusion (Ours)



Figure 12: Video generation with Wan2.1-T2V-1.3B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"In a serene forest setting, a brown and white dog with expressive eyes sits amidst a carpet of dry leaves. The tall trees create a peaceful and shaded ambiance, with sunlight filtering through the canopy. The dog is alert and attentive, as if listening to distant sounds, with its mouth slightly open revealing a playful expression. Its fur is fluffy and well-groomed, contrasting with the rough textures of the tree bark and forest floor. The surrounding tranquility enhances the focus on the dog, making it the central figure of this natural scene."*

Wan2.1-T2V-1.3B



JetDiffusion (Ours)



Figure 13: Video generation with Wan2.1-T2V-1.3B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"Sweeping over a breathtaking coastal landscape, the video captures a stunning aerial view of a secluded bay embraced by rugged mountains. The turquoise waters gently lap against expansive sandy shores that display intricate patterns shaped by the tides. Surrounding the bay, lush green mangroves and winding inlets create a patchwork of rich textures and vibrant colors. In the background, majestic mountains rise with jagged peaks casting shadows over the serene landscape. The clear blue sky above adds a layer of tranquility, accentuating the untouched beauty of this remote paradise. The scene is a harmonious blend of natural elements, portraying a tranquil and isolated place that invites exploration and reflection."*

D.2 WAN2.1-T2V-14B

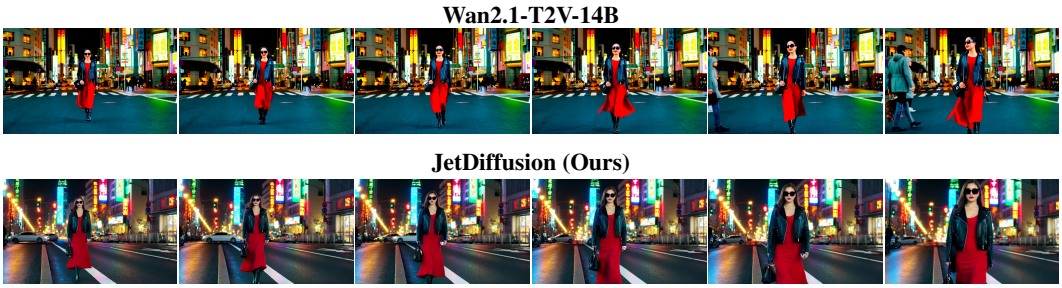


Figure 14: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. 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."*

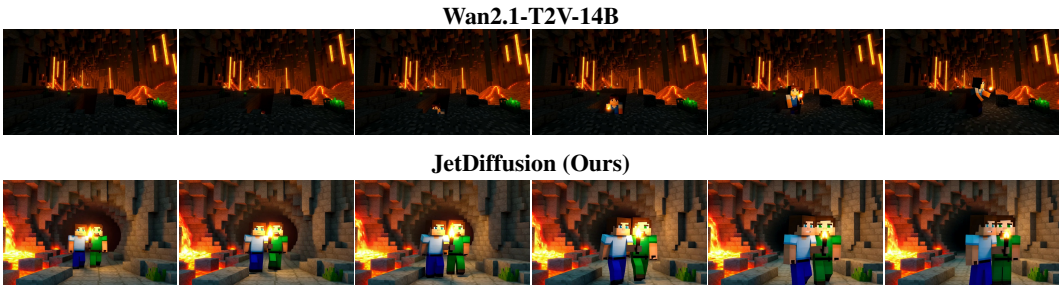


Figure 15: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A Minecraft player character holding a torch enters a massive underground cave. The torchlight flickers against jagged stone walls, illuminating patches of iron and diamond ores embedded in the rock. Stalactites hang from the ceiling, lava flows in glowing streams nearby, and the faint sound of water dripping echoes through the cavern."*

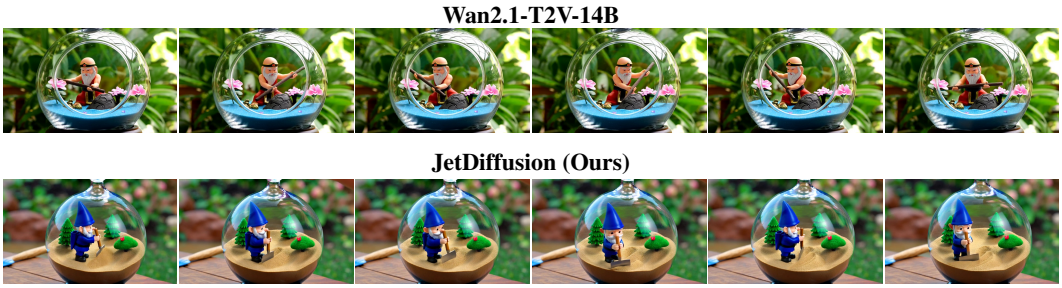
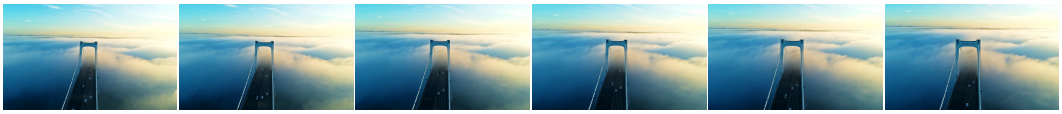


Figure 16: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A close up view of a glass sphere that has a zen garden within it. There is a small dwarf in the sphere who is raking the zen garden and creating patterns in the sand."*

Wan2.1-T2V-14B



JetDiffusion (Ours)



Figure 17: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A serene early morning scene with dense fog rolling over a long suspension bridge. The camera slowly pans along the length of the bridge, capturing the misty atmosphere as it gently moves across the structure. The bridge appears slightly ethereal, with only the occasional car light piercing through the fog. The rolling fog creates a dynamic and atmospheric effect, enhancing the sense of mystery and tranquility. Wide shot, slow horizontal pan."*

Wan2.1-T2V-14B



JetDiffusion (Ours)

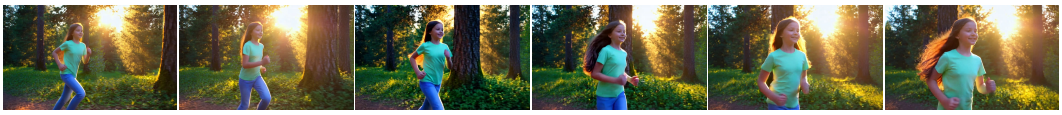


Figure 18: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A young girl with long brown hair running through a dense forest early in the morning. She is wearing a light green t-shirt and blue jeans, with bare feet, showcasing her adventurous spirit. The rising sun illuminates the scene, casting golden rays through the canopy of trees, creating a warm and vibrant atmosphere. The background shows tall trees with dappled sunlight filtering through the leaves. The girl's face is filled with joy and determination as she runs freely. Medium shot focusing on the girl from a side angle, capturing her motion and the surrounding nature."*

Wan2.1-T2V-14B



JetDiffusion (Ours)



Figure 19: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A hot air balloon drifting gracefully between lush green mountains, with cascading waterfalls visible in the distance. The balloon is brightly colored with vibrant patterns, contrasting against the serene backdrop of nature. The mountains are covered in dense forests, and mist rises from the waterfalls, creating a mystical atmosphere. The camera follows the balloon as it navigates through the valley, capturing the breathtaking scenery from a mid-shot perspective."*

Wan2.1-T2V-14B



JetDiffusion (Ours)



Figure 20: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"Cinematic UHD shot of a lone samurai walking gracefully through a dense forest bathed in dappled afternoon sunlight. The samurai is clad in traditional armor, with a katana sheathed at his side. His face is stern and focused, with a determined expression as he strides confidently along a moss-covered path. The forest around him is lush, with tall trees casting intricate patterns of light and shadow on the ground. The scene captures the tranquility and majesty of nature while highlighting the samurai's disciplined presence. Wide shot, static frame, emphasizing the interplay of light and shadow."*

Wan2.1-T2V-14B



JetDiffusion (Ours)



Figure 21: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"A firefighter in full gear moves dynamically towards a burning house, his helmet reflecting the intense flames. Snowflakes gently fall around him, creating a stark contrast between the cold winter environment and the raging fire. The house is engulfed in bright orange and red flames, with smoke billowing out from various windows and the roof. The scene captures the tension and urgency of the moment, with the camera focused closely on the firefighter as he advances, and the flames roaring behind him. Close-up view with a slight pan to follow the firefighter's movement."*

Wan2.1-T2V-14B



JetDiffusion (Ours)



Figure 22: Video generation with Wan2.1-T2V-14B. Top: the base model. Bottom: JetDiffusion. Text prompt: *"Aerial view of a peaceful village nestled between lush, rolling green hills. The village consists of quaint, colorful houses with red roofs, winding cobblestone streets, and small gardens filled with blooming flowers. Smoke gently rises from chimneys as the sun casts a warm glow over the scene. The landscape is dotted with fields of golden wheat and patches of dense forests. The camera slowly descends to a medium-altitude shot, capturing the tranquility and beauty of the rural setting. Wide shot."*