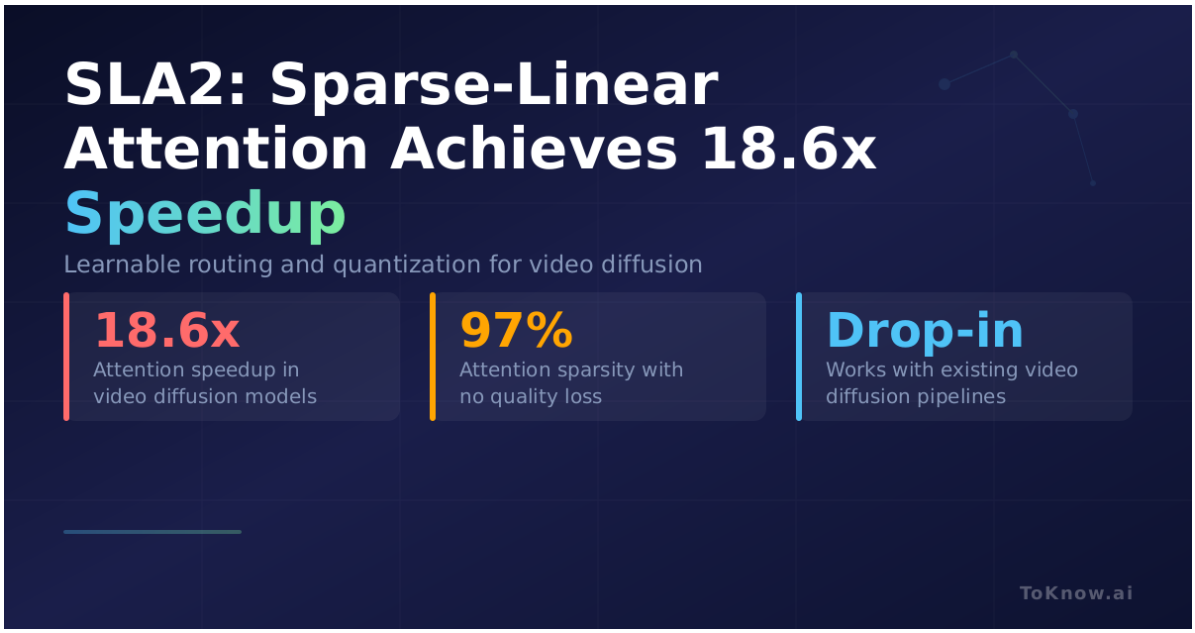
SLA2: Sparse-Linear Attention Achieves 18.6x Speedup in Video Diffusion Models

Kabui, Charles

2026-03-04



Researchers at Tsinghua University and UC Berkeley identified two problems with Sparse-Linear Attention (SLA), a technique that splits attention computations between a sparse branch (which only computes the most important attention scores) and a linear branch (which cheaply approximates the rest) to speed up diffusion models. First, SLA uses a fixed rule to decide which computations go to which branch, based on attention-weight magnitude. This

is often suboptimal. Second, the decomposition introduces a mathematical mismatch that degrades output quality. SLA2 fixes both. It introduces a learnable router that dynamically assigns each attention computation to the sparse or linear path during inference. It adds a learnable ratio to combine the two branches more faithfully. It also replaces the linear branch with low-bit quantized attention, meaning the model stores and computes attention values at reduced numerical precision, trained via quantization-aware fine-tuning so accuracy is preserved. The result: the model skips 97% of full-precision attention computations and delivers an 18.6x speedup on attention operations in video diffusion models, with no measurable loss in generation quality.

This matters for anyone running video generation at scale. Attention is the computational bottleneck in transformer-based diffusion models. An 18.6x speedup on attention translates directly to faster inference, lower GPU costs, and the ability to generate higher-resolution or longer videos within the same compute budget. The approach works as a drop-in improvement for existing video diffusion pipelines, since it modifies the attention mechanism without changing the overall architecture. For companies deploying models like Sora, Seedance, or Kling, this class of optimization determines whether video generation is economically viable at production volumes.

Sources:

- SLA2 Paper (arXiv:2602.12675)
- SLA2 on Hugging Face Papers
- SLA GitHub Repository
- SLA Original Paper (arXiv:2509.24006)
