DiFlash: A 13.70TOPS/W 156.05Pixels/mJ Video Generation Accelerator with Bubble-Free Online-Quantized FlashAttention and Query-Reuse Sliding-Tile Attention

Jun Liu[1,2,†], Shiwei Liu[3,†], Li Ding[1,†], Shuaiheng Li[1], Tianlang Zhao[1], Yuxuan Zhang[1], Xinhao Li[1], Shan Huang[1,4], Jinhao Li[1,4], Qi Luo[5], Yadong Dai[2], Jintao Li[2], Shulin Zeng[2], Chixiao Chen[5], Guohao Dai[1,2,4,‡]
[1]Shanghai Jiao Tong University, China [2]Infinigence-AI, China [3]Tongji University, China [4]SII, China [5]Fudan University, China
[†]These authors contributed equally to this work. [‡]Corresponding author: daiguohao@sjtu.edu.cn

***Abstract*—This work presents DiFlash, a 22nm 3D Diffusion Transformer accelerator for video generation. DiFlash features (1) a Fully-pipelined Quantized FlashAttention Core with bubble-free scheduling, boosting computation utilization to 93.6%, (2) a Reconfigurable Function Unit with dynamic skipping, eliminating >80% redundant non-linear operations for 1.94× energy reduction, and (3) a Sliding-Tile Management Unit implementing Query-Reuse Sliding-Tile Attention (STA), delivering 1.38× memory access reduction and 3.21× throughput gain over STA-only. Fabricated in 22nm CMOS, the prototype achieves 13.70 TOPS/W energy efficiency and 156.05 Pixels/mJ vision energy efficiency, outperforming prior works by 5.89–12.44×.**

***Index Terms*—Video Generation, 3D Diffusion Transformer, FlashAttention, Hardware Accelerator**

I. Introduction

3D Diffusion Transformers (3D DiTs), such as Wan2.1 [1] and CogVideoX [2], are emerging as the backbone of video generation. Their long token sequences and spatiotemporal redundancy make attention the dominant bottleneck, accounting for over 82.1% of inference overhead (Fig. 1). While FlashAttention (FA) [3] with online quantization and Sliding-Tile Attention (STA) [4] reduce algorithmic complexity, prior accelerators [5]–[8] target U-Net or DiT backbones with short attention sequences, where the score matrix fits within on-chip SRAM. Scaling to 3D DiT's long sequences (e.g., 16k tokens) inflates this matrix by over 250×, forcing all intermediate results off-chip to DRAM and leaving compute severely underutilized. Three critical challenges emerge when realizing them on 3D DiT hardware for long-sequence video generation (Fig. 2): (1) FA with online per-group quantization requires precision conversion between matrix multiplication (MM) and non-MM stages, introducing execution dependencies that cause pipeline bubbles and limit utilization to <27.9%. (2) The spatiotemporal similarity of activations renders over 80% of non-MM operations redundant, yet conventional hardware executes them indiscriminately, wasting dynamic energy on trivial outputs. (3) Although STA reduces computation by exploiting spatial locality, it does not reduce external memory access (EMA) accordingly: the sequential dataflow over 1D-flattened 3D tokens generates fragmented access patterns, causing EMA to dominate over 63.3% of total power.

To address these challenges, this work presents **DiFlash**, a 22nm 3D DiT accelerator for video generation with three key

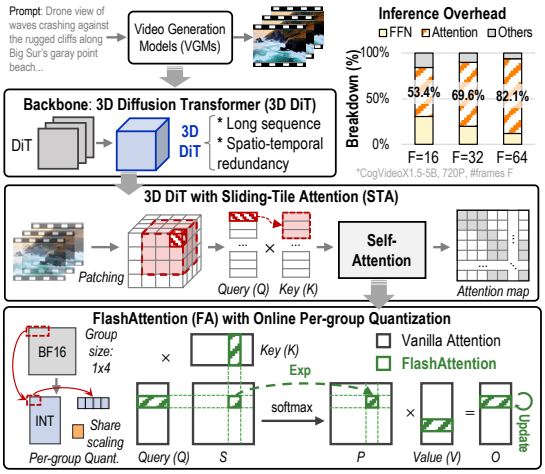

Fig. 1. Overview of 3D DiT-based video generation.

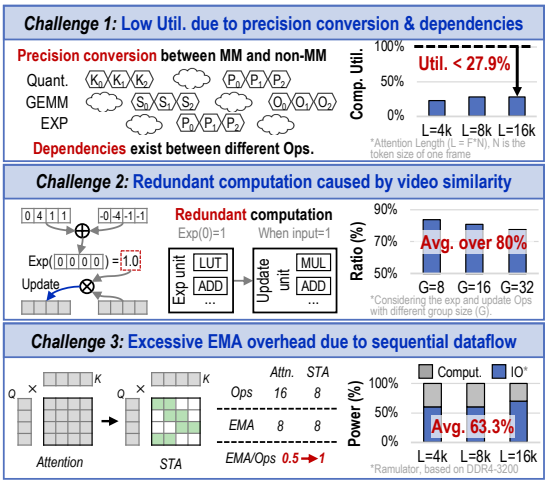

Fig. 2. Challenges of 3D DiT with FA and STA.

features: (1) a Fully-pipelined Quantized FA Core (FQFC) that fuses mixed-precision computation with out-of-order scheduling to eliminate bubbles, (2) a Reconfigurable Function Unit (RFU) that exploits temporal redundancy via dynamic skipping to reduce non-MM energy, and (3) a Sliding-Tile Management Unit (STMU) that restores spatiotemporal locality through

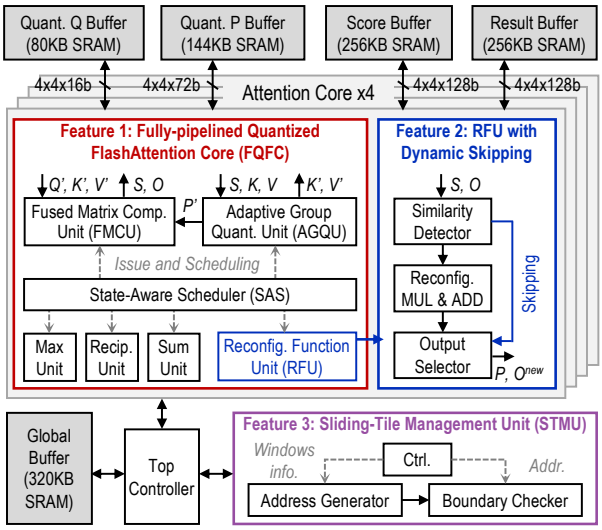

Fig. 3. Overall architecture of DiFlash and its key features.

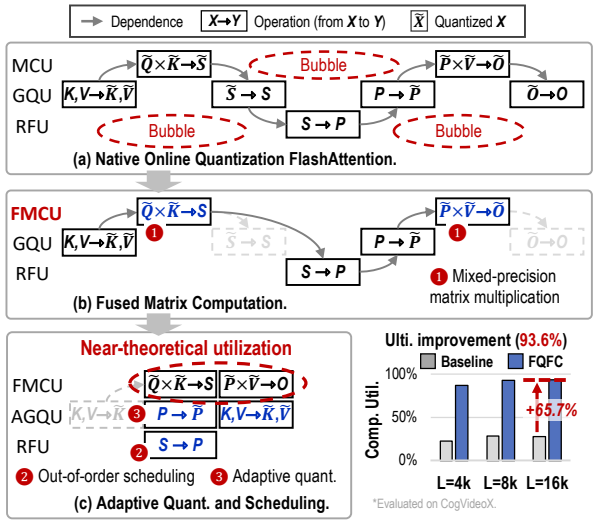

Fig. 4. Execution dataflow and orchestration of the FQFC.

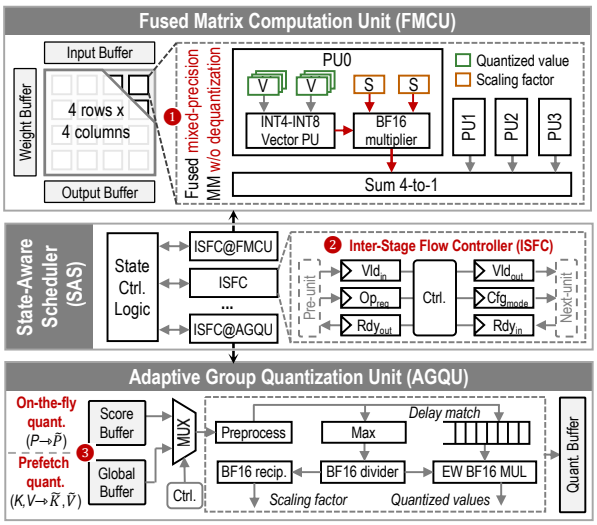

Fig. 5. Hardware architecture of the FQFC.

a Query-Reuse STA dataflow to minimize EMA. Fabricated in 22nm CMOS, the prototype achieves 13.70 TOPS/W and 156.05 Pixels/mJ, outperforming prior works by 5.89–12.44×.

II. Architectural Overview

Fig. 3 illustrates the overall architecture of DiFlash. The system integrates four parallel Attention Cores, an STMU, and a hierarchical memory system with a 320KB Global Buffer and distributed local buffers totaling 1056KB on-chip SRAM. Each Attention Core contains an FQFC (Feature 1) as the primary computation engine and an RFU (Feature 2) for non-MM operations with dynamic skipping. The STMU (Feature 3) manages data movement for the Query-Reuse STA method. The architecture supports INT4/8 group quantization for MM and BF16/FP32 for non-MM operations.

III. Fully-pipelined Online-Quantized FA Core

Fig. 4 illustrates the execution dataflow of the FQFC. As shown in Fig. 4(a), native online quantization FA suffers from pipeline bubbles because each tile must finish quantization statistics collection before the next MM can begin, limiting utilization to <30%. The FQFC eliminates these bubbles in two steps. First, fused mixed-precision computation (Fig. 4(b)) merges de-quantization scaling into the MM pipeline, removing the separate precision-conversion stage. Second, out-of-order scheduling with adaptive quantization (Fig. 4(c)) overlaps the current tile's MM with statistics collection for the next tile, hiding the quantization latency entirely and raising utilization to 93.6%.

Fig. 5 shows the hardware architecture of the FQFC. The Fused Matrix Computation Unit (FMCU) employs INT4-INT8 mixed-precision Vector PUs for high-parallelism dot products. The integer partial sums are immediately scaled by BF16 multipliers using quantization factors, restoring high precision without intermediate storage. The Adaptive Group Quantization Unit (AGQU) follows a Q4K8P8V4 scheme, adaptively supporting on-the-fly quantization for attention scores ($P \to \tilde{P}$) and prefetch quantization for keys/values ($K, V \to \tilde{K}, \tilde{V}$). A dual-path design derives scaling factors while an Input FIFO matches the data latency. The State-Aware Scheduler (SAS) coordinates these modules via Inter-Stage Flow Controllers (ISFCs) that monitor valid/ready handshakes and operation requests, enabling out-of-order dispatch across pipeline stages.

IV. RFU with Dynamic Skipping

Fig. 6 illustrates the reconfigurable datapath and dynamic skipping mechanism of the RFU. The non-MM stages in FA comprise two operations: *exp* for softmax normalization and *update* for output rescaling, both requiring FP32 arithmetic. Inspired by the Schraudolph approximation [9], we observe that *exp* ($e^z = 2^{z/\ln 2}$) can be reduced to a multiply-accumulate (MAC) via integer-to-float bit reinterpretation, enabling *exp* and *update* to share the same Sub-Mul-Add datapath. Fur-

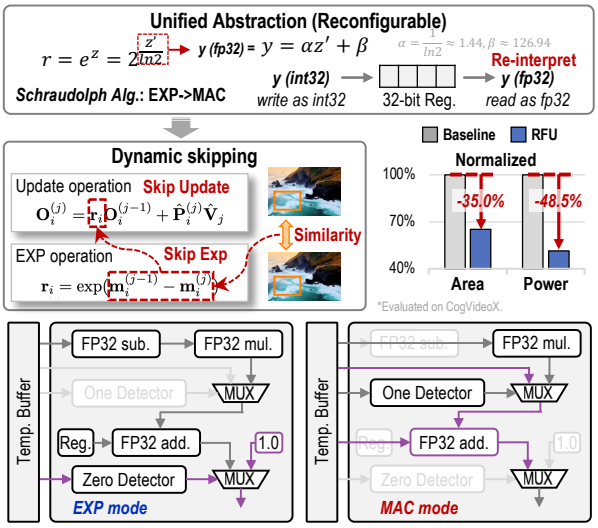

Fig. 6. Reconfigurable datapath and dynamic skipping of the RFU.

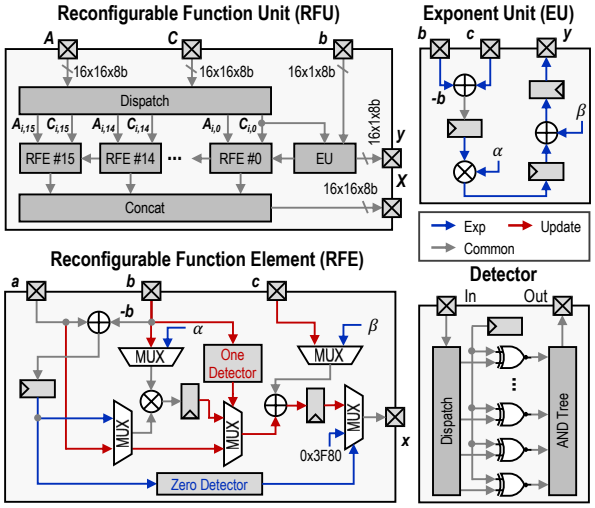

Fig. 7. Circuit design of the RFU.

thermore, high temporal similarity across video frames makes over 80% of these operations redundant: most *exp* inputs are near zero (e^0=1) and most *update* scaling factors are identity (r=1). A zero-detector bypasses *exp* when the input is negligible, and a one-detector clock-gates the multiplier in *update* when the scaling factor is unity. The RFU achieves 1.54× area reduction via datapath sharing, with 1.94× energy reduction via reconfiguration and dynamic skipping.

Fig. 7 shows the circuit design of the RFU. The unit comprises 16 parallel Reconfigurable Function Elements (RFEs) and one Exponent Unit (EU). A Dispatch module distributes operands ($\mathbf{A}$, $\mathbf{C}$, $\mathbf{b}$) across the RFE array, and a Concat module merges the element-wise results. Inside each RFE, a three-stage Sub-Mul-Add pipeline is shared between *exp* and *update* through per-stage MUXes: the subtractor computes ($\mathbf{a}-\mathbf{b}$), the multiplier takes either the constant α (*exp*) or the data operand $\mathbf{b}$ (*update*), and the adder appends the bias β (*exp*)

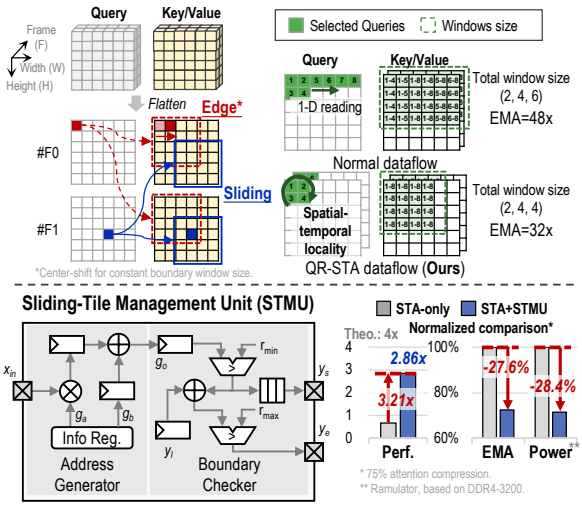

Fig. 8. QR-STA method and circuit design of the STMU.

or accumulates $\mathbf{c}$ (*update*). The zero-detector monitors the subtractor output and bypasses the remaining stages to force the output to 1.0 (0x3F80) when the output is negligible. The one-detector checks the multiplier operand and clock-gates the multiplier to directly route the data to the adder input when the operand is unity. Both detectors employ a parallel dispatch-and-AND-tree for single-cycle, per-element decisions. The EU reuses the same Sub-Mul-Add pipeline for computing the row-level rescaling vector $\mathbf{y}$.

V. Sliding-Tile Management Unit

Fig. 8 illustrates the QR-STA method and the STMU circuit design. STA reduces theoretical FLOPs, yet flattening 3D video tokens into a 1D sequence breaks spatiotemporal locality, generating irregular KV windows that inflate EMA and force token-level sequential processing with low PE utilization. QR-STA restores locality by classifying each query tile as Edge or Sliding. Edge tiles near frame boundaries apply center-shift clamping to keep a constant window size (2,4,4), enabling static KV buffering and cross-tile Q reuse that eliminates redundant DRAM fetches. Sliding tiles exploit spatial continuity: only the incremental KV slice is streamed from DRAM while the remaining KV and cached Q are served from the Global Buffer.

The STMU realizes this strategy in hardware circuit. The Address Generator stores tile shape parameters in an Info Register and computes burst-aligned DRAM addresses (g_o) through shift, multiply, and add operations from the input coordinate x_{in}. The Boundary Checker compares g_o against pre-loaded range bounds r_{min} and r_{max} to classify tile regions, producing Sliding (y_s) or Edge (y_e) control signals that steer the data-fetch pipeline. STA-only forces token-level sequential processing due to irregular KV windows, severely degrading PE utilization. With 75% attention compression, the STMU regularizes the dataflow to enable packed MM execution, delivering a 3.21× throughput gain over STA-only, while reducing EMA by 27.6% and DRAM power by 28.4%.

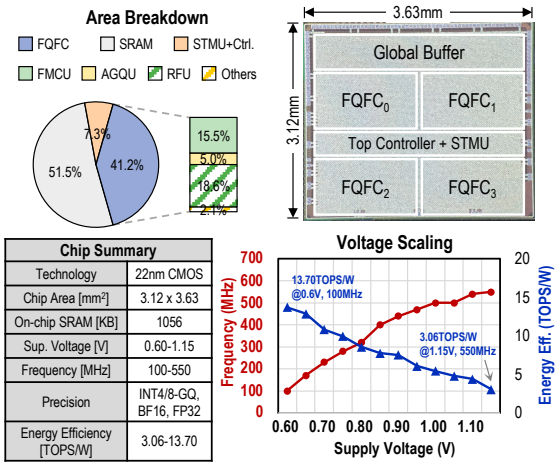

Fig. 9. Die photo and summary of the prototype.

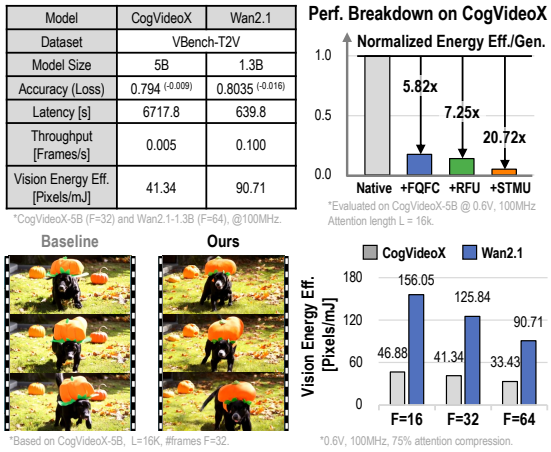

Fig. 10. Measurements with voltage scaling and breakdown.

VI. Measurement Results

Fig. 9 shows the die photo and chip summary. DiFlash is fabricated in 22nm CMOS with a die area of 11.33mm^2, integrating 1056KB on-chip SRAM (51.5%), four FQFCs (41.2%), and the STMU with top controller (7.3%). It operates at 0.60–1.15V and 100–550MHz, achieving 3.06–13.70 TOPS/W. Fig. 10 presents the measured results. DiFlash is evaluated on CogVideoX-5B and Wan2.1-1.3B using VBench-T2V [10] at 0.6V and 100MHz with 75% attention compression, showing negligible accuracy loss. It achieves vision energy efficiency of 46.88 Pixels/mJ on CogVideoX (F=16) and 156.05 Pixels/mJ on Wan2.1 (F=16). The feature-wise energy breakdown on CogVideoX shows cumulative gains from the proposed modules: FQFC improves efficiency by 5.82× through pipeline-bubble elimination, RFU further improves it to 7.25× via reconfiguration and redundancy skipping, and STMU raises the total gain to 20.72× through attention compression. Table I compares DiFlash with recent accelerators [5]–[8]. Prior works target U-Net or DiT backbones with attention lengths below 1k, whereas DiFlash targets 3D DiT video generation with sequences up to 16k tokens. It achieves 156.05 Pixels/mJ on Wan2.1 (F=16), outperforming prior works by 5.89–12.44×, consistent with its long-sequence 3D DiT-oriented design.

TABLE I
Comparison with state-of-the-art accelerators.

	ISSCC'25 [5]	ISSCC'25 EdgeDiff [6]	VLSI'25 [7]	ISSCC'26 Tiamat [8]	This work
Technology	22nm	28nm	28nm	16nm	**22nm**
Application	Image, 3D, Video Gen.	Image Gen.	Robot Control	Image Gen.	**Video Gen.**
Model Size [Billion]	0.44-0.75	2.6	0.025	0.9	**1.3-5.0**
Attention Length	0.9k	1k	10	256	**16k**
Support Model	SDv1.5, SVD, Wonder-3D	SDXL-Turbo	DiTAG	DiT-XL/2, PixArt-α	**CogVideoX-5B, Wan2.1-1.3B**
Backbone	U-Net	U-Net	DiT, U-Net	DiT	**3D DiT**
Chip Area [mm^2]	8.20	20.25	8.41	8.00	**11.33**
On-chip SRAM [KB]	448[*1]	1952	2208	1518	**1056**
Sup. Voltage [V]	0.6-0.9	0.68-1.0	0.6-1.2	0.66-1.05	**0.60-1.15**
Frequency [MHz]	180-478	50-250	250-1250	100-400	**100-550**
Precision	INT4/8/12, FP8, BF16	INT4/8/12/ 16-GQ	INT4/8, FP4	MXINT5	**INT4/8-GQ, BF16, FP32**
Peak Throughput[*2] [TOPS]	9.71	31.3 (INT4) 7.8 (INT8)	2-10	3.68	**0.82-4.51**
Energy Eff.[*3,4] [TOPS/W]	49.74-60.81[*5]	14.2-15.5	17.6[*6]	4.99-11.64	**13.70**
Vision Energy Eff. Task measured[*7]	Video Gen. unknown	Image Gen. 512x512	/	Image Gen. 256x256	**Video Gen. 480x832x{64,16}**
Vision Energy Eff. [Frames/J][*3]	0.06	0.05	/	0.40	**0.23, 0.39**
Vision Energy Eff. [Pixels/mJ][*3,8]	/	12.54	/	26.48	**90.71, 156.05**

1) 3Mb eDRAM + 64kB SRAM. 2) 1 MAC = 2 OPs. 3) 0.6V, 100MHz, 75% attention compression, excluded EMA. 4) Mixed-precision: INT4/8 (FMCU), BF16 (others), FP32 (RFU). 5) Considering CIM utilization. 6) 80% input activation sparsity w/ 50% weight sparsity. 7) Resolution: Height(H) x Width(W) x Frames(F). 8) Pixels/mJ = H x W x Frames/mJ.

Acknowledgment

This work was supported by Shanghai Rising-Star Program (No. 24QB2706200) and the National Natural Science Foundation of China (No. 62561160156).

References

[1] Wan Team, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.
[2] Z. Yang *et al.*, "CogVideoX: Text-to-Video diffusion models with an expert transformer," in *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
[3] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
[4] P. Zhang *et al.*, "Fast video generation with Sliding Tile Attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.
[5] Y. Jing *et al.*, "A 22nm 60.81TFLOPS/W diffusion accelerator with bandwidth-aware memory partition and BL-segmented Compute-in-Memory for efficient multi-task content generation," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 68, 2025, pp. 1–3.
[6] S. Kim *et al.*, "EdgeDiff: 418.4mJ/inference multi-modal few-step diffusion model accelerator with mixed-precision and reordered group quantization," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 68, 2025, pp. 1–3.
[7] S. Zhang *et al.*, "A 94Hz inference and 7.4mJ/epoch fine-tune edge SoC for diffusion-based robot manipulation with speculation and disturbance enhancement," in *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2025, pp. 1–3.
[8] P.-Y. Lu *et al.*, "Tiamat: A 98-to-134ms/step Transformer-based diffusion model processor supporting classifier-free guidance for image generation," in *2026 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 69, 2026, pp. 54–56.
[9] N. N. Schraudolph, "A fast, compact approximation of the exponential function," *Neural Computation*, vol. 11, no. 4, pp. 853–862, 1999.
[10] D. Zheng *et al.*, "VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness," *arXiv preprint arXiv:2503.21755*, 2025.