ACCELERATING CAUSAL VIDEO GENERATION WITH MIXED SPARSE ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal video diffusion has emerged as a streaming alternative to bidirectional models that denoise entire clips before playback. These autoregressive models use a key/value cache, but attention costs grow with visible history. However, causal masking changes attention support, so directly adapting bidirectional sparse attention can degrade quality. Prior work also leaves attention patterns across chunks, denoising passes, and horizons underexplored. We observe stable spatial-local heads, weakening recent-window support, and changing reads despite immutable cached keys. Based on these, we propose Mixed Sparse Attention (MSA), which combines static/content routing, horizon/cost gates, reusable statistics, and budgeting across passes. For 20-second Causal Forcing generation at 720p, MSA achieves up to 2.634× denoising and 1.550× end-to-end speedup over dense attention. Moreover, MSA achieves comparable VBench and VBench-Long scores to the matched sparse-attention baseline across six evaluated dimensions.

1 INTRODUCTION

Bidirectional video diffusion transformers generate high-quality videos but denoise entire clips before playback, limiting interactive usage (Wang et al., 2025; Kong et al., 2025; Yang et al., 2025c; Zheng et al., 2024). Causal video diffusion instead generates short chunks conditioned on a key–value (KV) cache of preceding frames. Combined with few-step distillation, this formulation supports streaming output, longer generation horizons, and prompt changes during generation (Yin et al., 2024b; Huang et al., 2025a; Zhu et al., 2026; Liu et al., 2025; Yang et al., 2025a). However, a full-history cache makes each new chunk attend to all preceding chunks, so total attention work grows quadratically. For Self Forcing (SF) at 720p, extending generation from 5 to 20 seconds increases dense query–key comparisons from 5.9 to 79.4 trillion, a 13.5× increase. Rolling windows bound history, but attention within the visible window remains costly, motivating sparse execution.

Existing methods reduce attention reads through fixed spatial or temporal patterns, or through content-dependent block selection (Zhang et al., 2025b; Li et al., 2025; Xi et al., 2025; Yang et al., 2025b; Xu et al., 2025). However, causal generation changes their operating conditions. Current queries evolve during denoising, whereas cached history contains committed clean representations. Moreover, context changes as chunks are appended or evicted, and attention dependencies can exceed the calibrated horizon. Autoregressive selectors such as LightForcing (LF) address this setting directly (Lv et al., 2026), but the behavior of attention across heads, denoising passes, and generation horizons remains insufficiently characterized.

Our evaluation illustrates this transfer difficulty (Figure 1). STA (Zhang et al., 2025b) and XAttention (Xu et al., 2025) exhibit substantial losses in subject consistency at several calibrated settings, while SVG2 (Yang et al., 2025b) also degrades. The autoregressive baseline also shows calibration sensitivity on Rolling Forcing (RF) and the long-video extension of Causal Forcing (CF-long). These failures depend on the configuration and do not establish universal sparsity limits. The reported baseline densities are not verified as the fraction of visible query–key pairs actually read, precluding equal-budget comparisons.

Dense-rollout probes reveal stable spatial-local heads, weakening recent-window support, and changing reads despite immutable cached keys (Section 3). These findings motivate Mixed Sparse Attention (MSA), a training-free framework that combines static patterns with content selection. Its global budget covers every attention call, including dense warm-up and cache-update passes, while

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.

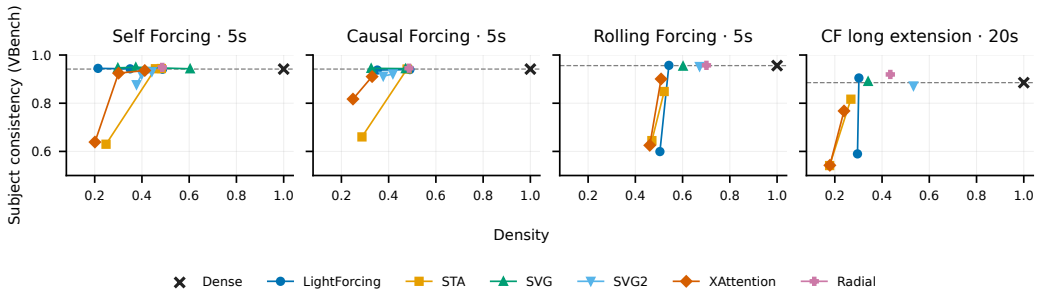

Figure 1: Subject consistency (VBench) of six sparse baselines versus density. The dashed line marks dense attention.

reusable history statistics preserve query-dependent selection. At 720p, MSA achieves up to 2.634× denoising and 1.550× end-to-end speedup over dense attention on CF-long, with comparable quality to the matched sparse-attention baseline on six VBench and VBench-Long dimensions. These gains reflect all-content execution; dense-quality tradeoffs and limited samples constrain broader claims.

Our contributions are summarized as follows.

- We investigate and characterize attention patterns across heads, denoising passes, and generation horizons for autoregressive video diffusion.
- We introduce Mixed Sparse Attention (MSA), combining calibrated head routing, runtime gates, reusable statistics, and budgeting over all executed passes.
- We conduct generation quality and system evaluations with matched baselines and achieved 2.634× denoising over the dense baseline and comparable quality on VBench and VBench-Long.

2 Background

Diffusion models and video diffusion transformers. Diffusion and flow models progressively transform Gaussian noise into data through repeated neural-network evaluations at decreasing noise levels (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Liu et al., 2022). Latent video models perform this process in a compressed representation: an autoencoder reduces spatial and temporal resolution, and a diffusion transformer operates on patches of the resulting latent grid (Rombach et al., 2022; Peebles & Xie, 2023; Wang et al., 2025; Kong et al., 2025; Yang et al., 2025c). In a bidirectional model, each video token can attend to tokens from both earlier and later frames, coupling appearance and motion across the clip.

For a single attention head with query and key dimension d_h, attention computes

$$A = \mathrm{softmax}\left(QK^\top/\sqrt{d_h} + M\right), \qquad O = AV, \tag{1}$$

where Q, K, and V are query, key, and value matrices, softmax acts over keys, and M is zero for visible pairs and $-\infty$ otherwise. With N_q queries, N_k keys, and value dimension d_h, the two matrix products cost approximately $4N_qN_kd_h$ FLOPs, counting a multiply–add as two operations. Bidirectional self-attention has $N_q = N_k = N$, giving quadratic scaling in video-token count.

Compression still leaves long sequences. Wan2.1-1.3B compresses space by 8×8 and time by four, then forms $1 \times 2 \times 2$ latent patches. An 81-frame video at 1280×720 yields 21 latent frames of 3,600 tokens ($N = 75{,}600$). Across 30 layers and 12 heads of dimension 128, dense $QK^\top$ and AV require $4 \cdot 30 \cdot 12 \cdot 128\, N^2 = 1.053 \times 10^{15}$ FLOPs per forward pass. This count excludes projections, feed-forward layers, and decoding. FlashAttention reduces memory traffic by tiling exact attention without materializing the full probability matrix, but still evaluates all visible query–key pairs (Dao et al., 2022). Reducing those pairs offers a complementary source of savings.

Causal video diffusion. Causal video diffusion generates successive chunks conditioned on preceding content, with bidirectional attention within each chunk and a KV cache supplying com-

pleted chunks; future chunks are masked out (Chen et al., 2024; Yin et al., 2024b). Denoising recomputes the current chunk's queries, keys, and values, and a clean-cache pass stores its finalized representations. Caching avoids recomputing history but retains its attention reads: for C chunks of n tokens, one pass per chunk reads $n^2C(C+1)/2$ pairs per head under full-history attention. Distribution-matching distillation reduces denoising evaluations (Yin et al., 2023; 2024a), while training on model-generated rollouts exposes the model to errors inherited from previous chunks (Huang et al., 2025a; Zhu et al., 2026). Neither removes the attention cost within each evaluation.

Our full-context Self Forcing configuration retains all preceding chunks, whereas Rolling Forcing jointly denoises a bounded rolling window and retains initial-frame KV states as an attention sink. Causal Forcing (CF) uses a chunkwise cache for 5-second generation and a rolling procedure for its long-video extension (Huang et al., 2025a; Liu et al., 2025; Zhu et al., 2026). Bounded windows limit per-chunk attention work but remove access to evicted content. These cache policies determine visible keys and attention layouts during initialization, denoising, and cache updates, which sparse execution must accommodate.

Sparse attention for diffusion transformers. Sparse attention selects a subset $\mathcal{S}_i$ of the visible keys $\mathcal{V}_i$ for each query i and normalizes attention over that subset. Selection changes both the available values and softmax normalization, making retained dependencies central to approximation quality. Block-sparse kernels skip omitted blocks while preserving efficient matrix operations within selected blocks; block alignment and selection overhead affect the resulting latency savings (Zhang et al., 2025b; Xu et al., 2025).

Static methods prescribe support using token coordinates: STA (Zhang et al., 2025b) uses hardware-aligned three-dimensional neighborhoods, while Radial Attention (Li et al., 2025) shrinks spatial neighborhoods with temporal distance. SVG (Xi et al., 2025) profiles sampled queries online to choose spatial or temporal patterns per head and rearranges temporal tokens for efficient access. Content-dependent methods offer more flexible support: SVG2 (Yang et al., 2025b) clusters and permutes tokens by semantic similarity, selecting cluster pairs through centroid attention and a cumulative-probability threshold, while XAttention (Xu et al., 2025) estimates block importance from sampled antidiagonal entries. This flexibility introduces scoring, clustering, or permutation overhead.

For autoregressive generation, LightForcing assigns larger budgets to earlier chunks and progressively increases sparsity. Pooled queries retrieve historical frames through frame-level key summaries, then select blocks within candidate frames (Lv et al., 2026). The original method includes additional training; our evaluation uses fixed pretrained weights. Evolving queries and committed history motivate analyzing pattern stability during causal rollout (Section 3).

3 How attention behaves in causal video diffusion

We examine attention across heads, denoising passes, and horizons in unmodified dense SF rollouts. For three 720p prompts, we sample chunks 0/2/6 at 5 seconds and 0/2/6/13/26 at 20 seconds, using 24 stratified queries across all 360 heads and five passes. FP32 softmax covers all visible keys. These 43,200 head–snapshot observations are correlated measurements from six videos (Appendix B). *Retained mass* denotes the dense attention probability captured by a candidate pattern. A separate two-prompt probe supports O3; O4 derives budget constraints (Figure 2).

O1: Spatial locality remains stable for a subset of heads. A subset of heads consistently concentrates attention within local spatial bands over the sampled horizons. Calibration assigns 34 SF heads to local row bands of radius one or two, 53 to recent windows ($m \in \{3, 4, 5\}$), and 273 to content selection (Section 4.1). Local patterns retain 0.971 attention mass at the 5-second endpoint and 0.970/0.966/0.957 at 21/42/81 visible frames during 20-second rollouts. Neither these probes nor the separate 480p temporal-offset study establish universal stationarity. Runtime gates reassess static candidates, retaining static execution only for SF at density 0.2 in our evaluation (Appendix B).

O2: Recent context becomes insufficient as history grows. Recent-window patterns calibrated on short histories require reassessment as generation extends beyond that horizon. Endpoint attention mass on history rises from 0.46 at 5 seconds to 0.69 at 20 seconds, beyond the training horizon.

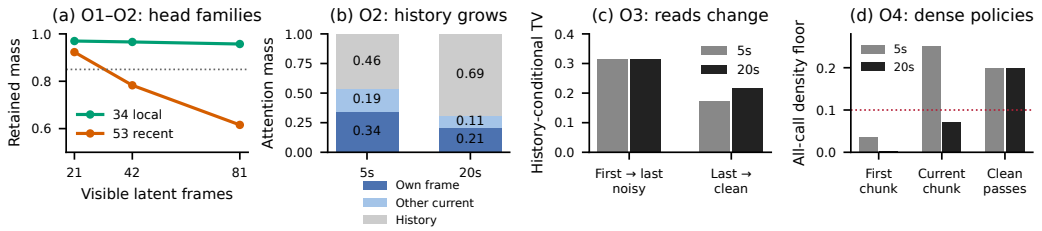

Figure 2: Attention measurements and budget constraints. (a) Local and recent-window retained mass at the final noisy pass of sampled chunks in 20-second full-history rollouts. (b) Endpoint history share at the final noisy pass. (c) History-conditional total variation between passes from a separate structural probe. (d) Analytic density floors for three dense-support policies, with a 0.1 budget marked by the dotted line. These floors overlap and cannot be added.

Frames covering 90% of mass increase from 8.8 to 36.5. Meanwhile, recent-window retained mass falls to 0.923/0.783/0.616 at 21/42/81 visible frames, and head-output normalized root-mean-square error (NRMSE) rises from 0.076 to 0.326. Of 87 static candidates, 5 fail the 0.85 retained-mass threshold at 5 seconds, versus 54 at 20 seconds.

Frame selection also limits support. Six historical frames plus the current chunk cover only 11% of an 81-frame cache, insufficient to fill a 0.2 budget. In fixed 20-second probe snapshots, unrestricted pooled-block selection retains 0.759 attention mass at density 0.1, versus 0.476 for the six-frame hierarchical selector; at 0.2, the values are 0.846 and 0.508. These retained-mass diagnostics motivate expanding eligible frames to accommodate the budget; they do not measure generation quality or latency.

O3: Fixed cached keys do not imply fixed attention. Cache immutability permits exact reuse of deterministic key statistics, but does not justify reusing attention masks. Committed keys and values remain fixed absent eviction or RoPE-coordinate changes, while queries evolve during denoising. History-normalized attention has first-to-last-pass total variation 0.31 at both durations, compared with 0.27–0.30 on the current chunk (Figure 2(c)). Mask reuse also reduces subject consistency in an ablation (Appendix B). For frame-aligned key means, repeated pooling over C chunks reads $C(C-1)/2$ history chunk-equivalents per layer, versus $C-1$ with incremental reuse. History-pooling work consequently falls by 3.5×/13.5× at 5/20 seconds, without equivalent latency guarantees.

O4: Attention budgets must include every executed pass. Reliable attention budgets require a complete, pair-weighted census of executed reads. For MSA and the matched study, we therefore define density as

$$\rho = \frac{\sum_{a,h,i} |\mathcal{S}_{ahi}|}{\sum_{a,h,i} |\mathcal{V}_{ahi}|}, \tag{2}$$

where $\mathcal{V}_{ahi}$ and $\mathcal{S}_{ahi}$ denote visible and read keys for query i, head h, and call a. Accounting covers every layer, noisy/clean pass, dense fallback, and partial tile. In 5-second full-history SF, dense first-chunk, current-chunk, and clean-pass policies individually impose floors of 0.036, 0.25, and 0.20. Local row bands read 9.7–13.9% of visible keys per tile; calibrated static patterns plus a dense first chunk consume 0.086 before content selection. Incomplete call averages can also misrepresent density. An RF LF log reports 0.469 after 512 calls, whereas the complete 660-call census gives $\rho = 0.199596$. We distinguish the reported density d in the baseline results from this audited measure. Memory contiguity also affects runtime (Appendix B).

4 Mixed Sparse Attention

MSA accelerates causal video attention without training through four complementary components (Figure 3). Calibrated head families capture spatial structure (O1), a horizon gate reassesses recency (O2), reusable statistics support refreshed selection (O3), and complete execution accounting enforces the attention budget (O4).

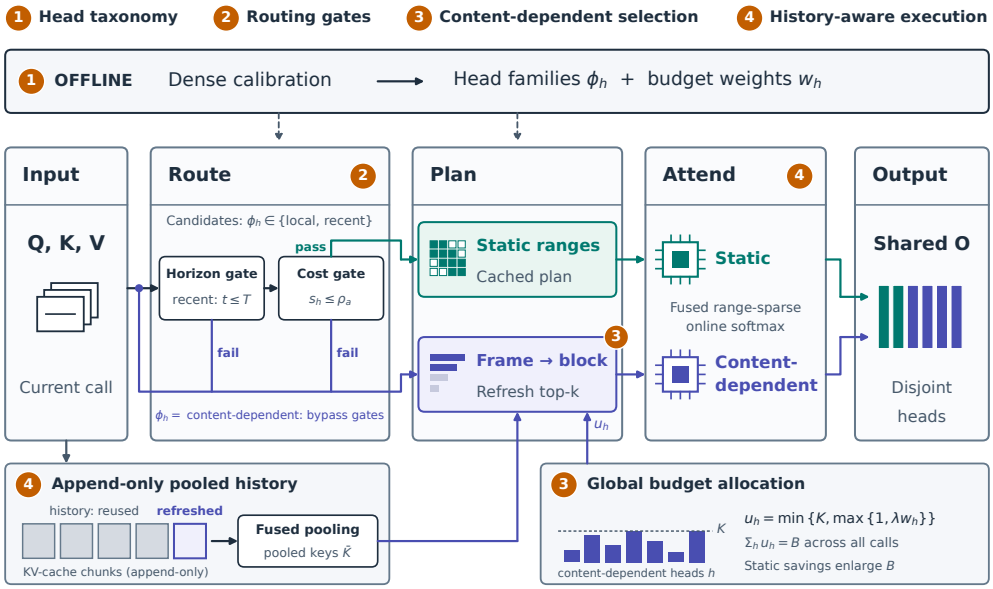

Figure 3: Overview of MSA. Badges mark its four components: (a) calibrated head taxonomy (Section 4.1), (b) runtime routing gates (Section 4.2), (c) content-dependent selection with a global budget (Section 4.3), and (d) history-aware execution (Section 4.4). Dashed arrows carry offline calibration; solid arrows show runtime flow.

4.1 Calibrated Head Taxonomy

Offline calibration assigns static patterns only to heads that meet attention-mass and output-error thresholds. Each head receives $\phi_h \in \{\text{local}(r),\ \text{recent}(m),\ \text{content}\}$, an offline analogue of SVG's online classification (Xi et al., 2025). Local heads read $2r+1$ grid rows around the query row in every visible frame; recent heads read the newest m frames, including the current chunk. Calibration chooses the cheapest static family retaining at least 0.85 attention mass on every calibration prompt with head-output NRMSE at most 0.2; other heads use content selection. It also assigns weights w_h increasing with measured hardness, defined by mass missed by refreshed top-20% block selection. Weights are normalized to model-wide mean one. Head assignments and weights are the only offline inputs.

4.2 Runtime Routing Gates

Runtime gates enforce horizon and cost constraints on each call. The *horizon gate* promotes recent-window heads to content selection beyond calibrated history T. With $T = 21$, it is inactive at 5 seconds and promotes SF's 53 recent heads during the last three quarters of a 20-second full-history rollout. The *executed-cost gate* requires the exact static read fraction s_h, computed from the range plan, to satisfy $s_h \le \rho_a$ for call target ρ_a. Savings $\sum_{h \in \text{static}} (\rho_a - s_h)$ fund content heads. A separate density census accounts for block rounding, partial tiles, and dense fallbacks.

The main RF/CF-long results evaluate content selection and execution with every sparse head assigned to content selection. This assignment, also used by CF at density 0.2, follows development tests in which transferred static patterns failed the quality or runtime criteria. In the evaluated SF/CF configurations at density 0.1, the cost gate promotes every static candidate; only SF at 0.2 retains static execution (Appendix B).

4.3 Content Selection with a Global Budget

Content selection combines hierarchical retrieval with a global allocation that respects available capacity. Following LightForcing, content heads score mean-pooled, frame-aligned 128-token blocks. The first stage retains sink, nearby, and current frames plus the top-k_f historical frames; the second

selects their top-k key blocks. Frame selection encourages contiguous reads. The density schedule keeps the first chunk dense and reduces attention reads in later chunks; calibration includes every attention call and meets the overall density target within ± 0.002. If the selected frames contain too few unmasked key blocks to fill the allocated budget, the selector admits additional historical frames. The allocator distributes a total budget of B key blocks among content heads according to their weights. Each head receives at least one block and at most K, the number of eligible blocks; λ is chosen to satisfy the total budget:

$$u_h = \min\{K, \max\{1, \lambda w_h\}\}, \qquad \sum_h u_h = B, \tag{3}$$

Integer allocations are obtained by rounding u_h down, then assigning remaining blocks to heads with the largest fractional remainders. Model-wide weights redistribute budget across layers and heads, while each call refreshes selections. The final census verifies executed blocks, partial tiles, and dense fallbacks because nominal top-k settings alone do not determine the all-call density.

4.4 History-aware execution

Append-only pooled history (O3). Incremental pooling reduces repeated processing of immutable history while preserving refreshed selection. It reuses history-key means and updates current-chunk means with a fused Triton reduction (Tillet et al., 2019). Buffers track rotary positional embedding (RoPE) coordinates and query/history boundaries and are invalidated after eviction or restart. Corrected pooling arithmetic for partially filled final blocks removes the observed differences in selected attention masks between the fused and reference pooling implementations; this agreement has only been checked on the tested inputs. History reuse is disabled in default SF and enabled in its pooling ablation and in CF, RF, and CF-long. Estimated key-pooling work, measured as the total number of latent frames processed by pooling across calls, falls from 153 to 93 at 5 seconds and from 1,443 to 393 at 20 seconds, excluding unchanged query pooling. Measured latency effects remain small and workload-dependent (Appendix B).

Two fused kernels. Static and content plans share range-sparse execution with online softmax, writing disjoint output heads without Q/K/V copies. Static plans are cached per layer and visible layout, using merged row-band intervals for local heads and a recent interval for window heads. Content plans are refreshed; reusing a static plan does not fix attention probabilities. During rolling warm-up, multiple chunks are denoised without cached history. MSA sparsifies these queries while retaining the first chunk dense, reaching density 0.2 on RF despite its native warm-up floor of 0.27.

5 Experiments

5.1 Setup

Models and workload. We evaluate short- and long-horizon causal generation through separate quality and timing experiments. We use Wan2.1-1.3B (Wang et al., 2025) with RF for 5-second generation (a 21-frame rolling window with a 3-frame sink) (Liu et al., 2025) and CF-long for 20-second generation using the official EMA checkpoint (Zhu et al., 2026). RF and CF-long generate 81 and 321 frames, respectively, at 1280×720 resolution and 16 fps. All experiments use BF16 on a single Nvidia H200 GPU. We included sparse attention methods designed for bidirectional diffusion models including STA(Zhang et al., 2025b), SVG(Xi et al., 2025), SVG2(Yang et al., 2025b), XAttention(Xu et al., 2025), and Radial Attention(Li et al., 2025). As for the sparse attention method for autoregressive generation, we adopted Light-Forcing(Lv et al., 2026). These methods are tested in different density tiers including 0.1, 0.2 and 0.3, where density is defined as the number of key tokens attended by query over the total number of key tokens in the video. Some methods such as SVG have inherent implementation limitations so that they cannot reach a specific density.

The comparisons use the same checkpoint and sparse execution backend. We implemented all these methods in SGLang Diffusion (Zheng et al., 2023), where all sparse methods use the same range-sparse Triton implementation, whereas dense attention uses its unmodified backend. We verify selector parity for these runtime reimplementations (Appendix A); LF is applied in a training-free manner to ensure fairness against our MSA.

Table 1: RF 5s generation quality evluation on VBench with seed 42. All six evaluated dimensions are shown on their prescribed prompt subsets.

Method (tier)	Density	Subject	Backgr.	Smooth.	Dynamic	Aesth.	Imaging
Dense	1.000	0.9561	0.9615	0.9794	0.417	0.6396	71.92
STA (0.2)	0.279	0.6449	0.7883	0.9232	1.000	0.3848	60.60
STA (0.3)	0.301	0.8486	0.8708	0.9619	0.514	0.5415	66.84
SVG (0.3)	0.399	0.9538	0.9629	0.9787	0.431	0.6383	71.76
SVG2 (0.3)	0.661	0.9523	0.9576	0.9803	0.444	0.6361	72.05
XAttention (0.2)	0.274	0.6252	0.7768	0.9130	1.000	0.3613	58.35
XAttention (0.3)	0.302	0.9007	0.8972	0.9597	0.847	0.6119	70.08
Radial (0.3)	0.542	0.9572	0.9623	0.9796	0.431	0.6394	71.80
LF (0.2)	0.200	0.8993	0.9301	0.9576	0.420	0.5559	60.68
LF (0.3)	0.300	0.9571	0.9627	0.9792	0.431	0.6360	71.80
MSA (0.2)	0.200	0.9664	0.9675	0.9807	0.444	0.6495	71.26

Table 2: CF-long 20s generation quality evluation on VBench-Long with seed 42. All six evaluated dimensions are shown on their prescribed prompt subsets.

Method (tier)	Density	Subject	Backgr.	Smooth.	Dynamic	Aesth.	Imaging
Dense	1.000	0.8854	0.8939	0.9766	0.861	0.5951	69.57
STA (0.2)	0.200	0.5413	0.7620	0.8882	1.000	0.3190	49.08
STA (0.3)	0.301	0.8167	0.8243	0.9537	1.000	0.4904	65.23
SVG (0.3)	0.299	0.8905	0.8876	0.9677	0.847	0.5989	71.52
SVG2 (0.3)	0.511	0.8716	0.8859	0.9721	0.931	0.5843	69.23
XAttention (0.2)	0.201	0.5423	0.7606	0.8786	1.000	0.3082	53.37
XAttention (0.3)	0.308	0.7679	0.7857	0.9259	1.000	0.5065	62.60
Radial (0.3)	0.354	0.9194	0.9119	0.9772	0.764	0.6174	70.74
LF (0.2)	0.200	0.8897	0.8864	0.9460	1.000	0.5513	63.54
MSA (0.2)	0.200	0.8987	0.9009	0.9783	0.8551	0.5885	67.41

Quality evaluation. The quality experiments evaluate six VBench dimensions on their prescribed subsets: subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality (Huang et al., 2023), among which the imaging quality is reported as the raw MUSIQ score (Ke et al., 2021). RF is evaluated with VBench throughout. For CF-long, MSA is evaluated with VBench-Long (Huang et al., 2025b).

The experiments control the prompts, random seeds, evaluators, and runtime configuration. The quality tables use a common density column for all methods.

System performance. We sampled 10 prompts from the VBench prompt pool and profile the wall time of each method. System performance is measured using warmed requests within a single session. An exclusive H200 hosts a resident model that is warmed once for each method, and the evaluation order is shuffled across methods. DiT and end-to-end (E2E) latency are averaged over 3 repeated runs. E2E latency additionally includes text encoding, VAE decoding, and output handling. Speedups are computed relative to the corresponding dense baseline. Separate probes verify the achieved density and output identity (Appendix A).

5.2 Generation Quality

Tables 1 and 2 show that generation quality depends strongly on both the sparse-attention method and its effective density. Among the methods designed for bidirectional attention, STA and XAttention incur substantial losses in subject consistency, background consistency, aesthetic quality, and imaging quality at their more aggressive tiers, despite retaining high dynamic-degree scores. Their quality generally recovers at the less aggressive tier, although several metrics remain below the dense baseline. SVG, SVG2, and Radial Attention remain closer to dense quality at their achieved densities; SVG and Radial Attention also match or exceed dense attention on selected CF-long metrics. Because these methods attain different effective densities, these results do not constitute a density-matched ranking of bidirectional sparse-attention designs.

Table 3: RF and CF-long system performance evluation at 720p on one H200 over 3 runs.

Setting	Method (target)	Density	DiT (s)	E2E (s)	DiT ×	E2E ×
RF 5s	Dense	1.0	9.675	17.365	1.000	1.000
	STA (0.2)	0.279	5.014	12.724	1.926	1.365
	STA (0.3)	0.301	5.258	12.894	1.836	1.347
	SVG (0.3)	0.399	7.494	15.107	1.288	1.149
	SVG2 (0.3)	0.661	10.748	18.264	0.898	0.951
	XAttention (0.2)	0.274	5.675	13.333	1.701	1.302
	XAttention (0.3)	0.302	6.028	13.605	1.602	1.276
	Radial (0.3)	0.542	8.075	15.701	1.196	1.106
	LF	0.1	4.353	11.848	2.223	1.466
	LF	0.2	5.486	12.999	1.764	1.336
	MSA	0.1	4.190	12.131	**2.309**	**1.431**
	MSA	0.2	5.360	12.943	1.805	1.342
CF-long 20s	Dense	1.0	36.210	63.650	1.000	1.000
	STA (0.2)	0.200	19.116	46.301	1.888	1.370
	STA (0.3)	0.301	23.647	50.865	1.526	1.247
	SVG (0.3)	0.299	27.239	54.396	1.325	1.166
	SVG2 (0.3)	0.511	39.831	66.925	0.906	0.948
	XAttention (0.2)	0.201	22.782	49.857	1.584	1.273
	XAttention (0.3)	0.308	27.794	54.832	1.298	1.157
	Radial (0.3)	0.354	25.851	53.349	1.396	1.189
	LF	0.1	14.502	41.776	2.497	1.524
	LF	0.2	19.086	46.223	1.897	1.377
	MSA	0.1	13.746	41.075	**2.634**	**1.550**
	MSA	0.2	18.445	45.787	1.963	1.390

Against this broader set of baselines, MSA provides a favorable quality profile at density 0.2. As shown in Table 1, MSA improves subject consistency, background consistency, motion smoothness, and aesthetic quality over dense attention for RF, while incurring only a modest reduction in imaging quality. MSA also achieves quality comparable to or better than LF across most reported RF and CF-long dimensions. Nevertheless, agreement with a sparse control does not imply dense-level fidelity: CF-long MSA remains below dense attention in dynamic degree, aesthetic quality, and imaging quality, as given in Table 2. More broadly, these results indicate that preserving temporal consistency does not necessarily preserve per-frame fidelity.

The temporal analysis provides additional context for the aggregate CF-long scores. Beyond the VBench-Long results given in Table 2, the temporal profiles show that CF-long imaging quality declines over time. Without a matched temporal profile for dense attention, however, this decline cannot be attributed solely to sparsity.

5.3 System performance

Table 3 reports both denoising (DiT) and end-to-end (E2E) latency for RF and CF-long. Within each method that is evaluated at multiple densities, the lower-density configuration consistently reduces DiT and E2E latency. The relative improvement is larger for DiT because E2E latency also includes text encoding, VAE decoding, and output handling, none of which is accelerated by sparse attention. The gap between DiT and E2E speedup therefore quantifies the effect of this fixed, non-denoising overhead.

The methods designed for bidirectional attention exhibit substantial variation in runtime. STA provides the largest speedups within this group, reaching $1.926\times$ DiT and $1.365\times$ E2E speedup for RF and $1.888\times$ DiT and $1.370\times$ E2E speedup for CF-long at its more aggressive tier. XAttention follows the same density-dependent trend but provides smaller gains. SVG and Radial Attention achieve more moderate acceleration at their higher effective densities, whereas SVG2 is slower than dense attention for both workloads. These results show that reducing attention density alone does not guarantee acceleration; the selector and sparse execution pattern must also map efficiently to the hardware.

At matched densities, MSA consistently reduces DiT latency relative to LF. For CF-long, this reduction also yields lower E2E latency at both evaluated densities. For RF, however, the E2E benefit is smaller: MSA provides a slight improvement at density 0.2 but is marginally slower than LF at density 0.1 despite its lower DiT latency. As shown in Table 3, the largest gains over dense attention occur for CF-long at density 0.1, where MSA reaches **2.634× denoising** speedup and **1.550× E2E** speedup.

5.4 Density sensitivity and scope

MSA and LF are evaluated at the same target densities, 0.1 and 0.2. Table 3 shows that the 0.1 target lowers both DiT and E2E time relative to 0.2 for each method, on RF and on CF-long. MSA is the faster denoiser at every one of these four settings. On CF-long, its DiT time is 13.746 s against LF's 14.502 s at density 0.1, and 18.445 s against 19.086 s at density 0.2. On RF, the corresponding DiT times are 4.190 s against 4.353 s and 5.360 s against 5.486 s. E2E time keeps this order on CF-long at both densities, 41.075 s against 41.776 s at 0.1 and 45.787 s against 46.223 s at 0.2, and on RF at 0.2, 12.943 s against 12.999 s. At RF density 0.1, MSA E2E time is 12.131 s and LF E2E time is 11.848 s.

We further conducted experiments on Self-Forcing model to evaluate the impact of the density on the 20s generation quality. MSA retain a 6.4% more fraction of oracle attention mass at density 0.2 compared to 0.1, however, speed at density 0.1 is 29% faster than at density 0.2. In terms of the VBench-Long scores, in particular, MSA loses 0.0205 subject consistency at density 0.1 compared to density 0.2.

6 Related work

Causal diffusion, distillation, and rollout training enable streaming generation (Chen et al., 2024; Yin et al., 2024b; Huang et al., 2025a; Zhu et al., 2026); windows and sinks extend its horizon (Liu et al., 2025; Yang et al., 2025a; Cui et al., 2025; ai et al., 2025; Kodaira et al., 2026). MSA targets the shared attention cost, complementing general sparse attention (Zhang et al., 2025a; Jiang et al., 2024; Yuan et al., 2025) with LF-based selection (Lv et al., 2026), routing gates, and all-call accounting. Sinks/windows (Xiao et al., 2024), compression (Samuel et al., 2026), trained sparse memory (Xu et al., 2026), recurrent memory (Li et al., 2026b), and omitted-block compensation (Zhou et al., 2026; Li et al., 2026a) change the attention approximation; exact cache-statistic reuse is complementary.

MSA preserves LF's hierarchical selector where content selection is needed and adds calibrated head families and runtime budget checks. The relevant distinction is between reusing immutable key statistics and reusing query-dependent masks: only the former follows from cache immutability. Likewise, exact pair accounting describes attention support, not a complete latency model; selection overhead, memory layout and non-attention work remain part of the measured system.

7 Conclusion

In conlusion, this work characterizes attention in causal video diffusion and shows that spatial locality persists for a subset of heads, whereas recent-window support deteriorates as history grows; moreover, immutable cached keys do not yield stationary attention, and reliable budgets must account for every executed pass. Motivated by these findings, we propose MSA, a training-free framework that integrates calibrated head routing, runtime gates, reusable history statistics, and a budget spanning all attention calls. For 20-second Causal Forcing generation at 720p, MSA achieves up to 2.634× denoising and 1.550× end-to-end speedup over dense attention while maintaining quality comparable to LightForcing. Nevertheless, these gains arise from content selection and efficient execution rather than static routing, and MSA remains below dense attention on several per-frame metrics for CF-long.

AI USE STATEMENT

Generative AI tools assisted with implementation, experiment protocol development and execution, analysis, figures, and manuscript drafting and revision. The authors are responsible for verifying the methods, results, citations, and final content, including AI-assisted artifacts.

REPRODUCIBILITY STATEMENT

Section 5.1 and Appendices A–B summarize the protocol and essential validation. The separate extended appendix preserves detailed diagnostics, controls and qualitative comparisons. The code will be open-sourced at a later stage.

ETHICS STATEMENT

Our experiments use publicly available models and datasets in accordance with their respective terms of use. We do not collect new personal data or conduct experiments involving human participants. The proposed method does not alter the safety mechanisms of the underlying models, although it does not independently prevent their misuse. Deployments should therefore retain appropriate safeguards, including content moderation, provenance or watermarking mechanisms, access controls, and compliance with applicable data and intellectual-property regulations. Our evaluation focuses primarily on generation quality and computational efficiency and should not be interpreted as a comprehensive assessment of model fairness, safety, or societal impact.

REFERENCES

Sand. ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. Magi-1: Autoregressive video generation at scale, 2025. URL https://arxiv.org/abs/2505.13211.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=yDo1ynArjj.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuan-Hao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *ArXiv*, abs/2510.02283, 2025. URL https://api.semanticscholar.org/CorpusID:281724666.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=H4DqfPSibmx.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=mSiN7i0BYH.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Minference 1.0: accelerating pre-filling for long-context llms via dynamic sparse attention. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer, 2021. URL https://arxiv.org/abs/2108.05997.

Akio Kodaira, Tingbo Hou, Ji Hou, Markos Georgopoulos, Felix Juefei-Xu, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation, 2026. URL https://arxiv.org/abs/2507.03745.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Haopeng Li, Yitong Li, Junsong Chen, Tian Ye, Haozhe Liu, Jincheng Yu, Duomin Wang, Ruihua Zhang, Zeke Xie, Enze Xie, and Song Han. Sol-attn: Accelerating video generation inference via on-the-fly attention sparsification, 2026a. URL https://arxiv.org/abs/2607.24027.

Kunyang Li, Mubarak Shah, and Yuzhang Shang. Attend locally, remember linearly: Linear attention as cross-frame memory for autoregressive video diffusion, 2026b. URL https://arxiv.org/abs/2605.16579.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, and Song Han. Radial attention: $\mathcal{O}(n\log n)$ sparse attention with energy decay for long video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=hYovE4nHTt.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Kunhao Liu, Wen-Bo Hu, Jiale Xu, Shan Ying, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *ArXiv*, abs/2509.25161, 2025. URL https://api.semanticscholar.org/CorpusID:281676207.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Chengtao Lv, Yumeng Shi, Yushi Huang, Ruihao Gong, Shen Ren, and Wenya Wang. Light forcing: Accelerating autoregressive video diffusion via sparse attention, 2026. URL https://arxiv.org/abs/2602.04789.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, Los Alamitos, CA, USA, October 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00387.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

Dvir Samuel, Issar Tzachor, Matan Levy, Michael Green, Gal Chechik, and Rami Ben-Ari. FAST-AR: Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=UxgMJP0JKY.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Philippe Tillet, H. T. Kung, and David Cox. Triton: an intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, MAPL 2019, pp. 10–19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367196. doi: 10.1145/3315508.3329973. URL https://doi.org/10.1145/3315508.3329973.

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, Xiaofeng Meng, Ningying Zhang, Pandeng Li, Ping Wu, Ruihang Chu, Rui Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tian-Yi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wen-Chao Zhou, Wente Wang, Wen Shen, Wenyuan Yu, Xianzhong Shi, Xiaomin Huang, Xin Xu, Yan Kou, Yan-Mei Lv, Yifei Li, Yi-Jing 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*, abs/2503.20314, 2025. URL https://api.semanticscholar.org/CorpusID:277321639.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, and Song Han. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=u8CA3qIS0V.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Boxun Xu, Yuming Du, Zichang Liu, Siyu Yang, Ziyang Jiang, Siqi Yan, Rajasi Saha, Albert Pumarola, Wenchen Wang, and Peng Li. Sparse forcing: Native trainable sparse attention for real-time autoregressive diffusion video generation, 2026. URL https://arxiv.org/abs/2604.21221.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. XAttention: Block sparse attention with antidiagonal scoring. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=KG6aBfGi6e.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *ArXiv*, abs/2509.22622, 2025a. URL https://api.semanticscholar.org/CorpusID:281658815.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, Jianfei Chen, Song Han, Kurt Keutzer, and Ion Stoica. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=WPU17d1l7R.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=LQzN6TRFg9.

Tianwei Yin, Michael Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2023. URL https://api.semanticscholar.org/CorpusID:265506768.

Tianwei Yin, Michael Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. *ArXiv*, abs/2405.14867, 2024a. URL https://api.semanticscholar.org/CorpusID:269982921.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, 2024b. URL https://api.semanticscholar.org/CorpusID:274610175.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23097, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1126. URL https://aclanthology.org/2025.acl-long.1126/.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. SpargeAttention: Accurate and Training-free Sparse Attention Accelerating Any Model Inference. In *International Conference on Machine Learning*, 2025a. URL https://arxiv.org/abs/2502.18137.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zheng Liu, and Hao Zhang. Fast video generation with sliding tile attention. *ArXiv*, abs/2502.04507, 2025b. URL https://api.semanticscholar.org/CorpusID:276236090.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chu-Yue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christoforos E. Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark W. Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. *Advances in Neural Information Processing Systems 37*, 2023. URL https://api.semanticscholar.org/CorpusID:266174771.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

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, 2026. URL https://arxiv.org/abs/2603.08982.

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. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02214.

A ESSENTIAL EVALUATION PROTOCOL

Our evaluation uses seed 42 and dimension-specific prompts: 72 prompts for subject consistency, motion smoothness and dynamic degree; 86 for background consistency; and 93 for aesthetics and imaging. The pinned VBench evaluator is `fd18b3d0`. CF-long VBench-Long splits 321 frames into eleven 2-second clips at 16 fps, with an overlapping padded tail and scene splitting disabled. Subject/background consistency includes the official DINOv2/DreamSim across-clip term. An adapter repairs underscore-based filename grouping and checks feature coverage. In our evaluation, baseline scores computed with whole-video VBench are distinguished from VBench-Long scores.

Matched validation. Calibration uses one development prompt at seed 42; confirmation uses two other prompts, each at seeds 42 and 123. These four prompt/seed pairs are excluded from calibration but were also used in our evaluation. The LF/MSA rerun contains 96 timed requests and 24 density probes across eight model/density settings, including SF controls. All probes meet the ± 0.002 density tolerance and match uninstrumented outputs; all 96 videos match their evaluated counterparts byte for byte. Timing excludes probes and warm-ups and uses same-session ratios of medians. The 825 BF16 CF-long transformer tensors match the official EMA checkpoint; its four-step schedule and flow shift are checked.

Native-baseline timings. Our RF/CF-long evaluation contains 80 timed requests and 20 output-identical density probes, with one excluded warm-up per configuration. Calibration on the same development prompt targets ± 0.005 density error, retaining discrete masks and the minimum densities imposed by each implementation when necessary. The original baseline quality-evaluation settings are unavailable; no quality scores transfer to these new configurations. Checkpoint hashes and all 175 runtime parity tests pass; selector parity does not establish upstream end-to-end equivalence.

Quality screens. Mean candidate-minus-LF differences must be at least -0.01 for subject/background consistency and smoothness, -0.02 for aesthetics, and -1 MUSIQ point; every pair must preserve LF's dynamic score. CF-long additionally evaluates frames 240–320 with ordinary VBench. All twelve full-video comparisons and both CF-long tail comparisons pass. These engineering thresholds do not establish perceptual noninferiority or dense-level quality.

B ESSENTIAL DIAGNOSTICS AND LIMITATIONS

Probes and accounting. Dense SF probes use three prompts at 5/20 seconds, seed 42, all heads/passes, and 24 stratified queries per snapshot. The six videos yield 43,200 correlated head–snapshot observations. Probes leave generation unchanged; all three checked 5-second outputs are bit-identical. A separate two-prompt FP32 structural probe yields 294 snapshots and 84,672 query/head observations. At 480p, temporal-offset correlations between chunks 3 and 13 are 0.87–0.98; this does not validate 720p spatial stationarity. Full-history first-chunk, current-chunk and clean-pass density floors are $2/[C(C+1)]$, $2/(C+1)$ and $1/5$, respectively; overlapping supports cannot be added. Rolling geometry is counted directly: RF has 22 calls per layer at 5s and CF-long 60 at 20s.

Controls and failures. Routing replay finds static execution only for SF/0.2 (1,125 head calls); RF/CF-long gains therefore do not establish static-routing benefits. Five-pair LF reuse comparisons, after two warm-ups per variant/density, give denoising ratios of 0.992–1.017 and E2E ratios of 0.996–1.005, with all 40 video pairs identical. These small effects do not isolate the contributions to the 1.078–1.146$\times$ LF denoising speedup in our evaluation, which combined fused pooling with history reuse. On SF at 20 seconds (the development prompt, seed 42), first-pass mask reuse loses 0.0205 subject consistency versus LF at density 0.1. Compact 4×16 spatial key tiles capture more mass per key but run at 0.70$\times$/0.67$\times$ LF speed at densities 0.1/0.2, illustrating the cost of the read layout. SF's late-motion failure remains unresolved. Missing original baseline videos and per-call density records and matched CF-long VBench-Long scores in our evaluation limit attribution and comparison.

Supporting artifacts. The separate extended appendix retains full tables, temporal profiles, ablations and five RF frame-comparison grids. Bundled records and scripts preserve configurations, per-video scores, timing samples, hashes and arithmetic checks. The paper reports all main RF/CF-long quality and performance results in Tables 1–3.