

A Derivations and Proofs

A.1 Complexity

In this section, we provide the detailed complexity derivation in Eq (5), which scales as $\mathcal{O}(n \log n)$. Since the computational cost of masked attention is proportional to the number of zeros in the attention mask M , we only need to derive an $\mathcal{O}(n \log n)$ upper bound for the latter.

Central band&attention sink. Firstly, recall from Fig 4(a) that we apply dense attention on these frame-to-frame attention blocks within the central band and attention sink. The attention sink refers to the pattern that every token attends to all tokens in the first frame. Using the same notation as in Sec 4, where n is the total number of tokens, s is the number of tokens per frame, and f is the number of frames (so $n = fs$), we define the attention mask for this region as $M^{(1)} \in \{-\infty, 0\}^{f \times f \times s \times s}$:

$$M_{i,j,k,l}^{(1)} = \begin{cases} 0, & \text{if } |i-j| \leq 1 \text{ or } j = 0 \\ -\infty. & \text{otherwise} \end{cases} \quad (8)$$

Here, $M_{i,j,k,l}^{(1)}$ indicates whether the k -th token in frame i is allowed to attend to the l -th token in frame j , with 0 denoting allowed attention and $-\infty$ indicating disallowed attention. Since the attention sink spans f blocks and the central band includes at most $3f$ blocks, the total number of nonzero entries in this region is bounded by:

$$\#\text{zeros in } M^{(1)} \leq 4 \cdot f \cdot s^2 = 4s^2f. \quad (9)$$

Bands with diagonal width ≥ 1 . The second part is those bands with diagonal width ≥ 1 , except the central band. The mask for this region can be defined as $M^{(2)} \in \{-\infty, 0\}^{f \times f \times s \times s}$:

$$M_{i,j,k,l}^{(2)} = \begin{cases} 0, & \text{if } 2^{\lfloor \log_2 \max(|i-j|, 1) \rfloor} \leq s \text{ and } |k-l| + 1 \leq \frac{s}{2^{\lfloor \log_2 \max(|i-j|, 1) \rfloor}} \\ -\infty. & \text{otherwise} \end{cases} \quad (10)$$

Thus, since there are at most $\lfloor \log_2 s \rfloor$ bands in this region, the number of zeros in these bands is bounded by:

$$\#\text{zeros in } M^{(2)} \leq \sum_{r=1}^{\lfloor \log_2 s \rfloor} \underbrace{2^{r+1}sn}_{\text{area bound for band } \pm r} \cdot \underbrace{2/2^r}_{\text{compute density bound of band } \pm r} \quad (11)$$

$$\leq \sum_{r=1}^{\lfloor \log_2 s \rfloor} \frac{2^{r+2}s^2f}{2^r} \quad (12)$$

$$= 4s^2f \cdot \lfloor \log_2 s \rfloor. \quad (13)$$

Bands with diagonal width < 1 . The last part is those bands with $\frac{s}{2^{\lfloor \log_2 \max(|i-j|, 1) \rfloor}} < 1$, where we reduce the frequency of diagonals. The mask for this region $M^{(3)} \in \{-\infty, 0\}^{f \times f \times s \times s}$ is given by:

$$M_{i,j,k,l}^{(3)} = \begin{cases} 0, & \text{if } |i-j| \bmod \lceil \frac{2^{\lfloor \log_2 \max(|i-j|, 1) \rfloor}}{s} \rceil = 0 \text{ and } k = l \\ -\infty. & \text{otherwise} \end{cases} \quad (14)$$

Since there are at most $(\lceil \log_2 f \rceil - 1) - (\lfloor \log_2 s \rfloor + 1)$ bands satisfying this condition, we have the number of zeros in these bands bounded by:

$$\#\text{zeros in } M^{(3)} \leq \sum_{r=\lfloor \log_2 s \rfloor + 1}^{\lceil \log_2 f \rceil - 1} \underbrace{2^{\lfloor \log_2 s \rfloor + 1}}_{\text{number of diagonals}} \cdot \underbrace{n}_{\text{area bound of each diagonal}} \quad (15)$$

$$\leq (\lceil \log_2 f \rceil - \lfloor \log_2 s \rfloor) 4s^2f. \quad (16)$$

Combining Eq 9, Eq 13, and Eq 16 together, we have the aggregate upper bound of the number of zeros in Aura Attention’s mask:

$$\# \text{ of zeros in } \mathbf{M} \leq 4s^2 f \cdot \lfloor \log_2 f \rfloor \leq 4s \cdot n(\log_2 n - \log_2 s), \quad (17)$$

which scales $\mathcal{O}(n \log n)$ with the number of frames f for long video generation.

A.2 Error Bound

The design of Aura Attention is inspired by the spatial-temporal structure in video. In this section, we formulate this intuition by theoretically bounding the asymptotic approximation error of Aura Attention with respect to the decay speed of the attention value in the spatial and temporal dimensions.

We focus on bounding the approximation error of a single row of the attention matrix. We fix a reference query token at position k_0 of frame i_0 , and write the unnormalized row of the attention matrix as

$$a_{j,l} = \exp(\mathbf{Q}_{i_0 s + k_0} \mathbf{K}_{j s + l}^\top).$$

where $\mathbf{Q}_{i_0 s + k_0}$ refers to the query vector at position k_0 in frame i_0 , $\mathbf{K}_{j s + l}$ refers to the key vector at position l in frame j , and s refers to the number of tokens per frame.

Assumptions

(A1) **Relative exponential decay.** To capture the intuition that the closer frames have a stronger correlation and each token typically attends to tokens in other frames at similar spatial positions, we assume there exist $C_{\text{rel}} > 0$ and $(\alpha, \beta) > 0$ such that

$$0 \leq a_{j,l} \leq C_{\text{rel}} e^{-\alpha|j-i_0|-\beta|l-k_0|} a_0, \quad a_0 := a_{i_0, k_0} > 0.$$

where α characterizes the temporal decay rate and β characterizes the spatial decay rate.

(A2) **Infinite temporal grid & finite spatial grid.** To conduct asymptotic analysis, we let $j \in \mathbb{Z}$ (temporal) but keep $l \in \{1, \dots, s\}$ (spatial). Extending j to \mathbb{Z} only enlarges the sums we bound.

Notation

$$Z := \sum_{j \in \mathbb{Z}} \sum_{l=1}^s a_{j,l}, \quad Z_{\text{keep}} := \sum_{(j,l) : \mathbf{M}_{i_0, j, k_0, l} = 0} a_{j,l}, \quad Z_{\text{out}} := Z - Z_{\text{keep}}.$$

Exact and masked softmax rows: $p_{j,l} = a_{j,l}/Z$, $\tilde{p}_{j,l} = a_{j,l} \mathbf{1}_{\{\mathbf{M}=0\}}/Z_{\text{keep}}$. The total variation error can be calculated as follows by standard algebraic argument,

$$\|\tilde{p} - p\|_1 = 2 \frac{Z_{\text{out}}}{Z}. \quad (1)$$

Because a_0 itself is in the sum, $Z \geq a_0$. Hence

$$\frac{Z_{\text{out}}}{Z} \leq \frac{Z_{\text{out}}}{a_0}. \quad (2)$$

Mask geometry For a temporal offset $\Delta t := |j - i_0| \geq 0$, define the bandwidth

$$w(\Delta t) := \frac{s}{2^{\lfloor \log_2 \max(\Delta t, 1) \rfloor}} \in \{1, 2, 4, \dots, s\}.$$

The mask keeps a spatial index l iff $|l - k_0| \leq w(\Delta t)$ and the frame is one of the sub-sampled frames; otherwise $\mathbf{M}_{i_0, j, k_0, l} = -\infty$.

Two kinds of approximation errors, therefore, appear:

(i) Spatial tails inside kept frames

For each Δt , the discarded spatial part satisfies

$$\sum_{d > w(\Delta t)} e^{-\beta d} \leq \frac{e^{-\beta(w(\Delta t)+1)}}{1 - e^{-\beta}}.$$

714 Because $w(\Delta t) \geq \frac{s}{2}$ when $\Delta t \leq s$,

$$T_1 := 2C_{\text{rel}}a_0 \sum_{\Delta t \geq 0} e^{-\alpha\Delta t} \sum_{d > w(\Delta t)} e^{-\beta d} \leq \frac{4C_{\text{rel}}a_0}{(1 - e^{-\alpha})(1 - e^{-\beta})} e^{-\beta(\frac{s}{2}+1)}. \quad (3)$$

715 (ii) Frames skipped by the subsampling rule

716 For $\Delta t > s$, only every $K(\Delta t) = \lceil 2^{\lfloor \log_2 \Delta t \rfloor} / s \rceil$ frame is kept; the remainder contributes

$$T_2 := 2C_{\text{rel}}a_0 \frac{1 + e^{-\beta}}{1 - e^{-\beta}} \sum_{\Delta t > s} e^{-\alpha\Delta t} \leq \frac{2C_{\text{rel}}a_0 (1 + e^{-\beta})}{(1 - e^{-\beta})(1 - e^{-\alpha})} e^{-\alpha(s+1)}. \quad (4)$$

717 **Total variation error** Combine all equations above:

$$\|\tilde{p} - p\|_1 \leq C_{\text{rel}} \left[\frac{8e^{-\beta(\frac{s}{2}+1)}}{(1 - e^{-\alpha})(1 - e^{-\beta})} + 4 \frac{1 + e^{-\beta}}{1 - e^{-\beta}} \frac{e^{-\alpha(s+1)}}{1 - e^{-\alpha}} \right] = O(C_{\text{rel}} e^{-\min\{\beta/2, \alpha\}s}).$$

718 This characterizes how the decay rates affect the approximation error.

719 B Additional Implementation Details

720 For text-to-video generation at the default length, we follow SVG [8] by retaining full attention
721 during the first 12 denoising timesteps for both HunyuanVideo [1] and Wan2.1 [7]. Additionally, for
722 Wan2.1, we keep full attention in the first DiT block to maintain generation quality.

723 For longer-video generation, we fine-tune HunyuanVideo [1] and Mochi 1 [12] at a global batch
724 size of 1 with sequence parallelism, and train Wan 2.1 [7] with a global batch size of 8. All tuning
725 experiments are conducted on 8 H100 GPUs. During training, we keep the first two DiT blocks with
726 full attention. Since there are 60, 48, and 40 blocks for HunyuanVideo, Mochi 1, and Wan2.1, this
727 only incurs negligible overhead. We train HunyuanVideo for 2× and 4× length video generation for
728 2400 and 1200 steps, respectively. We train Mochi 1 for 5000 steps for both 2× and 4× length video
729 generation. We train Wan2.1 for 2500 steps for 2× length video generation. The LoRA rank is 128
730 for all training tasks.

731 C Visualization of the generated videos

732 In this section we compare Aura Attention against various baselines in video quality, and list our
733 speedup in both training and inference. For the complete video comparisons, see our supplementary
734 materials and [this link](#).

735 C.1 Default Video Length

736 We provide a visual comparison between the original dense attention, STA [45], and our Aura
737 Attention on HunyuanVideo [1] and Wan2.1 [7]. We conduct experiments under 768p, 117 frames
738 settings for HunyuanVideo, and 768p, 69 frames settings for Wan2.1. As shown in Fig A and Fig B,
739 Aura Attention has higher PSNR compared to STA [45], effectively maintaining the high fidelity of
740 the original videos.

741 C.2 Extended Video Length

742 We provide a visual comparison between the aforementioned baselines and Aura Attention on
743 HunyuanVideo [1], Mochi 1 [12], and Wan2.1 [7]. We conduct experiments under the default
744 resolution settings, which are 720p for HunyuanVideo and Wan2.1, and 480p for Mochi 1. Moreover,
745 we generate videos at 4× longer length for HunyuanVideo (21 seconds, 509 frames) and Mochi
746 1(22 seconds, 667 frames), and 2× longer length for Wan2.1 (10 seconds, 161 frames). We use

747 Vision Reward [69] to evaluate the generated videos. Fig C, Fig D, and Fig E demonstrate that
748 Aura Attention achieves the highest average Vision Reward score compared to the baselines, well
749 preserving the video quality even at long-video settings.

750 D Broader Impacts

751 Aura Attention significantly reduces computational costs for video diffusion models, enabling
752 longer video generation with minimal fine-tuning efforts while maintaining quality. This paves
753 the way for high-quality video creation tools for education and creative arts. Since Aura Attention
754 accelerates self-attention to $\mathcal{O}(n \log n)$ complexity, it can accelerate video diffusion models and
755 decrease energy consumption, leading to greener AI applications. This also helps the popularization
756 of generative models. However, malicious users can misuse our method to create deepfakes and
757 spread misinformation. The technology may also exacerbate the digital divide between those with
758 and without access to the minimal necessary computational resources. To address these concerns,
759 we advocate for responsible deployment, adherence to ethical standards, and the development of
760 effective detection methods. We encourage the research community to continue advancing both
761 efficient generation techniques and safeguards to ensure these powerful tools benefit society while
762 minimizing potential harms. We will explicitly specify the usage permission of our code and models
763 with proper licenses.

764 E License

765 Here, we show all the licenses for our used assets. Wan 2.1 [7], Mochi 1 [12], Diffusers, and STA [45]
766 are under Apache-2.0 license. The license of HunyuanVideo [1] is here. SVG [8] and OpenVid-1M
767 do not have an explicit license.

Prompt: A shark is swimming in the ocean, featuring a steady and smooth perspective. Realistic, Natural lighting, Mysterious.

Original HunyuanVideo
TFLOPs: 612 Latency: 1649s
Speedup: 1.0×

STA(FA3)
PSNR: 29.8
TFLOPs: 331 Latency: 719s
Speedup: 2.3×

Aura Attention (Ours)
PSNR: **31.2**
TFLOPs: **339** Latency: **876s**
Speedup: **1.9×**



Prompt: Martial artists exchanging fluid, powerful strikes in a serene, ancient temple courtyard, dust clouds rising in slow motion from every footfall and impact.

Original HunyuanVideo
TFLOPs: 612 Latency: 1649s
Speedup: 1.0×

STA(FA3)
PSNR: 23.6
TFLOPs: 331 Latency: 719s
Speedup: 2.3×

Aura Attention (Ours)
PSNR: **26.1**
TFLOPs: **339** Latency: **876s**
Speedup: **1.9×**



Prompt: A couple in formal evening wear walks home and gets caught in a heavy downpour with umbrellas, surrealism style. Night lighting, Mysterious.

Original HunyuanVideo
TFLOPs: 612 Latency: 1649s
Speedup: 1.0×

STA(FA3)
PSNR: 24.2
TFLOPs: 331 Latency: 719s
Speedup: 2.3×

Aura Attention (Ours)
PSNR: **25.5**
TFLOPs: **339** Latency: **876s**
Speedup: **1.9×**



Prompt: A dancer spinning with explosive energy under a sharp spotlight, loose fabric and fine dust swirling around her in a whirlwind of motion and emotion.

Original HunyuanVideo
TFLOPs: 612 Latency: 1649s
Speedup: 1.0×

STA(FA3)
PSNR: 27.2
TFLOPs: 331 Latency: 719s
Speedup: 2.3×

Aura Attention (Ours)
PSNR: **28.8**
TFLOPs: **339** Latency: **876s**
Speedup: **1.9×**

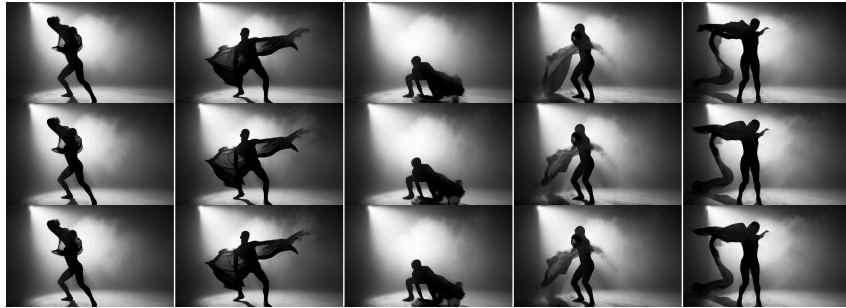


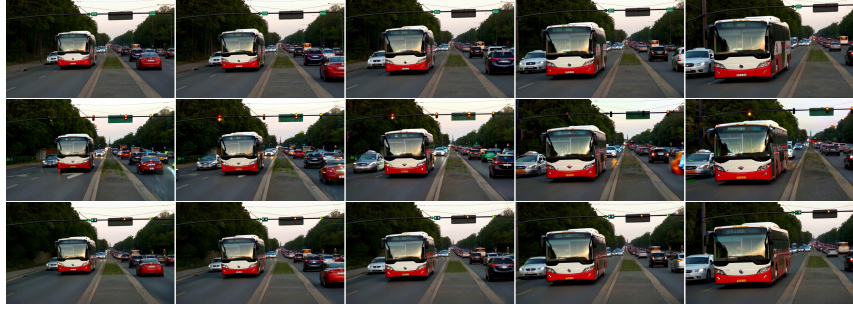
Figure A: Comparison of Dense Attention and Aura Attention on HunyuanVideo Text-to-Video generation at the default length (5 seconds, 117 frames, 768p).

Prompt: A bus is stuck in traffic during rush hour. Realistic, Natural lighting, Tense.

Original Wan2.1 14B
TFLOPs: 560 Latency: 1630s
Speedup: 1.0×

STA(FA3)
PSNR: 19.4
TFLOPs: 322 Latency: 812s
Speedup: 2.0×

Aura Attention (Ours)
PSNR: **22.2**
TFLOPs: **323** Latency: **917s**
Speedup: **1.8×**

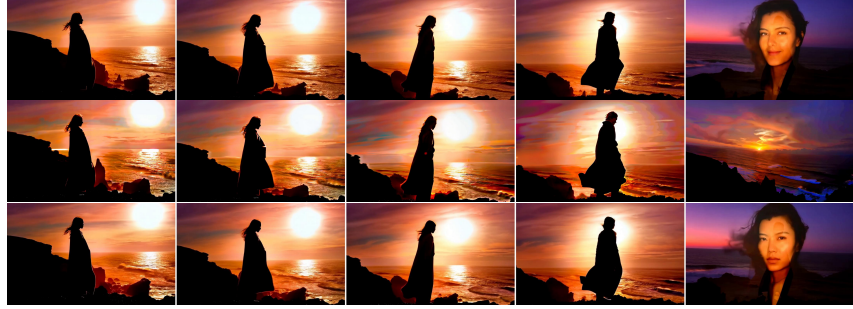


Prompt: A solitary figure stands on a windswept cliff, their silhouette framed by a dramatic sunset, wearing a long, flowing coat that billows in the breeze.

Original Wan2.1 14B
TFLOPs: 560 Latency: 1630s
Speedup: 1.0×

STA(FA3)
PSNR: 21.8
TFLOPs: 322 Latency: 812s
Speedup: 2.0×

Aura Attention (Ours)
PSNR: **23.6**
TFLOPs: **323** Latency: **917s**
Speedup: **1.8×**



Prompt: A breathtaking coastal beach in spring, with gentle waves lapping against the golden sand, is depicted in the vibrant, swirling brushstrokes of Van Gogh.

Original Wan2.1 14B
TFLOPs: 560 Latency: 1630s
Speedup: 1.0×

STA(FA3)
PSNR: 19.6
TFLOPs: 322 Latency: 812s
Speedup: 2.0×

Aura Attention (Ours)
PSNR: **22.1**
TFLOPs: **323** Latency: **917s**
Speedup: **1.8×**



Prompt: A close-up shot captures a cluster of plump, dewy grapes, glistening under soft studio lighting as they slowly rotate on a sleek, reflective table.

Original Wan2.1 14B
TFLOPs: 560 Latency: 1630s
Speedup: 1.0×

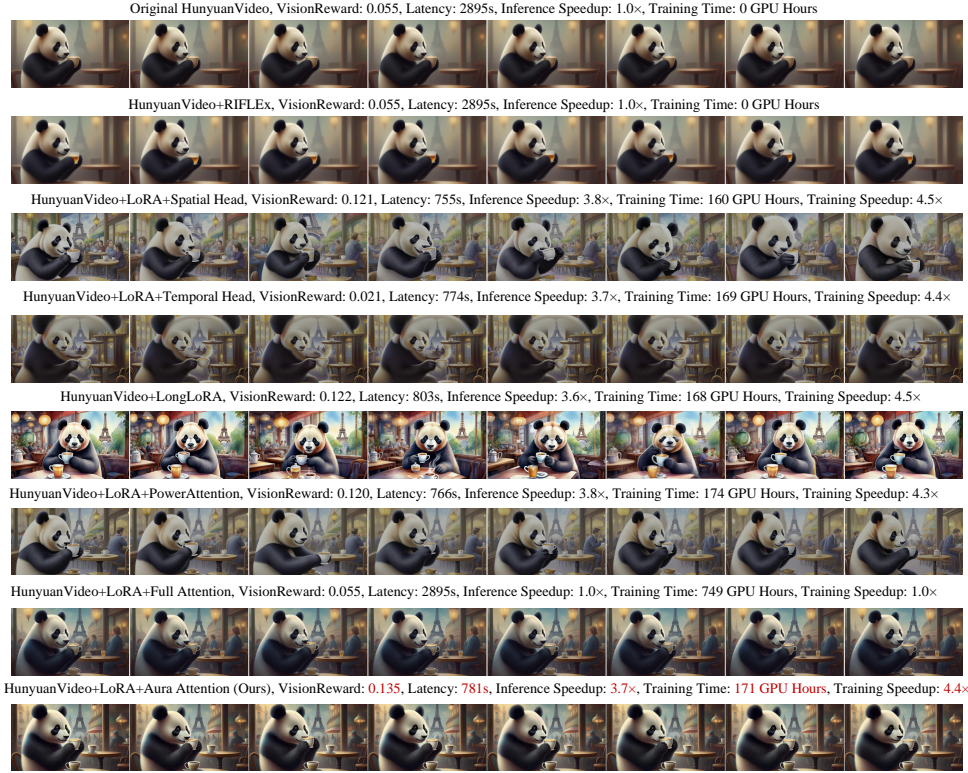
STA(FA3)
PSNR: 19.6
TFLOPs: 322 Latency: 812s
Speedup: 2.0×

Aura Attention (Ours)
PSNR: **25.0**
TFLOPs: **323** Latency: **917s**
Speedup: **1.8×**



Figure B: Comparison of Dense Attention and Aura Attention on Wan2.1 14B Text-to-Video generation at the default length (4 seconds, 69 frames, 768p).

Prompt: A gentle and curious panda, with soft, fluffy fur and large round eyes, is depicted in a charming watercolor painting. The panda sits at a cozy table in a quaint cafe located in the heart of Paris.



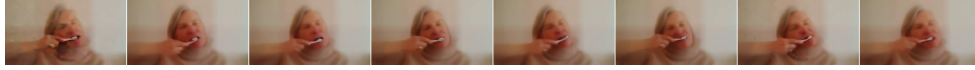
Prompt: A picturesque coastal beach in the enchanting spring season, where gentle waves lap rhythmically against the soft sandy shore. The scene captures the beauty of nature during this vibrant time of year.



Figure C: Comparison of all baselines and Aura Attention at 4 \times default length (21 seconds, 509 frames) Text-to-Video video generation from HunyuanVideo. Aura Attention achieves the best Vision Reward score with good visual quality and consistency. In contrast, Original HunyuanVideo and RIFLEx generate blurred videos with poor visual quality. Temporal Head generates distorting figures. Spatial Head, Long LoRA, and PowerAttention generate temporally inconsistent video backgrounds. Full Attention generates less dynamic videos.

Prompt: A simple yet detailed scene set in a cozy living room, a person, likely a middle-aged woman with gentle features and neatly styled silver hair, sits comfortably on a plush armchair.

Original Mochi 1, VisionReward: -0.041, Latency: 992s, Inference Speedup: 1.0×, Training Time: 0 GPU Hours



Mochi 1+LoRA+Spatial Head, VisionReward: 0.143, Latency: 382s, Inference Speedup: 2.6×, Training Time: 139 GPU Hours, Training Speedup: 2.8×



Mochi 1+LoRA+Temporal Head, VisionReward: -0.024, Latency: 393s, Inference Speedup: 2.5×, Training Time: 141 GPU Hours, Training Speedup: 2.8×



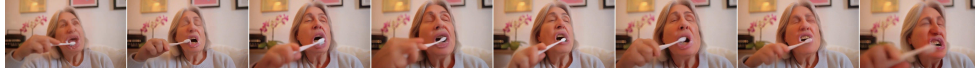
Mochi 1+LongLoRA, VisionReward: 0.037, Latency: 426s, Inference Speedup: 2.3×, Training Time: 152 GPU Hours, Training Speedup: 2.6×



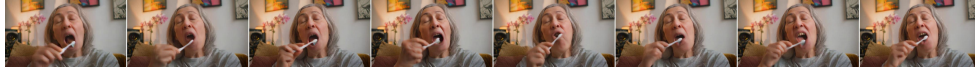
Mochi 1+LoRA+PowerAttention, VisionReward: 0.090, Latency: 381s, Inference Speedup: 2.6×, Training Time: 138 GPU Hours, Training Speedup: 2.8×



Mochi 1+LoRA+Full Attention, VisionReward: 0.130, Latency: 992s, Inference Speedup: 1.0×, Training Time: 394 GPU Hours, Training Speedup: 1.0×



Mochi 1+LoRA+Aura Attention (Ours), VisionReward: 0.182, Latency: 386s, Inference Speedup: 2.6×, Training Time: 139 GPU Hours, Training Speedup: 2.8×



Prompt: A breathtaking coastal beach in the vibrant spring season, waves gently lap at the golden sandy shores. In black and white, the scene captures the serene beauty of nature. A lone figure in a stylish beige windbreaker strolls along the edge of the water, casting occasional glances towards the horizon. Seagulls fly overhead, their silhouettes stark against the clear blue sky. Soft dunes rise behind them, blending seamlessly into the lush greenery of nearby trees.

Original Mochi 1, VisionReward: -0.024, Latency: 992s, Inference Speedup: 1.0×, Training Time: 0 GPU Hours



Mochi 1+LoRA+Spatial Head, VisionReward: 0.046, Latency: 382s, Inference Speedup: 2.6×, Training Time: 139 GPU Hours, Training Speedup: 2.8×



Mochi 1+LoRA+Temporal Head, VisionReward: 0.008, Latency: 393s, Inference Speedup: 2.5×, Training Time: 141 GPU Hours, Training Speedup: 2.8×



Mochi 1+LongLoRA, VisionReward: 0.006, Latency: 426s, Inference Speedup: 2.3×, Training Time: 152 GPU Hours, Training Speedup: 2.6×



Mochi 1+LoRA+PowerAttention, VisionReward: 0.097, Latency: 381s, Inference Speedup: 2.6×, Training Time: 138 GPU Hours, Training Speedup: 2.8×



Mochi 1+LoRA+Full Attention, VisionReward: 0.056, Latency: 992s, Inference Speedup: 1.0×, Training Time: 394 GPU Hours, Training Speedup: 1.0×

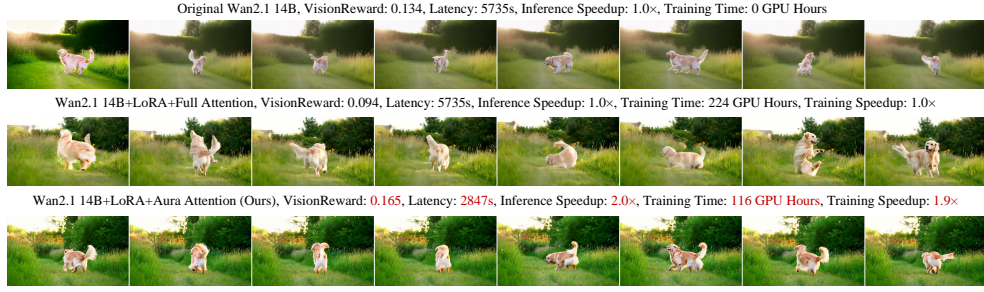


Mochi 1+LoRA+Aura Attention (Ours), VisionReward: 0.097, Latency: 386s, Inference Speedup: 2.6×, Training Time: 139 GPU Hours, Training Speedup: 2.8×

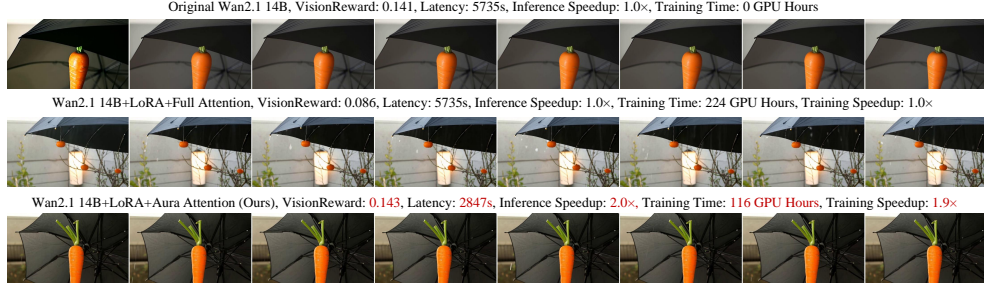


Figure D: Comparison of all baselines and Aura Attention at 4× default length (22 seconds, 667 frames) Text-to-Video video generation from Mochi 1. Aura Attention achieves the highest Vision Reward score because it has excellent visual quality and consistency. In contrast, Original Mochi 1 generates blurred videos with poor visual quality. Spatial Head, Temporal Head, Long LoRA, PowerAttention, and Full Attention generate videos with either inconsistent backgrounds or inconsistent figures.

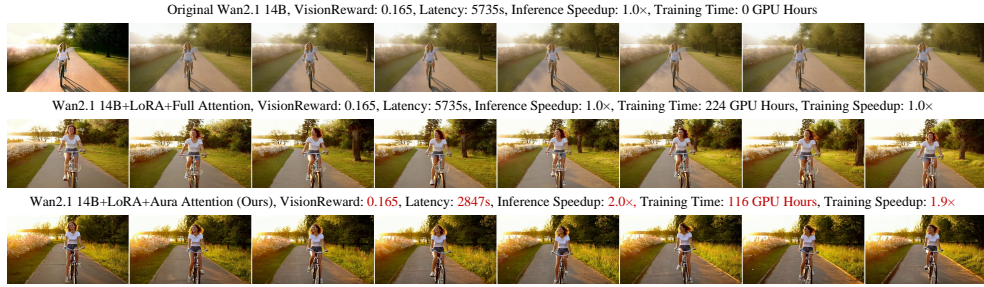
Prompt: A medium-sized golden retriever dog is sitting peacefully in a sunlit backyard, its tail wagging gently. Suddenly, it springs up and starts running in circles, tail wagging excitedly and ears flapping. The grass is lush and green, with wildflowers scattered around.



Prompt: A bright orange carrot and a black umbrella. Realistic, Bright lighting, Casual.



Prompt: A spirited individual rides a vintage bicycle along a sunlit, tree-lined path, wearing a casual outfit of a white t-shirt, denim shorts, and sneakers. The scene captures the golden hour, with sunlight filtering through the leaves, casting dappled shadows on the ground. The rider's hair flows freely in the breeze, and a joyful smile lights up their face.



Prompt: A solitary figure stands on a windswept cliff, their silhouette framed by a dramatic sunset, wearing a long, flowing coat that billows in the breeze. The sky is ablaze with hues of orange, pink, and purple, casting a warm glow on the scene.

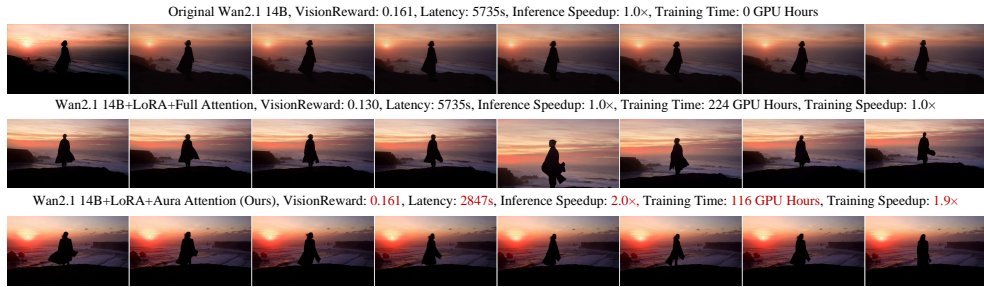


Figure E: Comparison of all baselines and Aura Attention at 2× default length (10 seconds, 161 frames) Text-to-Video video generation from Wan2.1 14B. Aura Attention achieves the highest Vision Reward score because it has excellent visual quality and consistency. In contrast, Original Wan2.1 14B generates blurred videos with poor visual quality. Full Attention generates videos with inconsistent figures.