A. Implementation Details

A.1. Spatial Scan Path in MM-DiM Block.

As shown in Figure 6, we follow [22] to apply eight type of scan paths along the spatial dimension for Mamba, which include:

(a) top-left to the bottom-right, following a "downward first, then rightward" direction.
(b) top-left to the bottom-right, following a "downward right, then downward" direction.
(c) bottom-left to the top-right, following a "upward first, then rightward" direction.
(d) bottom-left to the top-right, following a "rightward first, then upward" direction.
(e) bottom-right to the top-left, following a "upward first, then leftward" direction.
(f) bottom-right to the top-left, following a "leftward first, then upward" direction.
(e) top-right to the bottom-left, following a "downward first, then leftward" direction.
(f) top-right to the bottom-left, following a "leftward first, then downward" direction.

Following [22], we apply a single type of scan path per layer, while alternating the type across layers.

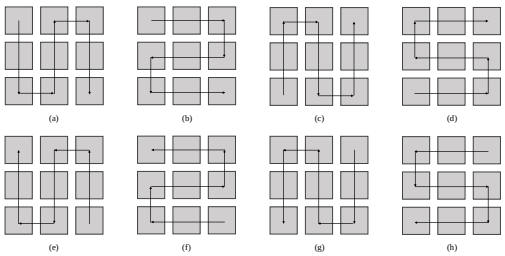


Figure 6. Sptial scan paths for Mamba.

A.2. Overall Architecture

We retain $L = 8$ *dual-stream blocks* from MM-DiT [9], which use separate parameter sets for different modalities. While replacing all dual-stream modules with Mamba blocks could theoretically reduce TFLOPs further (from 29.52 to 22.416), our preliminary experiments reveal that this substitution introduces significant latency during training (approximately a $1.5\times$ increase per iteration) as shown in Tab. 6. This overhead arises from the need to separately process text and video tokens for Mamba, involving additional reshaping, slicing, concatenation, and causal `conv1d` operations to prepare inputs for the state-space model (SSM). As a result, full replacement proves impractical for our architecture-level exploration.

Therefore, our design modifications focus exclusively on the *single-stream* blocks, where the M4V model employs MM-DiM blocks throughout. This choice ensures a balance between efficiency gains and manageable training costs. In future work, this limitation could potentially be mitigated through engineering optimizations, for instance, by developing custom PyTorch kernels to better utilize GPU resources. While such improvements would enhance runtime efficiency, they may also reduce flexibility for iterative architectural evolution.

Table 6. Computational analysis and ablation of the number of MM-DiT in *dual-stream blocks* (L).

Model	**TFLOPs** 241-frms	**Inference Time (s)** 241-frms	**Avg. Score**
$L = 0$	22.416	919	56.43
$L = 2$	24.192	842.5	57.01
$L = 4$	25.968	766	57.59
$L = 8$	29.52	613	58.75

A.3. Linear Scaling Temporal Corruption Noise

As introduced in Sec. 3.1, the prediction of a frame x^i is conditioned on

$$c^i = [K_{\downarrow_2}(x^0), \ldots, K_{\downarrow_2}(x^{i-3}), K_{\downarrow_1}(x^{i-2}), x^{i-1}]. \quad (7)$$

However, due to the error accumulation during autoregressive video generation, latter frames tend to have lower quality than previous ones. Besides, as indicated by [47], using clean latent during training would lead to training-inference inconsistency. Therefore, similar to PyramidFlow [24], we add corruption noise to the condition frames during training. However, different from PyramidFlow, we use a linear scaling corruption noise instead of a constent noise across frames. Specifically, given a corruption ratio η, we randomly sample the maximum corruption scale η_t^{max} in range $[0, \eta]$ and a minimum corruption scale η_t^{min} in range $[0, \eta_t^{max}]$. Then we add per-frame noise to the condition frames with noises $[\sigma_{\eta_t^{min}}, \ldots, \sigma_{\eta_t^{max}}]$ with linear intervals. This design brings slight improvement in convergence speed in our early training stages.

B. Experimental Settings.

B.1. Quantitative Evaluation Setting

In this work, we include various baseline methods for comparisons on VBench. Specifically, we include fully open-sourced methods including Open-Sora Plan [39], Open-Sora 1.2 [60], and PyramidFlow [24] for our major comparison. We also include approaches that uses proprietary data for reference, including Pika 1.0 [1], CogVideoX [58], Kling [26], Runway Gen-3 Alpha [43], and Hunyuan-Video [25]. VideoCrafter2 [6], T2V-Turbo [28], Vchitect-2.0 [10]. We directly source the results for all methods from the official leaderboard for comparison. All compared baseline methods are based on attention for spatialtemporal modeling, while we use Mamba instead.

VBench is an automatic benchmark designed for text-to-video generation models. It scores each submission along *sixteen* objective dimensions that jointly capture (i) low-level visual fidelity—e.g. absence of flicker, smooth motion and high aesthetic / imaging quality—and (ii) high-level semantic faithfulness such as correct object classes, actions, colours and scene composition. For every prompt the model must generate five clips; the per-metric scores are averaged, then linearly normalised with official *min–max* statistics and multiplied by a dimension weight (dynamic-degree is down-weighted to 0.5, all others to 1.0). The normalised scores are grouped into a Quality Score (7 metrics) and a Semantic Score (9 metrics). As summarised in Table 7, VBench reports the weighted mean of each block and finally fuses them with a 4:1 ratio so that perceptual quality carries four times the importance of semantic accuracy:

$$\text{Total Score} = \frac{4\,\text{Quality} + 1\,\text{Semantic}}{5}.$$

This single scalar is used for leaderboard ranking, while the two component scores still expose a model's individual strengths and weaknesses.

B.2. Human Preference Setting

In order to evaluate the performance of our method, we conducted a user study to compare it against four state-of-the-art (SOTA) models: PyramidFlow, CogVideoX, T2V-Turbo, and HunyuanVideo. The user study aimed to assess the generated video quality across three key aspects: aesthetic quality, motion smoothness, and semantic coherence. The design and methodology of the study are outlined as follows.

The study involved five methods: our proposed approach and the four SOTA models mentioned above. A total of 50 video prompts were selected for evaluation, sourced from both VBench and the Internet. These prompts were carefully chosen to cover a broad spectrum of content types and video scenarios, ensuring that the evaluation reflects a diverse range of real-world use cases.

Over 50 participants, including both experts and non-experts, took part in the study. Each participant was asked to rank the generated videos for each method based on three criteria: aesthetic quality, motion smoothness, and semantic coherence. The ranking scale used was from 1 to 5, with 1 representing the highest quality and 5 representing the lowest.

For each of the 50 prompts, the participants were shown videos generated by all five methods, and they were asked to assign a score to each model in the three evaluation categories. The win rates between ours and the compared methods were then aggregated across all participants.

B.3. Detail Results of User Study

We list the average ranking of all prompts in Fig. 7 and all prompts bellow:

1. A breathtaking coastal beach in spring, where gentle waves caress the golden sand in super slow motion. The scene captures the delicate dance of turquoise waters, each wave rolling gracefully and retreating with a soft whisper.

2. A bustling city street comes alive with vibrant energy, lined with towering skyscrapers and historic buildings. The scene captures the essence of urban life, with people of all ages and backgrounds walking briskly, some carrying shopping bags, others engaged in animated conversations.

3. A bustling hospital corridor, filled with the soft hum of activity, features doctors in white coats and nurses in scrubs moving purposefully. The walls are adorned with calming artwork and information.

4. A bustling train station platform comes to life in the early morning light, with commuters clad in winter coats and scarves, their breath visible in the crisp air. The platform is lined with vintage lampposts casting a warm glow, and a sleek, modern train pulls in, its doors sliding open with a soft hiss.

5. A charming panda, wearing a chef's hat and a red apron, stands in a cozy, rustic kitchen filled with wooden cabinets and colorful utensils. The panda carefully chops vegetables on a wooden cutting board, its furry paws moving with surprising dexterity.

6. A cheerful individual stands in a lush backyard, surrounded by vibrant greenery and blooming flowers, tending to a sizzling barbecue grill. They wear a red apron over a casual white t-shirt and jeans, with a chef's hat perched jauntily on their head.

7. A colossal, hyper-realistic spaceship descends gracefully onto the rugged Martian surface, its sleek metallic hull reflecting the crimson hues of the planet. Dust and small rocks scatter as the landing thrusters engage, creating a dramatic cloud of Martian soil.

8. A contemplative individual, dressed in a dark, hooded jacket, stands alone on a dimly lit urban street, the soft glow of streetlights casting long shadows. They lift a cigarette to their lips, the ember glowing brightly in the night.

9. A cozy, dimly-lit restaurant exudes warmth and charm, with rustic wooden tables adorned with flickering candles and fresh flowers. Soft, ambient music plays in the background, enhancing the serene atmosphere.

10. A cozy, dimly-lit restaurant with rustic wooden tables and chairs, adorned with flickering candles and fresh flowers in glass vases, creates an intimate ambiance. The walls are lined with vintage photographs and shelves filled with wine bottles, adding a touch of nostalgia.

11. A determined individual in a sleek, black athletic outfit

Table 7. Composition of the VBench headline scores.

Score	Included sub-metrics
Quality Score	subject consistency; background consistency; temporal flickering; motion smoothness; aesthetic quality; imaging quality; dynamic degree (0.5× weight)
Semantic Score	object class; multiple objects; human action; color; spatial relationship; scene; appearance style; temporal style; overall consistency
Total Score	$\text{Total} = \frac{4 \times \text{Quality} + 1 \times \text{Semantic}}{5}$

jogs along a winding forest trail, surrounded by towering trees and dappled sunlight filtering through the leaves. Their rhythmic strides create a sense of purpose and focus, with the soft crunch of leaves underfoot adding to the serene ambiance.

12. A determined individual, dressed in a red flannel shirt, blue jeans, and sturdy boots, pushes a weathered wooden cart along a narrow, cobblestone street. The scene is set in a quaint, old-world village with charming stone buildings and ivy-covered walls.

13. A drone captures a breathtaking aerial view of a festive celebration in a snow-covered town square, centered around a towering, brilliantly lit Christmas tree adorned with twinkling lights and ornaments.

14. A fluffy orange cat with striking green eyes sits calmly to the right of a large, friendly golden retriever, both facing the camera. The cat's fur is meticulously groomed, and it wears a small, elegant collar with a bell.

15. A golden retriever with a shiny coat strolls leisurely through a sun-dappled forest path, the morning light filtering through the trees casting a warm glow. The dog's tail wags gently as it sniffs the air, ears perked up, taking in the serene surroundings.

16. A grand, historic mansion stands majestically atop a hill, its stone facade adorned with ivy and intricate carvings, bathed in the golden light of a setting sun. The camera pans to reveal tall, arched windows reflecting the vibrant hues of the sky, while the meticulously manicured gardens, with their blooming flowers and ornate fountains, add a touch of elegance.

17. A joyful dog, a golden retriever, sits proudly in a vibrant yellow turtleneck, its fur contrasting beautifully against the dark studio background. The dog's eyes sparkle with happiness, and its mouth is open in a cheerful pant, showcasing its playful nature.

18. A joyful individual, bundled in a red winter coat, knitted hat, and gloves, stands in a snow-covered park, rolling a large snowball to form the base of a snowman. The scene is set against a backdrop of snow-laden trees and a serene, overcast sky.

19. A joyful, fuzzy panda sits cross-legged by a crackling campfire, strumming a small acoustic guitar with enthusiasm. The panda's black and white fur contrasts beautifully with the warm glow of the fire.

20. A lone adventurer, clad in a bright red life jacket and a wide-brimmed hat, paddles a sleek, yellow kayak through a serene, crystal-clear lake surrounded by towering pine trees and majestic mountains.

21. A lone astronaut, clad in a pristine white spacesuit with reflective visors, floats gracefully against the vast, star-studded expanse of space. As the camera pans left, the astronaut's movements are slow and deliberate, capturing the serene beauty of weightlessness.

22. A lone rider, clad in a sleek black leather jacket, matching helmet, and dark jeans, navigates a winding mountain road on a powerful motorcycle. The sun sets behind the peaks, casting a golden glow on the rugged landscape.

23. A lone stormtrooper, clad in iconic white armor, stands on a sunlit beach, holding a futuristic vacuum cleaner. The scene opens with the stormtrooper methodically vacuuming the golden sand, the ocean waves gently lapping in the background.

24. A majestic steam train, with its vintage black and red carriages, chugs along a winding mountainside track, enveloped in a cloud of white steam. The train's powerful engine, adorned with brass accents.

25. A playful panda, with its distinctive black and white fur, sits on a wooden swing set in a lush bamboo forest. The panda's eyes sparkle with joy as it grips the ropes tightly, swaying back and forth.

26. A playful squirrel, with its bushy tail flicking, sits on a

park bench, holding a miniature burger in its tiny paws. The scene is set in a vibrant, sunlit park with lush green grass and colorful flowers in the background.

27. A plump rabbit, adorned in a flowing purple robe with golden embroidery, ambles through an enchanting fantasy landscape. The rabbit's large, expressive eyes take in the vibrant surroundings, where towering mushrooms with glowing caps and bioluminescent flowers light up the path.

28. A plush teddy bear, with soft brown fur and a red bow tie, stands on a stool in a cozy, vintage kitchen. The bear's tiny paws are submerged in a sink filled with soapy water, bubbles floating around.

29. A pristine white bicycle stands alone on a cobblestone street, its sleek frame and vintage design catching the morning light. The bike is adorned with a wicker basket on the front, filled with fresh flowers, adding a touch of charm.

30. A pristine white cat with striking blue eyes lounges gracefully on a sunlit windowsill, its fur glistening in the warm afternoon light. The cat stretches luxuriously, its paws extending and tail curling elegantly.

31. A quaint bakery shop, bathed in warm, golden light, showcases an inviting display of freshly baked goods. The rustic wooden shelves are lined with an assortment of crusty baguettes, flaky croissants, and golden-brown pastries, each meticulously arranged.

32. A refined couple, dressed in elegant evening attire, navigates a bustling street under a heavy downpour. The man, in a tailored black tuxedo, and the woman, in a flowing crimson gown, both hold delicate paper umbrellas adorned with intricate patterns.

33. A serene cow with a glossy brown coat lies comfortably on a bed of fresh straw inside a rustic, sunlit barn. The gentle rays of the afternoon sun filter through the wooden slats, casting a warm, golden glow over the scene.

34. A serene individual sits in a cozy, sunlit nook, surrounded by shelves filled with books, wearing a soft, oversized sweater and glasses. They hold an old, leather-bound book, its pages slightly yellow.

35. A serene individual, dressed in a flowing white blouse and light blue jeans, stands at a rustic wooden table in a sunlit room filled with greenery. They carefully select vibrant blooms from a wicker basket, including roses, lilies, and daisies, and begin arranging them in a crystal vase.

36. A skilled artisan, wearing protective gloves and a welding mask, stands in a dimly lit workshop filled with tools and metal scraps. The person carefully heats a metal rod with a blowtorch, the orange flames casting a warm glow on their focused face.

37. A sleek Mars rover, equipped with advanced scientific instruments and cameras, traverses the rugged, reddish terrain of the Martian surface. The scene opens with a panoramic view of the barren landscape, featuring rocky outcrops and distant mountains under a dusty, pinkish sky.

38. A sleek, black motorcycle with chrome accents roars to life on an open highway, its rider clad in a black leather jacket, helmet, and gloves. The camera captures a close-up of the rider's gloved hand.

39. A sleek, modern train glides effortlessly along the tracks, its metallic exterior gleaming under the bright midday sun. The train's windows reflect the passing landscape of lush green fields and distant mountains, creating a mesmerizing blend of nature and technology.

40. A sleek, silver airplane glides gracefully through a clear blue sky, its wings cutting through the air with precision. As it descends, the sun glints off its polished surface, casting a radiant glow.

41. 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.

42. A young man with long, flowing hair sits on a rustic wooden stool in a cozy, dimly lit room, strumming an acoustic guitar. He wears a vintage denim jacket over a white t-shirt and faded jeans, his fingers skillfully moving across the strings.

43. A young person, dressed in a vibrant red jacket and black jeans, rides a sleek electric scooter through a bustling city street. The scene captures the energy of urban life, with towering skyscrapers and colorful storefronts lining the background.

44. A young person, wearing a cozy gray hoodie and black-rimmed glasses, sits in a dimly lit room, intensely focused on a video game. The glow from the TV screen illuminates their face, highlighting their concentration.

45. A young woman with glasses is jogging in the park wearing a pink headband.

46. A young woman with long, dark hair sits alone in a dimly lit room, her face illuminated by the soft glow of a nearby lamp. Tears stream down her cheeks, glistening in the light, as she clutches a crumpled letter in her trembling hands.

47. A young woman with long, flowing hair sits at a grand piano in a dimly lit room, her fingers gracefully dancing across the keys. She wears a flowing white dress that contrasts beautifully with the dark wood of the piano.

48. a child is playing the guitar in a flower garden.

49. a couple of friends is biking in a living room.

50. a group of school children is seen walking together, with smartphones.

C. Ablation Study Settings

C.1. Evaluation Protocol

To address the computational challenges of comprehensive evaluation, we employ a *customized VBench* as our primary assessment methodology. This choice is motivated by the fact that a full VBench evaluation requires over 160 hours per model variant on a standard NVIDIA A100 GPU, rendering full metric computation impractical for iterative ablation studies. Our customized protocol uses a carefully selected subset of 50 video-generation prompts. All ablation experiments strictly adhere to this fixed prompt set, ensuring direct comparability across architectural variants while reducing the average evaluation time to 4 hours per model.

For these ablation studies, we report seven metrics that can be computed with this prompt subset: two visual consistency metrics—*subject consistency*, *background consistency*; two motion-related metrics—*temporal flickering*, *motion smoothness*; two visual quality-related metrics—*aesthetic quality*, *image quality*; and one video-text alignment metric—*overall consistency*. In Tab. 4, the Avg. Score is the arithmetic mean of these metrics.

C.2. Experimental Design

Our ablation study of architecture design adopts a strategic weight initialization approach to enable efficient hypothesis testing. We first pre-train the model with attention operation and then initialize part of the Mamba layer's projection matrices the using pre-trained weights, following the technique in [54]. Subsequent training is constrained to 20,000 iterations with a learning rate 1e-4, using the same training dataset for all models. This design ensures that each architectural variant undergoes identical optimization conditions, with only the target module parameters being modified between experimental conditions.

D. Training Details.

D.1. Multi-stage Training

To efficiently pre-train the M4V model, we adopt a progressive training strategy. The process begins with text-to-image (T2I) training at 384p resolution. During the subsequent text-to-video (T2V) pre-training phase, we gradually increase the resolution from 384p to 768p and extend the video length from 57 to 121 and 241 frames, training with a combination of image and video data. This staged approach ensures stable adaptation and longer token sequences. For the T2I phase, we follow the training settings from [24], using our own image dataset.

T2V Training. Direct training on 5-second (121-frame) videos led to very slow convergence. To address this, we first trained on 2-second (57-frame) video data at 384p resolution. This stage utilized the WebVid10M, OpenSora-Plan 1M, and OpenVid1M datasets. The 2-second T2V training was conducted using a learning rate of 1×10^{-4}, 64 GPUs, and 40k steps. We then transitioned to training on 5-second (121-frame) videos using the same datasets but with extended frame lengths. This phase used the same learning rate, 64 GPUs, and 60k steps.

Upscaling to 768p. To further improve visual fidelity and motion smoothness, we performed training on 768p videos with lengths of 121 or 241 frames. This step significantly enhanced video clarity and extended generation duration. At this stage, only the OpenSora-Plan 1M and OpenVid1M datasets were used, due to the relatively lower quality of WebVid10M. The model was trained with a learning rate of 5×10^{-5}, using 128 GPUs for 20k steps.

Quality Tuning with Synthetic Data. We observed that existing datasets lacked sufficient motion diversity (e.g., walking, running), limiting generalization in dynamic scenarios. To alleviate this, we synthesized approximately 80,000 videos using models such as HunyuanVideo. These were generated from GPT-4o prompts focused on various subject motions. In the final training stage, we incorporated these generated videos alongside OpenSora-Plan 1M and OpenVid1M, training the model with a learning rate of 5×10^{-5}, using 128 GPUs for 30k steps.

Reward Learning. To further enhance aesthetic quality in later frames, we introduced a post-training phase after the main training. This phase employed two reward models: one for aesthetic scoring [32, 55] and another for text-image alignment [42]. We set the reward loss weight to 0.1 and sampled the final 8 latent frames for fine-tuning. This phase used a learning rate of 1×10^{-5}, 64 GPUs, and ran for 10k steps.

Effect of Different Image/Video Ratio. In our experiments, we observe that a higher image-to-video ratio in the training data often leads to subject deformation in later frames. This may be due to the training gradients from image supervision dominating the learning process, causing the autoregressive model to overemphasize the generation of the first frame. Therefore, selecting an appropriate image-to-video mixing ratio is critical for improving the performance of the autoregressive model, as illustrated in the left part of Fig. 8. We adopt an image:video ratio of 1:8 as our default setting.

D.2. Pre-training Initialization

To accelerate the learning process of the Mamba-based model, we first pre-train the model using attention operations, at the resolution of 384p. Following the strategy proposed

	Aesthetic Quality					Motion Smoothness					Semantic Coherence				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
Prompt ID															
1	0.966667	0.9	1.433333	2.066667	2.133333	1.966667	1.233333	0.9	1.3	2.1	1.533333	1.3	0.833333	1.7	2.133333
2	0.966667	0.833333	1.366667	2.2	2.133333	1.833333	1.1	0.966667	1.666667	1.933333	1.633333	0.9	0.966667	2.066667	1.933333
3	1.1	1.033333	1.633333	2.066667	1.666667	1.833333	1.133333	1.033333	1.3	2.2	1.433333	1.1	0.966667	1.8	2.2
4	1.166667	0.733333	1.433333	2.3	1.866667	1.7	1.1	1.033333	1.6	2.066667	1.366667	1.3	1.033333	1.633333	2.166667
5	0.933333	1.166667	1.666667	2.166667	1.566667	1.733333	1.133333	0.8	1.5	2.333333	1.433333	1.433333	0.966667	1.8	1.866667
6	1.1	0.8	1.433333	2.066667	2.1	2	1.166667	0.866667	1.433333	2.033333	1.533333	1.1	0.8	2	2.066667
7	1.4	0.866667	1.366667	2.033333	1.833333	1.866667	1.133333	0.866667	1.4	2.233333	1.533333	1.1	0.833333	2	2.033333
8	1	0.833333	1.633333	2.1	1.933333	1.833333	1.266667	0.866667	1.5	2.033333	1.433333	1.166667	0.933333	1.766667	2.2
9	1.066667	0.733333	1.533333	2.166667	2	1.833333	1.3	0.933333	1.166667	2.266667	1.766667	1.066667	0.866667	1.9	1.9
10	1.266667	0.866667	1.366667	2.1	1.9	1.7	1	0.933333	1.666667	2.2	1.466667	1.133333	0.9	1.733333	2.266667
11	1	0.866667	1.533333	2.166667	1.933333	2.033333	1.233333	0.933333	1.4	1.9	1.5	0.866667	0.866667	1.933333	2.333333
12	1	1	1.433333	2.133333	1.933333	1.866667	1.3	0.833333	1.4	2.1	1.433333	1.266667	0.933333	1.766667	2.1
13	1.2	1	1.433333	2.1	1.766667	1.933333	1.266667	0.9	1.333333	2.066667	1.633333	1	0.866667	1.966667	2.033333
14	1.133333	1.033333	1.6	1.966667	1.766667	1.733333	1.233333	1.233333	1.166667	2.133333	1.3	1.2	0.866667	2.133333	2
15	1.333333	0.833333	1.266667	2.233333	1.833333	2.066667	1.133333	0.533333	1.766667	2	1.733333	1.1	0.8	1.866667	2
16	0.933333	0.866667	1.666667	2.266667	1.766667	1.933333	0.933333	0.966667	1.6	2.066667	1.433333	1.066667	0.966667	1.9	2.133333
17	1.366667	1.2	1.333333	2	1.6	1.833333	1.133333	0.866667	1.433333	2.233333	1.2	1.266667	0.966667	2.1	1.966667
18	0.966667	1	1.566667	2.233333	1.733333	2.033333	1.2	1	1.333333	1.933333	1.633333	1.2	0.8	1.8	2.066667
19	1.4	0.8	1.433333	2.233333	1.633333	1.933333	1.233333	0.766667	1.433333	2.133333	1.7	1.2	0.7	1.9	2
20	1.166667	0.766667	1.6	1.966667	2	1.7	1.1	1.033333	1.533333	2.133333	1.5	1.3	0.733333	1.666667	2.3
21	1.233333	0.7	1.5	2.066667	2	1.833333	1.366667	0.766667	1.5	2.033333	1.5	1.1	0.933333	1.833333	2.133333
22	1.066667	0.933333	1.466667	2.1	1.933333	2	1.166667	0.833333	1.466667	2.033333	1.5	1.3	0.7	1.766667	2.233333
23	1.166667	1.033333	1.466667	2.133333	1.7	1.866667	1.066667	0.7	1.6	2.266667	1.666667	0.866667	0.833333	1.966667	2.166667
24	1.066667	0.833333	1.433333	2.3	1.866667	1.933333	0.933333	0.966667	1.5	2.166667	1.5	1.366667	0.9	1.8	1.933333
25	1.3	0.733333	1.533333	2.2	1.733333	1.8	1.033333	0.9	1.6	2.166667	1.5	1.066667	0.966667	1.666667	2.3
26	1.2	0.6	1.7	2	2	1.7	1.366667	1	1.6	1.833333	1.633333	1.233333	0.833333	1.766667	2.033333
27	1.033333	1	1.666667	2.066667	1.733333	1.8	1.466667	1	1.3	1.933333	1.533333	1.166667	0.766667	1.9	2.133333
28	1.033333	0.866667	1.6	2.033333	1.966667	1.7	1.2	0.9	1.6	2.1	1.4	1.1	1.1	1.933333	1.966667
29	1.033333	0.833333	1.5	2.166667	1.966667	1.933333	1.166667	1.033333	1.366667	2	1.466667	1.2	0.766667	1.966667	2.1
30	1.2	0.766667	1.666667	1.833333	2.033333	1.7	1.2	0.833333	1.466667	2.3	1.633333	1.2	0.633333	1.833333	2.2
31	1.133333	0.833333	1.5	1.966667	2.066667	1.666667	1.266667	0.766667	1.6	2.2	1.533333	1.066667	0.933333	1.933333	2.033333
32	1.266667	0.766667	1.433333	1.933333	2.1	1.9	1.233333	0.733333	1.6	2.033333	1.6	1.066667	0.8	1.866667	2.166667
33	1.166667	0.9	1.5	2.166667	1.766667	1.733333	1.333333	0.833333	1.533333	2.066667	1.366667	1.333333	0.833333	1.9	2.066667
34	1	0.866667	1.533333	2.066667	2.033333	1.966667	0.866667	1	1.5	2.166667	1.4	1.133333	0.8	2.1	2.066667
35	1.066667	0.866667	1.5	2.233333	1.833333	2.1	0.9	0.933333	1.433333	2.133333	1.6	1.566667	0.7	1.433333	2.2
36	1.2	1	1.466667	2.066667	1.766667	1.966667	1.166667	0.9	1.433333	2.033333	1.266667	1.2	0.766667	2.166667	2.1
37	1.033333	1.033333	1.5	2.2	1.733333	2	1.066667	1	1.3	2.133333	1.466667	1.133333	0.866667	2.033333	2
38	1	0.966667	1.433333	1.9	2.2	1.866667	1.433333	0.933333	1.433333	1.833333	1.7	1.1	1.033333	1.733333	1.933333
39	1.033333	0.733333	1.566667	2.1	2.066667	1.866667	1.166667	0.866667	1.433333	2.166667	1.766667	1.166667	0.7	1.666667	2.2
40	0.933333	1.033333	1.433333	2.166667	1.933333	1.833333	1.1	0.9	1.466667	2.2	1.366667	1.4	0.833333	2	1.9
41	1.066667	0.866667	1.733333	2.033333	1.8	1.833333	1.066667	0.866667	1.466667	2.266667	1.5	1.166667	1	1.766667	2.066667
42	1.033333	0.966667	1.566667	2.133333	1.8	1.933333	1.166667	0.833333	1.6	1.966667	1.6	1.233333	1	1.833333	1.833333
43	1	0.933333	1.9	1.9	1.766667	1.866667	1.166667	0.8	1.466667	2.2	1.5	1.233333	0.966667	1.666667	2.133333
44	1	0.866667	1.333333	2.366667	1.933333	1.866667	1.1	0.866667	1.633333	2.033333	1.4	1.266667	0.766667	2.033333	2.033333
45	0.933333	0.866667	1.6	2.233333	1.866667	1.933333	1.166667	0.9	1.533333	1.966667	1.566667	1.133333	0.933333	1.733333	2.133333
46	1.133333	0.7	1.433333	2.233333	2	1.766667	1.166667	1.066667	1.166667	2.333333	1.4	1.133333	0.9	2.1	1.966667
47	1.133333	0.766667	1.6	2.2	1.8	1.966667	1.033333	0.933333	1.533333	2.033333	1.366667	1.033333	1.166667	1.833333	2.1
48	1.166667	0.933333	1.6	2.066667	1.733333	1.666667	1.233333	0.733333	1.666667	2.2	1.833333	1.066667	0.866667	1.733333	2
49	1.133333	1.066667	1.3	2.133333	1.866667	2.1	1.133333	0.833333	1.466667	1.966667	1.566667	1.333333	0.733333	1.7	2.166667
50	1.2	0.766667	1.566667	2.133333	1.833333	1.966667	0.866667	0.833333	1.566667	2.266667	1.733333	0.966667	0.966667	1.866667	1.966667

Figure 7. Detail results of user study (A: PyramidFlow, B: ours, C: HuanyuanVideo, D: CogVideoX, E: T2V-Turbo)

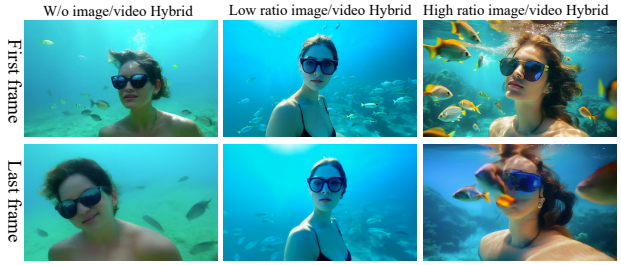


Figure 8. Ablation study of different image/video ratio.

in Mamba-in-LLaMA [54], we initialize the Mamba layers using the pre-trained attention weights. Specifically, we initialize the projection matrices B, C, and the input projection x in the Mamba block with the weights W_q, W_k, and W_v from the attention layer, respectively.

After this initialization, we perform additional fine-tuning to adapt the remaining uninitialized weights, such as A and Δ. This fine-tuning stage is conducted with a learning rate of 1×10^{-4}, using 64 GPUs for 20k steps at the resolution of 384p.

E. Addtional Visual Results.

E.1. Text-to-Video Results.

We show more visualisation results in Fig. 9. Our advanced design allows the model to create videos that are visually

consistent and aesthetically high-quality.

E.2. Image-to-Video Results.

Since our model is an autoregressive diffusion model, it is inherently suited for the task of generating videos from images. Specifically, by setting a given image as the initial frame, the model can autoregressively generate the subsequent frames. We show some results in Fig. 10.

E.3. Additional Improvements

In video generation, while the image quality, aesthetics, and motion of the earlier frames are generally good, the image quality of later frames tends to degrade. This is primarily due to the accumulation of errors from the previously generated frames, which impacts the subsequent ones. To prevent this error propagation, we propose the use of pure T2I-based correction. Specifically, when generating the first frame, the model effectively operates in an unconditional mode, similar to the unconditioned version of text generation. This allows us to leverage the model's strong T2I capabilities to guide the autoregressive generation of subsequent frames.

Our approach introduces a novel strategy for adjusting the flow prediction during the inference stage. Initially, at the final stage of the pyramid, the model predicts a low-quality, aesthetically suboptimal flow velocity v_l. We then compute the predicted x_0, and using forward noise addition, we re-input it back into the model to correct the quality. This process ensures that the autoregressive model's limitations in fitting high-quality video frames are mitigated by leveraging the model's capability in fitting high-quality image data during T2I tasks.

The self-quality guidance formula is defined as:

$$v_l = M(x_t, f, \varnothing) + w_p \cdot (M(x_t, f, p) - M(x_t, f, \varnothing)) \quad (8)$$

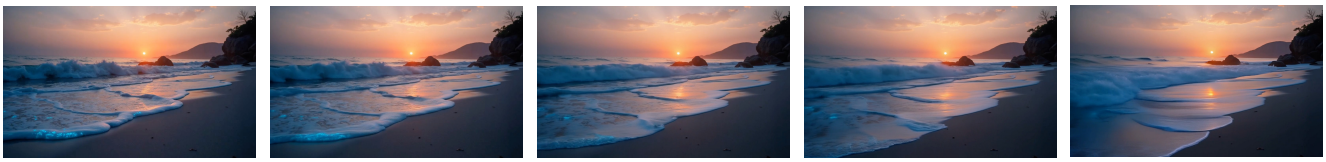
$$v_h = M(x_t, \varnothing, \varnothing) + w_p \cdot (M(x_t, \varnothing, p) - M(x_t, \varnothing, \varnothing)) \quad (9)$$

$$v_{\text{sqg}} = v_l + w_{\text{sqg}} \cdot (v_h - v_l), \quad (10)$$

where M represents our model, f denotes the condition frames, and p refers to the text prompt. Since the earlier frames are typically generated with higher quality, we only apply self-quality guidance starting from the 8th latent frame. Additionally, due to the early stages of the denoising pyramid not yet forming the overall structure and content of the image frames, we begin using self-quality guidance only at the later stages of the pyramid when the texture and content become clearer. This functionality is similar to conventional classifier-free guidance, and its hyperparameters can be adjusted during inference depending on the case. Note that for **all** results in our main paper, we do **not** use the self-quality guidance. We consider this additional improvement as an optional but useful plug-in, which users may choose to enable. Tab. 8 shows the effect of the self-guidance on our customized VBench, with our final-stage model.

E.4. The Use of Large Language Models (LLMs)

We utilize large language models only for grammar checking, style refinement, and language polishing. No LLMs are used for direct content generation or for research ideation.



(a) A tranquil beach at sunset, with bioluminescent waves gently lapping against the shore.

(b) Iron man, walks, on the moon.

(c) A young woman with glasses is jogging in the park wearing a pink headband.

(d) In the bustling heart of Amsterdam, a young man in a black beanie and jacket.

(e) A couple of friends are biking in a living room.

(f) A young Japanese woman standing waiting for a train outside station.

Figure 9. Visualization of text-to-video generation results which are generated at 5s, 768p, 24fps.

Table 8. Effect of self-guidance on *customized VBench* prompts.

	Sub-Cons	**BG-Cons**	**Temp-Flick**	**Motion-Smooth**	**Aes-Qual**	**Img-Qual**	**Overall-Cons**
w/o self-guidance	95.66	96.11	98.62	99.38	63.61	64.69	23.89
w/ self-guidance	95.60	96.12	98.61	99.38	63.89	66.38	26.62

(a) A person riding a bicycle on a wet road. The cyclist is wearing a white blouse with a black tie, a black skirt, black tights, and black shoes.

(b) A person sitting on the floor with their legs crossed. The individual has long, wavy hair in shades of purple and white.

(c) A photorealistic close-up video of two pirate ships battling each other as they sail inside a cup of coffee.

(d) A woman sitting in the driver's seat of a car. The woman has long dark hair and is wearing a black sleeveless top and orange tights.

Figure 10. Visualization of image-to-video generation results which are generated at 5s, 768p, 24fps.