M4V: Multi-Modal Mamba for Text-to-Video Generation

Jiancheng Huang, Gengwei Zhang, Zequn Jie, Siyu Jiao, Yinlong Qian, Ling Chen, Yunchao Wei, Lin Ma

January 14, 2026

In the bustling heart of NYC's Times Square, a life-sized teddy bear, dressed in a tiny leather jacket and sunglasses, sits behind a gleaming drum kit. The bear's furry paws expertly strike the drums.

Challenges in Diffusion-based Video Synthesis

- **Computational Bottleneck of Transformers**
  - High-resolution video generation involves a massive number of tokens.
  - Standard Attention mechanisms suffer from **quadratic complexity** $O(N^2)$, limiting the scalability for long-duration synthesis.
- **Mamba: Linear Complexity State Space Models (SSMs)**
  - Mamba (Selective SSM) achieves **linear scaling** $O(N)$ by utilizing a compressed hidden state to capture long-range dependencies.
  - *Limitation:* Conventional SSMs lack robust mechanisms for **cross-modal alignment** (e.g., precise text-to-video grounding).
- **Proposed Solution: M4V**
  - We introduce a specialized architecture that integrates Mamba with multi-modal conditioning to achieve high-fidelity video generation with significantly lower latency.

Related Work

- **Text-to-video Generation**
  - Transformer-based diffusion architecture (Sora) achieved remarkable high-fidelity video generation quality
  - substantial computational cost limit scalability in both training and deployment.
- **Mamba and Vision Mamba**
  - Mamba introduces time-varying parameters to enhance modeling capacity and a hardware-aware selective scan algorithm to maintain linear-time efficiency.
  - performance on-par with Transformer-based language models, and recently attract a few explorations in vision tasks.
  - Generation tasks still depends on cross-attention for multimodal interaction

What is related to my work? (Mamba-Colorization)

- **How can i improve color reference framework (Mamba)?**
  - pyramids autoregressive prediction
- **Substitute Cross-Attention $O(N^2)$ to Fully/Half-Mamba?**
  - **multi-model token re-composition** : mm-token re-composition
  - per-frame register
  - temporal branch

Preliminaries: PyramidFlow Framework

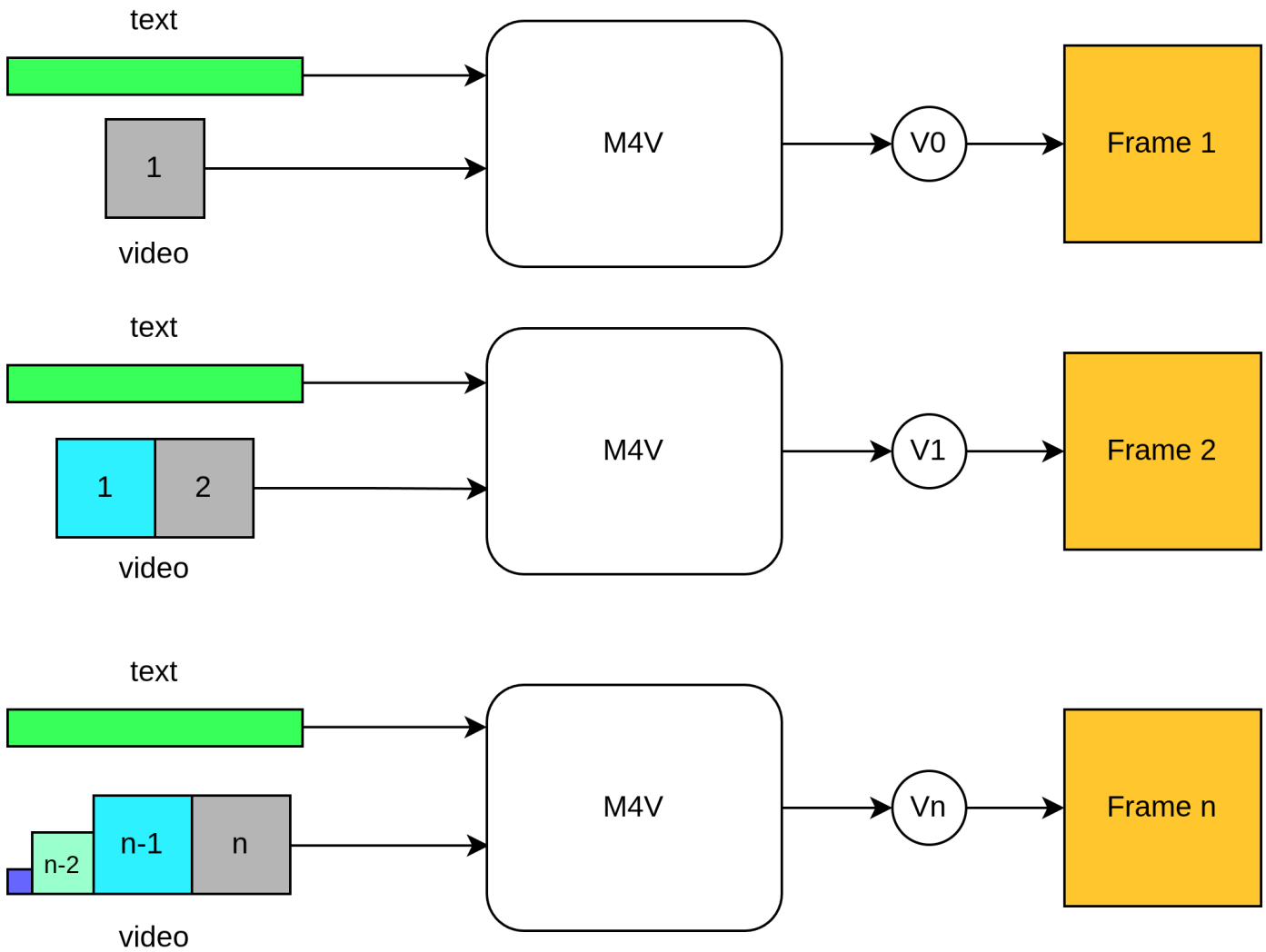
- **Autoregressive Video Generation**
  - M4V adopts the **PyramidFlow** paradigm, which models video sequences as a spatio-temporal autoregressive process.
  - The generation of frame x^i is conditioned on a compressed latent pyramid c^i from preceding frames:

$$c^i = [K_{\downarrow 2}(x^0), \ldots, K_{\downarrow 2}(x^{i-3}), K_{\downarrow 1}(x^{i-2}), x^{i-1}] \tag{1}$$

    where $K_{\downarrow k}(\cdot)$ signifies the k-th level spatial-temporal downsampling.

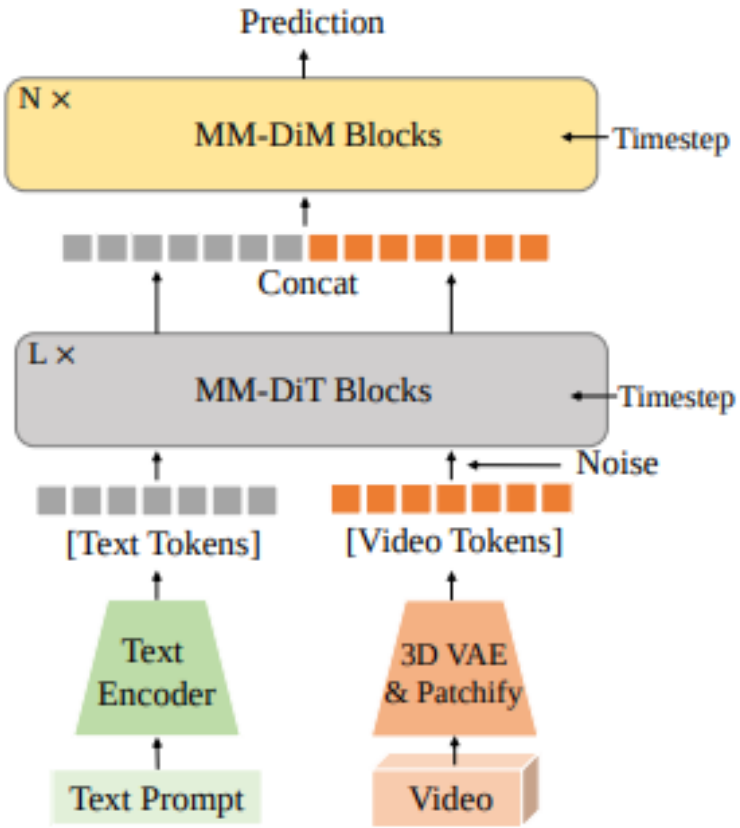
- **Flow Matching**
  - Flow-matching defines a linear probability path $p_t(x)$ between a noise distribution p_0 and a data distribution p_1
  - The model learns a **velocity field** $v(x_t, t; \theta)$ that predicts the direction and speed of the transformation.

Data flow



AutoRegressive architecture.

System Architecture: Overall Framework

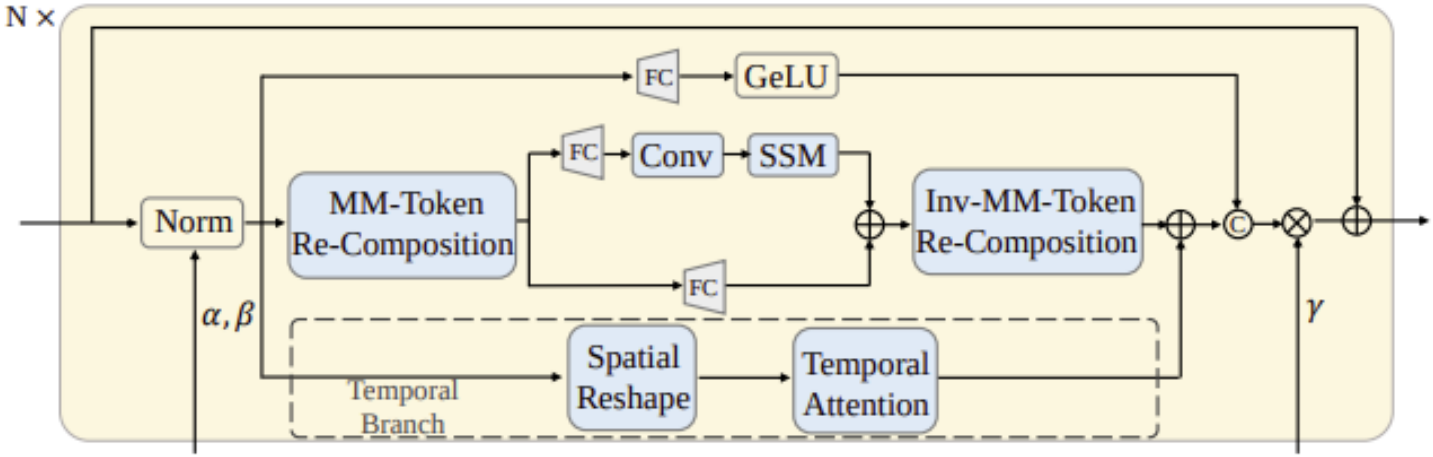


M4V Overall Pipeline.

Two-Stage Hybrid Backbone:

- **MM-DiT Stage ($L\times$):**
  - Acts as the *Introducer* layer.
  - Facilitates early-stage cross-modal interaction between text tokens and video patches.
- **MM-DiM Stage ($N\times$):**
  - The core generative backbone utilizing Mamba.
  - Replaces traditional Transformers to achieve **linear complexity** $O(N)$ for long-video synthesis.

Detailed Micro-Architecture: The MM-DiM Block

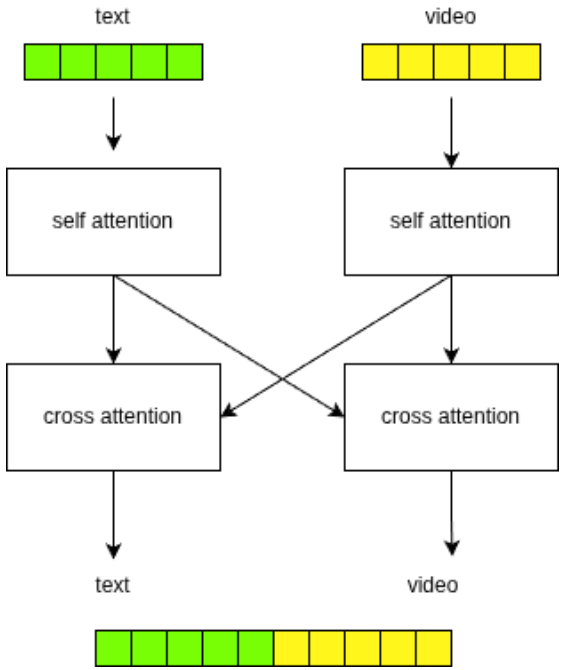


Data flow of the Multi-Modal Diffusion Mamba (MM-DiM) Block.

- **SSM Path**:
  - Processes the *Re-composed* token sequence.
  - Captures long-range dependencies via selective state scanning.
- **Temporal Branch**:
  - A parallel attention-based module (Spatial Reshape + Temporal Attention).
  - Ensures strict inter-frame coherence and motion stability.

Cross-Modal Initialization (MM-DiT)

- **Hybrid Initialization**: Utilizing an Multi-modal Diffusion Transformer (MM-DiT) layer to perform early fusion.
- **Mechanism**: Text embeddings and visual latents are projected into a shared latent space.
- **Objective**: Establishes a strong **semantic prior** before passing sequences to the Mamba backbone, ensuring strict adherence to the text prompt.



MM-DiT Block

MM-DiM: MM-Token Re-Composition

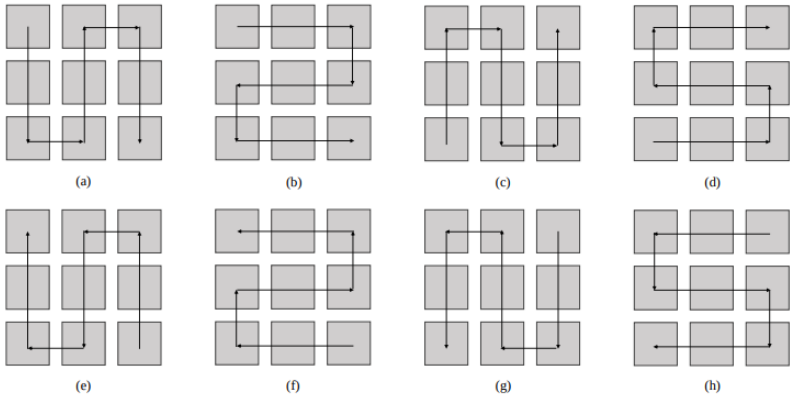
- **Text Token Re-Composition**: Text tokens are **prepended(zero-padding on left)** and **appended** to the visual token sequence.
- **Video Token Re-Composition**:
  - **Zigzag scanning strategy**
    - Map 2D spatial latents into 1D sequences suitable for Mamba's linear scan.
    - 8-directional Zigzag Scanning
  - **Per-Frame Registers**
    - Learnable tokens R(act as boundary anchors and "summary units") are interleaved between frame latents F_t:

$$S_{seq} = [T, R, F_1, R, F_2, \ldots, R, F_n, T] \tag{2}$$

- **Inv-MM-Token Re-Composition** operation removes the Per-Frame Registers and restores the original token order and structure.

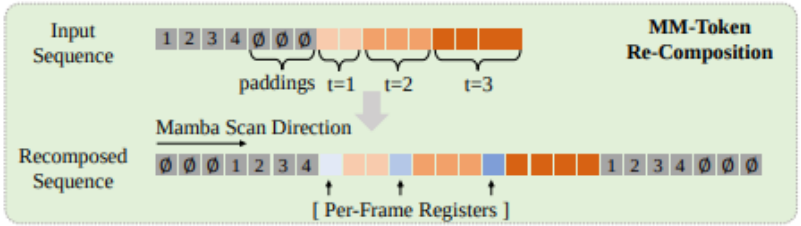
MM-DiM: Spatio-Temporal Token Serialization

Spatial: 8-Directional Zigzag Scanning



- Spatial Context
- Reduced Bias

Temporal: MM-Token Re-Composition



- Symmetric Conditioning
- Per-Frame Registers

MM-DiM: Temporal Branch

- **Temporal Drift**: Challenging for SSMs to model long-range motion consistency in the 3D latent space.
- **The Temporal Branch**: A dedicated module optimized for **inter-frame coherence**.
- **Mechanism**:
  - Consider conditioning latents: $x_s = [x^0, x^1, ..., x^{i-1}]$
  - **1.** Compress all conditioning frames to the smallest spatial resolution:

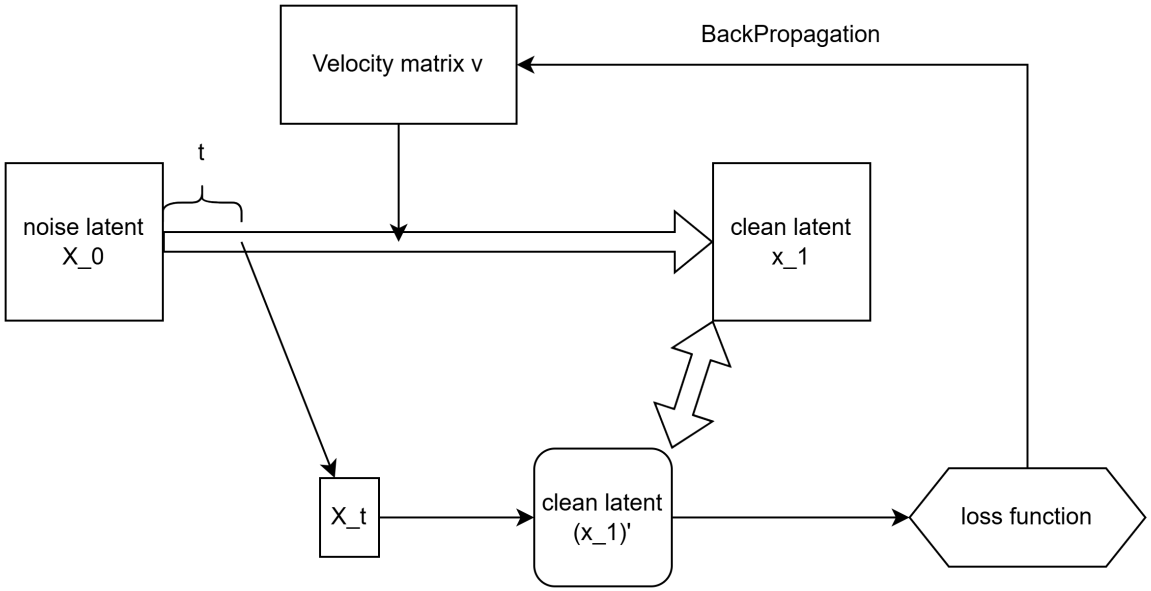
$$x_s \in \mathbb{R}^{\frac{H}{K_s} \times \frac{W}{K_s} \times c \times i} \tag{3}$$

  - **2.** Absorb spatial dimensions into channel dimension:

$$x_s \in \mathbb{R}^{i \times S}, S = c \times \frac{H}{K_s} \times \frac{W}{K_s} \tag{4}$$

  - **3.** noisy latent x^i split and reshape, concatenated with x_s
  - **4.** Apply Causal attention: $O(i^2)$

Autoregressive Reward Learning: Quality Insurance



- **Aesthetic Reward (r_1)**: Powered by HPSv2 to ensure visual realism.
- **Semantic Reward (r_2)**: Powered by CLIP to ensure strict prompt adherence.
- **Loss Function**: $\mathcal{L}_{reward} = -r_1(D(\hat{x}_1^i)) - r_2(D(\hat{x}_1^i))$.

Effect of Autoregressive Reward Learning

Experimental Settings

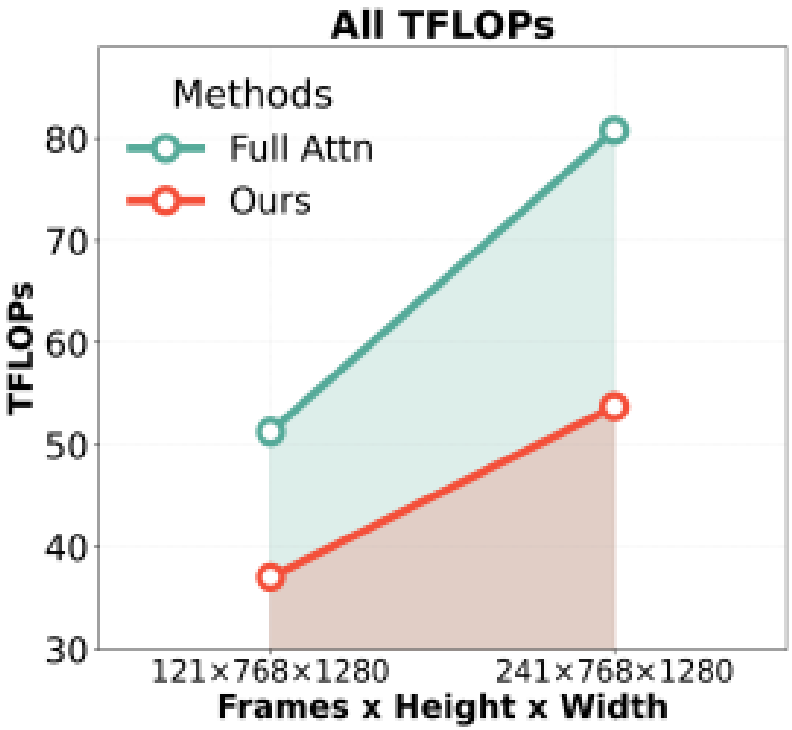
- Training Dataset: Approximately **10 million** single-shot video clips, **90 million** images.
- Implementation Details:
  - $8 \times 8 \times 8$ downsampling factor in 3D VAE to encode video
  - three pyramid stages (3 compression level)
  - progressive training strategy: image to video, low-to-high resolution, short-to-long video length
- Evaluation Metrics:
  - Total Score: weighted average of **Quality Score** and **Semantic Score**
  - **user study**: aesthetic quality, motion smoothness, and semantic coherence.
  - **ablation study**: subject consistency, background consistency, temporal flickering, motion smoothness, aesthetic quality, image quality, and overall consistency

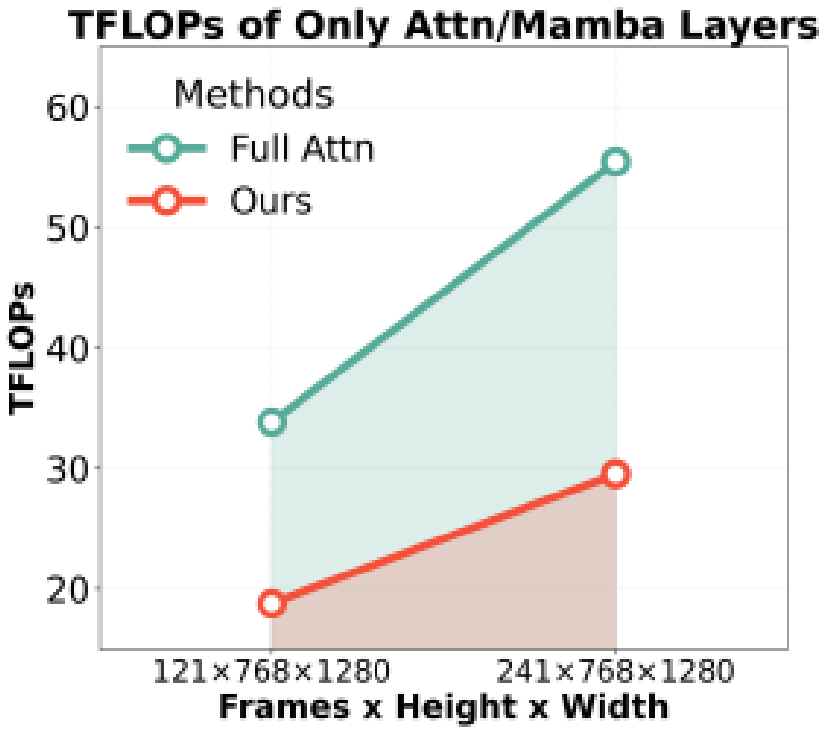
Experimental results on VBench

Model	Accessibility	Video Data Size	Total Score	Quality Score	Semantic Score	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
Gen-2	API	Proprietary	80.58	82.47	73.03	99.58	18.89	66.96	67.42
Pika 1.0	API	Proprietary	80.69	82.92	71.77	99.50	47.50	62.04	61.87
CogVideoX-2B	Inference code	Proprietary	80.91	82.18	75.83	97.73	59.86	60.82	61.68
CogVideoX-5B	Inference code	Proprietary	81.61	82.75	77.04	96.92	70.97	61.98	62.90
Kling	API	Proprietary	81.85	83.38	75.68	99.40	46.94	61.21	65.62
Gen-3 Alpha	API	Proprietary	82.32	84.11	75.17	99.23	60.14	63.34	66.82
VideoCrafter2	Inference code	Proprietary Pretrain+10M Public	80.44	82.20	73.42	97.73	42.50	63.13	67.22
T2V-Turbo	Finetune code	Proprietary Pretrain+10M Public	81.01	82.57	74.76	97.34	49.17	63.04	72.49
Vchitect-2.0-2B	Inference code	Proprietary	81.57	82.51	77.79	97.76	58.33	61.47	65.60
HunyuanVideo	Inference code	Proprietary	83.24	85.09	75.82	98.99	70.83	60.36	67.56
Open-Sora Plan v1.1	Pretrain code	25M Public	78.00	80.91	66.38	98.28	47.72	56.85	62.28
Open-Sora 1.2	Pretrain code	30M Public	79.76	81.35	73.39	98.50	42.39	56.85	63.34
Pyramidflow	Pretrain code	10M Public	81.72	**84.74**	69.62	99.12	64.63	63.26	65.01
Pyramidflow†	Pretrain code	10M Public	81.61	83.54	73.90	99.32	**66.66**	63.96	61.69
M4V	Pretrain code	10M Public	81.55	83.31	74.47	**99.33**	60.55	64.08	62.22
M4V⋆	Pretrain code	10M Public+80K Generated	**81.91**	83.36	**76.10**	99.25	55.55	**64.54**	**65.51**

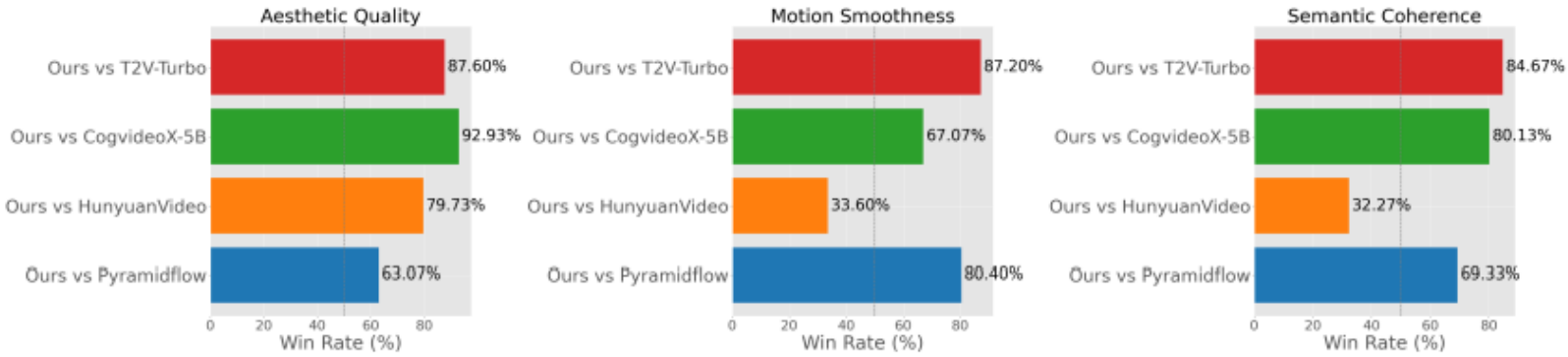
Star : Apply reward learning and including 80K generated samples in training stage.

Comparison of TFLOPS





User Study



User study between M4V, T2V-Turbo, CogvideoX, HunyanVideo and Pyramidflow

Model architecture

Setting	Sub-Cons	BG-Cons	Temp-Flick	Motion-Smooth	Aes-Qual	Img-Qual	Overall-Cons
Baseline	93.28	94.48	98.65	99.19	46.60	63.16	19.77
+ Per-Frame Registers	95.41	95.19	99.22	99.55	48.69	64.18	18.86
+ Text Re-Composition	92.19	93.98	98.76	99.38	45.39	54.83	21.23
+ MM Re-Composition	93.53	94.61	98.70	99.33	49.82	63.79	21.26
+ Temp-Branch	95.67	95.74	98.73	99.41	51.25	66.38	26.08

Ablation study of model architecture and training strategies

Reward Learning and Synthesized Dataset

Training Design		Total	Quality	Semantic
Reward Learning	Generated Data	Score	Score	Score
		81.55	83.31	74.47
✓		81.71	83.32	75.27
	✓	81.59	83.35	74.52
✓	✓	**81.91**	**83.36**	**76.10**

Ablation study of training improvements on official VBench

Visual Result

(a) A futuristic cityscape at dusk, with flying cars zipping between towering skyscrapers adorned with neon lights.

(b) A stylish woman walks down the streets of Tokyo, surrounded by warm neon lights and vibrant city signs. She wears a black leather jacket, a red long skirt, black boots, and carries a black purse.

(c) A determined individual in a sleek, black athletic outfit jogs along a winding forest trail, surrounded by towering trees and dappled sunlight filtering through the leaves.

Summary and Future Work

- **Contribution**: M4V successfully adapts the SSM architecture for multi-modal video diffusion, breaking the $O(N^2)$ complexity barrier.
- **Technical Core**: The combination of re-composition, latent registers, and reward-based alignment ensures both speed and fidelity.
- **Future Directions**: Exploring fully linear-time architectures and investigate methods to enhance the dynamic level of the generated videos

Thank You

感謝聆聽

Thank you for your attention.