

TempoMaster: Efficient Long Video Generation via Next-Frame-Rate Prediction

Supplementary Material

Training Stage	Training Steps	Batch Size	Warmup Steps	Weight Decay	EMA Weight
Single-Frame-Rate	15000	32	2000	1e-4	0.999
Multi-Frame-Rate	45000	32	2000	1e-4	0.999

Table 1. **Additional hyperparameters in training.** We enable ema for a more consistent training.

1. Training Details

Base Model We employ the high-noise model of Wan2.2 [5] as our base model, which can only denoise at an extremely high noise level. In contrast to its predecessor Wan2.1, this model eliminates the dependency on CLIP [4] features extracted from the first frame. This structural simplicity makes it more suitable for our Multi-Mask conditioning framework. Since the base model is inherently limited to extreme noise levels, we first adapt it through a training phase that enables denoising across all noise levels, as mentioned in this paper.

Hyperparameters We employ the AdamW optimizer with a consistent learning rate schedule, using learning rates of 5e-4 and 2e-5 for the respective training stages. We further provide additional hyperparameters during training in Tab. 1.

Noise Scheduling During training, we adopt the logit-normal distribution over t following [1]:

$$\pi_{\text{ln}}(t; m, s) = \frac{1}{s\sqrt{2\pi}} \frac{1}{t(1-t)} \exp - \frac{(\text{logit}(t) - m)^2}{2s^2} \quad (1)$$

Specifically, we set m to 0 and s to 1, and employ a sigma shift of 3 during training.

2. Training Data

Our data pipeline draws upon existing short video methods [2, 3, 5, 6]. To ensure the high quality of training data, we filter videos through multi-dimensional quantitative evaluation:

- **Aesthetic Assessment:** Video frames are evaluated through the average score from an image aesthetic model. This process guarantees that the training data possesses high aesthetic value, thereby enhancing the visual appeal of the generated content.

- **Clarity Detection:** This module primarily employs the Laplacian operator to quantify image sharpness. It focuses on monitoring detail fidelity in dynamic scenes, effectively identifying and filtering out blurred sequences to maintain high image quality across the training samples.
- **Motion Analysis:** The coherence and magnitude of motion are assessed by computing optical flow between consecutive frames. This allows for the effective exclusion of static frames and motion-distorted content, thereby improving the dynamic expressiveness of the dataset.

Handling Multi-Shot Videos In typical data curation pipelines for large-scale video training, raw videos are processed by detecting shot transitions and subsequently segmented into single-shot clips [3, 5, 6]. While in real-world scenarios, videos longer than several seconds often comprise multiple shot transitions. Training solely on single-shot clips thus increases the difficulty of collecting long video data, substantially hindering scalability. In our dataset, there are more than 300K videos curated from films, documentaries, and TV series, each with at least one shot transition.

To ensure caption accuracy, we first segment the videos into single-shot clips and remove blurred frames from transitions. We then captioned each clip in detail and concatenated its frames into new sequences to simulate cut transitions. To indicate temporal cuts, we connect captions with a `<Scene Cut>` token. Additionally, all multi-shot captions are prefixed with a `<MultiShot>` tag.

However, during the training phase, we observed that directly employing multi-mask training on multi-shot videos produced unsatisfactory performance. We attribute this failure to information leakage: the random frame sampling in multi-mask training allows the model to observe all shots and thereby weakens the need to learn shot transitions. Thus, for training on multi-shot videos, we randomly select shots and discard all conditioning frames within those shots. This augmentation forces the model to learn to generate new shots, thereby building an inherent capability for multi-shot video generation.

3. The Principle of Human Study

The human evaluation study is designed to assess the quality of generated long videos across multiple key dimensions, reflecting both aesthetic and functional performance. Each video is rated on a scale from 1 to 5 across the following four dimensions:

- **Aesthetic Quality:** Evaluates visual appeal through com-



Figure 1. **Qualitative comparisons.** We compare TempoMaster with representative long video methods. FramePack outputs are often characterized by a lack of dynamic motion, whereas the other compared methods exhibit marked degradation in visual quality.

position, clarity, lighting, and detail rendering. Penalties apply for overexposure, clutter, artifacts, or blurriness. Severe flaws receive the lowest score.

- **Semantic Alignment:** Measures fidelity to the text prompt, including background, action, and lighting. Minor deviations result in a 1-point deduction, while partial or complete failure to execute the described actions results in scores of 1–2.
- **Motion Quality:** Assesses movement amplitude, speed, and plausibility. Deductions occur for static frames, incoherent motion, implausible dynamics, or severe artifacts.
- **Content Consistency:** Tracks temporal degradation in subject appearance, motion degradation (e.g., incoherence or repetition), and visual decay (e.g., color shift or blurring). Significant drift or decay leads to a score of 1.

4. Additional Visualization Results

We provide more qualitative comparison results with prior works in Fig. 1. We also provide more videos in the compressed file.

References

- [1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024. 1
- [2] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 1
- [3] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, 2021*. 1
- [5] Team Wan, 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, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing 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 preprint arXiv:2503.20314*, 2025. 1
- [6] Yifu Zhang, Hao Yang, Yuqi Zhang, Yifei Hu, Fengda Zhu, Chuang Lin, Xiaofeng Mei, Yi Jiang, Zehuan Yuan, and Bingyue Peng. Waver: Wave your way to lifelike video generation. *arXiv preprint arXiv:2508.15761*, 2025. 1