Supplementary of "TokensGen: Harnessing Condensed Tokens for Long Video Generation"

Wenqi Ouyang[1], Zeqi Xiao[1], Danni Yang[2], Yifan Zhou[1],
Shuai Yang[3], Lei Yang[2], Jianlou Si[2], Xingang Pan[1]

[1]S-Lab, Nanyang Technological University, [2]SenseTime Research,
[3]Wangxuan Institute of Computer Technology, Peking University

{wenqi.ouyang, zeqi001, yifan006, xingang.pan}@ntu.edu.sg
williamyang@pku.edu.cn, sijianlou@gmail.com
{yanglei, yangdanni}@sensetime.com

Overview. The supplementary includes sections as follows:

- Details of Comparison Study (Sec. 1).
- Additional Comparisons (Sec. 2).
- Limitations and Discussions (Sec. 3).
- Additional Visual Results (Sec. 4).

1. Details of Comparison Study

1.1. Prompt Splitting

When comparing our approach with multi-prompt methods such as Video-Infinity [10], DiTCtrl [3], and Kling [1], we first divide the input text prompt into several chunks to guide the generation of individual clips. Specifically, for DiTCtrl and Kling, we employ GPT-4o [2] to split the provided prompt into 24 chunks for a 2-minute-long video or 13 chunks for a 1-minute-long video, using the following instructions:

Please split the prompt depicting a video into 24 separate prompts, each depicting a specific range of the duration of the video in order, and each should have the same style and length as the original prompt. Each prompt should be strictly aligned with the original prompt; if additional content is added, it should also be aligned with the scenery of the original prompt. Each prompt should occupy one line. Please do not insert a blank line between two prompts.

The output format is as follows:
<split prompt 1>
<split prompt 2>
<split prompt 3>
...
<split prompt 24>
The prompt needs to be split is:
<paste the input text prompt here>

For Video-Infinity, which is built on VideoCrafter2 [4] supporting text prompts of up to 77 tokens, we utilize its ability to perform parallel inference across 8 GPUs. To efficiently split text prompts for this method, we provide GPT-4o [2] with the following instructions:

Please split the prompt depicting a video into 8 separate prompts, each depicting a specific range of the duration of the video in order, and each should have the same style as the original prompt. Each prompt should be strictly aligned with the original prompt; if additional content is added, it should also be aligned with the scenery of the original prompt. Each prompt should have fewer than 55 words. Please do not insert a blank line between two prompts.

The output format is as follows:
<split prompt 1>

<split prompt 2>
<split prompt 3>
...
<split prompt 8>
The prompt needs to be split is:
<paste the input text prompt here>

Although we provided detailed instructions, we observed that this task remains highly challenging. GPT-4o often generates split prompts where each segment contains words with a different total number than the original prompt, deviating from the intended style and length. To ensure reproducibility and facilitate comparison, we include all the text prompts along with their corresponding split versions used in the study in the accompanying supplementary material.

1.2. User Study

We conduct a user study to further evaluate the effectiveness of our method. Test prompts are collected from MiraData [7]. For multi-prompt methods, we split the text prompts using the approaches described in the previous section. For A-FIFO+CogVideoX, the same input text prompt as our method is used. In total, we generate 12 video results for each method, with each video ranging from 1 to 2 minutes in length. The test categories include humans, cars, and natural scenes. All videos used in the user study are displayed on our webpage. To ensure an unbiased evaluation, the results are randomly shuffled and displayed to 24 participants. Participants are asked to evaluate the videos based on two aspects: text-visual alignment and motion and content consistency. Questions for each aspect are as follows:

- *Which one best aligned the given text?*
- *Which one keeps the best motion and content consistency in the long-range? For example, the video does not demonstrate scene disjoint, unreasonable content, or obvious quality degradation.*

Our method achieves the best performance across all aspects of the human evaluations, as presented in our main paper. These results highlight the superior long-term control capability of our proposed method, effectively demonstrating its ability to maintain text-visual alignment and ensure motion and content consistency over extended video durations.

2. Additional Comparisons and Analysis

Our expanded comparison includes more baseline methods evaluated with our standard settings, including StreamingT2V [5], FreeNoise [9], VideoTetris [11], and FIFO-VC2 [8], as shown in Fig. 1. StreamingT2V fails on longer videos, FreeNoise/FIFO+VC2 shows limited dynamics (static subjects), and VideoTetris has rich but illogical variations.

Figure 1. The qualitative comparison. We recommend readers refer to our webpage for video comparisons.

Table 1. Quantitative evaluation of comparison study.

	VBench						VBench-Long				Human Study	
Models	SC	BC	TF	MS	IQ	DD	SC	BC	MS	DD	TA	MC
Video-Inf	81.80	90.56	96.66	97.65	61.58	31.0	91.73	95.63	97.67	28.00	0.31%	0.93%
DiTCtrl	76.76	87.96	95.52	97.78	59.26	75.0	91.67	94.21	97.88	53.88	5.3%	4.98%
ST2V	67.71	85.18	93.51	94.40	42.53	34.0	86.10	93.69	94.39	25.42	0.93%	0.31%
FreeNoise	86.50	92.10	96.94	97.69	67.77	24.0	96.64	96.52	98.02	18.00	1.24%	2.18%
VideoTetris	69.27	85.86	94.60	97.04	55.95	96.0	86.86	92.84	94.73	97.12	0.93%	1.25%
FIFO+VC2	89.73	93.93	96.31	97.75	60.49	54.0	94.82	96.43	97.79	49.08	4.36%	3.12%
FIFO+CogX	86.22	92.89	94.78	97.48	64.10	78.57	93.78	95.42	97.43	66.53	23.36%	20.87%
Ours	84.57	92.21	95.41	98.08	63.31	78.95	94.20	95.52	98.40	68.58	63.57%	66.36%
TestSet	85.49	91.43	95.62	98.33	62.78	89.00	94.34	95.03	98.35	82.50	–	–

For quantitative evaluation of added baselines, we use our paper's setup, including a 26-participant human study (Tab. 1: first , second , subpar). We find that Subject and Background Consistency (**SC** & **BC**) and Temporal Flickering (**TF**) favor less dynamic videos, *e.g.*, FreeNoise/FIFO+VC2 ranks high in these but low in Dynamic Degree (**DD**). To further support this, we compute these metrics on MiraData's filtered test set, which features high-quality, continuous motion videos (bottom row). Some methods outperform TestSet on **SC**, **BC**, and **TF**, yet still significantly trail in **DD**. VideoTetris, with the highest **DD**, conversely shows lower **SC** & **BC**, indicating potentially disordered, abrupt motions. CogVideoX [12] and VBench++ [6] also report these metric limitations, as **SC**, **BC**, and **TF** assess quality based on neighboring frame similarity (DINO, CLIP, Mean Absolute Error), thus favoring static videos with higher inter-frame similarity. Therefore, reliable quality assessment requires considering both dynamic aspects and these consistency metrics, as also noted by VBench++. Recognizing these limitations, we also evaluate on VBench-Long, a benchmark for long-term consistency that analyzes keyframe similarity across video segments, overcoming the local metrics limitations. Filtering out methods with subpar **DD** and **SC**/**BC**, our method surpasses FIFO+CogX on all four metrics and all other baselines in human evaluations. The evaluation of long video generation quality is still a significant challenge that we will explore further in the future.

3. Limitations and Discussions

Despite the effectiveness of TokensGen in maintaining long-range consistency, it does not preserve all fine-grained details. Focusing on high-level semantics, tokens may cause gradual variations in foreground or background objects over extended sequences, as shown in Figs. 2 and 4.

Our current framework employs a tuning-free FIFO strategy to maintain short-term consistency during inference. While effective in many scenarios, FIFO can deliver suboptimal performance for cross-clip temporal consistency in some complex scenes. In such cases, the condensed tokens are also insufficient to capture intricate spatial-temporal cues, leading to performance limitations. We illustrate these failure cases in Fig. 3. Addressing these challenges will require more fine-grained tokenization and stronger short-term consistency strategies beyond tuning-free FIFO.

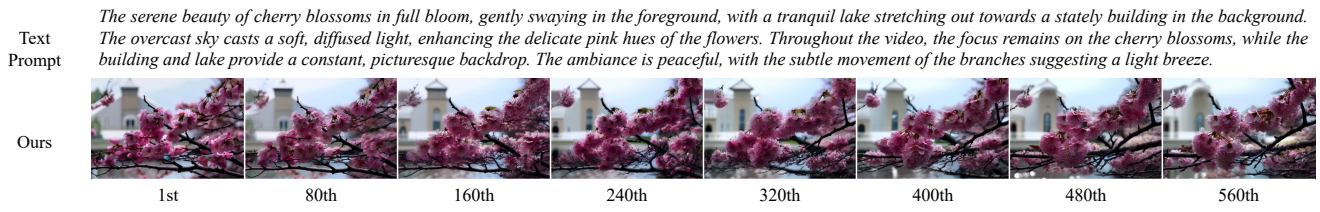

Figure 2. Gradual variations in foreground or background objects over extended sequences.

Our framework is trained and tested on a limited dataset of gameplay and landscape videos, but is scalable to larger datasets for broader applications. In future work, exploring multi-scale tokenization or hybrid representations could bolster fine-grained controllability, retaining subtle attributes while preserving the scalability and resource efficiency.

4. Additional Visual Results

The long video editing example is shown in Fig. 4. For more visual results, comparisons, and ablation studies, please refer to our webpage.

Figure 3. Both the FIFO strategy and the condensed tokens are insufficient to capture intricate spatial-temporal cues, leading to performance limitations.

Figure 4. Long Video Editing.

References

[1] Kling. https://kling.kuaishou.com/, 2024. 1

[2] Gpt-4o. chatgpt.com, 2025. 1

[3] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. *arXiv:2412.18597*, 2024. 1

[4] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 1

[5] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 2

[6] 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. *arXiv preprint arXiv:2411.13503*, 2024. 3

[7] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions, 2024. 2

[8] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *arXiv preprint arXiv:2405.11473*, 2024. 2

[9] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2023. 2

[10] Zhenxiong Tan, Xingyi Yang, Songhua Liu, and Xinchao Wang. Video-infinity: Distributed long video generation. *arXiv preprint arXiv:2406.16260*, 2024. 1

[11] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, Di Zhang, and Bin Cui. Videotetris: Towards compositional text-to-video generation. *arXiv preprint arXiv:2406.04277*, 2024. 2

[12] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3