Loong: Generating Minute-level Long Videos with Autoregressive Language Models

Anonymous Author(s)
Affiliation
Address
email

Figure 1: **One-Minute Videos Generated by Loong.** The images are sampled every 10 seconds. Loong is an autoregressive LLM-based model that can generate minute-level long videos with consistent appearance, large motion dynamics, and natural scene transitions.

Abstract

It is desirable but challenging to generate content-rich long videos in the scale of minutes. Autoregressive large language models (LLMs) have achieved great success in generating coherent and long sequences of tokens in the domain of natural language processing, while the exploration of autoregressive LLMs for video generation is limited to generating short videos of several seconds. In this work, we conduct a deep analysis of the challenges that prevent autoregressive LLM-based video generators from generating long videos. Based on the observations and analysis, we propose Loong, a pioneering autoregressive LLM-based video generator that can generate minute-long videos. Specifically, we model the text tokens and video tokens as a unified sequence for autoregressive LLMs. We propose progressive short-to-long training with a loss re-weighting scheme to mitigate the loss imbalance problem for long video training. We further investigate inference strategies, including video token re-encoding and sampling strategies, to diminish error accumulation during inference. Our proposed Loong can be trained on 10-second videos and be extended to generate minute-level long videos conditioned on text prompts, as demonstrated by extensive experiments.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

1 Introduction

Over the past few years, video generation models, including diffusion-based ones [1–8] and language model based approaches [9, 10], have shown impressive results in generating short videos of a few seconds. To capture more comprehensive content, it is desirable to generate long videos with consistent appearance, larger motion dynamics, and natural scene transitions. Despite recent works [11–13] to generate long videos with diffusion-based video generators, generating content-rich long videos on the scale of minutes remains largely underexplored and challenging.

Autoregressive large language models (LLMs) have shown remarkable success in generating long and coherent text sequences [14–19], demonstrating their ability to capture long-range dependencies and complex temporal patterns. Inspired by the success of autoregressive LLMs in other modalities and their flexibility in unifying various modalities and tasks, recent works [9, 10] have explored autoregressive language models for video generation. Those approaches map videos into discrete tokens and use text tokens as conditioning to generate the video tokens by next-token prediction with decoder-only transformers. State-of-the-art autoregressive LLM-based video generator [10] can generate high-quality 2-second short video clips and iteratively extend to 10-second coherent videos.

Despite demonstrating the ability of long sequence generation in NLP and being explored for short video generation, the potential of LLMs to generate minute-level, content-rich, and dynamic videos remains unexplored. In natural language processing, LLMs can be trained on long sequences and extended beyond the training length. However, we empirically observe that either training autoregressive LLMs on long video sequences or extending short video generators to generate long videos leads to unsatisfactory performance for minute-level video generation. A question arises: ***What restricts the capability of autoregressive language models for generating long videos?***

We hypothesize that the main obstacles are the large redundancy and strong inter-frame dependency among video tokens. The video tokens of the current frame depend heavily on the tokens of the previous frames, leading to two challenges for long video generation: (1) ***Imbalanced loss during training.*** When trained with the next-token prediction objective, predicting early-frame tokens from text prompts is much more difficult than predicting late-frame tokens based on the ground-truth tokens of previous frames. The imbalanced difficulty levels of tokens lead to imbalanced loss during training. The issue becomes more severe as the video length increases, where the accumulated loss of many easy tokens largely surpasses the loss of a few difficult tokens and dominates the gradient direction. (2) ***Error accumulation during inference.*** While the model predicts the next token conditioned on previous *ground-truth* tokens during training, it has to predict the next token conditioned on previous *predicted* tokens during inference. This training-inference discrepancy leads to error accumulation during inference. Because of the strong inter-frame dependency among video tokens and the large number of video tokens, such error accumulation is non-negligible and causes visual quality degradation for long video inference.

In this work, we propose **Loong**, aiming to unleash the power of autoregressive language models to generate long videos in the scale of minutes. Our autoregressive LLM-based video generator consists of two components: a video tokenizer that compresses videos into sequences of discrete video tokens, and an autoregressive LLM that models the unified sequence of text tokens followed by video tokens through next-token prediction. To mitigate the problem of imbalanced loss for long video training, we introduce a progressive short-to-long training strategy that gradually increases the training video length. We further propose loss re-weighting for early frames to prevent the model from being dominated by many easy tokens in the late frames. Moreover, we investigate inference strategies, including the video token re-encoding and sampling strategy, to further extend the video length by iteratively generating the next frames conditioned on previously generated frames. In order to enable training and inference with longer videos, we adopt low-resolution videos for the LLM-based video generator, and leverage a super-resolution and refinement module to further enhance the resolution and fine-grained details of the generated long videos.

In summary, we propose Loong, a novel autoregressive LLM-based video generator that can generate content-rich, coherent, and dynamic long videos in the scale of minutes. Based on our observations and analysis of the issues that limit the power of LLMs for long video generation, we propose progressive short-to-long training with a loss weighting scheme to enable model training on 10-second videos. We further investigate inference strategies to extend the 10-second videos to minute-level videos by autoregressive generation strategies designed for long video inference. Our model demonstrates its ability in generating minute-level long videos through extensive experiments.

2 Related Work

Video generation. The mainstream video generation methods can be categorized into GAN-based [20–22], Diffusion-based [23, 7, 24–26, 3, 6, 27, 28, 13] and language-model-based [29, 10, 30, 31]. Among them, Diffusion-based methods have recently gained the most popularity. Most Diffusion-based methods encode videos into latent space [32] for efficient training and utilize progressive inference strategies [25, 33, 34] to generate videos with high spatial-temporal resolution. With a new scalable Diffusion Transformer [35] architecture, Sora [13] has further pushed video generation to a new stage. Different from diffusion-based video generation methods, our work aims to explore and unleash the potentiality of language models for long video generation, as their ability for modeling long sequence and scaling up have been proved in NLP.

Image and video generation with language models. Language models have recently been explored for visual generation, with most works focusing on tokenizing visual data into a form that can be processed by these models. Quantization techniques like VQ-VAE [36, 37] are commonly used, and transformers are employed to model the resulting tokens. For image generation, autoregressive or masked transformers are prevalent [38–42]. In short video generation, image-level or video-level tokenizers are utilized, incorporating spatial-temporal compression and causal structures. Transformers model the spatial-temporal relationships, with various techniques proposed, such as sparse attention, spatial-temporal attention, large-scale pre-training, and improved tokenization [9, 43–46]. VideoPoet [10] stands out as a multimodal model using bidirectional attention for conditioning, while our method aligns better with the language model paradigm by using unidirectional attention for both text and video. However, these short video generation models focus on producing 1-5 second clips, limiting their ability to capture complex events and maintain consistency over longer durations.

Long video generation. Previous works have explored long video generation using various approaches. LongVideoGAN [47], NUWA-XL [48], and GAIA-1 [49] utilized GAN-based methods, diffusion-over-diffusion techniques, or world models but were limited to specific domains. More recently, video diffusion models have been extended for longer video generation. FreeNoise [50] and Gen-L [11] focus on sampling noise vectors and aggregating overlapping short video segments, respectively, while StreamingT2V [12] proposes an autoregressive approach with memory blocks for consistency and appearance preservation. In the language model domain, Phenaki [30] generates variable-length videos using a masked video transformer. Despite these advancements, generating long videos with rich motion dynamics, consistent appearance, and high visual quality in the open domain remains a challenge.

3 Method

We present Loong, an autoregressive LLM-based model for generating long videos in the scale of minutes. We introduce the overall framework, composed of the video tokenizer and the LLM-based video generator, in Sec. 3.1. We analyze the problem with long video training and propose the progressive short-to-long training with loss re-weighting scheme, enabling training on 10-second videos, in Sec. 3.2. We further investigate inference strategies to extend the generated video length to the minute level and post-processing techniques to enhance the spatial resolution of generated videos in Sec. 3.3.

3.1 Overall Framework

Inspired by previous work in LLM-based image generation and video generation models [38, 41, 46, 31, 10], Loong is designed with two components: a video tokenizer that efficiently compresses the videos into discrete tokens, and a decoder-only transformer that autoregressively predicts next video tokens based on text tokens.

Video Tokenizer. In order to enable spatial-temporal joint compression and joint modeling of images and videos, we leverage causal 3D CNN architecture for the tokenizer, inspired by MAGViT2 [31]. The encoded spatial-temporal features are quantized into discrete tokens with Clustering Vector Quantization (CVQ) [51], an improved version of VQGAN [37] designed to enhance codebook utilization. To extend the temporal coverage of videos within a limited number of tokens, we work with low-resolution videos and leave super-resolution for the post-processing in Sec. 3.3. The

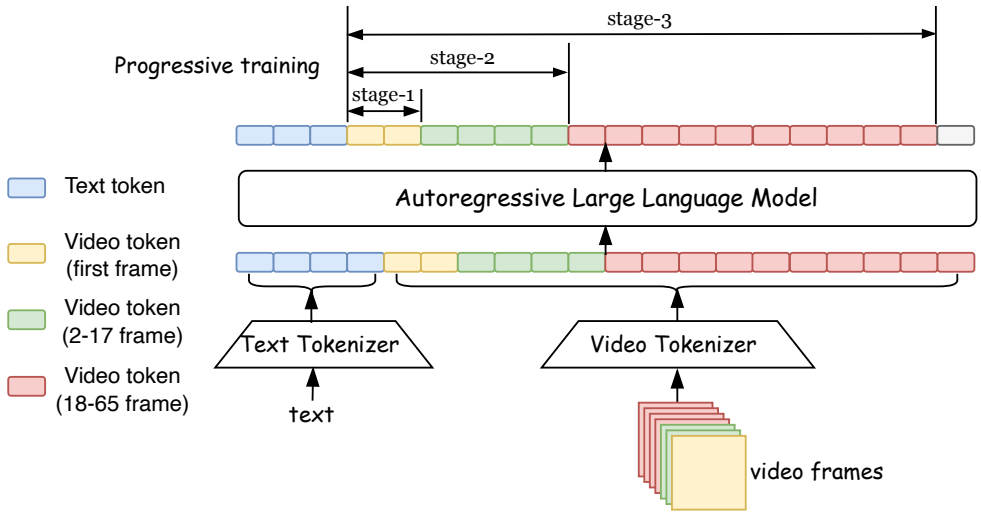

Figure 2: **Overall Framework and the Training process of Loong .** Given the input text tokens, the model predict video tokens autoregressively. All the text and video information is formulated into a unidirectional discrete token sequence, where the model predicts the next token based on the previous tokens. Video Tokenizer is utlized to convert video frames into discrete tokens. We use different color to represent first frame, short clip and long clip separately. We follow a progressive training pipeline to train on long videos.

tokenizer can compress a 10-second video (65 frames, 128×128 resolution for each frame) into a sequence of $17 \times 16 \times 16$ discrete tokens with a vocabulary size of 8192.

Autoregressive LLM-based Video Generation. With the video frames converted into discrete tokens, we can now model the text and video tokens as a unified sequence and formulate text-to-video generation as autoregressively predicting video tokens conditioned on the text tokens with decoder-only Transformers. The process is illustrated in Fig. 2. For simplicity, we omit the special separate tokens in the following formulation. Let $\mathbf{t} = \{t_1, t_2, \ldots, t_N\}$ represent the sequence of text tokens, where N is the number of text tokens. Similarly, let $\mathbf{v} = \{v_1, v_2, \ldots, v_L\}$ represent the sequence of video tokens, where L is the number of video tokens. The autoregressive LLM models the unified token sequence $\mathbf{s} = [\mathbf{t}; \mathbf{v}]$ and is trained with the next-token prediction loss for the video tokens.

$$\mathcal{L} = -\sum_{i=1}^{L} \log p(v_i \mid v_{<i}, \mathbf{t}) \tag{1}$$

where v_i denotes the i-th token in the video sequence $\mathbf{v}$, and $v_{<i}$ denotes all the video tokens preceding v_i.

Discussion. Different from VideoPoet [10], which encodes text with a pretrained T5 text encoder [52] and applies bidirectional attention for the input condition tokens and causal attention for the video tokens, our approach does not rely on a pretrained text encoder. Instead, we formulate the text tokens and video tokens as a unified token sequence and apply causal attention to all tokens. Our unified autoregressive modeling of text tokens and video tokens provides a simpler formulation that is consistent with modern GPT-style LLMs [16]. This design may lead to potential benefits in extending our model to multimodal LLMs that unify different modalities and different tasks for understanding and generation.

3.2 Progressive Short-to-Long Training with Loss Re-weighting

Most video generation models are trained on short video clips, typically no more than 4 seconds, which limits their ability to capture long-term dependencies and complex dynamics in longer videos. To address this limitation, it is essential to train these models on videos with longer durations, enabling them to learn and generate more coherent and contextually rich video content.

However, training directly on long videos leads to suboptimal performance, even when the model is trained for a large number of iterations. We illustrate the loss curve of different frame ranges when

training on 65-frame videos (with 4,356 tokens, covering 10 seconds) in Fig. 3. We empirically observe that tokens from early frames (frames 1-17) have larger losses than those from later frames (tokens from frames 50-65 have the smallest average loss). During training, the model learns through next-token prediction, where it is much easier to predict tokens of later frames given the previous ground-truth video and text tokens. In comparison, predicting early-frame tokens with little visual cues from previous frames is more challenging. The imbalanced loss is a severe problem for long-sequence training because the accumulated loss of the many easy-to-predict tokens from later frames (18-65) surpasses the loss of the few difficult-to-predict tokens from early frames (1-17) and dominates the gradient direction, leading to suboptimal visual quality in the generated videos.

To mitigate the aforementioned challenge of imbalanced video token difficulties, we propose a progressive short-to-long training strategy with loss re-weighting, demonstrated in the following.

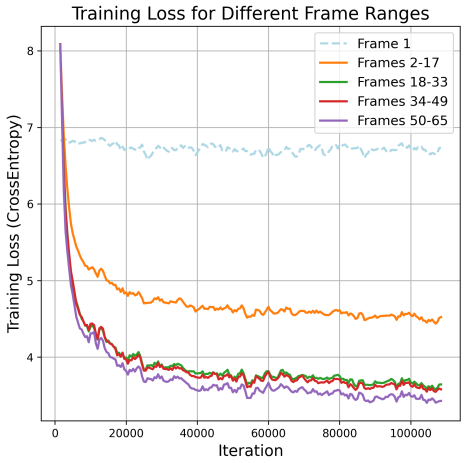

Figure 3: **Imbalanced Training Losses When Training Directly on Long Videos.** The training loss for late frames (18-65) is smaller than that of early frames (1-17), and the loss for the first frame remains relatively high, leading to suboptimal visual quality in the early frames(despite the model being pre-trained on text-to-image).

Progressive short-to-long training. In order to allow the model to first learn the text-conditioned appearance and motion of short videos, and then smoothly adjust to longer-range dependencies and more complex motion patterns in longer videos, we factorize training into three stages which gradually increases the training video length, as illustrated in the Fig. 2: (1) In *stage-1*, we pretrain the model with text-to-image generation on a large dataset of static images, which helps the model to establish a strong foundation for modeling per-frame appearance and structure. (2) In *stage-2*, we continue to train the model jointly on images and short video clips of 17 frames, where the model learns to capture short-term temporal dependencies and motion patterns while preserving the per-frame visual quality. (3) In *stage-3*, we increase the number of video frames to 65, covering a temporal range of 10 seconds, and continue joint training.

Loss re-weighting for early frames. To further strengthen the supervision of early frames and to prevent the model from forgetting the stage-1 and stage-2 priors, we propose a loss re-weighting scheme for stage-3. To be specific, we apply larger loss weights for the tokens of early frames, and the overall weighted loss is formulated as

$$\mathcal{L}_{\text{weighted}} = -(1+\lambda)\sum_{i=1}^{K}\log p(v_i \mid v_{<i}, \mathbf{t}) - \sum_{i=K+1}^{L}\log p(v_i \mid v_{<i}, \mathbf{t}), \tag{2}$$

where the first term denotes the loss for the K tokens corresponding to the early frames (the first 17 frames), and the second term denotes the loss for the $L-K$ tokens corresponding to the later frames (frames 18-65). λ is a positive value to strengthen the loss weight of early frames.

With the loss weighting and progressive training strategy, our model effectively mitigates the issues of long video training. As the model is trained on a temporal range of 10 seconds, it can generate videos of up to 10 seconds with improved temporal coherence and consistency while maintaining the strong appearance and motion priors learned from the image and short video clips.

3.3 Inference Strategies for Extending Video Length and Resolution

Large language models are proven to be length-generalizable, so we expect the LLM-based video generator trained on 10-second videos to be extended to generate longer videos autoregressively. However, generalizing beyond the training video duration is non-trivial and may lead to error accumulation and quality degradation. For instance, a one-minute video corresponds to approximately $26,112$ video tokens under our current settings, which is significantly longer than most text sequences typically encountered in language modeling tasks. The considerable length and the large inter-frame dependency among video tokens pose challenges for extending the LLM-based generator for

long video generation. In this subsection, we investigate inference strategies to generate minute-level videos and post-processing methods like video super-resolution and refinement to generate higher-quality videos.

Video token re-encoding. A natural way of extending videos beyond the training duration is to iteratively generate the tokens of the next video clip, conditioned on the text prompts and the previously generated tokens of the current video clip, exploiting the benefit of autoregressive language models. However, this strategy leads to severe video quality degradation for video frames beyond the training range. With further analysis, we find that this issue stems from the token misalignment caused by the causal video tokenizer. To be specific, the tokens from the last n frames in a video clip are derived based on the context of all previous frames, while the tokens from the first n frames in a new video clip are derived without the context of the previous video clip. Therefore, generating tokens for the new clip directly conditioned on previous tokens leads to distribution shift in the input features for LLMs. To address this issue, we decode the LLM-generated video tokens to the pixel-space videos and then re-encode the last n frames with the video tokenizer. The re-encoded video tokens and the text tokens serve as the conditions to generate the tokens of the next video clip.

Sampling strategy. Decoding video tokens with autoregressive language models is prone to *error accumulation* because of the autoregressive nature of the model and the strong inter-frame dependencies of video tokens. Errors in predicting one token can propagate and influence the generation of subsequent tokens, leading to a degradation in video quality as the length increases. To mitigate this issue, we draw inspiration from the Top-k sampling strategy commonly used in NLP tasks. During the token sampling process, we only sample from the Top-k most probable tokens, ensuring that the generated tokens are of high quality. By focusing on the most likely tokens, we reduce the influence of potential errors on subsequent token generation, effectively alleviating the error accumulation problem. On the other hand, we also observe that too small values of k ($k = 1$ degrades to greedy decoding) lead to almost static videos with little motion. To balance dynamic motion and error accumulation, we choose $k = 50$ for our model.

Super-resolution and refinement. As introduced in Sec. 3.1, our video tokenizer and LLM-based video generator operates on the low-resolution 128×128 videos. This design trades off spatial resolution for longer video sequences during training and inference. We apply off-the-shelf super-resolution and refinement models [53–56] on the LLM-genereated low-resolution videos. This module serves as a post-processing to enhance the spatial resolution and fine-grained visual details of videos, without affecting the content and motion of the generated videos.

4 Experiments

4.1 Implementation Details

Model Architecture. Our video generation model follows the same architecture as LLaMA [18], with a largest size of 7B parameters. We train the models from scratch, without using any text-pretrained weights. The vocabulary consists of 32,000 tokens for text, 8,192 tokens for video, and 10 special tokens, resulting in a total vocabulary size of $\mathbf{vocab_size} = 40,202$. For the video tokenizer, we attempt to reproduce the architecture of MAGViT2 [31], utilizing the Clustering Vector Quantization (CVQ) [51] method for quantization. The model compresses the spatial dimensions (width and height) by a factor of 8 and the temporal dimension by a factor of 4.

Training. Our models are trained on a combination of the CC3M [57] and LAION-2B [58] image datasets, as well as the WebVid-10M [59] video training set and 5.5M self-collected video clips. The training process follows the progressive strategy described in Sec. 3.2. We first pre-train the model on the combined image dataset for 200k iterations, followed by joint training on images and 17-frame video clips from the combined video dataset for another 200k iterations with a batch size of 512. We then jointly train on 65 frames (covering 10 seconds) for 100k iterations with a batch size of 256. The λ is set to 1.0 for the weighted loss of Eq. (2). In each stage, we use AdamW optimizer with a base learning rate of 1.0e-4. The learning rate is scheduled using a linear warmup for the first 10,000 iterations, followed by a cosine annealing decay until reaching the maximum iteration count. For the training of the tokenizer, we also use a progressive approach on the same dataset, increasing the video length from 1 to 17 to 65 frames while maintaining a resolution of 128×128, with a batch size of 64 and training for 400k iterations.

Figure 4: **Effectiveness of the Progressive Training with Loss Re-weighting.** We sample 4 frames from the 17 earlier frames of the video generation results, to show the performance of models trained with or without our training strategy. The top row shows results of the model trained directly on long video, the appearance of objects degrades largely. The bottom row shows the results model trained with our proposed training approach, the appearance preserves effectively.

Figure 5: **Effectiveness of Token Re-encoding during Video Extension.** For each sample, the left two images show the results before the extension process, and the right two images show the results after extension. Without token re-encoding, the extension fails to generate visually consistent content.

4.2 Ablation Study

In this section, we conduct ablation studies to evaluate the effectiveness of our main design choices. Unless otherwise specified, we use the 3B model with an output spatial resolution of 128×128, without any super-resolution and refinement module. To reduce computational cost, we train the models for half the number of iterations compared to the full setting described in Sec. 4.1. Due to the lack of a general long video generation benchmark, we build a custom one by selecting the top-1000 longest clips from the WebVid [59] validation set and slicing each to 27 seconds, the duration of the shortest among them. We employ two performance metrics on this benchmark: Fréchet Video Distance (FVD)[60] and Video-Text Matching (VTM) score calculated using CLIP (ViT-L/14)[61]. It is worth noting that these metrics serve as references, and human evaluation should be considered a more accurate quality assessment of the generated videos. We use the text prompt sets from prior works [4, 6, 62, 3, 13] to generate videos for visualization.

Model Scaling. Scalability is an important characteristic of LLMs. To study scaling behavior of our model, we evaluate performance of the models with different sizes. Tab. 1 presents the quantitative results of our models with 700M, 3B and 7B parameters on the custom benchmark. We observe that larger models achieve better FVD and VTM scores, demonstrating the scalability of our approach.

Progressive Training with Loss Re-weighting. To validate the effectiveness of our proposed training strategy, we compare our model with models directly trained on long videos. Both models are pretrained on images and then trained on long videos using different strategies. Fig. 4 compares the generated frames from a single generation stage without extension for each model. It is clear that the videos generated by the directly-trained models suffer from significant object appearance degradation, losing much of the structural information. In contrast, our model, trained with the progressive training approach, effectively preserves the appearance details.

Table 1: **Scalability of Loong .** The performance improves as the model size increases.

	$\text{FVD}_{\text{I3D}}\downarrow$	$\text{VTM}_{\text{c}}\uparrow$
700M	633	21.5
3B	572	22.8
7B	**432**	**24.1**

Video Token Re-encoding. Fig. 5 illustrates the importance of token re-encoding during the video extension process. Without proper token re-encoding, the model fails to maintain visual consistency when extending the video, resulting in abrupt changes in appearance and content. In contrast, by employing our token re-encoding technique, the extended frames seamlessly continue the video with coherent visual style and content.

Sampling Strategy for Inference. We compare three sampling strategies when predicting each token: greedy decoding ($k = 1$), top-k sampling, and multinomial sampling from the whole vocabulary (k equals video token vocabulary size). As shown in Fig. 6, greedy decoding generates stable results but lacks diversity, while multinomial sampling produces more dynamic content at the cost of quality. Top-k sampling ($k = 50$) balances stability and diversity. A smaller k value prioritizes stability, resulting in less diverse motion, while a larger k allows for more dynamic and varied content at the risk of introducing instability. In the process of video extension, selecting an appropriate k value is crucial for maintaining consistency and mitigating error accumulation over longer sequences.

4.3 Comparison to State-of-the-Art Methods

Table 2: **Comparison on zero-shot text-to-short-video benchmarks.**

Model	CogVideo[45]	MagicVideo[7]	ModelScopeT2V[63]	Show-1[28]	VideoPoet[10]	Loong
CLIPSIM	0.2631	-	0.2930	0.3072	0.3049	0.2903
FVD	1294	998	550	538	213	274

Zero-shot Text to Short Video Generation. Although our approach is not specifically designed for short video generation, we compare our performance on the MSR-VTT dataset [64] using CLIP similarity (CLIPSIM) [44] and FVD [60] metrics, evaluated on 16 frames. As shown in Tab. 2, our FVD score is the second-best, only slightly behind VideoPoet [10] (pretrained). However, our CLIPSIM score is lower compared to some other methods, which can be attributed to the fact that our approach is trained from scratch without utilizing any pre-trained text weights. In contrast, methods with higher CLIPSIM scores, such as VideoPoet, leverage pre-trained language models like T5 [52] for text encoding, while diffusion-based methods often employ CLIP [61] text embeddings, which are already trained on the CLIP dataset. Despite not using pre-trained text models, our method still achieves competitive performance, demonstrating its effectiveness in capturing the semantic relationship between text and video.

Figure 6: **Study on Sampling Strategies.** Results of three different inference sampling strategies. Greedy decoding produces stable results but lacks diversity between frames. Multinomial sampling generates more dynamic and diverse content but with lower quality. Top-k sampling achieves a balance between stability and diversity. k is set to 50 in this experiment.

User Study on Long Video Generation. We conduct a user study to compare our method with StreamingT2V [12], a state-of-the-art open-sourced long video generation method built on Stable Video Diffusion [26]. We use 50 text prompts from prior works [4, 6, 62, 3] to generate 1-min videos. In the study, users are presented with 2 videos generated by the two models, conditioned on the same text. They are asked to choose the preferred video based on visual text matching and content consistency. The videos are presented randomly, and users are not informed about the models. We collect 440 responses. As shown in Fig. 8, our model outperforms StreamingT2V in both content consistency (win rate 0.83 vs. 0.125) and visual text matching (win rate 0.65 vs. 0.19).

4.4 Visualization Results

In this section, we show the results of our model under different text-to-video generation scenarios.

Figure 7: **Generated videos from Loong across various text-to-video scenarios.**

Text to Short Video. In the top row of the figure, we show results of short video generation. As shown in the figure, our approach exhibits the capability to generate short videos with rich details and high fidelity while maintaining strong alignment with the given text descriptions.

Text to Long Video. The second row shows frames sampled from a long video generated by our model, conditioned on a concise text description. These examples demonstrate that our approach can generate long videos containing diverse content and larger dynamic changes compared to short video generation, while maintaining semantic alignment with the given text.

Dense Text to Long Video. Although not explicitly trained on dense captions, we found that our model can effectively adapt to long video generation in a zero-shot manner. Our model demonstrates the capability to generalize to dense caption conditions without requiring specialized finetuning. As illustrated in the last row of Fig. 7, The generated long videos maintain semantic alignment with the provided dense captions, showcasing rich content that corresponds to the detailed descriptions, including multiple characters, weather, scenery, and building information. However, we observe that the generated images appear slightly blurry. We attribute this to the low resolution of our transformer's output, which may result in blurriness when generating highly detailed content. We hope that future advancements in tokenizer compression capabilities and longer LLMs could help address this issue.

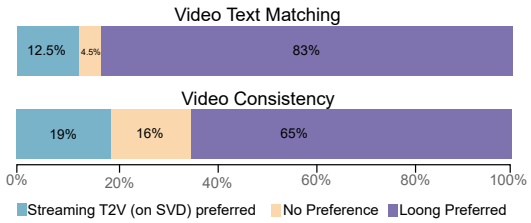

Figure 8: **User Study on 1-min videos.** Comparison with the StreamingT2V on SVD model. Our model is more preferred by human raters in terms of both visual text match and content consistency.

5 Conclusion and Discussions

We propose Loong, the pioneering autoregressive LLM-based video generation model that can generate minute-level long videos with consistent appearance, large motion dynamics, and natural scene transitions. We overcome the challenges of long video training with the progressive short-to-long training scheme with loss re-weighting. We also investigate inference strategies to extend generated videos beyond training duration. Our experiments demonstrate the effectiveness of our approach in generating minute-level long videos. We discuss the limitations in the Appendix.

Border impact. The model can be deployed to assist visual artists and film producers on video creation, enhancing their efficiency. It can also be deployed for entertainment purposes. On the other hand, it may be used for generating fake content and delivering misleading information. The

community should be aware of the potential social impacts. It is necessary to develop techniques to detect and watermark the videos generated by machine learning models.

References

[1] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[2] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 7346–7356, 2023.

[3] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *Proc. Int. Conf. Learn. Representations*, 2022.

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[5] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv:2311.10982*, 2023.

[6] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[7] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

[8] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024.

[9] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[10] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[11] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

[12] 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.

[13] OpenAI. Sora: Creating video from text. https://openai.com/sora, 2024.

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[15] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Sastry, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Proc. Advances in Neural Inf. Process. Syst.*, 2020.

[17] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 27, 2014.

[21] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Proc. Advances in Neural Inf. Process. Syst.*, 2016.

[22] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. Advances in Neural Inf. Process. Syst.*, 2020.

[24] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[25] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[26] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[27] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023.

[28] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Advances in Neural Inf. Process. Syst.*, 2017.

[30] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *Proc. Int. Conf. Learn. Representations*, 2022.

[31] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. In *Proc. Int. Conf. Learn. Representations*, 2024.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[33] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv:2211.13221*, 2022.

[34] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 22563–22575, 2023.

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 4195–4205, 2023.

[36] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Proc. Advances in Neural Inf. Process. Syst.*, 2017.

[37] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 12873–12883, 2021.

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proc. Int. Conf. Mach. Learn.*, pages 8821–8831, 2021.

[39] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[40] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Proc. Int. Conf. Mach. Learn.*, pages 4055–4075, 2023.

[41] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2022.

[42] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, et al. Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms. In *Proc. Advances in Neural Inf. Process. Syst.*, 2024.

[43] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *Proc. Eur. Conf. Comp. Vis.*, pages 102–118, 2022.

[44] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv:2104.14806*, 2021.

[45] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *Proc. Int. Conf. Learn. Representations*, 2022.

[46] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2023.

[47] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. In *Proc. Advances in Neural Inf. Process. Syst.*, 2022.

[48] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023.

[49] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[50] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *Proc. Int. Conf. Learn. Representations*, 2024.

[51] Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2023.

[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 10684–10695, 2022.

[54] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *Proc. Int. Conf. Learn. Representations*, 2023.

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3836–3847, 2023.

[56] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[57] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 35, 2022.

[59] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2021.

[60] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021.

[62] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[63] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[64] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5288–5296, 2016.