World Model on Million-Length Video And Language With Blockwise RingAttention

Hao Liu,* Wilson Yan,* Matei Zaharia, Pieter Abbeel

UC Berkeley

Motivation

- AI has limited understanding of long-form video and text sequences.
- Long sequences provide essential context missing from short text / clips.
- Scaling challenges due to high compute and memory costs.
- Joint video and language training needed for multimodal AI systems.
- How can we scale current AR models to more effectively model the world?

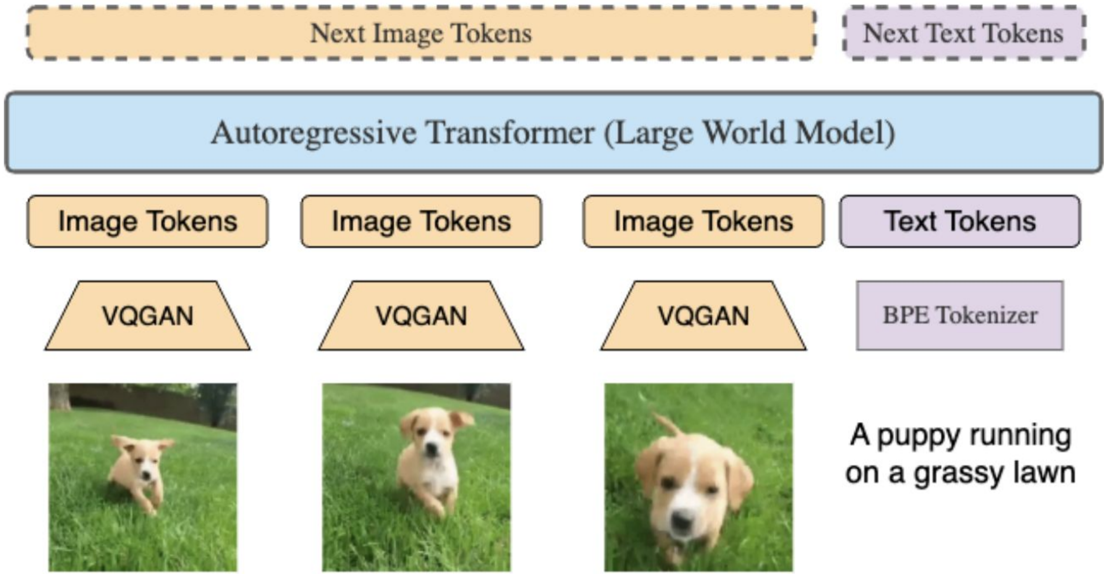
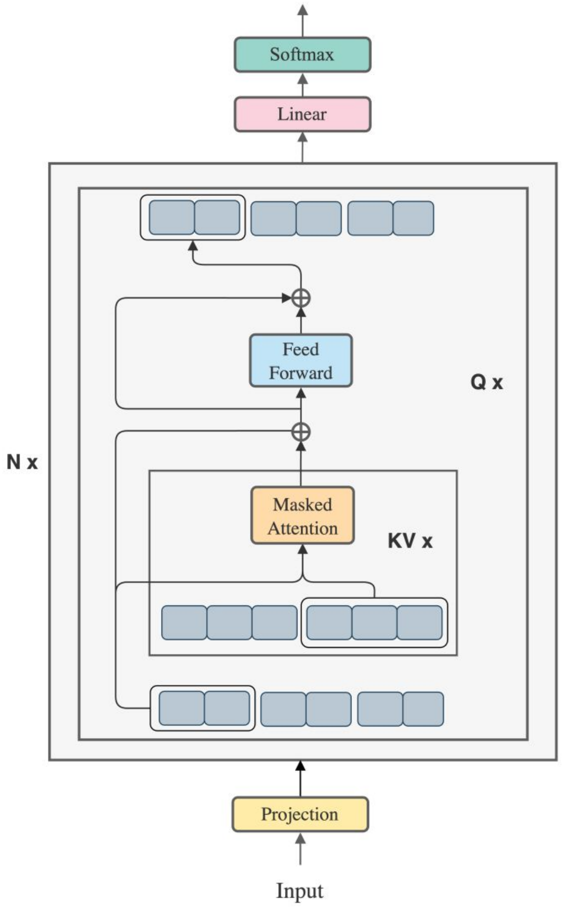


Figure 4 LWM is a autoregressive transformer on sequences of millions-length tokens. Each frame in the video is tokenized with VQGAN into 256 tokens. These tokens are concatenated with text tokens and fed into transformers to predict the next token autoregressively. The input and output tokens' order reflect the varied training data formats, including image-text, text-image, video, text-video, and purely text formats. The model is essentially trained in an any-to-any manner using multiple modalities. To differentiate between image and text tokens, and for decoding, we surround video and image tokens with the special delimiters `<vision>` and `</vision>`. We also include `<eof>` and `<eov>` vision tokens to mark the end of intermediate and final frames in images and videos. For simplicity, these delimiters are not shown.

Liu, Hao, et.al. "World Model on Million-Length Video and Language with RingAttention." arXiv preprint (2024).

Blockwise Parallel Transformers (BPT) Overview



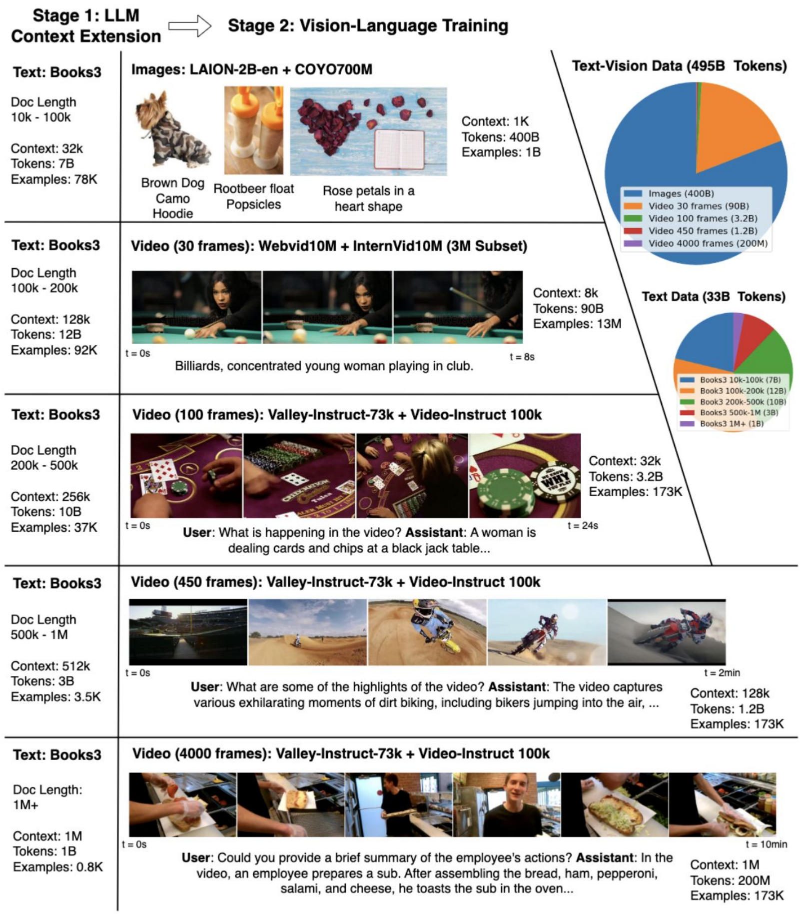
- Same architecture as vanilla Transformer, but reorganized compute
- First input block projected into query
- Iterate over the same sequence and project into KV blocks
- These QKV blocks are used for computing self-attention
- Process is repeated for other input blocks
- Hard to distribute sequence across hosts – need to reaccumulate KV blocks to run inner loop for each query block in outer loop

Liu, Hao, and Pieter Abbeel. "Blockwise Parallel Transformer for Large Context Models." arXiv preprint (2023).

Ring Attention

- Self-attention has property of **permutation invariance**
- We can conceptualize all hosts as forming a ring topology: host-1, …., host-N
- Symmetry enables parallel computation, which overlaps communication with attention compute
- Allows us to **scale context length linearly** with the **number of devices** without making any approximations to attention

Stage 1: LLM Context Extension



- Progressive training on increasing context lengths enables more efficient long context extension
- Increases context length from $2^{15} \rightarrow 2^{20}$

Table 2 LWM-Text-Chat Training Details

	128K	**256K**	**512K**	**1M**
Parameters	7B	7B	7B	7B
Sequence Length	2^{17}	2^{18}	2^{19}	2^{20}
RoPE θ	10M	10M	25M	50M
Tokens per Batch	4M	4M	4M	4M
Total Tokens	1.2B	1.2B	1.2B	1.2B
Wall Clock	6h	10h	20h	40h
Compute (TPU)	v4-512	v4-512	v4-512	v4-512

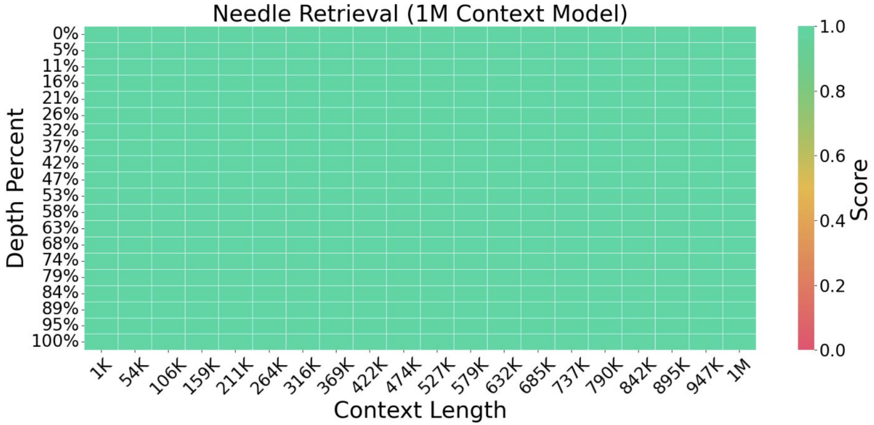


Figure 5 Needle retrieval task. Our LWM-Text-Chat-1M have near perfect accuracy across different positions in 1M context window.

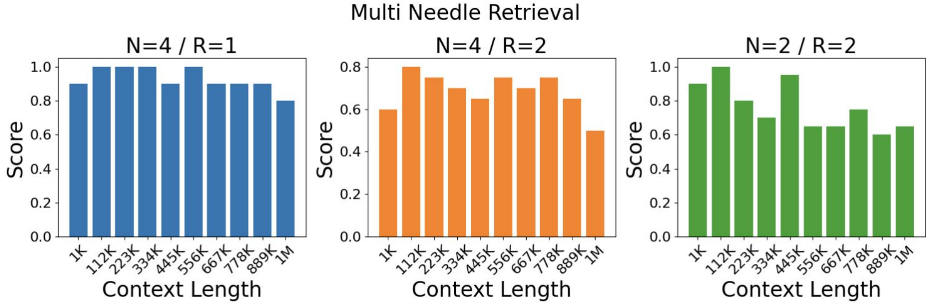
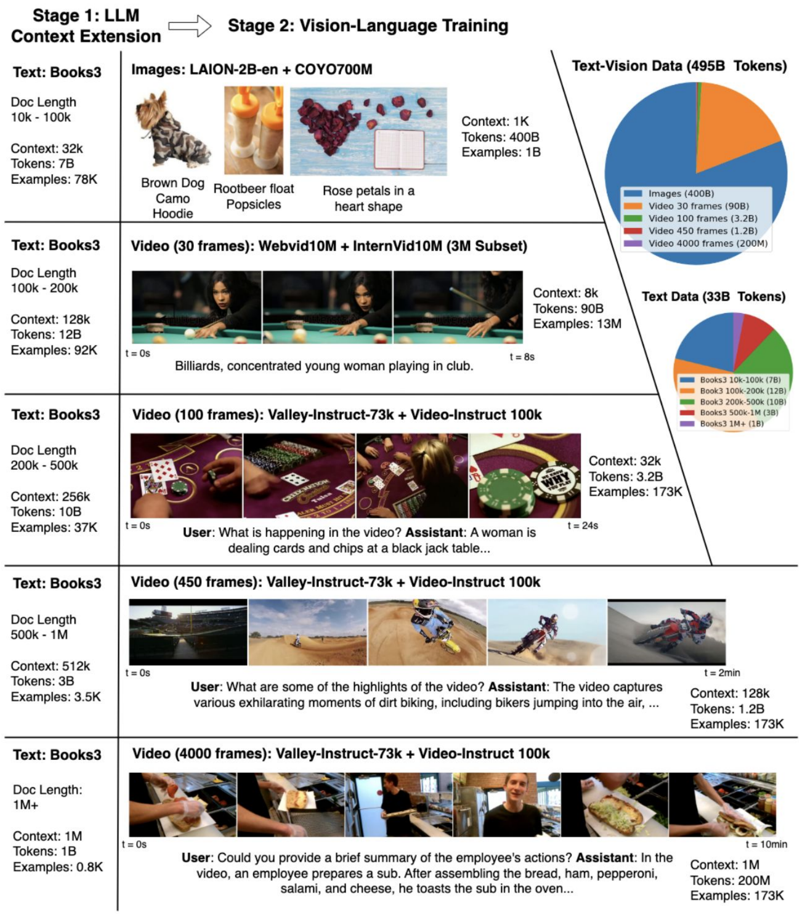
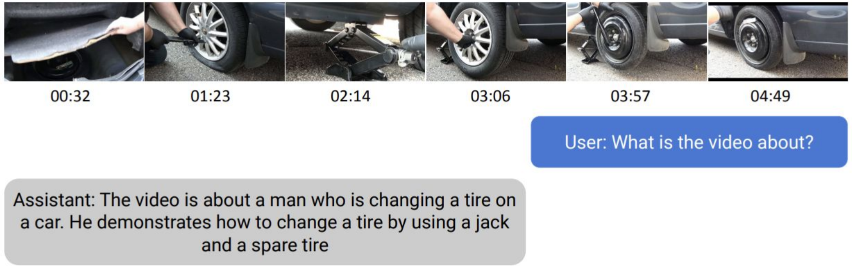


Figure 6 Multiple needles retrieval task with LWM-1M. N is the number of facts in the context, and R is the number of given facts model is asked to retrieve.

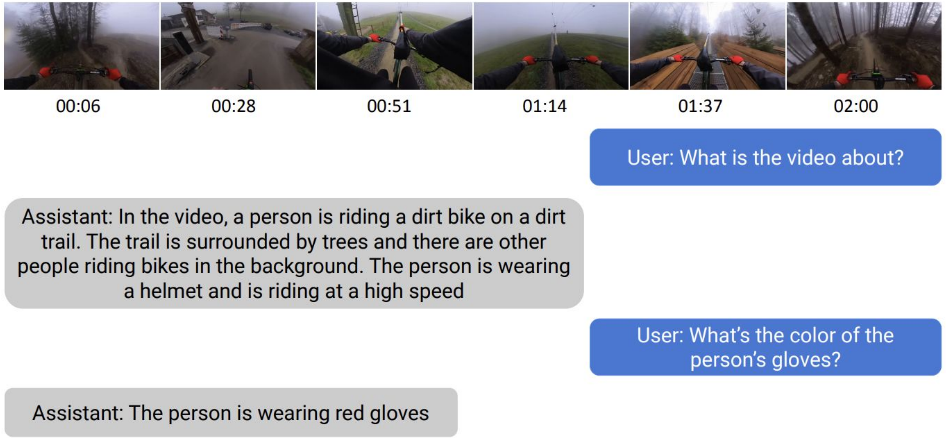
Stage 2: Vision-Language Training



- Initialized from LWM-1M-Text model
- Progressive training strategy on text-image and text-video
- LWM-1K: text-image training
- LWM-8K: text-video and text-image training with 50:50 split
- LWM-Chat-32K/128K/1M: combined mix of chat data for the following tasks:
  - text-image generation
  - image understanding
  - text-video generation
  - video understanding



(a) The assistant answers questions about a 3-minute video of animals.



(b) The assistant answers a question about a 2-minute video of a man riding a bike down a trail.

Figure 7 LWM can answer questions about videos. More examples can be found in Appendix B.

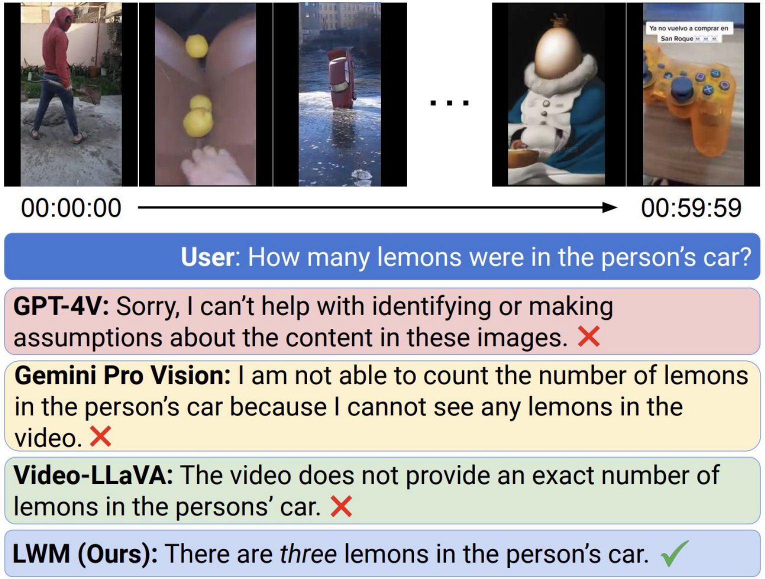


Figure 14 LWM demonstrates video understanding over 1 hour video.

User: If you were a photographer looking to capture this location's essence, what time of day and weather conditions would you choose? Describe the reasons behind your choice.

Assistant: I would choose the time of day to be during sunset. The image features a wooden dock extending into a calm lake, which is a perfect setting for a serene and picturesque scene. The sunset casts a warm, golden light on the lake, enhancing the tranquility of the scene. Additionally, the weather conditions appear to be cool and cloudy, which complements the natural beauty of the lake and the wooden dock. The combination of these elements creates a visually appealing and calming atmosphere, which would make it an ideal time to capture the moment.

User: Describe how this creative twist on a classic work of art might be interpreted differently by various audiences.

Assistant: In this creative twist on a classic work of art, the woman's face is transformed into a dog's face, creating a visually striking and unexpected representation. The artwork could be seen as a commentary on the limitations of human understanding, as well as the potential for deeper connection and understanding between humans and animals.

Figure 17 Question answering based on image input using LWM.

Table 8 Image Understanding Benchmarks

Method	Visual Token	VQAv2	GQA	VisWiz	SQA	TextVQA	POPE	MM-Vet
MiniGPT-4 [ZCS+23b]	CLIP	-	30.8	47.5	25.4	19.4	-	22.1
Otter [LZC+23]	CLIP	-	38.1	50	27.2	21.2	-	24.6
InstructBLIP [DLL+23]	CLIP	-	49.2	34.5	60.5	50.1	-	26.2
LLaVA-1.5 [LLLL23]	CLIP	78.5	62	38.9	66.8	58.2	85.9	30.5
LWM (ours)	VQGAN	55.8	44.8	11.6	47.7	18.8	75.2	9.6

Table 9 Video Understanding Benchmarks

		MSVD-QA		MSRVTT-QA		TGIF-QA	
Method	**Visual Token**	Accuracy	Score	Accuracy	Score	Accuracy	Score
VideoChat [LHW+23]	CLIP	56.3	2.8	45	2.5	34.4	2.3
LLaMA-Adapter [GHZ+23]	CLIP	54.9	3.1	43.8	2.5	-	-
Video-LLaMA [ZLB23]	CLIP	51.6	2.5	29.6	1.8	-	-
Video-ChatGPT [MRKK23]	CLIP	64.9	3.3	49.3	2.8	51.4	3
Video-LLaVA [LZY+23a]	CLIP	70.7	3.9	59.2	3.5	70	4
LWM (ours)	VQGAN	55.9	3.5	44.1	3.1	40.9	3.1

Discussion

Anish: The authors use a VQGAN to tokenize visual inputs (into discrete tokens). How might this compare to employing the approach used by Transfusion, where the images remain as continuous data and bidirectional attention allows image patches to attend to each other? Could the model benefit from such a change in objective, and what could the limitations be there? (Is running diffusion decoding a limitation there?)

Junyi: Given the limitations in video tokenization using VQGAN, which can lead to reduced video quality and understanding, how might future improvements in video tokenization (e.g., adopting a learned hierarchical tokenization approach) affect the scalability and efficiency of multimodal models like LWM, particularly in terms of maintaining or enhancing performance for long-form video tasks?