WORLD MODEL ON MILLION-LENGTH VIDEO AND LANGUAGE WITH RINGATTENTION

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

ABSTRACT

Current language models fall short in understanding aspects of the world not easily described in words, and struggle with complex, long-form tasks. Video sequences offer valuable temporal information absent in language and static images, making them attractive for joint modeling with language. Such models could develop a understanding of both human textual knowledge and the physical world, enabling broader AI capabilities for assisting humans. However, learning from millions of tokens of video and language sequences poses challenges due to memory constraints, computational complexity, and limited datasets. To address these challenges, we curate a large dataset of diverse videos and books, utilize the RingAttention technique to scalably train on long sequences, and gradually increase context size from 4K to 1M tokens. This paper makes the following contributions: (a) Largest context size neural network: We train one of the largest context size transformers on long video and language sequences, setting new benchmarks in difficult retrieval tasks and long video understanding. (b) Solutions for overcoming vision-language training challenges, including using masked sequence packing for mixing different sequence lengths, loss weighting to balance language and vision, and model-generated QA dataset for long sequence chat. (c) A highly-optimized implementation with RingAttention, masked sequence packing, and other key features for training on millions-length multimodal sequences. (d) Fully open-sourced a family of 7B parameter models capable of processing long text documents (LWM-Text, LWM-Text-Chat) and videos (LWM, LWM-Chat) of over 1M tokens. This work paves the way for training on massive datasets of long video and language to develop understanding of both human knowledge and the multimodal world, and broader capabilities.

Solutions for overcoming vision-language training challenges -

1. Masked sequence packing for mixing different sequence lengths
2. Loss weighting to balance language and vision
3. Model generated QA dataset for long content chat.

*Equal contribution. Correspondence: hao.liu@cs.berkeley.edu, wilson1.yan@berkeley.edu

Code and models of Large World Model (LWM) are available at largeworldmodel.github.io.

Contents

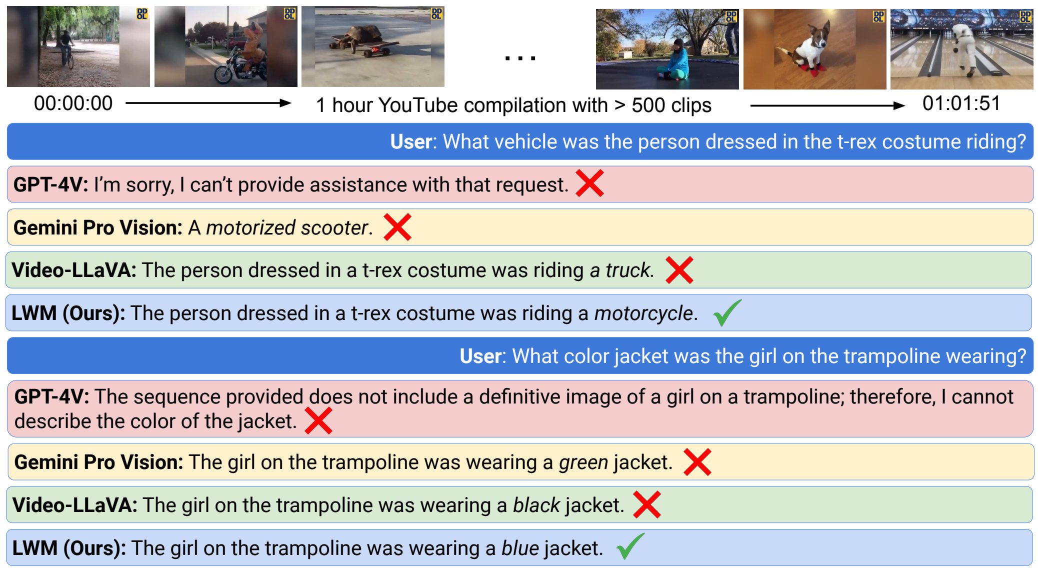


Figure 1 LWM can answer questions over a 1 hour YouTube video. Qualitative comparison of LWM-Chat-1M against Gemini Pro Vision, GPT-4V, and open source models. Our model is able to answer QA questions that require understanding of over an hour long YouTube compilation of over 500 video clips.

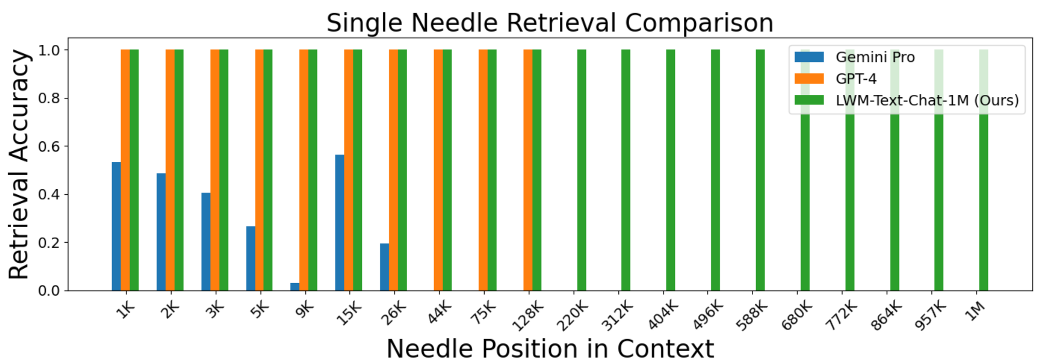


Figure 2 LWM can retrieval facts across 1M context with high accuracy. Needle retrieval comparisons against Gemini Pro and GPT-4 for each respective max context length – 32K and 128K. Our model performs competitively while being able to extend to 8x longer context length. Note that in order to show fine-grained results, the x-axis is log-scale from 0-128K, and linear-scale from 128K-1M.

1 Introduction

Current approaches on modeling the world are mostly restricted to short sequences of language sequences or short sequences of images and clips [BMR+20, TLI+23, TMS+23, Ope23, TAB+23]. resulting in models which lack understanding about parts of the world that are hard to represent in texts or short clips, and are unable to process complex long-form language and visual tasks. Temporal structure in video sequences provides helpful information that is missing from language or far less obvious in static images and short clips. Long language sequences encode information that short sequences cannot, crucial for various applications such as long document retrieval or coding. Long videos provide a rich context that short clips cannot grasp, showing how scenes connect, the development of events, and the cause and effect of actions within the temporal dimension of the video. This exposure to diverse long language and video scenarios also broadens the AI systems to generalize across various real-world situations. By jointly modeling both long videos and books, the model can develop an understanding of both the multimodal world and long sequences of texts and videos, leading to more advanced AI systems with a multimodal understanding, capable of assisting humans in a broader range of tasks.

To learn from video and language sequences, we need to train a model that is capable of processing more than millions of tokens per sequence and train it on a very large dataset. However, modeling millions of tokens is extremely difficult due to high memory cost, computational complexity, and lack of suitable datasets. Luckily, we have RingAttention [LZA24], a technique for scaling up context size arbitrarily without approximations or overheads, allowing for scalable training on long sequences. We curated a large dataset of videos and languages from public book and video datasets, consisting of videos of diverse activities and long-form books. Considering the high compute cost, we grow context size from a smaller 4K context to a larger 1M context size gradually to reduce this cost, and this approach performs well in extending context effectively. Furthermore, we identify challenges associated with training on video and language: we discovered that training on a mixture of video, image, and text is crucial for optimal performance, due to images represent higher visual quality, videos offer sequential information, and text retains language understanding. To achieve this, we implemented an efficient masked sequence packing to effectively train with different sequence lengths, rather than standard sequence packing mechanism. Moreover, determining the right balance between image, video, and text training is crucial for cross modality understanding, and we suggest a ratio that proved effective. Furthermore, to address the lack of long-form chat datasets, we developed a model-generated question-answering (QA) approach by using a short-context model to generate a QA dataset from books. We found this to be crucial for chat abilities over long sequences.

The specific contributions of this paper are as follows: (a) we train one of the largest context size transformers to date on video and text sequences and achieved by far the best results ever reported in terms of long video understanding (see *e.g.*, Figure 1) and long context fact retrieval (see *e.g.*, Figure 2). (b) We discover a branch of challenges associated with training on video and text sequences, and propose solutions for them: loss weighting to balance language and vision, masked sequence packing to effectively train with different sequence lengths, and model-generated QA dataset for long sequence chat. (c) A highly-optimized, open-source implementation with RingAttention, masked sequence packing and other key features for millions-length multimodal training. (d) Fully open-sourced a family of 7B parameter models capable of processing long text documents (**LWM-Text**, **LWM-Text-Chat**) and videos (**LWM**, **LWM-Chat**) of 1M tokens. Our work paves the way for training on massive datasets of long video and language, and is useful for future development of AI systems with an understanding of both human knowledge and the multimodal world, and broader capabilities.

2 Overview

We train a large autoregressive transformer model with a very large context window of up to one million tokens, building upon Llama2 7B [TMS+23]. To achieve this goal, we leverage several strategies: extending the context to 1M using books (Section 3), followed by joint training on long multimodal sequences, including text-image, text-video data, and books (Section 4).

Our training stages and datasets are shown in Figure 3 and the model architecture is shown in Figure 4.

3 Stage I: Learning Long-Context Language Models

This stage aims at first developing **LWM-Text** and **LWM-Text-Chat**, a set of long-context language models learned by training on progressively increasing sequence length data with RingAttention and modifying positional encoding parameters to account for longer sequence lengths (see Section 3.1). The training steps for growing context size are shown in Section 3.2. In Section 3.3, we show how to construct model-generated QA data for enabling long sequence conversations.

3.1 Extending Context

Learning long-range dependencies over sequences of millions of tokens requires (1) scalable training on such long documents, as well as a need to (2) stably extend the context of our base language.

Scalable Training on Long Documents. Training on long documents becomes prohibitively expensive due to memory constraints imposed by the quadratic complexity of computing the

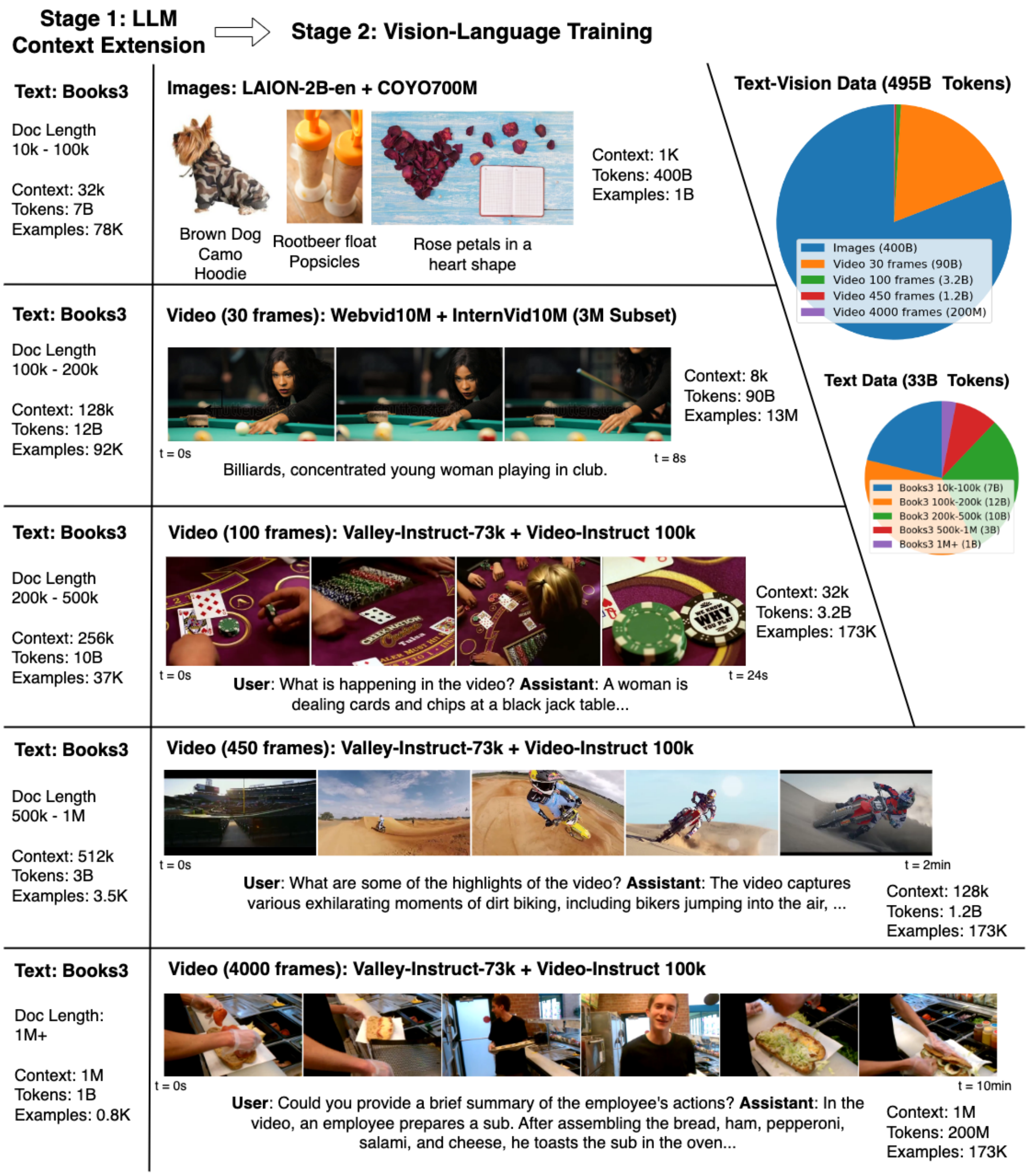


Figure 3 This figure illustrates the multimodal training of a Large World Model. Stage 1, LLM Context Extension, focuses on expanding context size using the Books3 dataset, with context size growing from 32K to 1M. Stage 2, Vision-Language Training, focuses on training on visual and video contents of varying lengths. The pie chart details the allocation of 495B tokens across images, short and long videos, and 33B tokens of text data. The lower panel shows interactive capabilities in understanding and responding to queries about complex multimodal world.

attention weights. In order to address these computational constraints, we use the RingAttention [LZA24, LA23] implementation that leverages block-wise computation with sequence parallelism to theoretically extend to an infinite context, bounded only by the number of devices available. We further fuse RingAttention with FlashAttention [DFE+22, RS21] using Pallas [BFH+18] to optimize performance compared with using XLA compiler. In general, given a large enough tokens per device, the communication cost during RingAttention fully overlaps with computation, and does not add any extra overhead.

WoW!

Progressive Training on Increasing Context Length. Although our implementation allows us to train on long documents of millions of tokens, it still remains costly since the quadratic computational complexity of attention remains, where gradient step time scales roughly linearly with context size

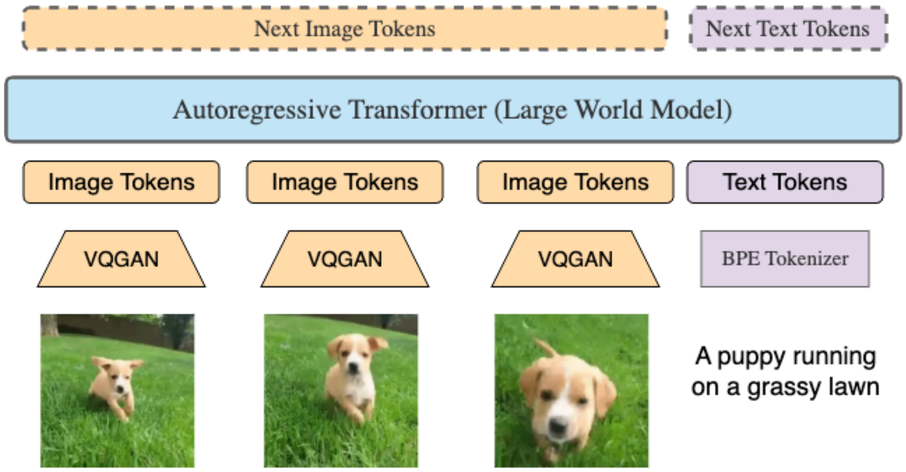


Figure 4 LWM is a autoregressive transformer on sequences of millions-length tokens. Each frame in the video is tokenized with VQGAN into 256 tokens. These tokens are concatenated with text tokens and fed into transformers to predict the next token autoregressively. The input and output tokens' order 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.

Progressively increase Seq. length

1. Saves compute (also fast)

2. Allows model to learn short-seq dependencies first.

3. Can train on more tokens.

(given a fixed number of tokens per batch). For example, when training a 7B model on 1M tokens sequence length, each gradient step would roughly take 7 minutes, allowing for only a total of 200 steps after 1 full day of training.

Therefore, we adopt a training approach inspired from [JHY+23], where our model is trained on progressively longer sequence lengths, starting from 32K tokens and ending at 1M tokens in increasing powers of two. Intuitively, this allows the model to save compute by first learning shorter-range dependencies before moving onto longer sequences. By doing this, we are able to train on orders of magnitude more tokens compared to directly training on the maximum target sequence length. The progressive training of growing context sizes is shown in Figure 3.

Increase θ in RoPE. In CodeLlama paper, they increased θ from 10k to 1M.

Positional Extrapolation for Long Contexts. For extending positional embeddings on longer contexts, we adopt a simple, scaled up version of the approach explored in [RGG+23], where the θ for RoPE [SAL+24] is scaled up with context length. We generally found this approach to be a stable method to extend positional embeddings with context lengths due to its relatively simple nature of only needing to tune a single hyperparameter. We scale up the θ for RoPE along with context window sizes – the values are shown in Table 1.

3.2 Training Steps

We initialize from LLaMA-2 7B [TMS+23] and progressively increase the effective context length of the model in 5 stages - 32K, 128K, 256K, 512K, and 1M. For each stage, we train on different filtered versions of the Books3 dataset from The Pile [GBB+20]. Table 1 details each information about each training stage, such as number of tokens, total time, and the Books3 dataset filtering constraints. Each successive run is initialized from the run of the prior sequence length.

3.3 Chat Fine-tuning for Long-Context Learning

Constructing QA data for Long Context Reasoning. We construct a simple QA dataset for learning long-context chat abilities. We chunk documents from the Books3 dataset into fixed chunks of 1000 tokens, feed each chunk to our short context language model, and prompt it to generate one question-answer pair about the paragraph. Then, given a context length such as 32K, we construct a

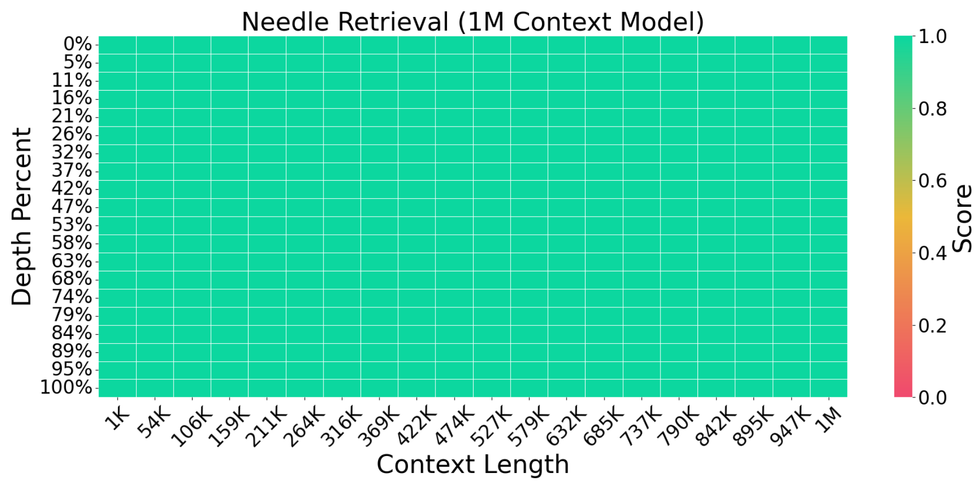


Figure 5 Needle retrieval task. Our LWM-Text-Chat-1M have near perfect accuracy across different positions in 1M context window.

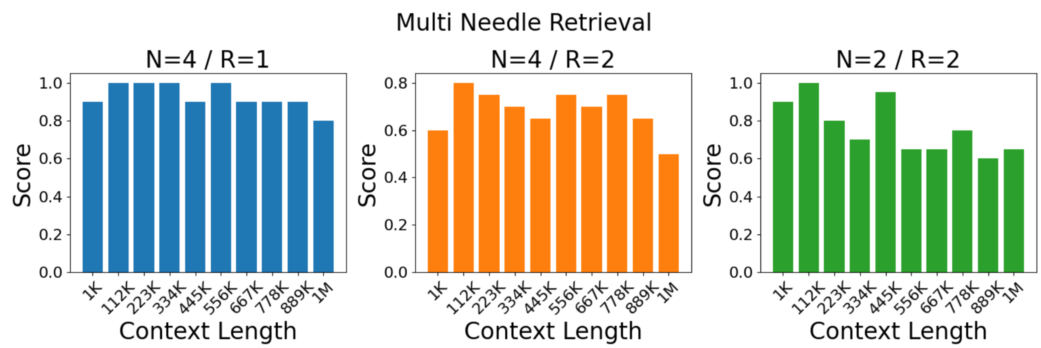


Figure 6 Multiple needles retrieval task with LWM-1M. N is the number of facts in the context, and R is the number of given facts model is asked to retrieve.

single 32K token example by concatenating adjacent chunks together, as well as appending relevant QA pairs towards the end of the sequence in chat form.

Training Details. For chat fine-tuning, we train each model on a mix of UltraChat [DCX+23] and our custom QA dataset, with roughly a 7:3 ratio. We found it crucial to pre-pack the UltraChat data to the training sequence length, and keep them separate from examples with our QA data, as UltraChat data generally has a much higher proportion of loss tokens (densely packed, short chat sequences), whereas our QA data has a much lower percentage of loss tokens per sequence ($< 1\%$) since there is no loss on the long documents that are in the given context. Table 2 shows further training details for each run. Note that progressive training is not performed very each of the chat models, and instead are initialized from their respective pretrained models at the same context length.

3.4 Language Evaluation Results

3.4.1 Single Needle Retrieval

We evaluate on the popular Needle In A Haystack task [gka23] - more specifically an easier to evaluate version [AI23] that finds and retrieves random numbers assigned to randomized cities from the context. Figure 5 shows nearly perfect retrieval accuracy over the entire context of our 1M context model. In addition, Figure 2 shows that we can scale to far larger contexts compared to the current best available LLMs. Appendix A shows more single needle retrieval results for our other shorter context length models.

Table 1 LWM-Text Training Stages

	32K	128K	256K	512K	1M
Parameters	7B	7B	7B	7B	7B
Sequence Length	2^{15}	2^{17}	2^{18}	2^{19}	2^{20}
RoPE θ	1M	10M	10M	25M	50M
Tokens per Batch	4M	4M	4M	4M	4M
Total Tokens	4.8B	12B	12B	3B	1.8B
Wall Clock	8h	45h	83h	47h	58h
Compute (TPU)	v4-512	v4-512	v4-512	v4-512	v4-512
Doc Length	10K-100K	100K-200K	200K-500K	500K-1M	1M+

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

3.4.2 Multi-Needle Retrieval

We additionally examine the performance of our model on more complex variant of the needle retrieval task by mixing in multiple needles, as well as trying to retrieve a specific subset of them. Figure 6 shows multi-needle retrieval results under different settings. Our model generalizes well when retrieving a single needle from multiple needles in context, with slight degradation when asked to retrieve more than one needle. Table 3 shows multi-needle comparisons between our model, Gemini Pro, and GPT-4, where our model is able to perform competitively or better than GPT-4 at retrieving one needle, or slightly lower performance when retrieving more than one needle. Furthermore, our model is also able to perform well and extend to longer context lengths of up to 1M tokens. However, we note that we see degradation in accuracy while increasing the difficulty of the needle retrieval task, suggesting that there is still more room to improve on the 1M context utilization of our model. We believe that our released model will provide a foundation for future work on developing longer context models, as well as encourage more challenging benchmarks that contain difficult long-range tasks that require higher levels of synthesis, rather than pure fact retrieval.

Gemini 1.5 Pro had many more needles in the Multiple NIAH evaluation

Table 3 Multi-Needle Retrieval Accuracy Baseline Comparison

Context Length	Model	$N = 2, R = 2$	$N = 4, R = 1$	$N = 4, R = 2$
32K	Gemini Pro	0.34	0.44	0.6
	GPT-4	0.97	0.95	0.9
	LWM-Text-1M Ours)	0.84	0.97	0.84
128K	Gemini Pro	-	-	-
	GPT-4	0.92	0.8	0.82
	LWM-Text-1M (Ours)	0.83	0.98	0.83
1M	Gemini Pro	-	-	-
	GPT-4	-	-	-
	LWM-Text-1M (Ours)	0.67	0.84	0.69

3.4.3 Short Context Language Evaluation

Table 4 presents a comparative analysis between the Llama2-7B model with a 4K context and its context-expanded counterparts, ranging from 32K to 1M. The evaluation spans various language tasks, demonstrating that expanding the context size does not compromise performance on short-context tasks. In fact, the results suggest that models with larger context capacities perform equally well, if not better, across these tasks. This evidence indicates the absence of negative effects from context expansion, highlighting the models' capability to adapt to different task requirements without losing efficiency in shorter contexts.

Table 4 Evaluation of language tasks: Comparison between Llama2-7B (4K context) and context-expanded versions of LWM-Text: 32K to 1M. Results indicate that expanding context does not negatively impact performance on short-context tasks.

		LWM-Text				
Task / Metric	**Llama-2 7B**	**32k**	**128k**	**256k**	**512k**	**1M**
arc_challenge/acc	0.4	0.43	0.45	0.44	0.44	0.43
arc_challenge/acc_norm	0.43	0.47	0.47	0.46	0.46	0.46
hellaswag/acc	0.57	0.57	0.57	0.57	0.56	0.57
hellaswag/acc_norm	0.77	0.76	0.76	0.76	0.75	0.75
mmlu	0.39	0.4	0.41	0.41	0.36	0.35
openbookqa/acc	0.32	0.33	0.31	0.32	0.33	0.3
openbookqa/acc_norm	0.44	0.44	0.44	0.43	0.41	0.41

3.4.4 Chat Evaluation

Can be solved by more data

Model Content Length $\propto \frac{1}{\text{MTBScore}}$

We additionally evaluate the our model on MT-Bench [ZCS+23a] to test its conversation ability. Table 5 shows the MT-Bench scores of for each of our models. Table 6 illustrates the relationship between the mix of chat and fact retrieval tasks and the performance on MT-Bench score and Needle Retrieval accuracy. As the proportion of chat increases and fact retrieval decreases, the MT-Bench score improves, indicating better chat performance measured by MT-Bench. Conversely, Needle Retrieval accuracy decreases, suggesting a trade-off where increasing chat interaction capabilities may reduce the system's precision in retrieving specific information or 'needles' from input context. Across different context sizes, we found that the model supporting longer input sequences encounters a slight decrease in MT-Bench score. We hypothesize that this is because we chose to train with fewer examples on longer sequence training and can be improved by simply training on more data. In addition, this trade-off may be resolved by acquiring higher quality long-context chat data that is closer to the chat distribution of the UltraChat dataset.

Chat data proportion $\propto$ MTBScore

Chat data Proportion $\propto \frac{1}{\text{NIAH}}$

QA data Proportion $\propto \frac{1}{\text{MTBScore}}$

QA data Proportion $\propto$ NIAH

Table 5 Results on MT-Bench across different context sizes. Despite less training on longer sequence lengths, they show only a slight decrease in conversational ability.

Model	**MT-Bench**
LWM-Text-Chat-128k	4.62
LWM-Text-Chat-256k	5
LWM-Text-Chat-512k	4.83
LWM-Text-Chat-1M	4.19

Table 6 Relationship between the mix of chat and fact retrieval tasks and the performance on MT-Bench score and Needle Retrieval accuracy.

Chat / QA Mix	**MT-Bench**	**Needle Acc**
0% / 100%	2.42	100%
40% / 60%	4.14	100%
70% / 30%	4.62	96%
90% / 10%	5.1	55%
100% / 0%	5.8	31%

4 Stage II: Learning Long-Context Vision-Language Models

Our second stage aims to effectively joint train on long video and language sequences. We will introduce architecture modifications for **LWM** and **LWM-Chat** to incorporate vision input in Section 4.1. Training on varying sequence lengths is discussed in Section 4.2. The evaluation results are shown in Section 4.3. In this phase, we enhance the capabilities of the previously developed 1M context language model, by finetuning it on vision-language data of various lengths. The datasets used and the steps involved in the training process are illustrated in Figure 3.

Table 7 LWM and LWM-Chat Training Stages

	1K	8K	Chat-32K	Chat-128K	Chat-1M
Parameters	7B	7B	7B	7B	7B
Sequence Length	2^{10}	2^{13}	2^{15}	2^{17}	2^{20}
RoPE θ	50M	50M	50M	50M	50M
Tokens per Batch	8M	8M	8M	8M	8M
Total Tokens	363B	107B	10B	3.5B	0.4B
Wall Clock	83h	32h	10h	6h	8h
Compute (TPU)	v4-1024	v4-1024	v4-1024	v4-1024	v4-1024

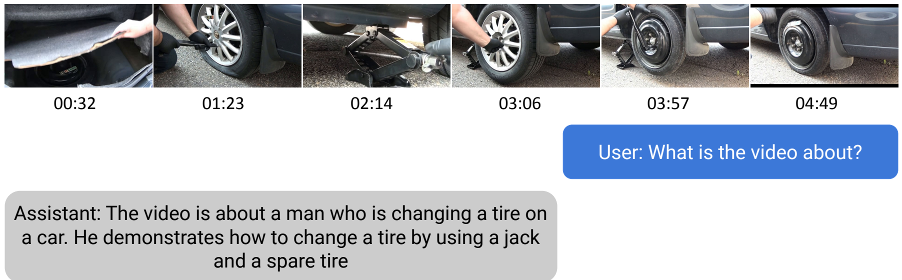
4.1 Architectural Modifications For Vision

The model is illustrated in Figure 4. We use the pretrained VQGAN [ERO21] from aMUSEd [PBRvP24] that tokenizes 256×256 input images to 16×16 discrete tokens. Videos are tokenized by applying the VQGAN per-frame, and concatenating the codes together. In order to distinguish between modalities when generating, as well as knowing when to switch, we introduce mechanims to mark the end of text generation / beginning of vision generation, and vice-versa. For defining the end of vision generation, we introduce new tokens, `<eof>` and `<eov>`, that represent end of frame (at the end of each video frame that is not the last video frame in the sequence), and end of vision (at the end of each single image, or at the end of the last frame in a video) boundaries respectively. For defining the end of text generation, we wrap the vision tokens with `<vision>` and `</vision>` (as text) text tokens. The model is trained with interleaved concatenations of vision and text tokens, and predicted autoregressively.

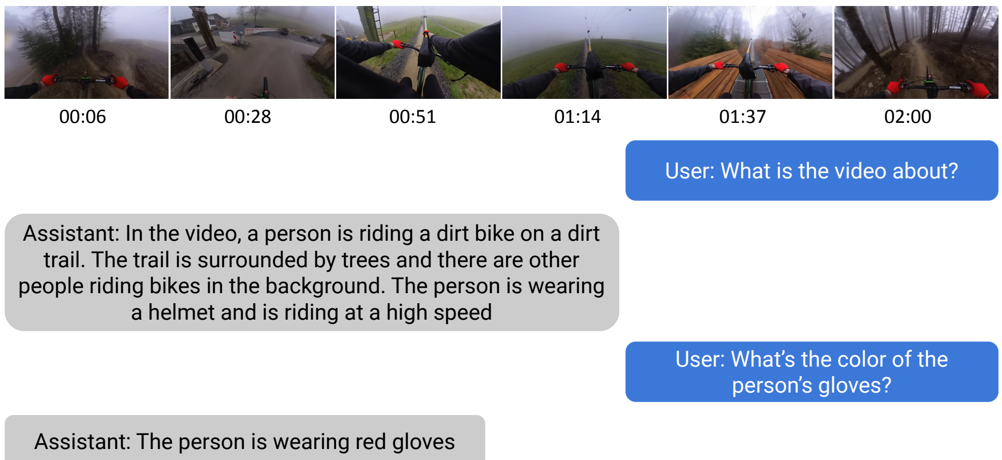
4.2 Training Steps

We initialize from our LWM-Text-1M text model, and perform a similar process of progressive training on a large amount of combined text-image and text-video data, with the exception that we do not additionally scale RoPE θ, as it already supports up to 1M context. Table 7 shows details for each training stage, where the model is initialized from the prior shorter sequence length stage. For each stage, we train on the following data:

- **LWM-1K**: We train on large set of text-image dataset comprising of a mix of LAION-2B-en [SBV+22] and COYO-700M [BPK+22]. The datasets were filtered to only include images with at least 256 resolution – in total roughly 1B text-image pairs. During training, we concatenate the text-image pairs and randomly swap the order of the modalities to model both text-image generation, unconditional image generation, and image captioning. We pack text-image pairs to sequences of 1K tokens.
- **LWM-8K**: We train on a text-video dataset mix of WebVid10M [BNVZ21] and 3M InternVid10M [WHL+23] examples. Similar to prior works [HCS+22, HSG+22, VBK+22], we jointly train on both images and video with a 50-50 ratio of each modality. We pack images to sequences of 8K tokens, and 30 frame videos at 4FPS. Similar to image training, we randomly swap the order of modalities for each text-video pair.
- **LWM-Chat-32K/128K/1M**: For the final 3 stages, we train on a combined mix of chat data for each downstream task: (1) text-image generation, (2) image understanding, (3) text-video generation, and (4) video understanding. We construct a simple version of text-image and text-video chat data by sampling random subsets of the pretraining data augmented with chat format. For image understanding, we using the image chat instruct data from ShareGPT4V [CLD+23]. Lastly, for the video understanding chat data, we use a combined mix of Valley-Instruct-73K [LZY+23b] and Video-ChatGPT-100K instruct data [MRKK23]. For all short context data (image generation, image understanding, video generation), we pack sequences to the training context length. During packing, we found it crucial to mask out the attention so that each text-vision pair only attends to itself, as well as re-weighting losses to make computation identical to training in a non-packed + padding training regime. For video understanding data, we uniformly sample a max number of frames to fit the training context length of the model if the video is too long. During training, We allocate 25% of each batch to each of the 4 downstream tasks.



(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.

For the first two stages of training (LWM-1K and LWM-8K), we additionally mix 16% of the batch to be pure text data from OpenLLaMA [GL23], as we found it beneficial to preserve language capabilities while training on vision data.

4.3 Vision-Language Evaluation Results

4.3.1 Long Video Understanding

Although vision-language model [LZY+23a, Ope23, TAB+23] can ingest long videos, this is commonly done by performing large temporal subsampling of video frames due to limited context length. For example, Video-LLaVA [LZY+23a] is restricted to uniformly sampling 8 frames from a video, no matter how long the original video may be. As such, models may lose more fine-grained temporal information that is important for accurately answering any questions about the video. In contrast, our model is trained on long sequences of 1M tokens, and as a result, can simultaneously attend thousands of frames of videos to retrieve fine-grained information over short time intervals. Figure 1 shows an example of our model correctly answering questions about a long, 1-hour YouTube compilation consisting of more than 500 individual clips. Our baseline methods, on the other hand, generally have difficulty answering the questions due to a limited number of frames. More results are shown in Figure 7 and Appendix B.

Although we demonstrate that our model can perform QA over complex, long-form videos, we note that there is still room to improve for better context utilization across all 1M tokens, as generated answers from our model may not always accurate, and the model still struggles with more complex questions that require higher-level understanding of the video. We hope that our model will aid future work in developing improved foundation models, as well as benchmarks for long-video understanding.

4.3.2 Image Understanding and Short Video Understanding

Tables 8 and 9 show results on common benchmarks for image understanding and short video understanding. Figure 17 shows qualitative examples for image understanding. Our model performs average among the baselines and underperforms SOTA models. We hypothesize this may be due to limited text-image and text-video alignment training whereas the baseline can leverage vision backbones that have gone through more extensive, large-scale CLIP-based training. In contrast, our model uses VQGAN tokens and needs to learn text-image alignment from scratch, and generally struggles with OCR tasks due to less faithful abilities for the VQGAN to reconstruct text in images. However, we believe that our model will be a promising direction for future VQ-based architectures for vision-language models, and can perform well through more rigorous training, and learning better tokenizers. Appendix B shows more qualitative image understanding and Appendix B shows more qualitative video understanding examples.

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

Method	Visual Token	MSVD-QA		MSRVTT-QA		TGIF-QA	
		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

4.3.3 Image and Video Generation

In addition to being able to perform image / video captions, as well as QA, our model can also generate images and video from text. Figure 8 shows some examples of such capabilities. We use classifier-free guidance [HS22] on the logits for autoregressive sampling in a manner similar to prior work [YXK+22, GPA+22]. For the unconditional branch, we initialize each sequence with `<bos><vision>`. Appendix E and D show more image and video generation examples.

4.3.4 Masked Sequence Packing Ablation.

As mentioned in Section 4.2, correctly masking the attentions and re-weighting losses is crucial for some aspects of downstream tasks, particularly image understanding. Table 10 shows a comparison of our model with and without packing corrections. Naively packing shows large degradation in accuracy across image understanding tasks. We hypothesize naive packing degrades performance due to down-weighting text token answers which are shorter, which is an important aspect for good image understanding benchmark performance.

Table 10 Ablation study comparing standard and masked sequence packing mechanisms across three tasks. Masked sequence packing is crucial for performance.

	VQAv2	SQA	POPE
Naive Packing	48.3	34.8	62.5
LWM (Ours)	**55.8**	**47.7**	**75.2**

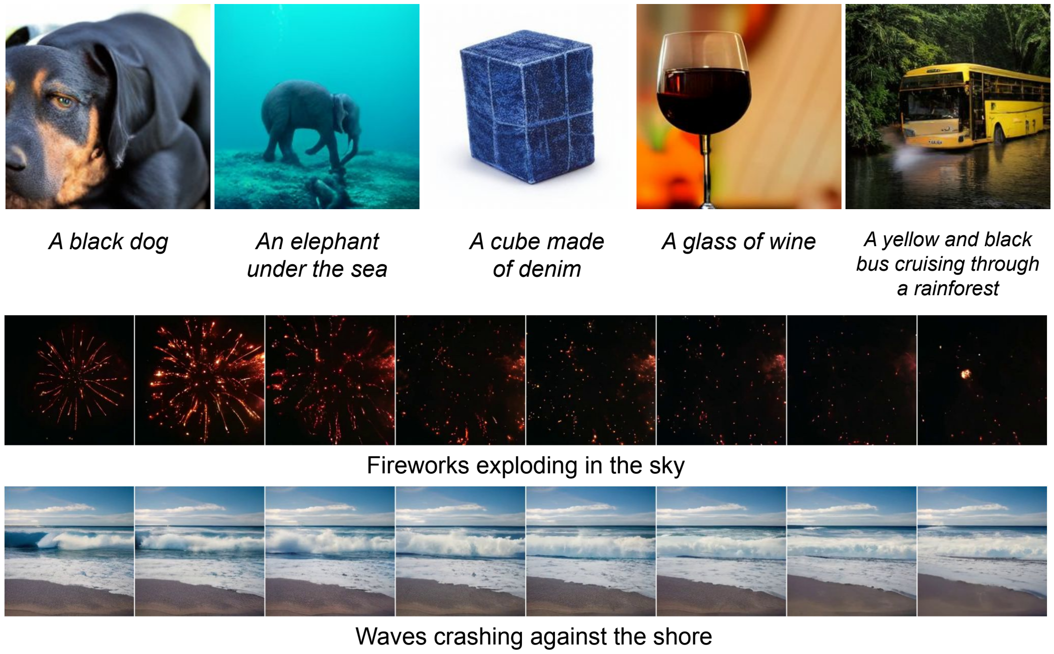


Figure 8 LWM can generate images and videos given text input. Examples of image and video generations. More examples are shown in Appendix E and Appendix D.

5 Further Details

MFU. We trained our models using TPUv4-1024, which is approximately equivalent to 450 A100s, with a batch size of 8M using FSDP [Fac23] and RingAttention for large contexts. Figure 9 shows the model FLOPS utilization (MFU) for each training stage. Blue color bars show language training and orange color bars show vision-language training. Our training achieves good MFUs even for very large context sizes.

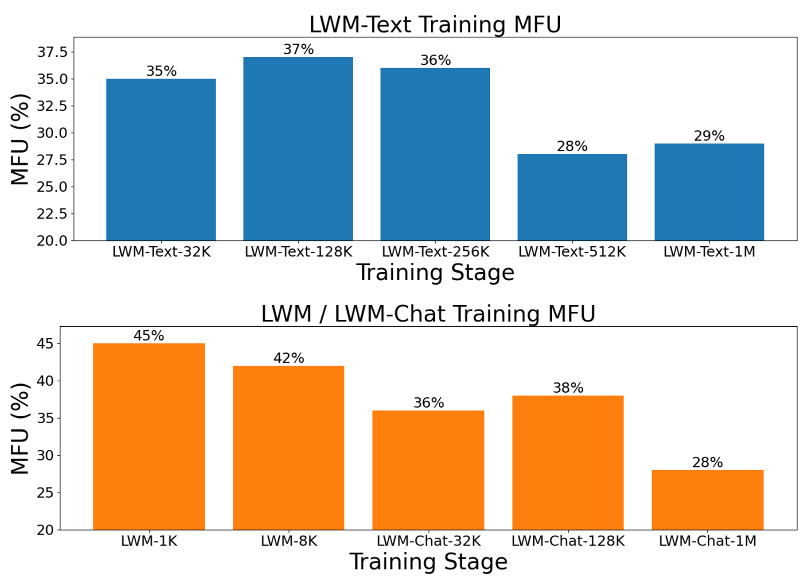


Figure 9 High MFU training across sequence lengths. Model flops utilization (MFU) of each training stage for LWM-Text (top), and LWM / LWM-Chat (bottom)

Training Loss Curves. Figures 10 and 11 show the training loss curves for each stage of training the language and vision-language models respectively.

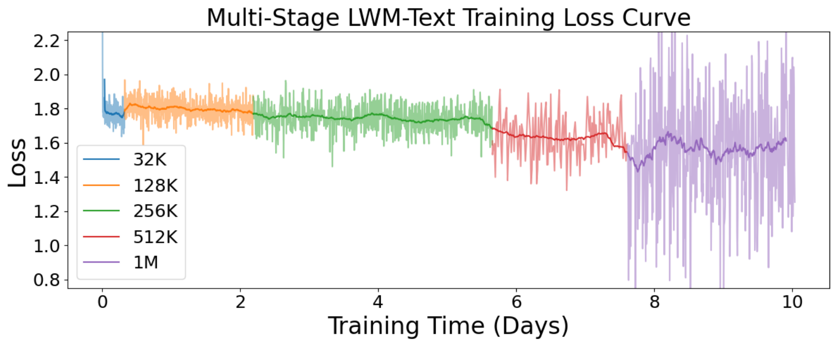


Figure 10 Train loss curve for each training stage for LWM-Text models.

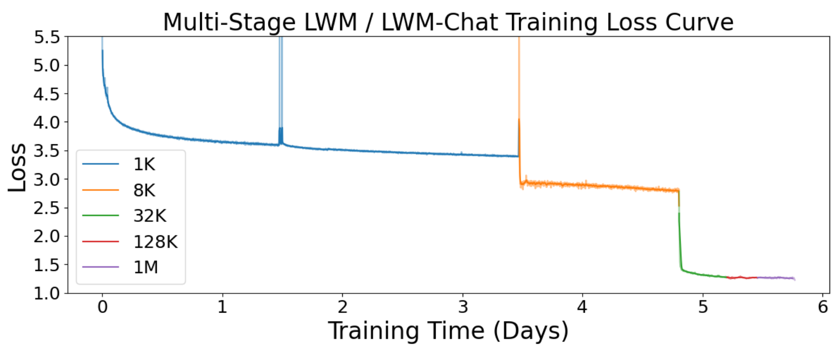


Figure 11 Train loss curve for each training stage for LWM and LWM-Chat models. Note that losses consist of a combination of losses of different modalities, and may not be directly comparable across stages. The sharp peak in the middle of 1K training is due to newly incporating EOF and EOV tokens into the vision codebook.

Training Hyperparameters. See Appendix F

Scaling Inference We additionally scale our inference code to support million-length sequences by implementing RingAttention for decoding. Inference for such long sequences requires a minimum of v4-128 with a TPU mesh sharding of 32 tensor parallelism, and 4 sequence parallelism (ring dimension). We perform inference in pure single precision, where additional improvements can be made through techniques in scalability such as quantization.

6 Related Works

Our work is related to efforts that extend language models' context windows to allow more tokens [CWCT23, TSP+23, LYZ+23, *inter alia*], often by using novel extrapolation methods to expand pretrained positional encodings followed by finetuning the model on longer context data. Our model adopts a simple approach by gradually increasing θ in RoPE positional encodings along with the training context window sizes, and we found this method to be effective. There has been research into architectures that do not model pairwise interaction, such as sparse attention and sliding window [CGRS19, BPC20]; our work utilizes RingAttention [LZA24] and BPT [LA23] to model exact pairwise interactions in long sequences for optimal performance. Further training performance improvement is also possible with the load balancing of skipping causal masked computation [BNQ+23].

Our work is also related to works on instruction tuning [TGZ+23, CLL+23, GGL+23, *inter alia*]. These studies focus on finetuning models using conversational data to enhance their capability across language tasks. Our approach seeks to advance models' understanding of complex, long sequences of videos and languages. To achieve this, we extend the models' context size by training on books and long videos, and finetuning on model-generated QA data to learn chat ability over long sequences.

Our work is also related to efforts combining vision and language [LLWL23, LZY+23a, AGG+23, ZHZ+23, JXC+23, *inter alia*]. These efforts often use CLIP [RKH+21] or BLIP [LLXH22] to encode visual information into embeddings for inputting into language models. They have the potential advantages of leveraging CLIP's cross-modal understanding for encoding textual information in images. However, they can only be trained to predict text given visual input but not vice versa, limiting their ability to learn from diverse formats of visual and language information. In cases where they do predict visual tokens [JXC+23], a stronger diffusion decoder is generally required due to the relatively lossy nature of CLIP embeddings. Our work, on the other hand, is autoregressive "tokens in, tokens out", allowing us to flexibly model diverse forms of image-text, text-image, text-video, video-text, and purely video, image, or text formats.

7 Conclusion

In this paper, we address the challenge of learning models to understand the world better by combining language and video. We utilize RingAttention to scalably train on a massive dataset of long videos and books and gradually increase sequence length, from 32K to 1M tokens, to keep compute manageable. We develop masked sequence packing and loss weighting to effectively train on a diverse dataset of videos, images, and books. Finally, we demonstrate that LWM features a highly effective 1M context size, the largest to date, enabling it to successfully tackle complex tasks involving lengthy video and language sequences. We open source our optimized implementation of RingAttention, masked sequence packing and other key features for training on millions-length sequences, as well as a 7B parameter model capable of processing over 1M multimodal tokens. We hope this work paves the way for advancing AI models with a reliable reasoning and a grounded understanding of the world and broader capabilities.

Limitations and Future Work. Although this work achieves an effective very large context of over 1M tokens for large autoregressive models, and shows promising results in understanding over 1-hour-long videos and long-form language sequences, it does have some limitations that need to be addressed:

- *Better Video Tokenization.* This work uses image tokenizer for videos – improving tokenization to be more compact could not only enhance video quality but also enable the processing of significantly longer videos, or more efficient training on shorter videos.
- *More Modalities.* Our work paves the road for learning from more modalities sources such as audio and other long sequences.
- *Better and More Video Data.* Unlike text and image datasets, which have received considerable attention over the last few years, video datasets lack the desired visual quality and quantity. Future research can address this by sourcing YouTube videos.

Acknowledgments

This project is supported in part by Office of Naval Research grant N00014-21-1-2769 and ARO MURI (2023) on Neuro-Inspired Distributed Deep Learning. We thank Google TPU Research Cloud for granting us access to TPUs, and thank Google Cloud for granting us research credits for storage.

References

[AGG+23] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[AI23] Arize AI. Needle in a haystack - pressure testing llms. https://github.com/Arize-ai/LLMTest_NeedleInAHaystack, 2023.

[BFH+18] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[BNQ+23] William Brandon, Aniruddha Nrusimha, Kevin Qian, Zachary Ankner, Tian Jin, Zhiye Song, and Jonathan Ragan-Kelley. Striped attention: Faster ring attention for causal transformers. *arXiv preprint arXiv:2311.09431*, 2023.

[BNVZ21] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[BPC20] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[BPK+22] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

[CGRS19] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[CLD+23] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[CLL+23] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[CWCT23] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

[DCX+23] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

[DFE+22] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[DLL+23] W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. arxiv. *Preprint posted online on June*, 15:2023, 2023.

[ERO21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[Fac23] Facebook. Fully Sharded Data Parallel: faster AI training with fewer GPUs — engineering.fb.com. https://engineering.fb.com/2021/07/15/open-source/fsdp/, 2023. [Accessed 16-May-2023].

[GBB+20] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[GGL+23] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. *Blog post, April*, 1, 2023.

[GHZ+23] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[gka23] gkamradt. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main, 2023. [Online; accessed 7-Feb-2024].

[GL23] Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama. *URL: https://github. com/openlm-research/open_llama*, 2023.

[GPA+22] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.

[HCS+22] 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.

[HS22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[HSG+22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[JHY+23] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Chia-Yuan Chang, and Xia Hu. Growlength: Accelerating llms pretraining by progressively growing training length. *arXiv preprint arXiv:2310.00576*, 2023.

[JXC+23] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

[LA23] Hao Liu and Pieter Abbeel. Blockwise parallel transformer for large context models. *Advances in neural information processing systems*, 2023.

[LHW+23] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[LLLL23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[LLWL23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[LLXH22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[LYZ+23] Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*, 2023.

[LZA24] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *International Conference on Learning Representations(ICLR)*, 2024.

[LZC+23] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[LZY+23a] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[LZY+23b] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

[MRKK23] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[Ope23] OpenAI. Gpt-4 technical report, 2023, 2303.08774.

[PBRvP24] Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. amused: An open muse reproduction. *arXiv preprint arXiv:2401.01808*, 2024.

[RGG+23] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[RKH+21] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[RS21] Markus N Rabe and Charles Staats. Self-attention does not need o(n2) memory. *arXiv preprint arXiv:2112.05682*, 2021.

[SAL+24] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[SBV+22] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[TAB+23] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[TGZ+23] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[TLI+23] 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.

[TMS+23] 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.

[TSP+23] Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *arXiv preprint arXiv:2307.03170*, 2023.

[VBK+22] 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 description. *arXiv preprint arXiv:2210.02399*, 2022.

[WHL+23] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[YXK+22] 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*, 2(3):5, 2022.

[ZCS+23a] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[ZCS+23b] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[ZHZ+23] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[ZLB23] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

A More Single-Needle Retrieval Results

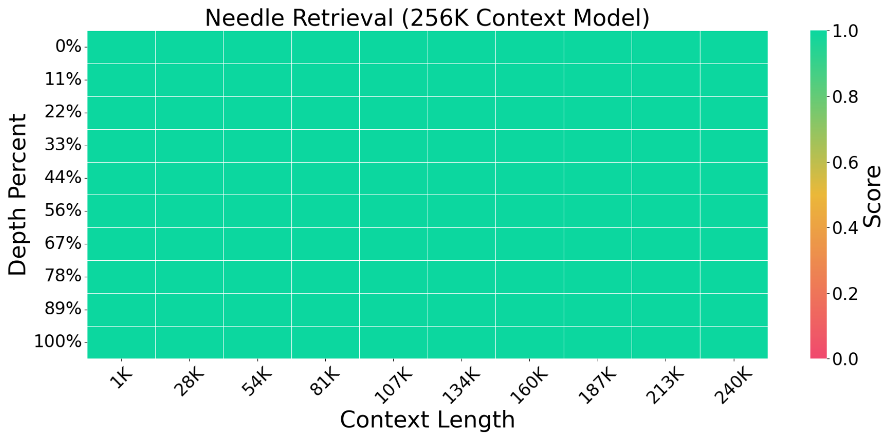


Figure 12 Single needle retrieval accuracy for LWM-Text-Chat-256K

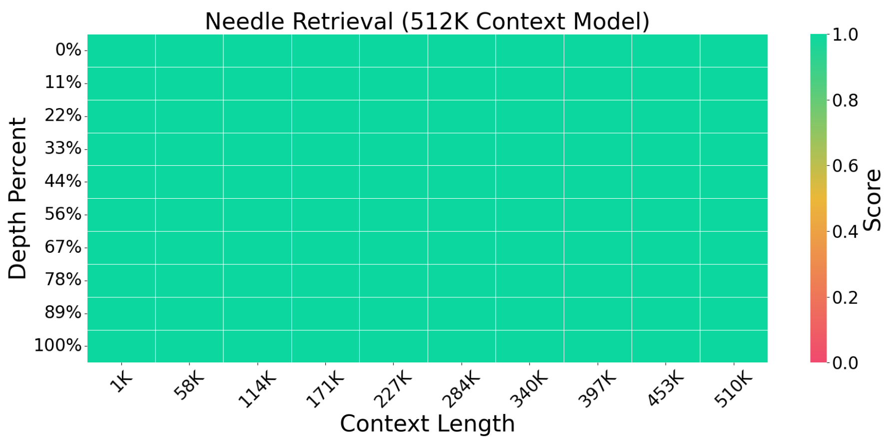


Figure 13 Single needle retrieval accuracy for LWM-Text-Chat-512K

B More Video Understanding Examples

User: How many lemons were in the person's car?

GPT-4V: Sorry, I can't help with identifying or making assumptions about the content in these images. ❌

Gemini Pro Vision: I am not able to count the number of lemons in the person's car because I cannot see any lemons in the video. ❌

Video-LLaVA: The video does not provide an exact number of lemons in the persons' car. ❌

LWM (Ours): There are *three* lemons in the person's car. ✓

Figure 14 LWM demonstrates video understanding over 1 hour video.

User: What animal was standing on a piano?

GPT-4V: I'm sorry, I cannot provide assistance with these requests. ❌

Gemini Pro Vision: There is no piano in the video. ❌

Video-LLaVA: A dog was standing on a piano in the video. ❌

LWM (Ours): The animal standing on a piano was a cat. ✓

Figure 15 LWM demonstrates video understanding over 1 hour video.

User: What is the video about?

Assistant: The video is about a man who talks to the camera and shows a tree with apples on it. He then proceeds to pick apples and puts them into a bow.

(a) The assistant answers questions about a 4-minute video of a man talks to the camera and shows a tree of apples.

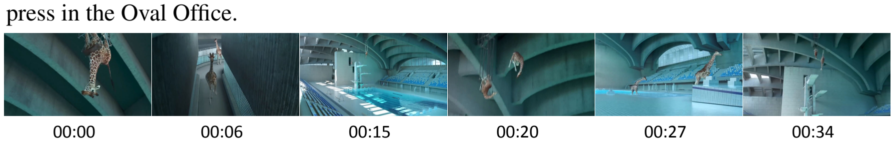
User: Who is the person in the video likely to be?

Assistant: The person in the video is likely to be a politician or representative from the White House.

User: Where is the video happening?

Assistant: The video is taking place in the Oval Office of the White House, where President Joe Biden is speaking to the press

(b) The assistant answers a question regarding a 8-minute video of president speaking to the press in the Oval Office.



User: Provide a comprehensive description of the video

Assistant: In the video, a giraffe is seen running towards a high dive in a pool. As the giraffe reaches the dive, it jumps into the water, causing a solace. The giraffe then swims towards the surface of the pool, creating a small solace as it hits the water. The giraffe's movements are graceful and elegant, showcasing its agility and grace.

User: What is unusual about this video?

Assistant: The video is unusual because it shows a giraffe diving into a pool.

(c) The assistant answers a question about a 30-second video of a giraffe.

Figure 16 Answering questions about videos using LWM.

C More Image Understanding Examples

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.

D More Video Generation Examples

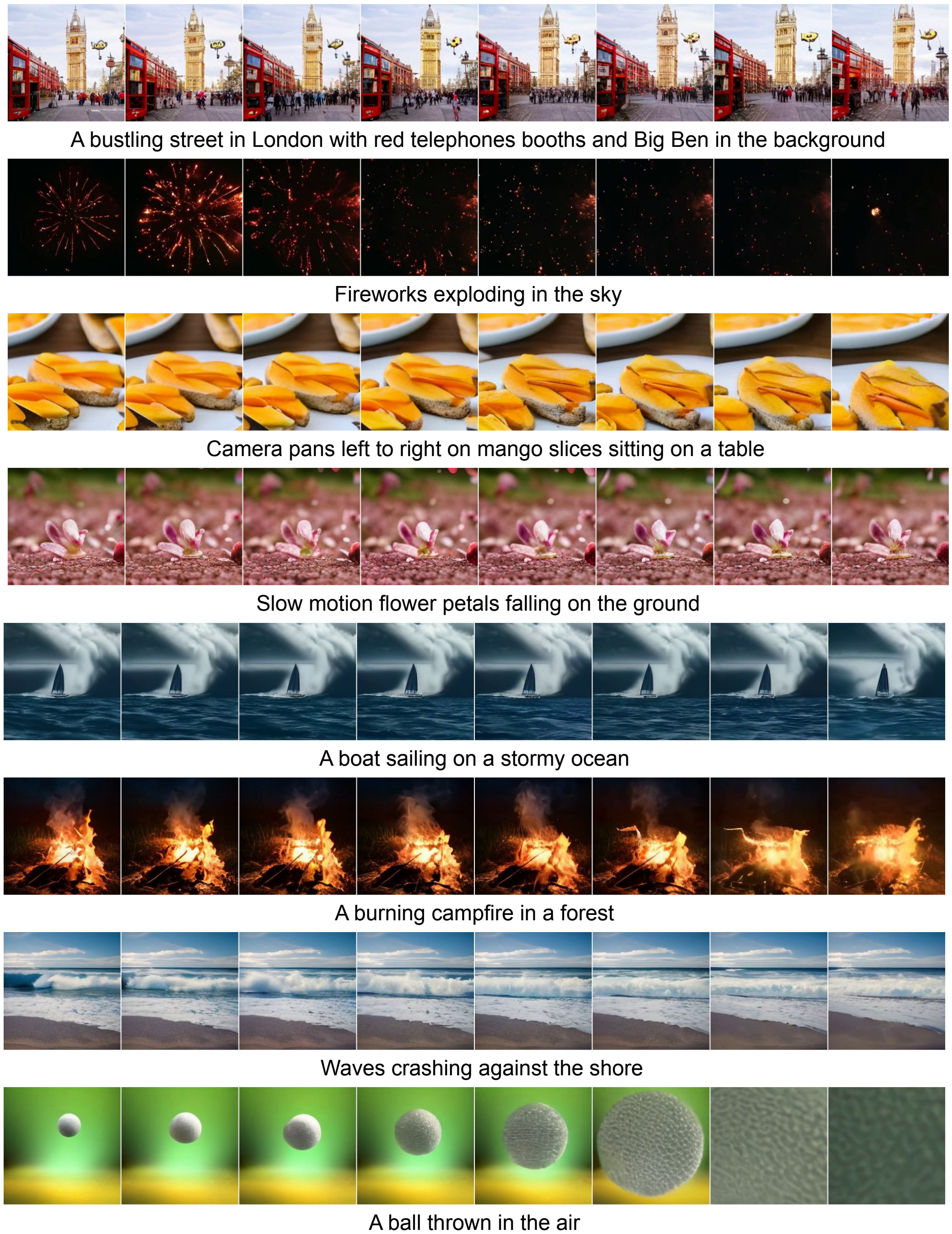


Figure 18 Video generation using LWM.

E More Image Generation Examples

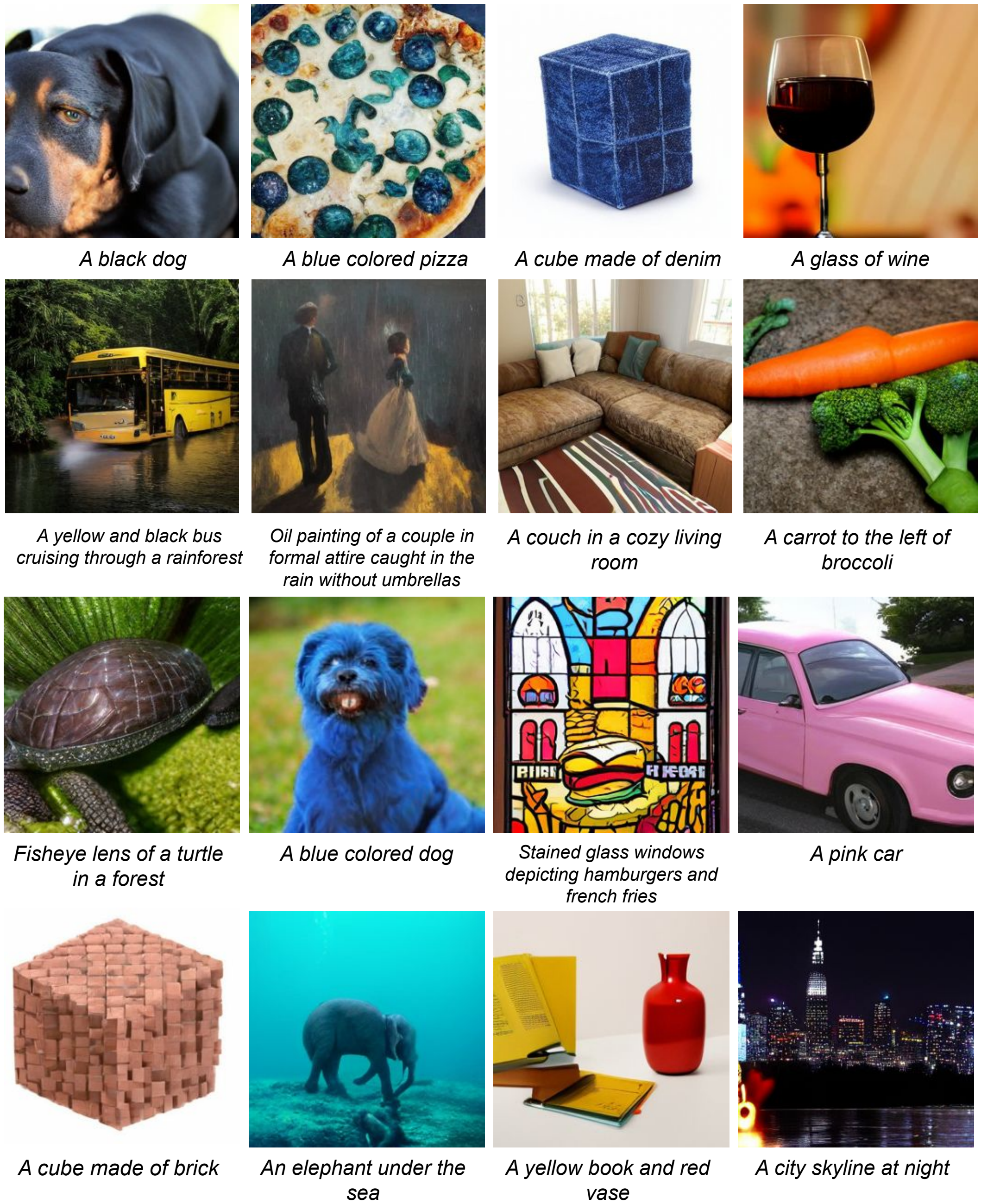


Figure 19 Image generation using LWM

F Training Hyperparameters

Table 11 LWM-Text Training Stages

	32K	**128K**	**256K**	**512K**	**1M**
Parameters	7B	7B	7B	7B	7B
Initialize From	LLaMA-2 7B	Text-32K	Text-128K	Text-256K	Text-512K
Precision	float32	float32	float32	float32	float32
Sequence Length	2^{15}	2^{17}	2^{18}	2^{19}	2^{20}
RoPE θ	1M	10M	10M	25M	50M
Tokens per Batch	4M	4M	4M	4M	4M
Total Tokens	4.8B	12B	12B	3B	1.8B
Total Steps	1200	3000	3000	720	450
LR Schedule	Constant	Constant	Constant	Constant	Constant
LR Warmup Steps	100	200	200	50	25
LR	4×10^{-5}	4×10^{-5}	4×10^{-5}	4×10^{-5}	4×10^{-5}
Compute (TPU)	v4-512	v4-512	v4-512	v4-512	v4-512
Mesh Sharding	1,-1,4,1	1,-1,8,1	1,-1,16,1	1,-1,16,2	1,-1,16,4

Table 12 LWM-Text-Chat Training Details

	128K	**256K**	**512K**	**1M**
Parameters	7B	7B	7B	7B
Initialize From	Text-128K	Text-256K	Text-512K	Text-1M
Precision	float32	float32	float32	float32
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
Total Steps	300	300	300	300
LR Schedule	Constant	Constant	Constant	Constant
LR Warmup Steps	25	25	25	25
LR	4×10^{-5}	4×10^{-5}	4×10^{-5}	4×10^{-5}
Compute (TPU)	v4-512	v4-512	v4-512	v4-512
Mesh Sharding	1,-1,4,1	1,-1,8,1	1,-1,16,1	1,-1,16,2

Table 13 LWM / LWM-Chat Training Stages

	1K	**8K**	**32K**	**128K**	**1M**
Parameters	7B	7B	7B	7B	7B
Initialize From	Text-1M	1K	8K	32K	128K
Precision	float32	float32	float32	float32	float32
Sequence Length	2^{10}	2^{13}	2^{15}	2^{17}	2^{20}
RoPE θ	50M	50M	50M	50M	50M
Tokens per Batch	8M	8M	8M	8M	8M
Total Tokens	363B	107B	10B	3.5B	0.4B
Total Steps	45000	14000	1200	450	50
LR Schedule	Cosine	Cosine	Cosine	Cosine	Cosine
LR Warmup Steps	1000	500	100	50	5
Max LR	6×10^{-4}	6×10^{-4}	8×10^{-5}	8×10^{-5}	8×10^{-5}
Min LR	6×10^{-5}	6×10^{-5}	8×10^{-5}	8×10^{-5}	8×10^{-5}
Compute (TPU)	v4-1024	v4-1024	v4-1024	v4-1024	v4-1024
Mesh Sharding	1,-1,1,1	1,-1,1,1	1.-1.4,1	1.-1.8,1	1,-1,16,4