Appendix.

7. Extended Related Work

Vision Language Model Architecture Vision language models (VLMs) primarily adopt two paradigms to process visual input. The first paradigm freezes language model weights and integrates visual information via cross-attention mechanisms [1], while the second paradigm utilizes a pre-trained image encoder, such as CLIP [50] or SigLIP [75], to convert images into tokens. These tokens are then concatenated with text tokens and input into the language model [13, 35, 40]. This approach can be naturally extended to video understanding by treating videos as sequences of images processed by the vision encoder [9, 29]. To enhance video processing, some works introduce specialized video encoders. For instance, InternVideo [64, 65] uses VideoMAE [58] as a video encoder, while Kangaroo [41] integrates depth-wise 3D convolution for fusing video tokens. In this work, we retain SigLIP as the vision encoder and focus on enhancing long video understanding by incorporating a linear-complexity temporal module in the Mamba [18] architecture. Positioned between the SigLIP vision encoder and the language model, this module efficiently improves spatial-temporal modeling effectiveness.

Long Video Understanding Understanding long videos with VLMs presents significant challenges in both accuracy and efficiency. Previous approaches have employed long-context language models trained on short-context video data to enable long video comprehension [76]. However, these methods lack sufficient long video training data and incur high computational costs during both training and inference as the number of frames increases. LongVILA [69] addresses these challenges through a multi-modal sequence parallelism system that directly handles long video data during training and inference, but this approach requires customized system implementations tailored for multi-GPU setups. Another line of research focuses on token reduction to shorten input sequences, thereby enabling efficient inference for long videos [33, 51, 53, 54, 66, 71]. For instance, VoCO-LLaMA [71] and VideoXL [54] use recursive KV cache compression learnt in an end-to-end manner, and LongVU [53] leverages DINO features for frame selection and inter-frame similarity to reduce tokens. Despite these diverse strategies, direct pooling along the temporal or spatial dimensions often performs sufficiently well, with additional gains being marginal. In this paper, we apply temporal and spatial pooling for token reduction, achieving superior performance when combined with our temporal projector.

Mamba for Video Understanding Recent advances in linear state space models such as Mamba [10, 18] have sparked extensive exploration in applying them to video understanding tasks. Due to their sub-quadratic computation complexity, Mamba models achieve significant efficiency improvements compared to transformer-based architectures while still delivering competitive performance [10, 18, 59, 60, 82]. These properties make Mamba particularly suitable for video processing as the models are required to process long sequence inputs. For example, VideoMamba in [31] and VideoMamba (identical model naming) in [48] use Mamba-based visual backbone in video models and demonstrates the model's strong ability to capture both local redundancy and long-term spatiotemporal dependencies. VideoMambaPro [43] proposes to improve Mamba's video understanding ability by applying masking and residual connection during the backward scan. The Video Mamba Suite [5] further explores various architectures to integrate Mamba into existing video understanding models, demonstrating favorable efficiency-performance trade-offs for long sequence inputs. Mamba-ND [32] aims to improve Mamba's performance on multi-dimensional data by investigating design choices such as SSM layer structure and scanning order within and across dimensions. However, unlike our approach, these works do not directly apply Mamba for Multimodal LLMs. More importantly, they primarily focus on replacing traditional backbones with Mamba architectures without explicitly leveraging Mamba's unique ability to summarize historical information for reducing video redundancy and enabling visual token compression. Our paper addresses this research gap by proposing STORM, which proves to be both effective and efficient for video understanding while significantly reducing computational demands.

Concurrent Work Recently, BIMBA [20] explores a similar architecture for long-video understanding, reporting similar benefits through empirical evaluation. We are encouraged to see these independent findings further support our hypothesis.

8. Qualitative Results

We present comprehensive qualitative evaluations in Figure 7 to Figure 11, which are segmented into three subsections

1. **Effective Long Video Understanding**: Demonstrating STORM's ability to effectively utilize long video inputs by comparing it with existing long-video LLMs.
2. **Importance of Long Video Context**: Highlighting the need for long video inputs by showcasing scenarios where 128-frame inputs (with token compression) enable accurate predictions, whereas 32-frame inputs fail.

3. **Showcase of Video Understanding Abilities**: Illustrating STORM's capabilities in various aspects such as OCR, spatial perception, temporal reasoning, and so on.

Effective Long Video Understanding. We compare our proposed STORM + Temporal Sampling with LongVILA and LongVU, both designed for long video understanding. We use a short film depicting a "moonfall disaster" from the VILA webpage [1]. The models are prompted to provide a narrative description of the video. The short film was chosen for its engaging and dramatic storyline that spans various interconnected scenarios, all contributing to a cohesive narrative. Understanding this video requires the models to comprehend each individual scene and effectively integrate temporal events to grasp the complete story. Both STORM and LongVILA use 128 input frames, while LongVU output was obtained from its online demonstration which uses 1fps input.

As shown in Figure 7, STORM delivers the most detailed and coherent summary of the video's narrative, effectively capturing key events and transitions throughout the entire film. Its response showcases a comprehensive understanding of the content, highlighting its ability to connect temporal events across different scenes. In contrast, the baseline models LongVILA and LongVU focus on some of the events but fail to cover all critical moments that contribute to the overall storyline. Their responses also highlight specific scenes without integrating them into the full context. Moreover, we observed that the baseline models often generate redundant content, repeating the same sentences with minimal new information, which reveals their limitations in handling open-ended queries. Notably, our STORM with Temporal Sampling is also computationally more efficient. By applying temporal sampling, we reduce the number of tokens to the equivalent of processing 32 frames. This comparison showcases STORM's superior ability to leverage long video inputs for in-depth visual understanding.

Importance of Long Video Context. We further demonstrate the significance of incorporating long video context by providing qualitative examples where a 128-frame input yields more accurate predictions than a 32-frame input, as shown in Figure 8. Using samples from the VideoMME benchmark, we compare two configurations of our STORM: one with a 32-frame input without compression, and another with a 128-frame input employing a temporal sampling ratio of 4. In both settings, the number of tokens fed into the LLM remains the same; however, the STORM with temporal sampling encodes additional information into the compressed tokens due to the extended frame sequence.

[1] https://vila.mit.edu/

The inclusion of more frames allows the model to capture richer temporal dynamics and contextual information. For example, the 128-frame input enables the model to develop a stronger understanding of the video's narrative (Figure 8 top). It also allows the 128-frame model to capture additional events that the 32-frame model misses (Figure 8 center). Finally, the additional information further improve model's ability to reason through different temporal events across the entire video to form a coherent understanding (Figure 8 bottom). These example demonstrate the crucial role of long video context in tasks that require detailed temporal reasoning and comprehensive content understanding.

Showcase of Video Understanding Abilities. Finally, we conclude our qualitative evaluation by showcasing the diverse video understanding capabilities of STORM, including OCR, attribute perception, spatial perception, information synopsis, and temporal reasoning. Results are shown in Figure 9 to Figure 11. We use the same setting of STORM + Temporal Sampling with 128-frame input and sampling ratio of 4. Utilizing videos from the VideoMME benchmark, we designed a more challenging assessment to thoroughly evaluate the model's proficiency. Instead of providing the model with multiple-choice questions accompanied by predefined answer options, we transformed these tasks into open-ended queries that require the model to generate answers in raw text form without any given choices. This modification significantly increases the task's difficulty, as it demands a precise understanding of the content and the ability to accurately locate and extract specific information from the video input.

Our qualitative results demonstrate that STORM provides strong performance in these scenarios. Despite the increased complexity, the model effectively interprets intricate visual details, recognizes textual information within videos, and provides coherent summaries of temporal events. This showcases STORM's robust ability to handle various aspects of video understanding.

9. Additional Results

Ablation on Token Budget and Token Compression Strategies. Table 8 extends Table 5 in the main text by providing a comprehensive comparison of different compression method combinations across various token budgets during training. Overall, considering both compression ratio and inference latency, we find that STORM with temporal pooling (STORM + T. Pooling) is the most efficient and effective approach. Additionally, test-time temporal sampling offers a lossless way to further enhance inference efficiency in inference time.

Task-Level Analysis on VideoMME Table 11 shows the VideoMME results with different video lengths. The short is less than 2 minutes, the medium is up to 15 minutes, and the long is up to 60 minutes. Overall, our STORM with token compression outperforms the VILA baseline and STORM with no token compressions for all video lengths. Figure 5 compares the VideoMME results by task categories. We find that STORM with temporal pooling especially improves the object reasoning task accuracy, and STORM with test-time temporal sampling improves the attribute perception accuracy. Both token compression methods improve the temporal perception task accuracies compared to VILA and STORM. It indicates that the temporal perception task requires a longer video context, and our token compression methods are effective for such tasks.

Effect of Dataset Composition Table 10 shows how dataset composition affects model performance during 128 frames fine-tuning. We compare using the full LLaVA-Video dataset ($\sim$ 1.35M samples) versus only its longest 25% videos with at least 128 frames ($\sim$ 360K samples). Interestingly, while STORM improves with the larger dataset across all benchmarks, the baseline model actually performs worse on several benchmarks when trained on the full dataset.

Two key differences between these datasets are size and video length distribution, where the full dataset contains more data but with a mixture of short and long videos, whereas the long-video subset exclusively consists of longer videos. Since larger, more diverse datasets typically improve performance (assuming similar data quality), we attribute the baseline model's unexpected performance drop to its limited ability to generalize from shorter to longer videos. More specifically, when trained predominantly on shorter clips from the full dataset, the baseline overfits and can not effectively handle long contexts at inference time. Training solely on longer videos dataset variant better aligns with test conditions, partially addressing this limitation.

In contrast, STORM shows consistent performance gains in all benchmarks when trained on the larger and more diverse data set. This suggests that STORM is more robust in handling longer sequences and is capable of using a wide range of video lengths to enhance its overall performance.

10. Architecture Details

STORM is built on a standard multimodal pipeline but introduces key modifications for improved reasoning ability and token efficiency. Figure 4 illustrates the detailed composition of our models. Instead of an MLP projector, STORM uses a linear layer followed by a Mamba-based temporal module projector which integrates spatiotemporal information into visual tokens.

STORM incorporates three main components: (1) The Mamba-Based Temporal Projector captures and propagates spatiotemporal information within visual tokens. (2) **Temporal Token Compression Module** applies compression on temporal dimension using training-based average pooling and/or training-free sampling (applied only at test time). (3) **Spatial Token Compression** further reduces token number by performing training-based frame-level spatial average pooling. Both spatial and temporal compression methods—whether training-free or training-based—are independently applicable. Notably, spatial and temporal pooling can be applied in parallel after the Mamba module, while temporal sampling is performed separately at test time. These components enable STORM to process longer sequences more efficiently before passing them to the LLM.

11. Training Details

We utilize the pre-trained SigLIP [75] from PaliGemma [2] and Qwen2-VL [62], respectively, and fine-tune them to adapt to our video datasets. The temporal projector is initialized with random weights. Each image is always resized to a 448×448 resolution. In the first stage, known as the Alignment Stage, we freeze both the image encoder and the LLM, training only the temporal projector using a small image-text dataset [79], containing 95K image-text pairs. Note that the Mamba layers perform not only temporal scan but also spatial scan within images, so video inputs are not strictly required to train it. For alignment stage, we find it sufficient to use only image-text pairs to pretrain the temporal projector. In the second stage, the supervised fine-tuning stage (SFT), we fine-tune all three components using a large and diverse dataset that includes text-only, image-text, and video-text data. There are around 12.5M samples in our SFT data mixture. due to space constraints. At this stage, we use 32 frames for each video input. For models with training-time token compression , we use a compression ratio of $4\times$ — temporal pooling models compress 32 frames to 8 frames while spatial pooling models compress 256 tokens per image into 64 tokens. Moreover, for models with training-time token compression , we further employ a long video fine-tuning stage using 128-frames long-video inputs from the LLaVA-Video dataset [78]. We provide further details about the full SFT dataset and long video fine-tuning dataset in the Appendix Section 12.

12. Datasets Information

For all models, we begin with an alignment stage to align the multi-modal projector using the LLaVA-CC3M-Pretrain-595K dataset [40]. Following this, we proceed to the visual instruction fine-tuning stage, experimenting with two different training data mixtures. These SFT mixtures incorporate both image and video data, encompassing three

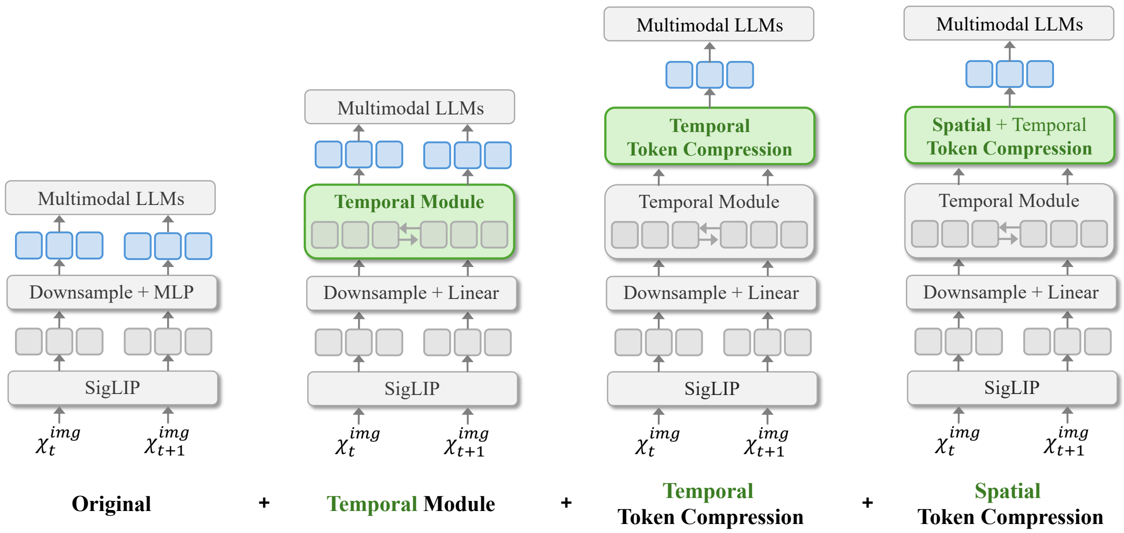


Figure 4. **Breaking Down the STORM Architecture.** We begin with a standard multimodal pipeline that uses a pixel-shuffle downsampling layer and an MLP projector. In STORM, we replace the MLP with a linear layer and introduce our Mamba-based temporal module on top. Since the Mamba layer propagates spatiotemporal information in each visual tokens, the model can then perform temporal and spatial token compression of these tokens before passing them to the LLM, allowing STORM to handle longer sequences more efficiently.

task types: captioning, open-ended question answering, and multiple-choice question answering. Further details are provided in the following:

- **SFT Data:** For most of our main experiments, we construct an expanded mixture by incorporating additional high-quality image datasets—such as Cambrian-1375K [57], Idefics2-SFT [26], and LLaVA-OneVision-Images-SFT [28]—along with video datasets including M4-Instruct-Video [77] and Youtube [81]. This enlarged dataset is used to scale up training and enhance overall performance. Detailed compositions of these mixtures are provided in Table 7.
- **Long Video SFT Data:** In the long video fine-tuning stage, our goal is to adapt models initially trained on full SFT data with 32-frame inputs to handle 128-frame inputs. Because processing 128-frame input incurs significant computational cost, we reduce training time by using a smaller dataset at this stage. Specifically, we select videos from only the LLaVA-Video dataset [78] that contains data amount roughly 11% of the full dataset (approximately 1.35M video-text pairs).
- **Long Video 25% Data:** As an ablation study to investigate the impact of dataset composition in the long video fine-tuning stage, we introduce an additional dataset derived from the LLaVA-Video dataset [78]. This subset consists exclusively of long videos with at least 128 frames, comprising approximately 25% of the full dataset (around 360K video-text pairs). Unlike the Long Video SFT Data, which includes both short and long videos, this dataset contains only long videos. Our experiments in Section 9 and Table 10 reveal distinct behaviors in our models and baselines in the composition of the dataset.

13. Mamba Temporal Module Latency

In this section, we compare the latencies of the vanilla VILA architecture and STORM across varying numbers of frames without token compression and provide a breakdown of the percentage contribution of the multi-modal projector. All experiments are conducted on a single NVIDIA DGX A100-80G. The results, shown in Figure 6, demonstrate that STORM incurs negligible overhead compared to the vanilla VILA architecture, with the introduced Mamba Temporal Module accounting for no more than 3% of the total latency.

14. Inference Details

Table 15 summarizes the number of frames used for inference. We evaluate all models between 8 and 512 frames and select the number of frames with the best accuracy overall for each task and setup.

Support for Streaming/Online Settings. To evaluate the applicability of our method in streaming scenarios, we replaced the default bi-directional Mamba with a uni-directional variant, allowing the model to reuse prior states for constant-time computation as new frames arrive. The results are provided in Table 13. We find that both uni-directional and bi-directional variants significantly outperform the baseline without temporal modeling. No-

Datasets
LLaVA-SFT [40], Idefics2-SFT [26]
MSR-VTT [68], Image Paragraph Captioning [25], ShareGPT4V-100K [8]
CLEVR [21], NLVR, VisualMRC [56]
ActivityNet-QA [74], LLaVA-OneVision-Images-SFT [28],
iVQA [70], MSRVTT-QA, STEM-QA [52]
DVQA [22], ST-VQA [3], SynthDoG-en [24], TextOCR-GPT4V [4], MTWI
ScienceQA-train [44], VQAv2-train, ViQuAE [27], Visual Dialog [12],
GQA-train [19], ChatQA [46], Geo170K [17],
LRV-Instruction [37], RefCOCO-train [72],
DocVQA [47], GeoQA [6], KVQA [45], Cambrian-1375K [57]
AI2D [23], Shikra [7], Unimm-Chat [73]
LRV-Instruction [36], SVIT [79], MMC-Instruction [38],
M4-Instruct-Images [39], M4-Instruct-Video [77] , WIT [55], Youtube [81], etc

Table 7. SFT data mixture.

Models	**Size**	**Comp. Ratio (%)**	**Latency (s)**	**#Frames (train)**	**# Frames (test)**	**MVBench test**	**MLVU dev**	**LongVideoBench val**	**VideoMME (w/o sub.)**
Duration						16 sec	3∼120 min	8 sec∼60 min	1∼60 min
Token Budget: 8K									
VILA Baseline	7B	100	4.31	32	256	69.5	70.2	55.9	60.1
STORM	7B	100	4.47	32	256	70.3	71.1	54.5	62.5
STORM + S. Pooling	7B	25	1.82	128	256	63.9	67.9	54.5	57.5
STORM + T. Pooling	7B	25	1.82	128	256	**71.3**	72.5	59.5	**63.4**
STORM + T. Sampling *	7B	50	2.50	32	256	70.1	70.8	54.8	**63.1**
STORM + S. Pooling + T. Sampling *	7B	12.5	1.51	128	256	65.2	68.3	55.0	57.6
STORM + T. Pooling + T. Sampling *	7B	12.5	1.51	128	256	70.6	**72.9**	**60.5**	62.4
Token budget: 2K									
STORM + S. Pooling	7B	25	1.82	32	256	68.9	69.2	56.0	61.1
STORM + T. Pooling	7B	25	1.82	32	256	**70.4**	71.0	54.2	61.2
STORM + S. Pooling + Sampling *	7B	12.5	1.51	32	256	68.9	69.5	56.3	60.9
STORM + T. Pooling + Sampling *	7B	12.5	1.51	32	256	68.9	69.5	56.3	60.9
Token budget: 0.5K									
STORM + S. Pooling + T. Pooling	7B	6.25	1.36	32	256	68.5	68.2	53.7	60.2

* 2x additional compression at test time.

Table 8. **Ablation on Token Budget and Token Compression Strategies.** Both spatial and temporal pooling are with $4\times$ compression. The number of frames used during testing is consistent across models but can differ across tasks. The # frames is the maximum number of frames during testing. We summarize the details of the number of frames for each task in Table 15. The temporal token sampling is with $2\times$ additional compression.

tably, the uni-directional Mamba performs competitively and even slightly surpasses the bi-directional counterpart when trained on 32-frame inputs. On the other hand, the bi-directional model demonstrates stronger performance as video length increases. These results highlight the critical role of the Mamba module and suggest its potential in enabling future designs that support streaming video input for video-LLMs.

14.1. STORM vs Other Temporal Fusion Strategies.

In this section, we present early explorations of temporal fusion strategies and show that the simple Temporal-Pooling design used in our final model is surprisingly effective when combined with the Mamba module. Specifically, we experimented with more complex fusion approaches, including TSM [34] and SlowFast [14], using LLaMA3-8B + SigLip. TSM incorporates temporal information by shifting tokens from neighboring frames into the image-based visual encoder. On the other hand, SlowFast encodes video using two token streams: one with high temporal but low spatial resolution, and the other with the opposite configuration.

The results of these experiments are shown in Table 14. These evaluations were conducted during the early stages of our study, and as such, the number of input frames was not

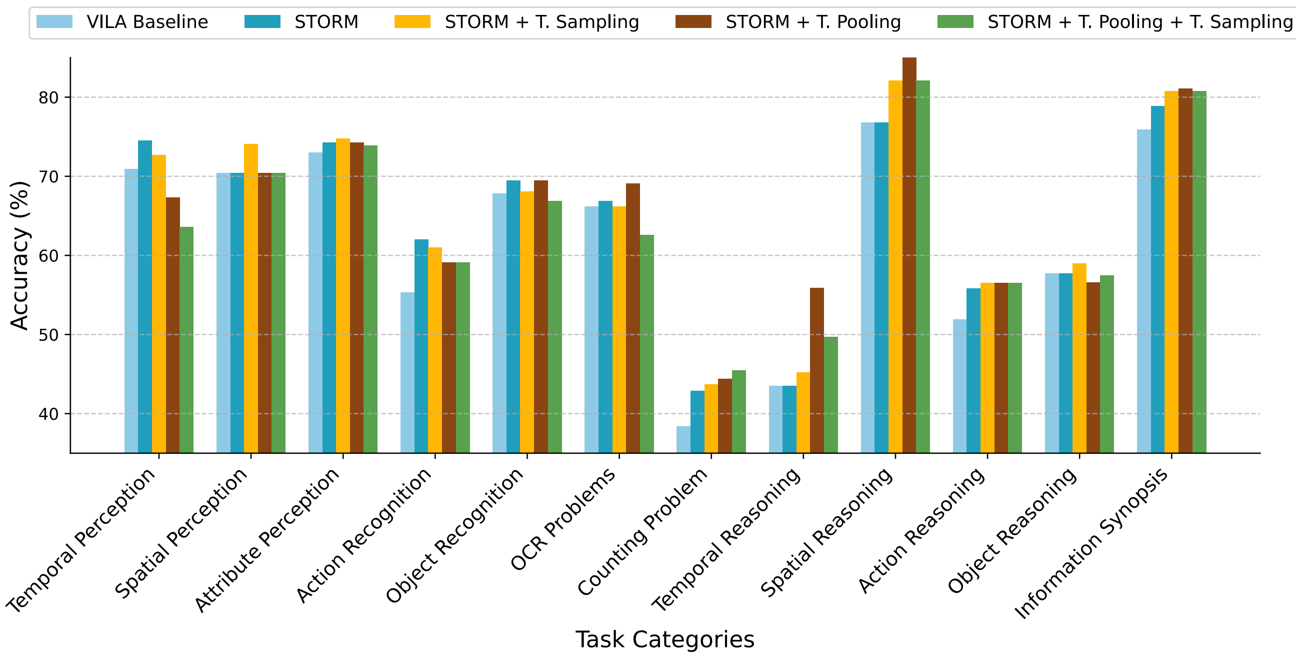


Figure 5. **VideoMME Results by Task Categories.**

Models	8F	32F (T. Pooling)	128F (T. Pooling)
MVBench			
VILA (w/o Mamba)	67.9	68.7	68.1
STORM (w/ Mamba)	**68.8**	**70.4**	**71.3**
MLVU			
VILA (w/o Mamba)	**67.7**	**71.0**	69.9
STORM (w/ Mamba)	66.8	**71.0**	**72.5**
LongVidBench			
VILA (w/o Mamba)	**52.4**	**55.4**	57.4
STORM (w/ Mamba)	50.6	54.2	**59.5**
VideoMME			
VILA (w/o Mamba)	60.0	58.9	61.7
STORM (w/ Mamba)	**60.2**	**61.2**	**63.4**
Avg			
VILA (w/o Mamba)	**62.0**	63.5	64.3
STORM (w/ Mamba)	61.6	**64.2**	**66.7**

Table 9. **Detailed Comparison across Benchmarks for Table 4.** STORM consistently improves performance across all benchmarks as the input video length increases from 8F to 32F to 128F. In contrast, the baseline VILA exhibits diminishing gains with longer inputs and even experiences performance degradation on certain benchmarks when extending from 32F to 128F. These results highlight the critical role of the Mamba module in effectively leveraging long-video inputs to enhance model performance.

consistently matched across variants, making exact comparisons difficult. However, the settings are generally favorable

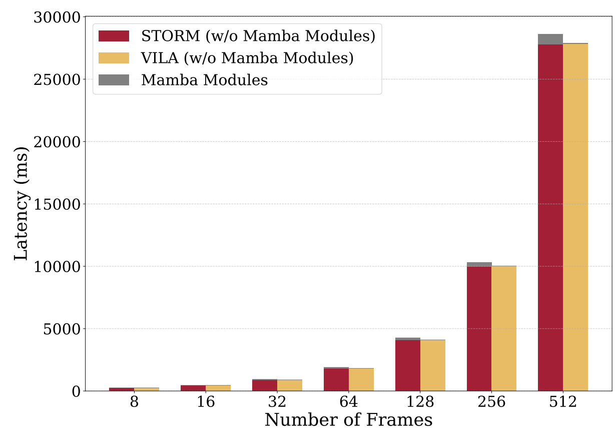


Figure 6. Latency Comparison: VILA vs STORM. The multimodal projector in VILA is a 2-layer MLP, while it is the Mamba Temporal Module in STORM.

to the TSM and SlowFast variants, as they use the same or more frames than the baseline, which does not perform any temporal fusion. Despite this, both fusion methods fail to yield meaningful improvements over the VILA baseline. In contrast, our STORM design achieves a significant gain, outperforming all other variants.

Models	Dataset type	MVBench test	MLVU dev	LongVideoBench val	VideoMME (w/o sub.)
Duration		16 sec	3∼120 min	8 sec∼60 min	1∼60 min
VILA Baseline + T. Pooling	long-video only (25% of full)	67.2	71.4	59.2	62.2
VILA Baseline + T. Pooling	full LLaVA-Video [78]	68.1 (+0.9)	69.9 (-1.5)	57.4 (-1.8)	61.7 (-0.5)
VILA Baseline + T. Pooling + T. Sampling *	long-video only (25% of full)	64.5	71.4	59.2	61.0
VILA Baseline + T. Pooling + T. Sampling *	full LLaVA-Video [78]	67.8 (+3.3)	70.1 (-1.3)	57.7 (-1.5)	59.3 (-1.7)
STORM + T. Pooling	long-video only (25% of full)	69.4	71.7	57.6	63.2
STORM + T. Pooling	full LLaVA-Video [78]	**71.3** (+1.9)	72.5 (+0.8)	59.5 (+1.9)	**63.4** (+0.2)
STORM + T. Pooling + T. Sampling *	long-video only (25% of full)	68.8	72.7	59.2	62.6
STORM + T. Pooling + T. Sampling *	full LLaVA-Video [78]	70.6 (+1.8)	**72.9** (+0.2)	**60.1** (+0.9)	62.4 (-0.2)

* 2x additional compression at test time.

Table 10. **Effect of Dataset Composition on 128-frame Fine-Tuning.** We compare models trained on two variants: the "full LLaVA-Video" dataset [78] (∼1.35M video-text pairs) versus the "long-video only" subset using top 25% longest videos (minimum of 128 frames) (∼360K pairs). Values in parentheses show performance differences between training on the long-video subset vs the full dataset. STORM consistently benefits from the larger, more diverse dataset across benchmarks, while the baseline VILA model degrades on several benchmarks when trained on the full dataset.

Models	# frames (train)	Short < 2 min	Medium 4∼15 min	Long 30∼60 min	Avg.
Token Budget: 8K					
VILA Baseline	32	73.0	58.0	49.2	60.1
STORM	32	**75.6**	60.9	51.1	62.5
STORM + T. Sampling*	32	75.2	60.8	53.2	63.1
STORM + T. Pooling	128	72.4	**64.4**	**53.4**	**63.4**
STORM + T. Pooling + T. Sampling*	128	72.9	60.9	**53.4**	62.4

* 2x additional compression at test time.

Table 11. **Break Donw of VideoMME Results by Input Video Length.**

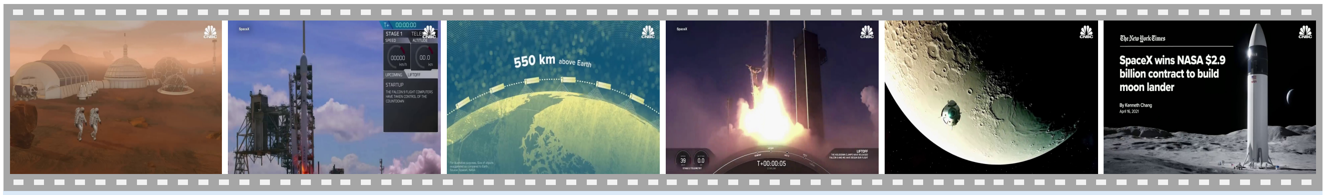
# Frames	Compression Ratio	Overall (ms)	llm (ms)	vision_tower (ms)	mm_projector (ms)
4	1	162.92	103.80	52.41	6.71
8	1	270.87	174.61	85.11	11.15
16	1	486.37	321.73	144.47	20.17
32	1	933.99	623.41	269.49	41.09
64	1	1910	1310	515.17	82.41
128	1	4270	3090	1020	163.34
256	1	10340	7960	2030	348.22
512	1	28620	23710	4090	811.31
32	4	486.97	175.75	269.96	41.26
64	4	920.10	322.22	515.82	82.06
128	4	1800	622.29	1010	163.23
256	4	3680	1310	2020	348.52
512	4	7950	3080	4060	811.84
64	8	772.18	175.27	514.59	82.32
128	8	1500	322.73	1020	163.09
256	8	3000	622.56	2030	348.71
512	8	6200	1310	4070	815.08

Table 12. **Full Latencies on Various Compression Ratios and Input Frames.**

Q: Please provide a description of the narrative of the video.

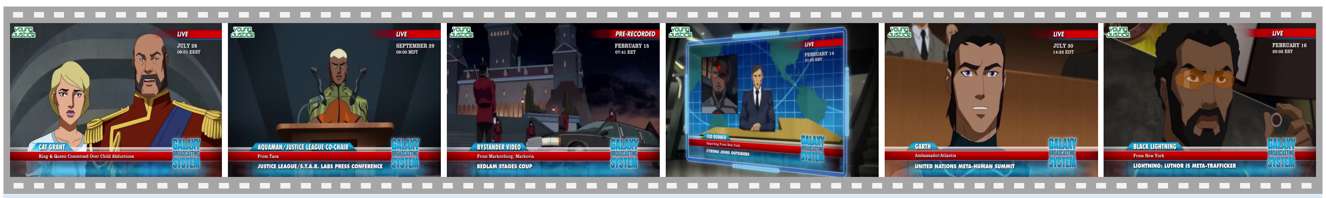
- **Ours:** The video begins with a person playing with a ball on a rooftop, which is disrupted by a large, glowing moon-like object in the sky. The person looks up in awe as the moon-like object grows larger and more detailed. The scene transitions to a close-up of the moon-like object, which then explodes, causing chaos on the rooftop. The person and a child run away from the explosion, and the scene shifts to a view of the moon-like object in space, surrounded by debris. The person and the child watch the moon-like object on a television, which then explodes, causing a massive fireball. The video then shows a cityscape with buildings exploding and debris flying everywhere, followed by a view of a city under attack with spaceships and debris in the sky. The scene transitions to a chaotic cityscape with buildings collapsing and debris flying everywhere. A person in a red shirt is seen running through the debris, followed by a view of a city under attack with spaceships and debris in the sky. The scene shifts to a close-up of a person in a spacesuit inside a spacecraft, looking out at the chaotic scene outside. The video concludes with a black screen displaying the text 'BUY TICKETS NOW IN THEATERS AND IMAX FRIDAY' along with a rating of 'PG-13' and the logo of the production company, Lionsgate.
- **LongVU:** The video starts with a scene of a man throwing a ball to another man who hits it back. The ball then flies towards the camera and hits the lens. The scene then shifts to a man looking through binoculars as a flock of birds flies overhead. The man then throws a ball that hits a bird, which then falls to the ground. The scene then shows a man throwing a ball that hits a bird, which then explodes into pieces. The video then shows a man throwing a ball that hits a bird, which then explodes into pieces…
- **LongVILA:** The video begins with a man in black short sleeves throwing a ball at a target on a rooftop, followed by a scene of a large explosion. Subsequently, a man in a spacesuit is seen flying through the air, transitioning to a scene of a car driving on a road. The video then shows a man and a boy on a balcony, the man pointing at the sky. The scene shifts to a large white object in the sky, with a man and a boy observing birds, and then to a man and a woman on the balcony. The video continues with a man and a boy on a staircase, the man holding a rope, and a large object falling from the sky. The video then shows a large object falling, a man and a boy on a balcony, and a man and a woman looking at birds. The scene changes to a man and a boy on a balcony, a man holding a wooden box, and a large object falling from the sky…

Figure 7. **Effective Long Video Understanding.** We compare STORM + Temporal Sampling with existing long video LLMs. Reults show that STORM delivers a more detailed and coherent summary, effectively capturing key events and transitions throughout the film. The example video is included in the supplementary material (moon.mp4).



Q: What does this video tell? The best answer is:
A. The process of building a starship.
B. Why Starship is the holy grail for SpaceX.
C. Why Starlink is crucial to SpaceX's success.
D. How SpaceX could Win The Space Race. **(GT answer)**

- **STORM (32-Frame Input, 8K Visual Tokens, No Compression) :** B. Why Starship is the holy grail for SpaceX.
- **STORM +Token Compression (128-Frame Input, 8K Visual Tokens, 4x Compression) :** D. How SpaceX could Win The Space Race.



Q: How many times do news segments appear in this video? The best answer is:
A. 2. B. 4. C. 6. **(GT answer)** D. 8.

- **STORM (32-Frame Input, 8K Visual Tokens, No Compression) :** B. 4.
- **STORM +Token Compression (128-Frame Input, 8K Visual Tokens, 4x Compression) :** C. 6.

Q: What is unique about the last performance? The best answer is:
A. The last performance features background dancers, distinguishing it from the others.
B. The stage lighting is noticeably dimmer in the last performance compared to the others.
C. Musical instruments are incorporated into the last performance, unlike the others. **(GT answer)**
D. The last performance showcases a collective group of performers, while the others consist of individual acts.

- **STORM (32-Frame Input, 8K Visual Tokens, No Compression) :** D. The last performance showcases a collective group of performers, while the others consist of individual acts.
- **STORM +Token Compression (128-Frame Input , 8K Visual Tokens, 4x Compression) :** C. Musical instruments are incorporated into the last performance, unlike the others.

Figure 8. **Importance of Long Video Context.** We compare STORM with a 32-frame input to STORM + Temporal Sampling using a 128-frame input. Both configurations have negligible differences in computational cost; however, the latter encodes additional information into compressed tokens due to the extended frame sequence. The examples illustrate that processing more frames allows the model to capture richer temporal dynamics and contextual information. This leads to a stronger understanding of the video's narrative, reduces information loss, and enhances the ability to reason through temporal events across the entire video.

a. OCR Problem

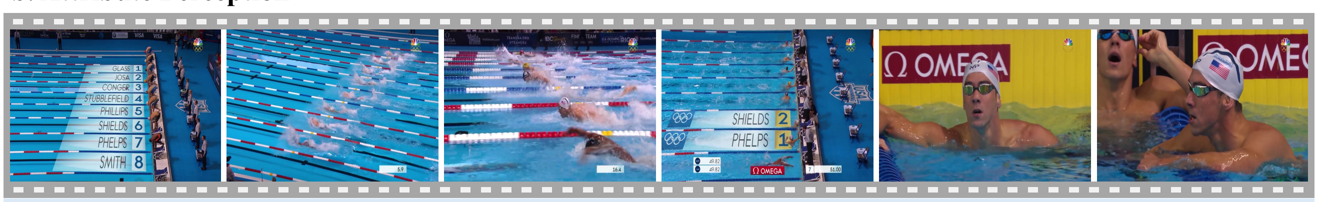
Q: Whom is the poem in the video written by?

Ours: The poem in the video is written by Edgar Allan Poe.

Q: What are the red texts recognized on the white wall?

Ours: The red texts recognized on the white wall are 'GO BOWLING'.

b. Attribute Perception



Q: Which country does the swimmer who was given a close-up in the video hail from?

Ours: The swimmer who was given a close-up in the video hails from the USA.

Q: What color are his gloves in the video?

Ours: His gloves are black.

Figure 9. **Showcase of Video Understanding Abilities in Various Task Categories.** We provide additional examples to showcase model's video understanding capabilities in different aspects. This is done by providing the models with open-ended queries that require the model to generate answers in raw text form without any given choices. Part 1. The OCR problem example videos are included in the supplementary material (ocr.mp4).

c. Spatial Perception

Q: Where is this video most likely shot?

Ours: The video is most likely shot inside a birdhouse.

Q: Which hand did the person in the video wear a glove on?

Ours: The person wore a glove on their right hand.

d. Information Synopsis

Q: According to the video, why is it very difficult to capture footage of sand cats?

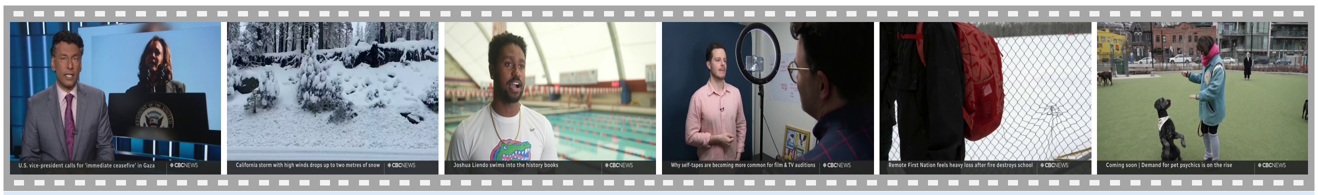
Ours: Because their fur blends in with their surroundings and their tracks are barely visible.

Q: What is this video mainly about?

Ours: This video is mainly about demonstrating how to fold a shirt. The man in the video shows how to fold a shirt in a step-by-step manner, and he also demonstrates how to fold a t-shirt and a polo shirt.

Figure 10. **Showcase of Video Understanding Abilities in Various Task Categories.** Continue 2.

e. Temporal Reasoning



Q: What is the fourth-to-last news item in this news video?

A. Josh Liendo swims into the history books. B. Coming soon | Rising demand for pet psychics.

C. U.S. vice president calls for Gaza ceasefire. D. California storm drops 2 meters of snow.

Ours: A. Josh Liendo swims into the history books.

Q: In which order do the six tips are introduced in the video?

(a) Clip coupons. (b) Eat at home. (c) Freeze leftovers. (d) Cook once, eat twice. (e) Meal plan. (f) Buy in bulk.

A. b e d f a c. B. b e f c a d. C. e c d a b f. D. c e b f a d.

Ours: A. b e d f a c.

Figure 11. **Showcase of Video Understanding Abilities in Various Task Categories.** Continue 3.

Models	**32F (T. Pooling)**	**128F (T. Pooling)**
VILA	58.9	61.7
Uni-dir STORM	**62.2**	62.5
Bi-dir STORM	61.2	**63.4**

Table 13. **Support for Streaming/Online Settings.** We evaluate a uni-directional variant of STORM designed for streaming video inputs. Results show that the Uni-dir STORM consistently outperforms the VILA baseline, highlighting the potential of our design to support streaming scenarios.

Table 14. Comparison of temporal fusion strategies.

16 frames			**32-64 frames**	
Baseline	**Baseline (T.pool)**	**TSM (T.pool)**	**SlowFast (T.pool)**	**STORM (T.pool)**
52.0	50.0	49.0	51.3	**56.8**

Models	**MVBench**	**MLVU**	**LongVidBench**	**VideoMME**
8-frame-models	8	64	32	64
+ Temporal Sampling	16	256	256	128
32-frame-models	16	64	64	64
+ Temporal Sampling	32	256	256	128
32-frame-models + Temporal Pooling	32	64	128	64
+ Temporal Sampling	64	256	256	128

Table 15. **The Number of Frames Used for Inference.** We evaluate all models for [8, 16, 32, 64, 128, 256, 512] frames and select the best overall for each task and setup.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 2022. 13

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 4, 15

[3] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4290–4300. IEEE, 2019. 17

[4] Jimmy Carter. Textocr-gpt4v. `https://huggingface.co/datasets/jimmycarter/textocr-gpt4v`, 2024. 17

[5] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024. 1, 13

[6] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online, 2021. Association for Computational Linguistics. 17

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *ArXiv preprint*, abs/2306.15195, 2023. 17

[8] Lin Chen, Jinsong 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. 17

[9] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 13

[10] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 13

[11] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024. 2

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society, 2017. 17

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*. PMLR, 2023. 13

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *ICCV*, 2018. 17

[15] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 4

[16] Daniel Y. Fu, Tri Dao, Khaled Kamal Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. 2023. 2

[17] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model, 2023. 17

[18] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. 1, 2, 13

[19] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019. 17

[20] Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. Bimba: Selective-scan compression for long-range video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29096–29107, 2025. 13

[21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society, 2017. 17

[22] Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. DVQA: understanding data visualizations via question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5648–5656. IEEE Computer Society, 2018. 17

[23] Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016. 17

[24] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 17

[25] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3337–3345. IEEE Computer Society, 2017. 17

[26] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv: 2405.02246*, 2024. 16, 17

[27] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G. Moreno, and Jesús Lovón-Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3108–3120. ACM, 2022. 17

[28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5, 16, 17

[29] 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. 1, 13

[30] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multimodal video understanding benchmark. *Computer Vision and Pattern Recognition*, 2023. 4

[31] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024. 1, 13

[32] Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024. 1, 13

[33] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *European Conference on Computer Vision*, 2023. 13

[34] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3, 17

[35] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *CVPR*, 2024. 1, 4, 13

[36] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 17

[37] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2023. 17

[38] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310, Mexico City, Mexico, 2024. Association for Computational Linguistics. 17

[39] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 17

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 2024. 1, 13, 15, 17

[41] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv: 2408.15542*, 2024. 4, 5, 13

[42] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 4, 5

[43] Hui Lu, Albert Ali Salah, and Ronald Poppe. Videomambapro: A leap forward for mamba in video understanding. *arXiv preprint arXiv:2406.19006*, 2024. 1, 13

[44] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 17

[45] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019. 17

[46] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 17

[47] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 17

[48] Jinyoung Park, Hee-Seon Kim, Kangwook Ko, Minbeom Kim, and Changick Kim. Videomamba: Spatio-temporal selective state space model. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1, 13

[49] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. 2023. 2

[50] 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*. PMLR, 2021. 1, 13

[51] Michael S. Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Silvio Savarese, Ran Xu, Caiming Xiong, and Juan Carlos Niebles. xgen-mm-vid (blip-3-video): You only need 32 tokens to represent a video even in vlms. *arXiv preprint arXiv: 2410.16267*, 2024. 13

[52] Jianhao Shen, Ye Yuan, Srbuhi Mirzoyan, Ming Zhang, and Chenguang Wang. Measuring vision-language stem skills of neural models, 2024. 17

[53] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv: 2410.17434*, 2024. 1, 4, 5, 13

[54] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv: 2409.14485*, 2024. 4, 5, 13

[55] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2021. 17

[56] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13878–13888. AAAI Press, 2021. 17

[57] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv: 2406.16860*, 2024. 16, 17

[58] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35, 2022. 13

[59] Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, et al. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024. 13

[60] Junxiong Wang, Daniele Paliotta, Avner May, Alexander Rush, and Tri Dao. The mamba in the llama: Distilling and accelerating hybrid models. *Advances in Neural Information Processing Systems*, 37:62432–62457, 2024. 13

[61] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 3

[62] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv: 2409.12191*, 2024. 4, 5, 15

[63] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *arXiv preprint arXiv: 2409.02889*, 2024. 4, 5

[64] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1, 13

[65] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 1, 13

[66] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2025. 13

[67] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv: 2407.15754*, 2024. 4

[68] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society, 2016. 17

[69] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv: 2408.10188*, 2024. 1, 3, 4, 5, 13

[70] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1666–1677. IEEE, 2021. 17

[71] Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, Ying Shan, and Yansong Tang. Voco-llama: Towards vision compression with large language models. *arXiv preprint arXiv:2406.12275*, 2024. 13

[72] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. 17

[73] Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*, 2023. 17

[74] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134. AAAI Press, 2019. 17

[75] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1, 4, 13, 15

[76] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 1, 4, 5, 13

[77] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv: 2404.01258*, 2024. 16, 17

[78] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 15, 16, 19

[79] Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *ArXiv preprint*, abs/2307.04087, 2023. 4, 15, 17

[80] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv: 2406.04264*, 2024. 4

[81] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *International Conference on Learning Representations*, 2023. 16, 17

[82] Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. Falcon mamba: The first competitive attention-free 7b language model. *arXiv preprint arXiv:2410.05355*, 2024. 13