VAMBA: Understanding Hour-Long Videos with Hybrid Mamba-Transformers

Supplementary Material

7. Additional Implementation Details

We use 8 NVIDIA A800 80G GPUs to train our models for both ablation study and full-scale training. For ablation studies, the learning rate is set to 1e-5 for pretraining and 1e-7 for instruction-tuning. We further conduct a hyperparameter search and find that setting the learning rate to 5e-6 during instruction-tuning works the best for VAMBA across multiple benchmarks. We therefore set the learning rate to 1e-5 for pretraining and 5e-6 for instruction tuning for full-scale training. We employ a cosine learning rate schedule for all training stages in both ablation studies and full-scale training. The training batch size is set to 128. We employ training optimization methods such as Flash-Attention 2 [16], DeepSpeed ZeRO-3 [52] and gradient checkpointing [13] to reduce the training cost, and apply sequence parallelism to pack multiple samples into one sequence during training in both stages.

8. Model Evaluation Details

In this section, we provide more details for benchmarking VAMBA and our selected baseline models.

8.1. Baseline Models

Qwen2-VL [63] is an LMM that uses the Qwen2 LLM as its backbone and a DFN-derived Vision Transformer with 2D RoPE positional embedding. It is pretrained on a vast 1.4T-token multimodal corpus composed of image-text pairs, OCR text (images of text), interleaved image–text web articles, visual QA data, video dialogues, and image-based knowledge datasets. The pre-training is staged: 600B tokens for vision-language alignment followed by 800B tokens mixing richer image–text content and VQA/multitask data, alongside continued pure text to maintain language skills. Finally, Qwen2-VL is instruction-tuned via ChatML-format dialogs that span multiple modalities, e.g. document parsing, comparisons of two images, long video understanding, and even agent-oriented visual tasks.

LLaVA-Mini [74] is a compact multimodal model built on a 7–8B Vicuna LLM with a CLIP ViT-based vision encoder. It uses the same training data as LLaVA-1.5 [43]: about 558K image–caption pairs for initial vision–language pre-training and 665K image-grounded instruction examples for fine-tuning. The pre-training stage aligns visual features to text using caption datasets like COCO [14] and VisualGenome-based [32] captions, while the instruction-tuning stage uses multimodal dialogues. An enhanced variant of LLaVA-Mini further incorporates 100K video-based instruction samples from Video-ChatGPT [49] and other open sources, combined with the original 665K image instructions (total 3M training instances) to extend its capability to video understanding.

LongLLaVA [65] extends LLaVA [43] to handle very long visual contexts by using a hybrid Transformer–Mamba architecture with a Jamba-9B backbone for language. It follows a three-stage training process, including a single-image feature alignment on Allava-Caption [10] and ShareGPT4V [11], a single-image instruction fine-tuning on LLava-1.5 and Mantis-Single [30], and multi-image instruction fine-tuning on VideoChat2 [38] and ShareGPT4Video [12]. By progressively increasing the number of images per sample, LongLLaVA learns temporal and spatial dependencies and can efficiently handle input sequences up to around 1000 images.

LongVU [55] is a multimodal model geared toward long video understanding. It first learns from 3.2 million image–text pairs via a single-image training stage using the LLaVA-OneVision dataset [35]. It then leverages a subset of VideoChat2-IT [38] that contains around 0.55M videos, 1K video-classification clips from Kinetics-710 [7], and about 85K multimodal video instruction dialogues from the ShareGPT4Video [12]. Additionally, the MovieChat long-video dialogue data [58] is used to provide hour-length conversational examples. This rich training mix enables LongVU to handle extended videos by adaptively compressing frames while preserving essential visual details.

Video-XL [56] employs an LLaMA-based 7B language model and a CLIP ViT-L vision encoder, and it is trained entirely on image-based data despite targeting long videos. Its two-stage training first performs projection-layer pretraining on 2M image–text pairs from Laion-2M [54] to align visual features with the text space. It then undergoes visual instruction tuning on roughly 695K image-grounded instruction samples from Bunny-695k [27], where the model learns to follow image-based instructions. The training approach lets Video-XL handle hour-long videos in context by compressing visual tokens, achieving strong results on benchmarks for long video comprehension.

8.2. Evaluation Benchmarks

LVBench [64] is a benchmark designed to test the ability of video LMMs to comprehend extremely long videos. It contains 1,549 question-answer pairs, with an average video length of 4,101 seconds. The evaluation focuses on six fundamental aspects: temporal grounding, which involves identifying specific moments in a video; video summarization, which assesses the model's ability to condense key information; video reasoning, which tests logical inference from video content; entity recognition, which identifies people, objects, or places; event understanding, which captures the sequence and significance of events; and key information retrieval, which ensures the model extracts crucial details. The full test set is used for evaluation.

HourVideo [8] is a benchmark dataset for long-form video-language understanding, focusing on videos up to one hour in length. It consists of 500 carefully selected first-person videos sourced from the Ego4D [23] dataset, with each video ranging from 20 to 120 minutes in duration. The dataset includes 12,976 human-annotated multiple-choice questions covering four major task categories: summarization, perception, visual reasoning, and navigation. HourVideo is designed to challenge models in long-context reasoning and multimodal comprehension across extended video timelines. Benchmark results reveal that existing multimodal models, such as GPT-4 and LLaVA-NeXT, perform only marginally better than random chance, while human experts achieve an accuracy of 85.0%. This highlights the dataset's difficulty and the current gap in long-video understanding capabilities.

Video-MME [21] is a benchmark specifically designed to evaluate how well LMMs can analyze video content. It features a dataset of 900 videos and 2700 questions, covering six different visual domains. The questions are grouped based on video length into short, medium, and long categories, with median durations of 26 seconds, 164.7 seconds, and 890.7 seconds, respectively. The benchmark supports two evaluation methods: (1) the "w/ subtitle" setting, where both subtitles and questions are provided as text inputs, and (2) the "w/o subtitle" setting, which relies only on raw video inputs alongside the questions. Our study primarily focuses on the "w/o subtitle" format to enhance long video comprehension by leveraging video-based augmentation rather than textual cues from subtitles.

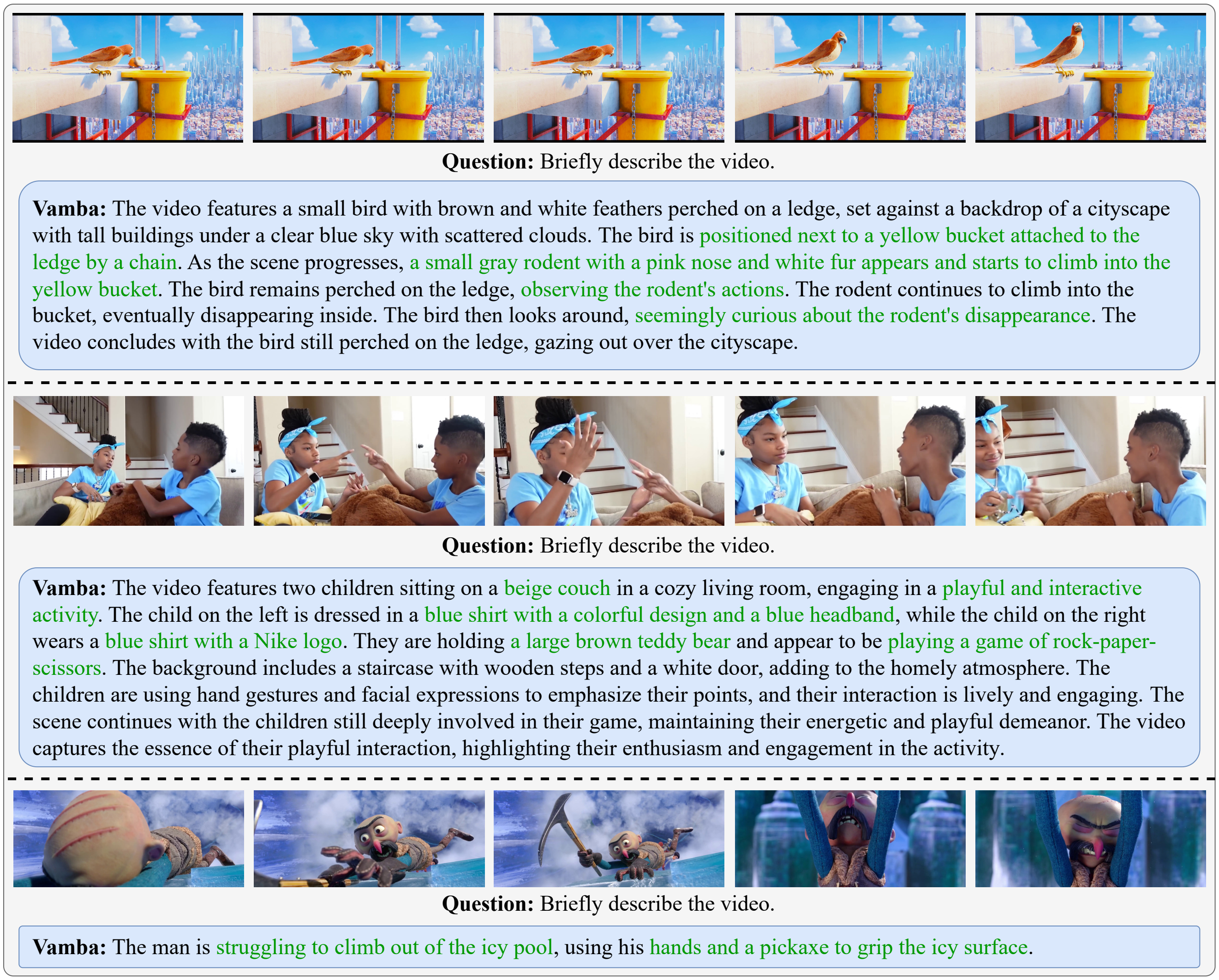
MLVU [76] is a benchmark designed to assess long video understanding across various tasks and video genres. It includes two types of questions: multiple-choice and freeform generation. The evaluation framework measures LMM performance in three key aspects: (1) holistic video understanding, which requires comprehending the entire video for global context; (2) single-detail video understanding, which focuses on recognizing key moments or short segments; and (3) multi-detail video understanding, which involves drawing connections between multiple short clips within the video. Our paper specifically reports accuracy scores for multiple-choice questions from the MLVU development set.

LongVideoBench [67] is a question-answering benchmark designed for interleaved long video-text input. It includes 3,763 videos and 6,678 human-annotated multiple-choice questions covering 17 fine-grained categories. The benchmark supports two evaluation formats: (1) the standard format, where video tokens are processed first, followed by the question descriptions, and (2) the interleaved video-text format, where subtitles are inserted between video frames. We evaluate all baseline models and our VAMBA using the standard input format. The reported results are based on the validation split.

NExT-QA [68] is a video question-answering benchmark designed to evaluate reasoning-based video understanding. It consists of 5,440 videos and approximately 52,000 human-annotated question-answer pairs, covering a diverse range of real-world activities. The dataset includes two types of question formats: multiple-choice questions and open-ended free-form questions. NExT-QA emphasizes causal and temporal reasoning, requiring models to understand event sequences, cause-effect relationships, and interactions within videos. The dataset is divided into three categories: causal, temporal, and descriptive questions. The dataset is split into training (3,870 videos), validation (570 videos), and test (1,000 videos), ensuring standardized benchmarking.

MVBench [40] is a comprehensive multimodal video understanding benchmark. The dataset introduces a novel static-to-dynamic task transformation, converting existing static image tasks into video-based challenges, assessing a model's ability to perform both low-level perception and high-level cognitive reasoning over time. MVBench automatically converts annotations from 11 publicly available video datasets into unified multiple-choice question-answer pairs, covering diverse scenarios ranging from first-person to third-person perspectives and indoor to outdoor environments. Each question presents five answer choices, ensuring standardized evaluation through human-verified ground truth responses.

DREAM-1K [62] is a video description dataset designed for fine-grained event and motion understanding. It contains 1,000 short videos, each averaging 9 seconds, and covers a diverse set of real-world and cinematic scenarios. Unlike question-answering datasets, DREAM-1K requires models to generate detailed multi-sentence descriptions that capture all key actions, interactions, and temporal sequences within each video. The dataset includes videos from five different sources—live-action movies, animated films, stock footage, long YouTube videos, and short-form social media clips—ensuring broad coverage of visual styles. DREAM-1K prioritizes event-based reasoning, expecting models to understand sequential actions, motion cues, and interactions rather than just static descriptions. Evaluation is conducted using AutoDQ (Automatic Description Quality), which measures how well generated descriptions align with reference descriptions by comparing extracted events.



Question: Briefly describe the video.

Vamba: The video features a small bird with brown and white feathers perched on a ledge, set against a backdrop of a cityscape with tall buildings under a clear blue sky with scattered clouds. The bird is positioned next to a yellow bucket attached to the ledge by a chain. As the scene progresses, a small gray rodent with a pink nose and white fur appears and starts to climb into the yellow bucket. The bird remains perched on the ledge, observing the rodent's actions. The rodent continues to climb into the bucket, eventually disappearing inside. The bird then looks around, seemingly curious about the rodent's disappearance. The video concludes with the bird still perched on the ledge, gazing out over the cityscape.

Question: Briefly describe the video.

Vamba: The video features two children sitting on a beige couch in a cozy living room, engaging in a playful and interactive activity. The child on the left is dressed in a blue shirt with a colorful design and a blue headband, while the child on the right wears a blue shirt with a Nike logo. They are holding a large brown teddy bear and appear to be playing a game of rock-paper-scissors. The background includes a staircase with wooden steps and a white door, adding to the homely atmosphere. The children are using hand gestures and facial expressions to emphasize their points, and their interaction is lively and engaging. The scene continues with the children still deeply involved in their game, maintaining their energetic and playful demeanor. The video captures the essence of their playful interaction, highlighting their enthusiasm and engagement in the activity.

Question: Briefly describe the video.

Vamba: The man is struggling to climb out of the icy pool, using his hands and a pickaxe to grip the icy surface.

Figure 6. Additional qualitative results for VAMBA.

9. Comparison with Contemporary Work

Several contemporary works also investigate hybrid Mamba-Transformer models for long video understanding. For example, STORM [31] and BIMBA [28] utilize Mamba blocks between the vision encoder and LMM decoder as additional processing and compression modules for video tokens, achieving high performance in long video understanding. However, different from VAMBA, the overall architecture of the LMM remains unchanged in these methods, with the decoder still relying on full causal self-attention layers for both text and video tokens. As a result, the model architectures proposed in these methods offer limited gains in training and inference efficiency, with any speedup in video processing still primarily attributed to token reduction. In comparison, VAMBA directly employs Mamba-2 layers in the LMM decoder and bypasses the self-attention updates for video tokens. This design enables highly efficient video processing even without reducing the number of tokens.

Table 6. Quantitative results for VAMBA with token reduction.

Models	GPU Mem (MB)	LVBench	VideoMME	MLVU
VAMBA	45791	42.4	57.4	65.9
VAMBA-TR	33847	41.6	56.9	66.5

10. Combining VAMBA and Token Reduction

As mentioned in the paper, we expect VAMBA to be compatible with token reduction, and combining VAMBA and token reduction can potentially result in similar performance and even higher efficiency. We provide some preliminary results for combining VAMBA and token reduction in this section. As shown in Table 6, we can simply uniformly drop 50% of the video tokens during inference (denoted as VAMBA-TR) and achieve little performance drop across multiple benchmarks with better efficiency. We believe finetuning VAMBA with token reduction can further preserve its capacity and leave this as a future work.

11. Additional Qualitative Results

In this section, we showcase more qualitative results from our VAMBA for detailed video captioning and video event understanding. The results are shown in Figure 6.