VISTA: Enhancing Long-Duration and High-Resolution Video Understanding by VIdeo SpatioTemporal Augmentation

Supplementary Material

7. HRVideoBench Details

As detailed in Section 3, our HRVideoBench consists of 200 questions covering 10 question types and 10 video types. That is, we collect two questions for each combination of video type and question type. The 10 video types are:

- POV driving videos
- Egocentric sports videos
- Sportscast videos (broadcasting of sports events)
- Public event recordings
- Surveillance camera/CCTV footage
- Wildlife stock videos
- Aerial videos/Drone videos
- Factory and industrial stock videos
- Public transport videos
- Product review videos

For each question in the benchmark, we ensure the video duration falls between 3 to 10 seconds. This relatively short duration is chosen to maximize the likelihood of the frames relevant to the question getting sampled by the models. The final dataset has an average video duration of 5.4 seconds and an average resolution of 3048×1699. Example questions and answers from our HRVideoBench are shown in Figure 4.

8. Model Training and Evaluation Details

In this section, we provide additional details for training and benchmarking our selected baseline models.

8.1. Baseline Models

VideoLLaVA [26] is a video LMM jointly pretrained on image and video data. It uses the pretrained Vicuna v1.5 model as its LLM backbone and LanguageBind as its image and video encoder. The model is pretrained on 558K image-text pairs from LAION-CC-SBU [37, 40, 41] and 702K video-text pairs from Valley [31]. During the instruction-tuning stage, it incorporates 665K image-text pairs from LLaVA-1.5 [27] and 100K video-text pairs from Video-ChatGPT [32].

Mantis-Idefics2 [13] is an LMM specialized in processing inputs with multiple interleaved images. It is initialized from Idefics2 [18] and continually pretrained on Mantis-Instruct, a dataset comprising 721K interleaved image-text instruction-tuning examples. This dataset focuses on enhancing multi-image understanding across four dimensions: co-reference, comparison, reasoning, and temporal understanding. Mantis-Idefics2 achieves state-of-the-art performance on various multi-image benchmarks and excels on short video understanding benchmarks, such as MVBench [23].

LongVA [64] is a long-context LMM designed for understanding long video content. It first performs continual pretraining using a Qwen2 [57] model to support up to 224K context length. Following this, it uses this modified Qwen2 model as the backbone for visual instruction tuning. LongVA is instruction-tuned on pure image data, using the same training data as LLaVA-1.6 [27]. It introduces the UniRes strategy, which divides an image into multiple grids and encodes each grid independently using the vision encoder. During inference, these grids are replaced by different frames from the input video, enabling effective processing of long video sequences.

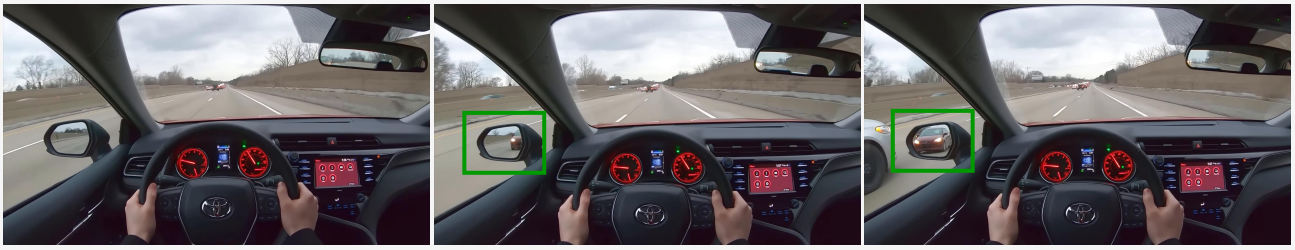
8.2. Additional Implementation Details

For all three models, we conduct full-finetuning for one epoch using 8 Nvidia H800 GPUs. The total training time for ∼400K data is around one day. We use the Adam [16] optimizer with a batch size of 128 during training. The learning rate is set to 5e-6 for VideoLLaVA and 1e-7 for LongVA and Mantis-Idefics2, with a cosine learning rate scheduler and a warm-up ratio of 0.03 applied to all models. We employ Flash-Attention 2 [7] and DeepSpeed ZeRO-3 [38] to accelerate training.

8.3. Evaluation Benchmarks

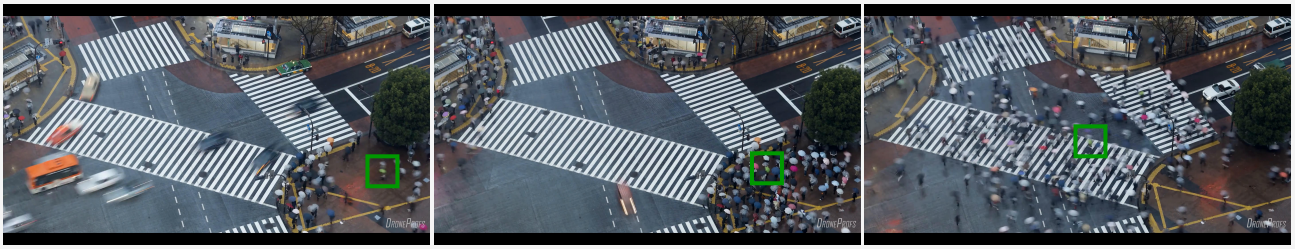
Video-MME [10] is a comprehensive benchmark designed to evaluate the video analysis capabilities of LMMs. It includes 900 videos and 2700 questions across six visual domains. The questions are categorized based on video durations into short, medium, and long video questions, with median durations of 26s, 164.7s, and 890.7s, respectively. The median duration values for short, medium and long video questions are 26s, 164.7s, and 890.7s, respectively. Video-MME supports two evaluation formats: (1) the "w/ subtitle" format, which includes both the video subtitles and questions as text inputs, and (2) the "w/o subtitle" format, which uses only the raw video and questions as inputs. In the main paper, we focus on the "w/o subtitle" format to emphasize improving the long video understanding capabilities of video LMMs through video augmentation, rather than relying on additional subtitle information. For completeness, we provide results for the "w/ subtitle" format in Section 9.

MLVU [67] is a long video understanding benchmark en-



Question: What is the color of the car in the left rearview mirror? A. red, B. green, C. yellow, D. orange.

Answer: A. red.



Question: A person is in the bottom right of the video using a green umbrella. In which direction is this person going? A. bottom left, B. top right, C. bottom right, D. top left.

Answer: D. top left.

Figure 4. Example questions from our HRVideoBench. Zoom in for better visualizations.

compassing diverse tasks and video genres. It features two types of questions: multiple-choice questions and freeform generation questions. The benchmark evaluates LMMs across three dimensions: holistic video understanding, requiring global information from the entire video; single-detail video understanding, focused on short and salient moments within the video; and multi-detail video understanding, involving connections across multiple short clips in the video. In this paper, we report the accuracy scores for the multiple-choice questions from the development set of MLVU. In the paper, we report the accuracy scores for the multiple-choice questions from the dev set of MLVU.

LVBench [48] evaluates the comprehension capabilities of video LMMs for extremely long videos. It consists of 1549 QA pairs, with an average video duration of 4101 seconds. The benchmark assesses video LMMs across six core aspects: temporal grounding, video summarization, video reasoning, entity recognition, event understanding, and key information retrieval. We use the full test set for evaluation.

LongVideoBench [52] is a question-answering benchmark featuring interleaved long video-text input. The dataset contains 3763 videos and 6678 human-annotated multiple-choice questions spanning 17 fine-grained categories. LongVideoBench supports two evaluation formats: the standard input format, where video tokens are processed first followed by question descriptions, and an interleaved video-text format, where subtitles are inserted between video frames. Although Mantis-Idefics2 supports interleaved image-text input, as our VISTA-400K does not include training examples in such format, we still evaluate Mantis-Idefics2 and the finetuned VISTA-Mantis using the standard format. We report the results of the validation split.

MVBench and NExT-QA [23, 53] are short video understanding benchmarks, focusing on videos under one minute in duration. MVBench includes 4,000 multiple-choice questions derived from 3,641 video clips, with an average video duration of 16 seconds. NExT-QA comprises 8,564 questions (both multiple-choice and open-ended) sourced from 1,000 videos, averaging 40 seconds in length. In our experiments, we evaluate the models on the full MVBench dataset and the MCQ split of NExT-QA.

MSVD-QA, MSRVTT-QA, TGIF-QA and ActivityNet-QA [12, 54, 59] are open-ended QA benchmarks designed to evaluate the response generation capabilities of video LMMs. These benchmarks consist of short videos and assess the ability of video LMMs to produce simple, coherent answers. For all four benchmarks, we follow Video-ChatGPT [32] and use GPT-3.5-Turbo to evaluate the accuracy and quality of the responses. Specifically, GPT is prompted with the ground truth answer and the model's response to determine if the answer is correct (yes/no) and to assign a quality score between 1 and 5. Following Video-ChatGPT, we evaluate the models on the validation sets of MSVD-QA, MSRVTT-QA and ActivityNet-QA, and use the FrameQA split from TGIF-QA's test set for evaluation. Since GPT-3.5-Turbo's API version has changed and the

Table 5. Comparison between the baseline VideoLLaVA model, VideoLLaVA finetuned on VISTA-400K and VideoLLaVA finetuned on VISTA-400K + 300K VideoChat2-IT data (VISTA-VideoLLaVA in the main paper) on long video understanding benchmarks. "SFT" indicates supervised finetuning.

Models	Long Video Understanding						
	Video-MME w/o subtitles				MLVU	LVBench	LongVideoBench
	avg	short	medium	long	m-avg	test	val
VideoLLaVA	39.9	45.3	38.0	36.2	45.0	29.3	39.1
VideoLLaVA (SFT on VISTA-400K)	43.6	47.3	43.8	39.8	48.7	32.6	41.0
Δ - VideoLLaVA	**+3.7**	**+2.0**	**+5.8**	**+3.6**	**+3.7**	**+3.3**	**+1.9**
VideoLLaVA (SFT on VISTA-400K + 300K VideoChat2-IT)	43.7	48.2	43.9	38.9	49.5	33.8	42.3
Δ - VideoLLaVA (SFT on VISTA-400K)	**+0.1**	**+0.9**	**+0.1**	**-0.9**	**+0.8**	**+1.2**	**+1.3**

Table 6. Comparison between the baseline VideoLLaVA model, VideoLLaVA finetuned on VISTA-400K and VideoLLaVA finetuned on VISTA-400K + 300K VideoChat2-IT data (VISTA-VideoLLaVA in the main paper) on HRVideoBench. "SFT" indicates supervised finetuning.

Models	High-Resolution Video Understanding		
	HRVideoBench		
	avg	object	action
VideoLLaVA	32.5	36.0	27.9
VideoLLaVA (SFT on VISTA-400K)	44.0	42.1	46.5
Δ - VideoLLaVA	**+11.5**	**+6.1**	**+18.6**
VideoLLaVA (SFT on VISTA-400K + 300K VideoChat2-IT)	47.5	50	44.2
Δ - VideoLLaVA (SFT on VISTA-400K)	**+3.5**	**+7.9**	**-2.3**

older API versions are no longer accessible, we are unable to reproduce the results for some baseline models. In the paper, we report all scores based on our evaluation script.

9. Additional Experimental Results

9.1. Training Data Ablations for VideoLLaVA

As mentioned in Section 4.1, unlike Mantis-Idefics2 and LongVA, we fine-tune VideoLLaVA using a combination of our VISTA-400K and 300K short video samples from VideoChat2-IT to preserve its short video understanding capabilities. In this section, we examine how this additional training data impacts the model's performance on long and high-resolution video understanding tasks after finetuning. To assess this, we finetune another VideoLLaVA model exclusively on our VISTA-400K and compare the results against the combined training approach in Table 5 and Table 6.

As shown in Table 5, finetuning VideoLLaVA exclusively on our VISTA-400K results in consistent improvements across all long video understanding benchmarks. On the other hand, incorporating an additional 300K short video samples does not yield further significant gains in long video understanding. Notably, the Video-MME results indicate that adding short video data slightly detracts from the model's performance on long videos, underscoring the importance of our dataset for enhancing long video understanding capabilities.

For high-resolution video understanding, according to Table 6, finetuning VideoLLaVA on our data leads to a significant improvement (+11.5%) on HRVideoBench. While adding additional short video data further enhances model performance, the improvement is less substantial. These findings suggest that our dataset remains the primary driver of performance gains in high-resolution video understanding. Moreover, incorporating VideoChat2-IT training data leads to a decline in performance on action-related questions, highlighting the superior effectiveness of our dataset for tasks requiring temporal understanding.

9.2. Video-MME w/ Subtitles Results

We show the results for Video-MME w/ subtitles in Table 7. In this evaluation setting, the video's subtitles are provided as part of the question input to the model. The results indicate that both baseline models and our VISTA-finetuned models can be further enhanced by providing extra subtitle information. Similar to Video-MME w/o subtitles results, our VISTA-finetuned models consistently

Table 7. Comparison between VISTA-finetuned models and baseline models on Video-MME w/ subtitle benchmark.

Models	Video-MME w/ subtitles			
	avg	short	medium	long
VideoLLaVA	41.6	46.1	40.7	38.1
VISTA-VideoLLaVA	45.1	50.2	45.7	39.3
Δ - VideoLLaVA	**+3.5**	**+4.1**	**+5.0**	**+1.2**
Mantis-Idefics2	49.0	60.4	46.1	40.3
VISTA-Mantis	50.9	61.8	48.6	42.3
Δ - Mantis-Idefics2	**+1.9**	**+1.4**	**+2.5**	**+2.0**
LongVA	54.3	61.6	53.6	47.6
VISTA-LongVA	59.3	70.0	57.6	50.3
Δ - LongVA	**+5.0**	**+8.4**	**+4.0**	**+2.7**

achieve better performances compared to the baseline models. This shows that our synthetic data provides consistent and model-agnostic enhancements to the long video understanding capability of video LMMs.

10. Discussion on Method Validity

Our method leverages similar techniques to CutMix/Mixup for video augmentation, but the purpose of applying these types of video augmentation is different. In CutMix/Mixup, the original image is perturbed by the overlayed/mixed images such that the model learns to extract localized features from the resulting blend. In our case, the inserted/overlayed videos are still complete, and the model learns to focus and retrieve the inserted/overlayed videos to enhance long-duration/high-resolution video understanding capabilities. In our spatial or temporal NIAH tasks, our method can be ineffective if (1) in spatial NIAH tasks, the needle region becomes too small to recognize; (2) in temporal NIAH tasks, there are too few needle frames that the model missed sampling these frames. We carefully choose the minimum resolution/duration for the needle videos to avoid such scenarios. Specifically, in spatial NIAH tasks, we enforce the width or height of the overlayed videos to be at least 20% of the original video to make the overlayed videos large enough for the video LMMs to recognize. In temporal NIAH tasks, we set the ratio between the inserted videos and the original videos to be at least 1:16 to increase the probability of sampling the inserted frames during training.

11. Limitations

Our method exhibits a few limitations. First, since we generate instruction data based on video captions, and most public video-caption datasets contain simple captions for video clips, our synthesized data often contain short responses, leading to a shorter response from the finetuned models. This issue could be addressed by recaptioning the raw video data using high-capacity video captioning models. Second, while our synthesized augmented video data have been shown to enhance long and high-resolution video understanding, the current video augmentation paradigm does not fully align with real-world video distributions. Addressing this limitation would require more advanced video combination and blending techniques, such as leveraging segmentation maps to isolate specific regions from one video and seamlessly integrating them into another to create more natural and realistic augmented video samples.

12. Instruction Synthesis Prompt Templates

In this section, we list the Gemini prompts we used to synthesize instruction data below.

Freeform QA Generation Prompt

User:
Given a short paragraph of caption describing a video clip, can you try to extract relevant information from the caption and come up with a question-answer pair that could possibly reflect the facts of some local and fine-grained scenes in the video?
The caption of the video is as follows:
<Video Caption>

Please try not to come up with questions that you cannot answer. Please also note that the caption will not be presented in the actual training data. Return only the question and the answer. Format your output as:
###Question###
<your question>
###Answer###
<your answer>

Assistant:
<Synthesized Freeform QA pairs>

Long Video Caption Generation Prompt

User:
Given multiple short captions, each representing a short chunk of video in a longer video, create a detailed caption by combining the short captions such that the detailed caption describes the whole video. Note that because the short captions are from the same video, you can combine entities with slightly different descriptions in different captions, as they most likely represent the same thing. Return only the caption.

The short captions (in chronological order) are listed below:
Caption 1. <Caption 1>
Caption 2. <Caption 2>
...
Caption N. <Caption N>

Assistant:
<Synthesized Long Video Caption>

MCQ Generation Prompt

User:
Given the following Question-Answer pair, turn this short answer question into a multiple-choice question by synthesizing three additional incorrect options. Assume the correct option is <Random Option between A to D>.

Question:
<Question>
Answer:
<Freeform Answer>

Your output should be in the format of a python list:
[
"A. <answer1>",
"B. <answer2>",
"C. <answer3>",
"D. <answer4>"
]

Assistant:
<Synthesized MCQ pairs>

Event Relationship QA Generation Prompt

User:
Given multiple short captions, each representing a short chunk of video in a longer video, generate a question-answer pair related to the order of the events in the video. Note that because the short captions are from the same video, you can combine entities with slightly different descriptions in different captions, as they most likely represent the same thing. Format the output using the following format:
###Question###
<Your question>
###Answer###
<Your answer>

For example, given captions like:
Caption 1: A squirrel is sitting on a tree branch in a forest, surrounded by pine trees and blue sky.
Caption 2: A cartoon squirrel is holding an egg in a tree.
Caption 3: A cartoon squirrel is standing next to an egg.
Your output can be:
###Question###
What happens after the squirrel sits on a tree branch?
###Answer###
The squirrel holds an egg.

Try to be creative with your question and answer.

The short captions (in chronological order) are listed below:
Caption 1. <Caption 1>
Caption 2. <Caption 2>
...
Caption N. <Caption N>

Assistant:
<Synthesized Event Relationship QA pairs>