

# BIMBA: Selective-Scan Compression for Long-Range Video Question Answering

## Supplementary Material

Our supplementary materials contain additional implementation details (Section S1), additional quantitative results (Section S2), and qualitative results (Section S3).

### S1. Additional Implementation Details

**BIMBA-LLaVA** is based on the image-pretrained MLLM LLaVA-NeXT [44], which utilizes CLIP [59] as the vision encoder and Vicuna-7B [11] as the LLM. It processes 64 video frames at a resolution of  $336 \times 336$ , dividing each frame into  $14 \times 14$  patches, yielding  $64 \times 24 \times 24$  spatiotemporal tokens. These tokens are compressed to  $16 \times 12 \times 12$  before being fed into the LLM. In this variant, the vision encoder remains frozen, while the multimodal projector (a linear layer), spatiotemporal token selector, and LLM are trained using LoRA [23].

**BIMBA-LLaMA** is based on the image-pretrained MLLM LLaMA-3.2 [52], incorporating Meta-CLIP [74] as the vision encoder and LLaMA-3.2-LLM-8B as the LLM. It processes 64 video frames at a higher resolution of  $560 \times 560$ , dividing each frame into  $14 \times 14$  patches, resulting in  $64 \times 40 \times 40$  spatiotemporal tokens. These are compressed to  $16 \times 20 \times 20$  before being passed to the LLM. Unlike the other variants, both the vision encoder and multimodal projector remain frozen, with only the spatiotemporal token selector and LLM trained using LoRA.

**Training Details.** We employ standard cross-entropy loss for autoregressive text generation and train the model for 1 epoch with a batch size of 128 and a learning rate of  $2e-5$ . The AdamW [48] optimizer is used, along with a cosine learning rate scheduler and a warm-up ratio of 0.03.

### S2. Additional Quantitative Results

#### S2.1. Performance as a Function of Video Length

In this section, we evaluate the performance of our model on videos of varying lengths from the NextQA [73] dataset, with results presented in Figure S1. Figure S1 (left) shows the relative performance improvement over the PLLaVA [75] baseline for different video durations. We observe that as video duration increases, the relative performance improvement over the baseline becomes more pronounced. This demonstrates the effectiveness of our proposed Mamba-based token compression technique compared to pooling-based methods, particularly for long-range videos.

Similarly, the Figure S1 (right) illustrates the relative performance improvement of BIMBA-LLaVA over the LLaMA-3.2 (video) baseline for varying video durations.

Here, too, we observe that the relative performance gap widens as video duration increases, showcasing the advantages of our model over the vanilla LLaMA-3.2 (video) baseline, which does not use any compression mechanism.

#### S2.2. Computation Cost of BIMBA-LLaMA

In this section, we compare the computational cost of our model with other baselines in terms of GPU memory usage (Figure S2, left) and runtime (Figure S2, right). Our analysis shows that self-attention incurs quadratic costs for both memory and runtime, resulting in out-of-memory errors for inputs longer than 8 frames (12,800 tokens). In contrast, all other methods maintain low memory and runtime costs. Despite having computational efficiency similar to that of the other baselines, our method achieves superior performance, as demonstrated in the previous section.

### S3. Qualitative Results

Our qualitative results include open-ended video question answering (Section S3.1), multiple choice video question answering (Section S3.2), importance of question conditioning (Section S3.3), and significance of bidirectional Mamba and interleaved queries (Section S3.4).

#### S3.1. Open-Ended Video Question Answering

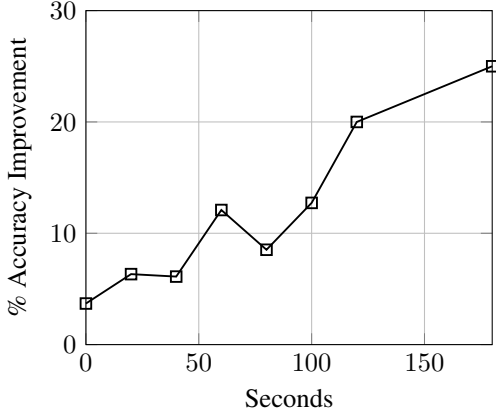
In Figure S3, we provide examples of our model’s performance in open-ended video question answering. The results showcase the model’s ability to handle diverse video understanding tasks, including generating detailed descriptions, recognizing objects and interactions, identifying fine-grained activities, and inferring high-level goals. These examples illustrate the model’s effectiveness in general-purpose video understanding.

#### S3.2. Multiple Choice Video Question Answering

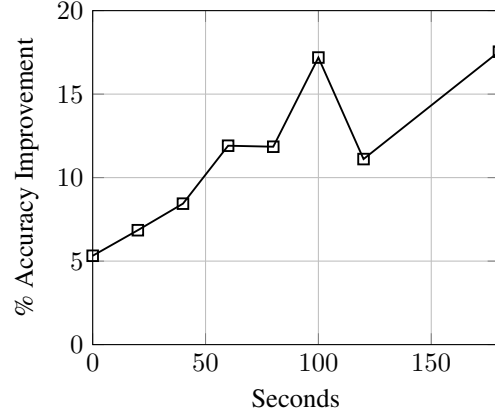
We show qualitative examples of video question answering of our model and other baselines on NextQA (Figure S4) and EgoSchema (Figure S5) datasets. Both BIMBA-LLaVA and BIMBA-LLaMA generate the correct answers while other baselines fail, demonstrating the effectiveness of our model for this task.

#### S3.3. Importance of Question Conditioning

In Figure S6, we showcase example predictions from our model with and without question-conditioned token selection on the NextQA (Figure S6 (a)) and EgoSchema (Figure S6 (b)) datasets. In both cases, incorporating question tokens into our spatiotemporal token selector enables the

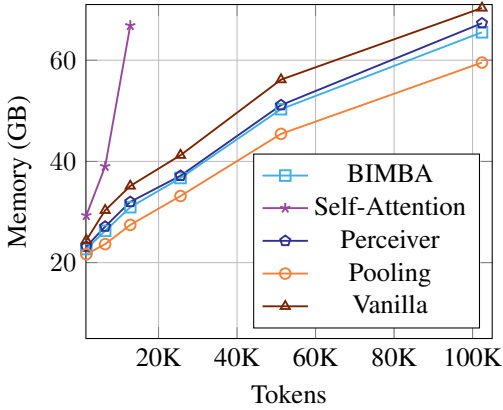


(a) LLaVA Backbone on NeXT-QA.

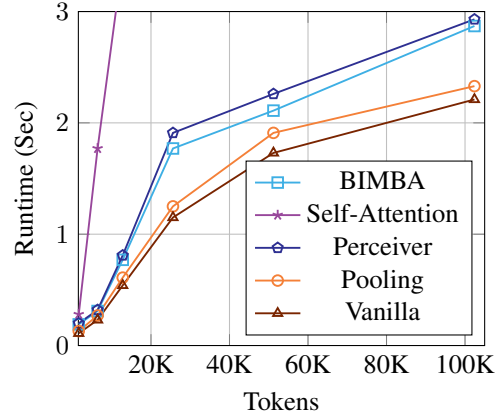


(b) LLaMA Backbone on NeXT-QA.

Figure S1. Relative performance improvement of (left) BIMBA-LLaVA over PLaVA baseline and (right) BIMBA-LLaMA over LLaMA-3.2 (video) baseline for different video durations on NextQA dataset. Our model achieves larger gains as the video length increases.



(a) Memory Usage of LLaMA-3.2 Backbone.



(b) Runtime of LLaMA-3.2 Backbone.

Figure S2. Computation cost of BIMBA-LLaMA and baseline models in terms of memory usage (left) and runtime (right). Self-attention runs out of memory for longer sequences. All other baselines, including our model, maintain low memory and runtime.

model to produce the correct answer. This exhibits the ability of our token selector to leverage question tokens effectively, selecting relevant visual tokens to enhance question-answering performance.

### S3.4. Bidirectional Mamba and Interleaved Queries

In this section, we visualize the effect of bidirectional Mamba and interleaved queries in Figure S7. We calculate a response for each frame as follows: first, we take the hidden states of each token after the spatiotemporal token selector and compute a dot product with the query tokens. Then, we apply max pooling to the dot product values of tokens within each frame to obtain a response for that frame. This response value reflects the weight of each frame in the compressed query representations.

Figure S7 (a) shows that using bidirectional scans and interleaved queries enables our model to capture critical information across the entire video and generate the correct answer. In contrast, (b) with bidirectional Mamba and standard queries, the model focuses mainly on the beginning

and end of the video, and (c) with unidirectional Mamba and standard queries, the model focuses only on the latter part of the video. Both designs are suboptimal, as they miss critical information and produce incorrect answers.



Describe the video in detail.

The video is a short film that features a group of people playing a game of paintball. The game is taking place in a forest, and the players are wearing camouflage clothing and carrying guns. The objective of the game is to eliminate the opposing team by shooting them with dye-filled, breakable, oil and gel paintball markers, shot from a paintball marker. The players are seen running through the forest, shooting at each other, and hiding behind trees.



What is the man wearing at the end of the video?

The man is wearing a purple jacket at the end of the video.



What vehicle is seen in the video?

A car is seen in the video.



Why is the video funny?

The video is funny because it shows a group of people playing a game of paintball, which is a competitive team shooting game. The participants are covered in paint, and they are running around in the woods. The game is intense, and the participants are trying to eliminate each other. The video captures the action and the humor of the game, as the participants are making funny faces and acting silly.



(a) Example 1 of open-ended video question answering.



Describe the video.

A man is in the kitchen. He is preparing his meal. He mixed the ingredients and cooked the food.



What is the man cooking?

The man is cooking chicken.



What are the key steps the man is performing for cooking?

The key steps the main is performing are:  
1. He gathered the ingredients for the meal.  
2. He mixed the ingredients in a bowl.  
3. he put the ingredients in the oven to cook.



Why the man uses the tap at the end of the video?

The man uses the tap to wash utensils at the end of the video.



(b) Example 2 of open-ended video question answering.

Figure S3. Qualitative Results on Open-Ended Video Question Answering. Our model demonstrates the ability to answer a wide range of questions about videos, including detailed descriptions, high-level goals, and fine-grained activities.

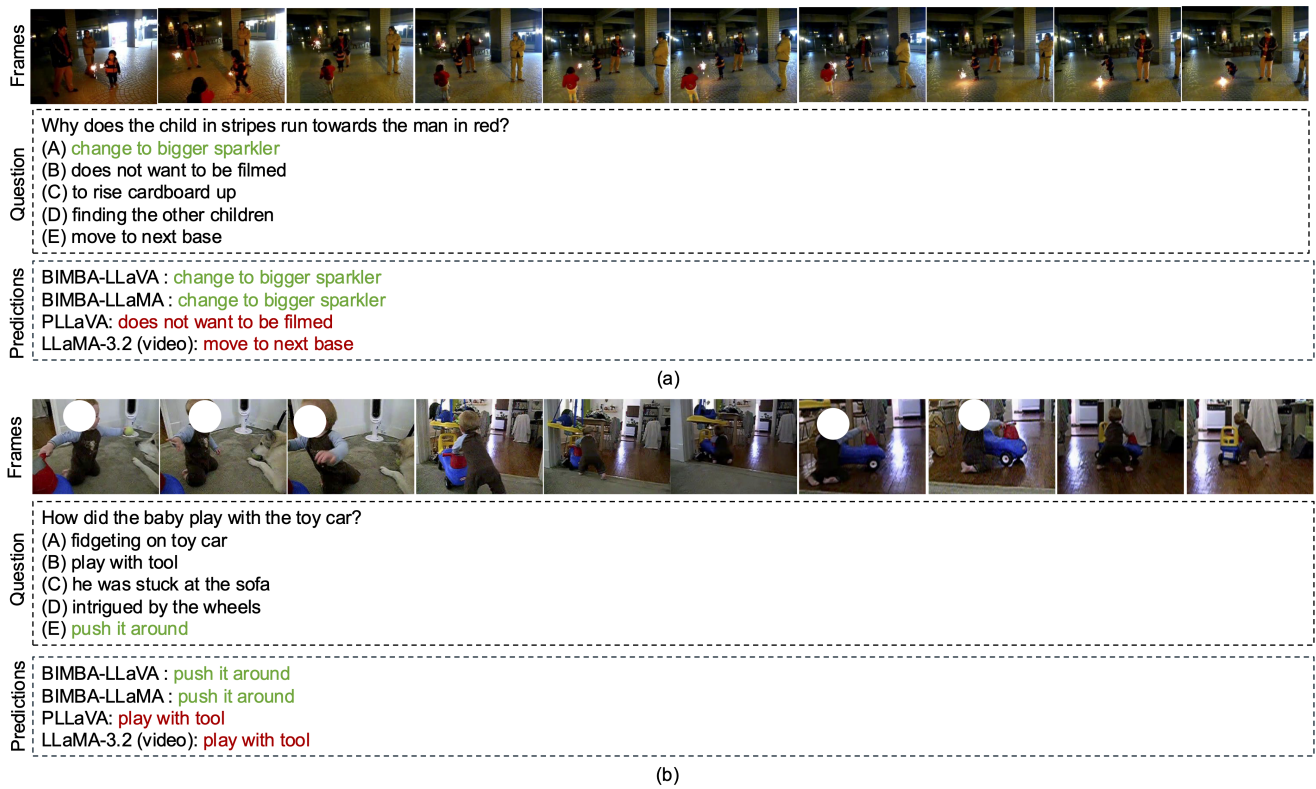
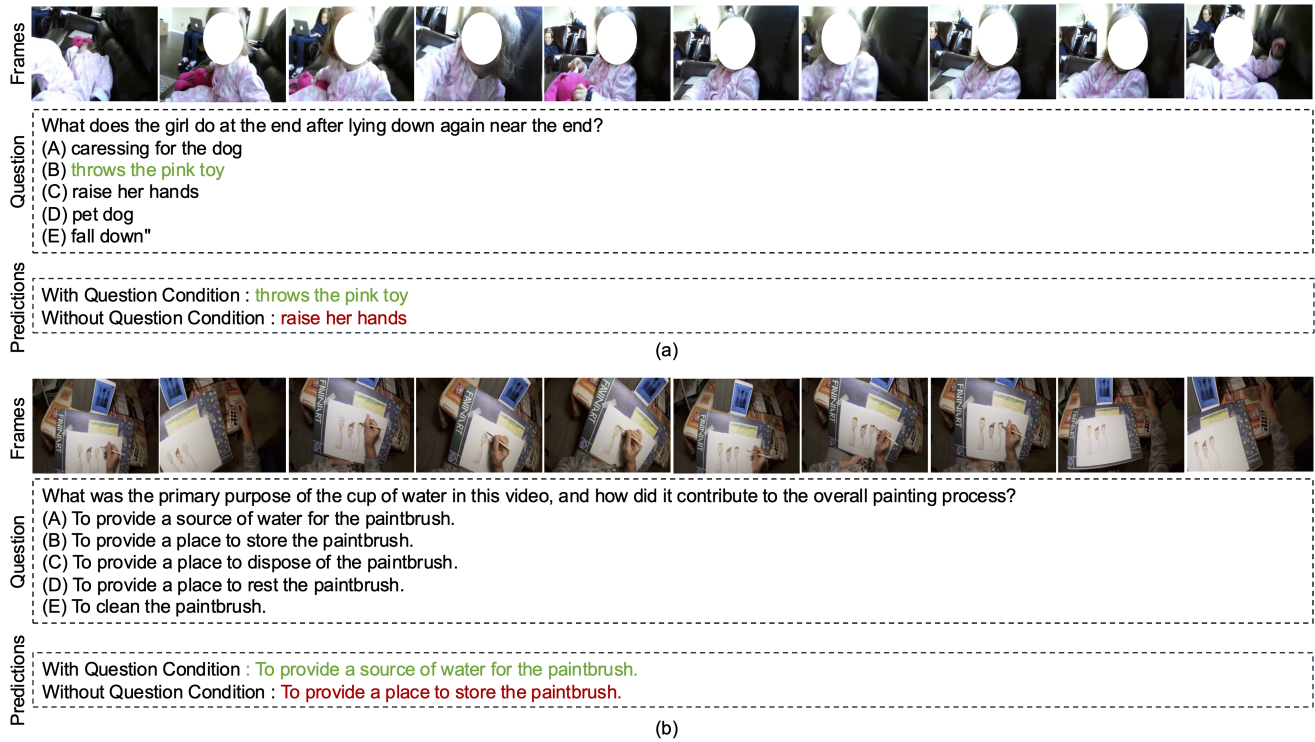


Figure S4. Qualitative Results on NextQA. Our model generates the correct answer while both PLLaVA and LLaMA-3.2 (video) baselines fail.





Figure S5. Qualitative Results on EgoSchema. Our model generates the correct answer while both PLLaVA and LLaMA-3.2 (video) baselines fail.



Frames

Question

What was the primary purpose of the cup of water in this video, and how did it contribute to the overall painting process?

(A) To provide a source of water for the paintbrush.  
 (B) To provide a place to store the paintbrush.  
 (C) To provide a place to dispose of the paintbrush.  
 (D) To provide a place to rest the paintbrush.  
 (E) To clean the paintbrush.

Predictions

With Question Condition : To provide a source of water for the paintbrush.  
 Without Question Condition : To provide a place to store the paintbrush.

Figure S6. Qualitative Results on Question Conditioned Token Selection on (a) NextQA and (b) EgoSchema datasets. Incorporating question tokens into our spatiotemporal token selector leads to the correct answer in both examples. Using the information from the questions allows our spatiotemporal selection module to focus on the most relevant video parts for answering the question.

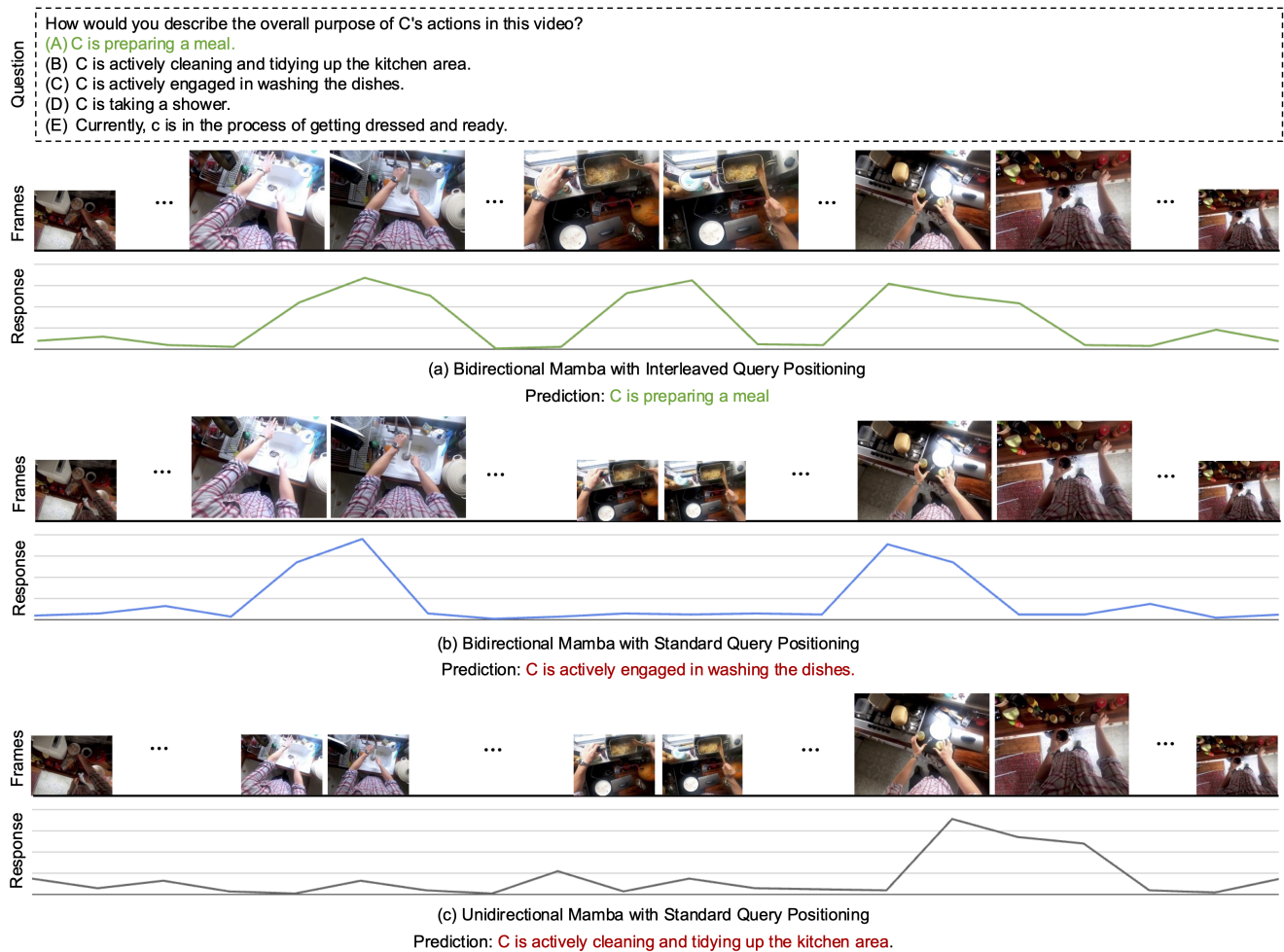


Figure S7. Visualization of Bidirectional Mamba and Interleave Queries. Utilizing bidirectional Mamba and interleaved queries leads to the correct answer, while the unidirectional Mamba and standard queries fail.