Query-based Cross-Modal Projector Bolstering Mamba Multimodal LLM

Anonymous ACL submission

Abstract

The Transformer's quadratic complexity with input length imposes an unsustainable computational load on large language models (LLMs). In contrast, the Selective Scan Structured State-Space Model, or Mamba, addresses this computational challenge effectively. This paper explores a query-based cross-modal projector designed to bolster Mamba's efficiency for vision-language modeling by compressing visual tokens based on input through the cross-attention mechanism. This innovative projector also removes the need for manually designing the 2D scan order of original image features when converting them into an input sequence for Mamba LLM. Experimental results across various vision-language understanding benchmarks show that the proposed cross-modal projector enhances Mamba-based multimodal LLMs, boosting both performance and throughput.

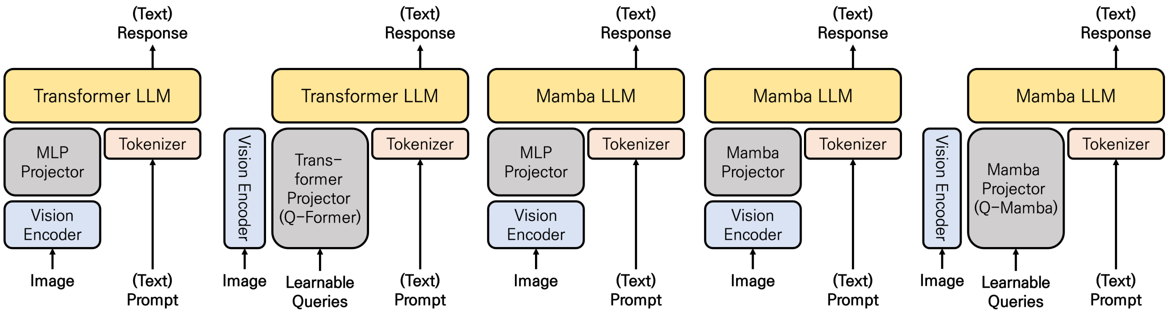
1 Introduction

Multimodal Large Language Models (MLLMs) aim to extend the capabilities of Large Language Models (LLMs) to various modalities, including text and images. By fusing visual information into the textual domain, MLLMs effectively leverage the powerful language generation and logical reasoning abilities of text-only pre-trained LLMs. This integration has demonstrated significant potential in solving real-world vision-language problems, with diverse applications such as visual question answering (VQA) and multimodal dialogue response generation.

The core element behind this advancement lies in the Transformer (Vaswani et al., 2017), an architecture defined by stacked layers of attention mechanisms capable of scaling up to over 100 billion parameters. Due to its capability and flexibility to capture long-term dependencies, the Transformer can better represent different modalities, serving as a foundational model for MLLMs. Unfortunately, the Transformer also inherits intrinsic bottlenecks due to its defining attention mechanism. The computational and memory complexities of self-attention increase quadratically with sequence length, imposing a limit on the input sequence length. Recent efforts have focused on extending the Transformer's context window to overcome this limitation, but the challenge of computational burden remains.

To address this issue, the state-space model (SSM) (Gu et al., 2021, 2022a,b; Fu et al., 2023) has been studied as an alternative architecture for efficiently capturing long-range dependencies. The SSM can be viewed as combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), enabling parallelizable training and fast inference. The latest advancement in SSMs is Mamba (Gu and Dao, 2023), which incorporates an input-dependent gating mechanism that enables selective scanning, along with a hardware-aware algorithm for efficient computation. Mamba matches or even surpasses the performance of advanced Transformers while achieving faster training and inference speeds, leading to applications in various domains, including image (Zhu et al., 2024; Liu et al., 2024b), speech (Jiang et al., 2024; Li and Guo, 2024), and video processing (Li et al., 2024). The utilization of Mamba architecture for MLLM foundation models has been considered (Qiao et al., 2024; Zhao et al., 2024) but not extensively explored. Moreover, there remains a limited understanding of the most effective methods for aligning visual information within the textual domain using Mamba.

Building upon the previous architecture, we introduce a non-trivial Mamba-based architecture for cross-modal projection to connect the pre-trained vision encoder and Mamba-based LLM. Inspired by Querying Transformer (Q-Former) (Li et al., 2023a), we utilize learnable queries to project vision information from image features into 1D



(a) LLaVA (b) BLIP-2 (c) Cobra (d) VL-Mamba (e) Ours

Figure 1: Model comparison between (a) LLaVA (Liu et al., 2023), (b) BLIP-2 (Li et al., 2023a), (c) Cobra (Zhao et al., 2024), (d) VL-Mamba (Qiao et al., 2024), and (e) ours. The key differences stem from the choice of LLM backbone architecture, the design of the projector architecture, and the incorporation of learnable queries for flexibility.

causal tokens by interleaving the Mamba sequence modeling layer and cross-modal attention. Our architectural design is motivated by three key objectives: (1) eliminating the heuristic choice of 2D visual scan order, (2) effectively and dynamically downsampling the projected visual feature sequence length, and (3) enhancing text-image alignment by adopting a structure tailored for Mamba-based multimodal modeling. We further propose MLLM with a pre-trained Mamba LLM backbone connected to the vision encoder using the proposed projector. The overall comparison between the previous models and ours is depicted in Figure 1.

Our contributions can be summarized as follows:

- We propose Querying Mamba, the multimodal connector based on the Mamba module, and the cross-modal attention for adaptive flexibility in downsampling the visual token lengths.
- We propose MLLM based on Querying Mamba and pre-trained Mamba LLM. We meticulously explore a range of choices regarding the components that integrate these models to boost Mamba's effectiveness in multimodal modeling.
- We carry out comprehensive experimental evaluations using multimodal comprehension benchmarks to assess the performance and robustness of our proposed models.

2 Related Works

2.1 State-Space Models (SSMs) and Mamba

Current state-space models are inspired by classical state-space models, which represent continuous systems that map a 1-dimensional function or sequence through an implicit latent state. The Linear State Space Layer (LSSL) (Gu et al., 2021) was one of the earliest attempts at deep SSMs, aiming to enhance sequence modeling performance by stacking multiple SSM layers. Although LSSL demonstrated the potential of deep SSMs for addressing long-range dependencies, its high computational and memory costs rendered it impractical.

The Structured State-Space Model (S4) (Gu et al., 2022a) tackled this bottleneck by re-parameterizing the latent matrix through decomposition into low-rank and normal terms. This innovation led to several variant architectures, such as the Diagonalized State-Space (DSS) (Gupta et al., 2022) and S4D (Gu et al., 2022b), which enabled more efficient and simplified computation via diagonalization. However, S4 and its variants can not remember specific past tokens or compare tokens across the sequence—capabilities crucial for language modeling. Hungry Hungry Hippos (H3) (Fu et al., 2023) aimed to overcome these shortcomings of S4 by incorporating 1-dimensional convolution along the sequence, allowing SSMs to compare and remember past tokens by shifting the input sequence.

The latest work, Mamba (Gu and Dao, 2023), further refines S4 by introducing a selective mechanism that utilizes input-dependent latent state parameters, making the model content-aware and enabling it to selectively focus on relevant information. Mamba also incorporates 1-dimensional convolution shifting from H3 and a gating mechanism similar to Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), which enhances its ability to handle long sequences with increased robustness and flexibility. With parallel associative scanning and a hardware-aware implementation, Mamba achieves efficient training and inference, matching or surpassing the capabilities of advanced Transformers.

The success of Mamba has led to various adaptations across different domains. For instance, several attempts have been made to apply Mamba in speech separation (Li and Guo, 2024; Jiang et al., 2024). In computer vision, Vision Mamba (Vim) (Zhu et al., 2024) and V-Mamba (Liu et al., 2024b) employ bidirectional SSMs to process two-dimensional image data with one-dimensional sequence modeling in Mamba. SiMBA (Patro and Agneeswaran, 2024) further enhances this by incorporating a channel-mixing layer into the Mamba block, analogous to the role of the feedforward network in the Transformer block.

2.2 Multimodal Large Language Models

With the introduction of ChatGPT (Ouyang et al., 2022), also referred to as InstructGPT, Large Language Models (LLMs) have emerged as a dominant approach for real-world natural language processing tasks. These models, typically featuring billions of parameters and trained on extensive corpora, are not only proficient in generating language responses but also in tasks requiring logical comprehension and reasoning. Although InstructGPT has not been publicly released, the research community has been actively developing open-source LLMs (Touvron et al., 2023; Gunasekar et al., 2024; Li et al., 2023c; Zhang et al., 2022), which have shown performance on par with InstructGPT. This progress has led to various adaptations and modifications of pre-trained LLMs for diverse applications.

A notable advancement is the development of Multimodal Large Language Models (MLLMs), which leverage pre-trained LLMs to process multimodal data. This extends beyond the original text-only domain, integrating capabilities to understand both textual and visual inputs. Models like LLaVA (Liu et al., 2023), BLIP(Li et al., 2022, 2023a), and GPT-4(OpenAI, 2024) have shown robust performance in tasks requiring nuanced vision-language integration. These models utilize transformer-based frameworks known for handling long-range dependencies effectively. However, the innate characteristic of high computational demands and slow inference rates of these transformer-based frameworks have started to become a target for recent research, leading to the adoption of the more efficient Mamba architecture in MLLMs. This initiative has given rise to models like Cobra(Zhao et al., 2024) and VL-Mamba(Qiao et al., 2024), which demonstrate promising pathways for enhanced efficiency in MLLM deployment.

Cobra (Zhao et al., 2024) employs a state-space model for multimodal tasks, leveraging the linear scalability of the Mamba architecture. It introduces an innovative approach to vision encoding by merging outputs from DINOv2 (Oquab et al., 2024) and SigLIP (Zhai et al., 2023), thereby generating visual representations that capture both spatial and semantic properties effectively. These outputs are then processed through a learnable projector module, which aligns the visual and textual features by adjusting the dimensions of the visual representations to match those of the Mamba LLM via a multi-layer perceptron. This approach enables Cobra to deliver the same volume of output tokens in just 30% of the time required by comparable 3B transformer-based LLMs, such as TinyLLaVA (Zhou et al., 2024) or MobileVLM v2 (Chu et al., 2024).

Similarly, VL-Mamba (Qiao et al., 2024) builds upon a pretrained Mamba framework and introduces a novel MultiModal Connector (MMC) architecture. This connector features a Vision Selective Scan (VSS) module and two linear layers, which enhance the causal relationships among image blocks from the vision encoder. Furthermore, this paper assesses the performance difference between the Bidirectional-Scan Mechanism (BSM), which scans the image blocks in both forward and backward directions, and the Cross-Scan Mechanism (CSM), which scans both from forward to backward and top to bottom. This paper suggests a preference for the simple BSM, as the two scanning methods show comparable efficacy.

However, the previous projector modules used in Cobra and VL-Mamba have limitations in that these connectors have no flexibility in vision token number, causing longer vision token input, and require manual scan mechanisms that grant causality between image blocks.

3 Method

In this section, we first review the preliminary concepts of structured state-space models and Mamba (Sec. 3.1). Then, we describe the details of the Cross-modal Mamba projector, which extracts

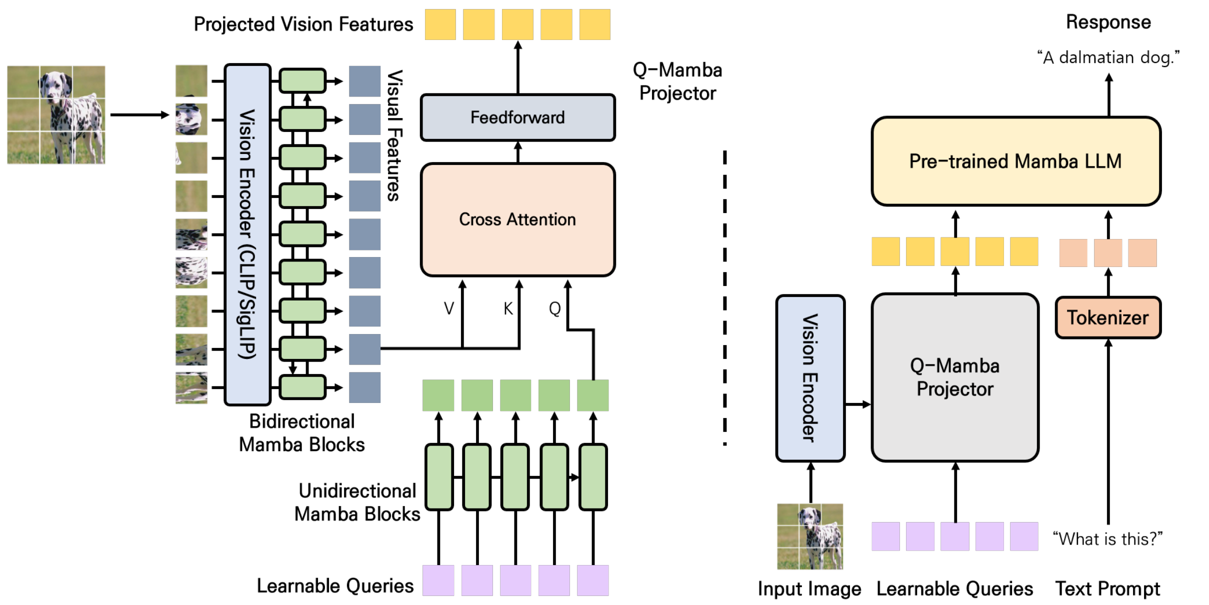


Figure 2: Overall architecture of Querying Mamba (left) and the Multimodal Mamba LLM (right) based on the proposed design. Querying Mamba projects the visual information, which is encoded by a pre-trained vision encoder with an additional bidirectional Mamba layer, into the learnable queries with causal Mamba prior via cross attention. The projected vision features work as vision token inputs for pre-trained Mamba LLM.

the 2-dimensional vision information into a 1-dimensional causal token sequence (Sec. 3.2). Lastly, we describe the two-stage fine-tuning of the multimodal Mamba with our proposed Q-Mamba (Sec. 3.3).

3.1 Preliminaries

State-Space Models (SSMs) (Gu et al., 2021, 2022a; Smith et al., 2023) represent linear time-invariant systems that map a continuous 1-dimensional function or a sequence $x(t) \in \mathbb{R}$ to a corresponding response $y(t) \in \mathbb{R}$, via a hidden state $h(t) \in \mathbb{R}^N$ with N latent dimensions. These systems are characterized by four parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$, which define the system dynamics and outputs as follows:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t) \end{aligned} \tag{1}$$

Typically, the parameter $\mathbf{D}$ is omitted as it can be interpreted as a skip connection, which is computationally straightforward to implement.

In practice, to deal with discrete-time input sequences, SSMs are discretized with matrices $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$. One common discretization method is the Zero-Order Hold (ZOH) method, outlined as:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\mathbf{\Delta A}) \\ \overline{\mathbf{B}} &= (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot (\mathbf{\Delta B}) \end{aligned} \tag{2}$$

where the parameter $\mathbf{\Delta}$ specifies the discretization step size. The reformulated discretized system is given by:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t \\ y_t &= \mathbf{C}h_t \end{aligned} \tag{3}$$

Structured State-Space Model (S4) (Gu et al., 2022a) operates as a time-invariant system, meaning its defining parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{\Delta})$ remain constant across all time-steps. Mamba (Gu and Dao, 2023) addresses this constraint by making $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{\Delta}$ input-dependent, enabling a dynamic gating mechanism based on the input sequence. This allows Mamba to selectively focus on pertinent information, significantly enhancing its language modeling capabilities.

3.2 Cross-Modal Mamba Projector

We propose the cross-modal projector, named Q-Mamba, which integrates the Mamba architecture with cross-attention. The architecture of Q-Mamba is shown in the left side of Figure 2. Q-Mamba comprises stacked Q-Mamba blocks, each containing a Mamba layer, cross-attention, and a feedforward network. The Mamba layer functions as a sequence mixer, while the feedforward network functions as a channel mixer. The set of learnable query embeddings is utilized as the input sequence of the Q-Mamba. The queries form causal dependencies

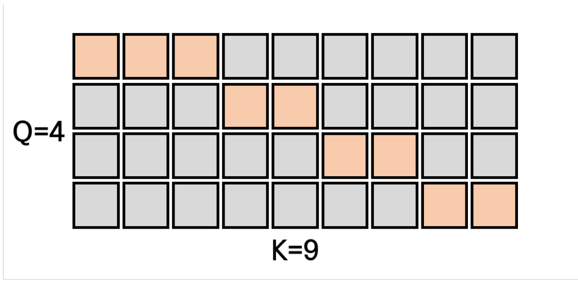


Figure 3: Example of local attention mask applied in the cross-attention layer inside Querying Mamba with 4 queries (Q) and 9 keys (K). Each query attends exclusively to K/Q keys, enabling the focused extraction of information from distinct visual components.

through the sequential Mamba layers, then interact with vision features from the frozen pre-trained vision encoder through cross-attention layers. For the cross-attention, we found that applying local attention mask as in Figure 3 empirically enhances the model performance.

This design offers three key advantages for cross-modal projection. The first advantage is its independence from visual scan order. Previous Mamba-based vision encoders relied on heuristic choices of visual scan order, such as bidirectional or cross-directional scans (Qiao et al., 2024; Zhu et al., 2024; Liu et al., 2024b). Q-Mamba eliminates this dependency by using cross-attention to project vision information from arbitrarily ordered image features onto a one-dimensional query sequence. The second advantage is the flexible choice of query sequence length. Direct application of Mamba on vision feature sequences typically yields projected features of equivalent length, which may be too extensive even for Mamba LLM. Our design, however, facilitates effective downsampling of the vision feature length. Finally, the architecture's resemblance to the Q-Former (Li et al., 2023a) from transformer-based MLLMs ensures proper alignment of text-image features.

We explore several architectural variants to identify the optimal configuration for Q-Mamba. Our investigation includes the use of bidirectional Mamba for preprocessing visual features, the incorporation of a feedforward network for channel mixing, and determining the optimal length of the learnable query sequence. The findings are detailed in Section 4.3.

3.3 Multimodal Mamba Language Model

We introduce the MLLM based on our querying cross-modal projector (Q-Mamba). As shown in Figure 3.2, the overall architecture consists of a pre-trained vision encoder, our cross-modal projector, and a pre-trained Mamba LLM. Initially, visual features are extracted from the input image using the vision encoder. These features are then processed by our projector, which outputs queries embedded with projected visual information. Subsequently, this output sequence is combined with a tokenized text prompt and fed into the Mamba LLM, which generates the corresponding text response.

Training We adopt a two-stage training scheme from LLaVA (Liu et al., 2023), where the initial stage involves aligning the projected features within the frozen LLM using a filtered visual instruction-following dataset. The subsequent stage entails end-to-end fine-tuning of both the projector and the LLM using an extensive visual instruction-following dataset.

4 Experiments

4.1 Settings

Datasets For the fine-tuning stage, we follow the existing two-stage training paradigm and dataset based on LLaVA (Liu et al., 2023) with additional datasets. For the alignment stage, we use a filtered dataset from CC3M with 595K image-text pairs. For the end-to-end fine-tuning stage, we use the combined dataset consisting of LLaVA v1.5 mixed dataset (Liu et al., 2023) with 655K visual conversations, LVIS-Instruct-4V (Wang et al., 2023) dataset with 220K context-aware visual instruction pairs, and LRV-Instruct dataset (Liu et al., 2024a) with 400K visual instruction pairs aimed for hallucination mitigation.

Models For the pre-trained vision encoder, we employ pre-trained SigLIP (Zhai et al., 2023), which encodes vision features for each patched image. We utilize a ViT structure with 400 million parameters. The input image resolution is configured at 384×384, and the total number of visual features is 729. We also attached a bidirectional multimodal connector from trained VL-Mamba (Qiao et al., 2024) to the vision encoder. The output of the multimodal connector is used as a vision feature input for the Q-Mamba projector.

The backbone of our model is the pre-trained Mamba (Gu and Dao, 2023) LLM, which consists

Name	Query Length	VQA^{v2}	GQA	VizWiz	VQA^{T}	POPE	MMB	sec/iter
Cobra*	-	75.38	58.16	49.22	44.9	87.6	56.2	-
VL-Mamba*	-	74.38	56.69	51.66	48.7	83.9	57.0	7.26
+ forward scan only	-	72.34	51.92	29.17	45.6	85.9	56.7	-
+ backward scan only	-	72.06	52.42	34.92	45.1	86.1	55.9	-
Ours	128	74.51	57.59	51.03	47.1	**87.9**	57.2	**5.52**
Ours	256	75.01	58.10	50.53	48.8	86.9	57.7	6.54
Ours	512	75.42	58.37	48.90	50.2	86.5	57.6	5.94
Ours	729	**75.62**	**58.33**	**49.30**	**51.2**	86.8	**58.0**	7.54

Table 1: Comparison with Multimodal Mamba LLMs on 6 benchmarks: VQA^{v2} (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018), VQA^{T} (TextVQA) (Singh et al., 2019), POPE (Li et al., 2023b), and MMB (MMBench) (Yuan Liu, 2023). * indicates the results were reproduced. We also examined variants of the previous Multimodal Mamba LLMs: + forward scan only and + backward scan only indicate the visual scanning order of multimodal connector inside VL-Mamba (Qiao et al., 2024). We also report the time consumed per fine-tuning iteration in seconds.

of 2.8 billion parameters. This model was initially pre-trained on the SlimPajama datasets (Soboleva et al., 2023) for 600 billion tokens, instruction-tuned on the UltraChat 200K dataset (Ding et al., 2023), and then fine-tuned on the UltraFeedback dataset (Cui et al., 2023) using Direct Preference Optimization (DPO) (Rafailov et al., 2023).

For the Q-Mamba projector, we stack 24 blocks with an inner dimension of 768. This choice of hyperparameter is to copy the pre-trained weights of Mamba (Gu and Dao, 2023) with the size of 130M parameters.

Training We train the model using four NVIDIA A100 80GB GPUs. During training, we leverage the PyTorch Fully Sharded Data Parallel (Zhao et al., 2023) framework, utilizing automatic mixed-precision with FP32 and BF16 for efficient distributed training. The batch sizes are set to 256 for the alignment stage and 128 for the end-to-end fine-tuning stage. We employ the Rectified Adam (RAdam) optimizer (Liu et al., 2020), coupled with a cosine decay learning rate scheduler. The learning rates are set at 1×10^{-4} for the alignment stage and 2×10^{-5} for the end-to-end fine-tuning, both with a warmup ratio of 0.03. Each training stage is conducted in a single epoch.

Evaluation To validate the performance of our model, we benchmarked it against five different datasets: VQA-v2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018), Text-VQA (Singh et al., 2019), POPE (Li et al., 2023b) and MMBench (Yuan Liu, 2023). Each dataset offers unique challenges and measures different aspects of the model's capabilities:

- VQA-v2 (Goyal et al., 2017) evaluates the model's general ability to reason over Vision-Question pairs.
- GQA (Hudson and Manning, 2019) extends VQA-v2 by testing the model's reasoning skills across a broader spectrum, incorporating spatial understanding and multi-step inference along with various reasoning skills.
- VizWiz (Gurari et al., 2018), similar to VQA-v2, includes unanswerable questions, thereby assessing the model's ability to identify when a question cannot be answered.
- Text-VQA (Singh et al., 2019) specifically measures the model's proficiency in recognizing text within images and answering related questions.
- POPE (Li et al., 2023b) differentiates itself by focusing on the model's susceptibility to hallucination problems. It provides a score based on the probability of the given answer, hence evaluating the likelihood that the model avoids generating incorrect information.
- MMBench (Yuan Liu, 2023) evaluates the multi-modal capabilities of vision-language models across 20 distinct abilities, including object localization, social reasoning, and fine-grained perception. It introduces a novel CircularEval strategy, ensuring comprehensive evaluation through multiple passes of QA to reduce biases and improve reliability.

Attention	VQA^{v2}	GQA	VizWiz	VQA^{T}	POPE
Global	73.12	52.87	49.09	44.0	85.1
Local	75.01	58.10	50.53	48.8	86.9

Table 2: Comparison between global attention and local attention for cross-attention layer inside our cross-modal Mamba projector. We used 256 learned queries for both models.

Bi-directional Mamba	VQA^{v2}	GQA	VizWiz	VQA^{T}	POPE
From Scratch	74.22	56.30	53.12	48.0	86.4
From Trained	75.01	58.10	50.53	48.8	86.9

Table 3: Comparison between using bidirectional multimodal connector inside vision encoder from scratch or from trained VL-Mamba (Qiao et al., 2024). We used 256 learned queries and local attention for both models.

Visual Scan Order	VQA^{v2}	GQA	VizWiz	VQA^{T}	POPE
Forward Only	76.58	58.44	50.00	50.0	86.9
Bidirectional	75.01	58.10	50.53	48.8	86.9

Table 4: Comparison between using raster scan only or bidirectional multimodal connector inside vision encoder from trained VL-Mamba (Qiao et al., 2024). We used 729 learned queries and local attention for both models.

4.2 Results

As presented in Table 1, our model consistently outperforms previous state-of-the-art Mamba-based multimodal models across all benchmarks. Specifically, the Q-Mamba with 729 queries achieves the highest overall performance, demonstrating significant improvements in tasks that require nuanced vision-language integration. Notably, our model shows remarkable gains in the VizWiz and Text-VQA datasets, which assess the model's ability to understand and interpret textual information within images.

The results indicate that increasing the number of queries generally improves performance. For instance, moving from 128 to 256 queries results in substantial performance gains across all benchmarks, highlighting the importance of having a sufficient number of queries to capture detailed visual information. Further increasing the number of queries to 512 and 729 continues to improve performance, though the gains are less pronounced compared to the initial increase. This suggests that while more queries help in capturing more information, there is a point of diminishing returns where additional queries contribute less to overall performance.

Compared to Cobra and VL-Mamba, the Q-Mamba design proves to be more effective in dynamically downsampling visual token sequences and eliminating the need for manual visual scan orders, contributing to higher throughput and better alignment of visual and textual information. The flexibility in choosing the query sequence length allows for a tailored balance between computational efficiency and model performance, making Q-Mamba adaptable to various application requirements.

4.3 Ablation Studies

In our ablation study, we meticulously analyzed various configurations to determine how different components within Q-Mamba affect model performance. Our initial investigations focused on the type of cross-attention mechanism employed, with results detailed in Table 2. These findings demonstrate that local attention significantly outperforms global attention in enhancing model performance. We then evaluated the effect of utilizing pre-trained weights for the bidirectional Mamba connector within the vision encoder, with outcomes presented in Table 3. The results confirm that leveraging weights from a trained VL-Mamba model leads to performance improvements. Finally, we explored the influence of the visual scan order in the bidirectional Mamba connector, as shown in Table 4. Interestingly, our data indicate that although the model is trained with a bidirectional scan setting, employing only a forward Mamba for inference does not decrease performance and can even enhance it.

5 Conclusion

This paper presents a query-based cross-modal projector designed to enhance Mamba's efficiency in multimodal vision-language modeling. By using the cross-attention mechanism between the learnable queries and the outputs of the visual encoder within a Mamba architecture, the proposed multimodal projector dynamically compresses visual tokens based on an input image context, eliminating the need for manually designing of the 2D scan order of image features. Experimental results on diverse vision-language understanding benchmarks demonstrate that the proposed cross-modal projector boosts the effectiveness of Mamba-based MLLMs.

Limitations

Despite the promising results, our approach has several limitations that need to be addressed in future work. The primary limitation is related to the amount and quality of the dataset used for training and fine-tuning the model.

For the alignment process, we used the LLaVA-LLVIS dataset, and for the fine-tuning process, we used the LLaVA-1.5 dataset. Both of these datasets are filtered and curated to ensure quality, but their limited size compared to the vast datasets typically used in training large language models (LLMs) can restrict the model's ability to generalize across diverse vision-language tasks. Specifically, we ran one epoch for each stage of our training process, whereas other models in the same domain were fine-tuned for two epochs instead of one. This difference in training duration can result in less robust model performance, as the additional epochs in other models allow for more comprehensive learning and fine-tuning of the parameters.

Additionally, the Mamba architecture are liable to "forget." The hidden states of the Mamba model take input and output sequentially, similar to how hidden states within the RNN would, where the current state depends on the previous inputs and hidden state outputs. This sequential dependency can potentially result in forgetting issues that plagued the RNN/LSTM-based models, if the input would be long enough.

It would be also necessary to pretrain the proposed Q-Mamba more thoroughly including contrastive learning as used in Q-Former based on image-text pair datasets. In addition, the parameters of Q-Mamba can be initialized by the pretrained compact Mamba LLM. Also, it would be helpful to perform more in-depth analysis on the resulting attention map for each query according to different input images.

Potential Risk

This paper presents a new architecture of a Large Language Model with over a billion parameters, which can cause potential discrimination in the use of these methods due to the disparity in access to computational resources. Also, the hallucination of Large Language Model can cause potential bias or harm when generating response.

References

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *Preprint*, arXiv:2402.03766.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

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

Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. 2023. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, Los Alamitos, CA, USA. IEEE Computer Society.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

Albert Gu, Karan Goel, and Christopher Re. 2022a. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. 2022b. On the parameterization and initialization of diagonal state space models. *ArXiv*, abs/2206.11893.

Albert Gu, Isys Johnson, Karan Goel, Khaled Kamal Saab, Tri Dao, Atri Rudra, and Christopher R'e. 2021. Combining recurrent, convolutional, and continuous-time models with linear state-space layers. In *Neural Information Processing Systems*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Conti Kauffmann, Gustavo Henrique de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Behl, Xin Wang, Sebastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2024. Textbooks are all you need.

Ankit Gupta, Albert Gu, and Jonathan Berant. 2022. Diagonal state spaces are as effective as structured state spaces. In *Advances in Neural Information Processing Systems*.

Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.

Xilin Jiang, Cong Han, and Nima Mesgarani. 2024. Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. *arXiv preprint arXiv:2403.18257*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Kai Li and Chen Guo. 2024. Spmamba: State-space model is all you need in speech separation. *arXiv preprint arXiv:2404.02063*.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videomamba: State space model for efficient video understanding. *Preprint*, arXiv:2403.06977.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023c. Textbooks are all you need ii: phi-1.5 technical report. *Preprint*, arXiv:2309.05463.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024b. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Badri N. Patro and Vijay S. Agneeswaran. 2024. Simba: Simplified mamba-based architecture for vision and multivariate time series. *Preprint*, arXiv:2403.15360.

Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. 2024. Vl-mamba: Exploring state space models for multimodal learning. *Preprint*, arXiv:2403.13600.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. 2023. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *ArXiv*, abs/2311.07574.

Yuanhan Zhang Bo Li Songyang Zhang Wangbo Zhao Yike Yuan Jiaqi Wang Conghui He Ziwei Liu Kai Chen Dahua Lin Yuan Liu, Haodong Duan. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. 2023. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, Los Alamitos, CA, USA. IEEE Computer Society.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. 2024. Cobra: Extending mamba to multi-modal large language model for efficient inference. *Preprint*, arXiv:2403.14520.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Preprint*, arXiv:2304.11277.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *Preprint*, arXiv:2402.14289.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *Preprint*, arXiv:2401.09417.