VL-Mamba: Exploring State Space Models for Multimodal Learning

Yanyuan Qiao[1], **Zheng Yu**[1], **Zijia Zhao**[2,3], **Sihan Chen**[2,3]
Mingzhen Sun[2,3] , **Longteng Guo**[2], **Qi Wu** [1]*, **Jing Liu**[2,3]
[1]Australian Institute for Machine Learning, The University of Adelaide
[2]Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
{yanyuan.qiao, zheng.yu, qi.wu01}@adelaide.edu.au
{sihan.chen, longteng.guo, jliu}@nlpr.ia.ac.cn {zijia.zhao, mingzhen.sun}@ia.ac.cn

Project URL: https://yanyuanqiao.github.io/vl-mamba

Abstract

Multimodal large language models (MLLMs) have gained considerable attention due to their ability to integrate visual and textual information, enhancing understanding and providing context for complex tasks. While Transformer-based architectures have been the dominant framework for MLLMs, recent studies suggest that state space models (SSMs) like Mamba can achieve competitive or even superior performance. However, no prior research has investigated the potential of SSMs to replace Transformers in multimodal tasks, which are inherently more challenging due to the heterogeneity of visual and language data and the complexities of aligning these modalities. In this paper, we introduce VL-Mamba, the first study to explore the application of state space models in multimodal learning tasks. VL-Mamba leverages a pretrained Mamba language model as its core, and we propose a novel MultiModal Connector (MMC) that incorporates a Vision Selective Scan (VSS) module to improve visual sequence modeling. We empirically explore how to effectively apply the 2D vision selective scan mechanism for multimodal learning and the combinations of different vision encoders and variants of pretrained Mamba language models. Our experiments across multiple multimodal benchmarks demonstrate that VL-Mamba achieves competitive performance against small MLLMs of similar size, and in some cases, surpasses larger models such as the 7B and 13B versions of LLaVA-1.5. These results suggest that state space models have the potential to serve as an alternative to Transformers in multimodal learning tasks.

1 Introduction

Multimodal large language models (MLLMs) have recently gained significant attention in the research community, building on the advanced capabilities of large language models (LLMs) such as powerful language expression and logical reasoning. By integrating both visual and textual information, MLLMs enhance the understanding of visual content and provide a more comprehensive context for language understanding and generation. These models have demonstrated significant potential in addressing real-world visual problems, with applications spanning various fields of vision and language, including image captioning (1; 2), referring expression comprehension (REC)(3; 4), and visual question answering (VQA)(5; 6). Leveraging transformer-based architectures (7) and large-

[*] Corresponding author: Qi Wu

4th NeurIPS Efficient Natural Language and Speech Processing Workshop (ENLSP-IV 2024).

scale web-sourced datasets, MLLMs have become a cornerstone of modern artificial intelligence research.

Although Transformer has become the mainstream framework of MLLM due to its effectiveness, recent studies have shown that state space models (SSM) such as Mamba can achieve or even exceed the performance of Transformer in some aspects, becoming an alternative possibility to replace the transformer structure. Since the attention mechanism in the Transformer architecture requires quadratic complexity calculation, which brings a huge computational burden, more studies have extended Mamba from natural language processing (NLP) to other fields (8; 9; 10). For example, in the field of computer vision, Vision Mamba (Vim) (11) integrates Mamba into the Vision Transformer (ViT) framework, using bidirectional SSM for data-dependent global visual context modeling and position encoding, enabling position-aware visual understanding. VMamba (12) introduced a cross-scanning mechanism to connect one-dimensional array scanning with two-dimensional plain traversing. In the biomedical image segmentation task, U-Mamba (13) introduced a hybrid CNN-SSM architecture capable of capturing both fine local details and long-range dependencies within images.

Though these works have achieved remarkable results in vision tasks, however, to the best of our knowledge, no research has explored whether state-space models (SSMs) such as Mamba can be used as an alternative to Transformer for multimodal tasks. Multimodal tasks are more challenging than unimodal tasks, mainly due to the heterogeneity of information in different modalities and the complexity of alignment. Modalities such as vision and language have completely different characteristics: visual data is continuous and spatially correlated, while language data is discrete and symbolic. To effectively fuse the information of these two modalities, the model needs to be able to handle the huge differences between them.

In this paper, we present VL-Mamba, the first study to explore the use of multiple state space models for multimodal learning tasks. To be specific, as illustrated in Fig. 1, we leverage the pre-trained Mamba language model as our backbone language model instead of conventional Transformer-based language models such as LLama (14) or Vicuna (15). Furthermore, we empirically explore the way to apply 2D vision selective scan mechanisms for VL-Mamba and introduce a novel MultiModal Connector (MMC) architecture, comprising a Vision Selective Scan (VSS) module and two linear layers, tailored to enrich the 2D-causal modeling of visual sequences. For the VSS module, we explore two distinct scan mechanisms: the Bidirectional-Scan Mechanism (BSM) and the Cross-Scan Mechanism (CSM). The BSM conducts scans of visual sequences in both forward and backward directions, while the CSM extends scanning capability to four directions. In addition, we study the combinations of different vision encoders, variants of pretrained Mambe language models, and multimodal connectors to find the effect of different components for VL-Mamba.

Extensive experiments are conducted on various multimodal learning benchmarks to verify the effectiveness of VL-Mamba. Our model achieves competitive performance with other small MLLMs of similar size and even outperforms large MLLMs (e.g., 7B and 13B versions of LLaVA-1.5 (16)) on some popular benchmarks.

The contributions of this study are summarized as follows:

- We propose VL-Mamba, the first work to explore and apply state space models to multimodal learning tasks, offering a novel alternative framework for multimodal large language models beyond Transformer-based architectures.
- We empirically investigate the impact of various components within VL-Mamba and introduce a novel MultiModal Connector, which includes a Vision Selective Scan (VSS) module, enhancing the model's representational capacity.
- We conduct extensive experiments on a wide range of multimodal learning benchmarks, demonstrating that VL-Mamba achieves competitive performance compared to existing multimodal large language models.

2 Related Work

2.1 Multimodal Large Language Model

With the development of the powerful Large Language Models (LLMs) (14; 17; 18), many studies (19; 20; 21; 22; 23; 24; 25) extend LLMs to multimodal domains by combining visual input with

LLM to build the multimodal large language model (MLLM). Flamingo (26) freezes pre-trained visual encoders and large language models and fuses visual and language modalities with gated cross-attention, demonstrating excellent few-shot learning performance. BLIP (27) uses a dataset bootstrapped from large-scale noisy image-text pairs to pre-train a multi-modal mixture of encoder-decoder models by injecting different synthetic captions and removing noisy captions. Based on this, BLIP-2 (28) uses Querying Transformer (Q-Former) to bridge the modal gap. InstructBLIP (29) further proposes an instruction-aware visual feature extraction mechanism that can flexibly and effectively extract visual information features according to the given instructions. LLaVA (16; 30) leverages advanced LLMs (*i.e.* LLaMA (14) and Vicuna (15)) as the language model and CLIP (31) as the visual encoder, it transforms visual tokens into language tokens with a simple MLP layer. MiniGPT-4 (32) directly aligns visual information with the language model to accomplish diverse vision-language tasks without using external vision models. Usually, the training of MLLMs contains two stages, of which the first stage is to pretrain the model on a large collection of image-text pairs to acquire the alignment of vision-language knowledge, and the second stage is to finetune the model with a smaller but high-quality multimodal instruction tuning dataset with a designed conversational template. These MLLM works have greatly advanced research in the fields of computer vision and natural language processing. However, since the main framework of these models relies on Transformers, the attention mechanism in Transformers inherently has high computational complexity in inference for long sequences. In this paper, we propose the VL-Mamba, which is based on the state space model. To be specific, we utilize pretrained Mamba (33) language model as our backbone language model, rather than Transformer-based LLMs such as LLama (14) or Vicuna (15). Instead of directly using a simple MLP layer, we propose a MultiModel Connector (MMC) that contains a Vision Selective Scan (VSS) module and two linear layers. Moreover, we empirically explore the effective application of 2D selective scan mechanism in the multimodal VL-Mamba and the combination of different vision encoders and variants of Mamba language models.

2.2 State Space Models

Modern state space models (SSMs) are derived from the classical state space model (34) and have become an efficient building block for constructing deep networks, thereby achieving cutting-edge performance in analyzing continuous long-sequence data. They particularly excel at capturing long-range dependencies (LRDs) and leveraging parallel training methods to increase efficiency. Initiated by a HiPPO matrix (35), Linear State Space Layer (LSSL) (36) combines the advantages of continuous-time models (CTMs), RNNs, and CNNs, which demonstrates the potential of deep SSMs to solve long-range dependencies. However, the practical feasibility of LSSL is hampered by the large computational and memory requirements imposed by the state representation. Subsequently, the Structured State Space (S4) (37) addresses the main computational bottleneck in prior research. This is achieved through novel parameterizations catering to continuous-time, recurrent, and convolutional views of the state space model, thereby effectively modeling long-range dependencies. S4 has subsequently seen some variants (38; 39; 40), such as the Diagonal State Space (DSS) model (39), which forces the state matrix to be a diagonal matrix, making it easier to formulate, implement, and analyze, and can be proven to be as expressive as a general state space, while S4D (40) provides a new mathematical analysis for DSS initialization, making it simpler and more efficient. A recent work, named Mamba (33), further improves S4 with a selection mechanism that incorporates time-varying parameters into SSM, allowing the model to select relevant information in an input-dependent manner. It proposes a hardware-aware algorithm to achieve efficient training and inference. Mamba's superior scaling performance shows that it is a promising alternative to the Transformer in long-sequence modeling. Many works extend Mamba from Natural Language Processing (NLP) to other fields (8; 9; 10). Vision Mamba (Vim) (11) applies Mamba to the Vision Transfomer (ViT) architecture, and combines bidirectional SSM for data-dependent global visual context modeling and position embedding for location-aware visual understanding. Visual State Space Model (VMamba) (12) designs a cross-scan mechanism to bridge the gap between 1-D array scanning and 2-D plain traversing. U-Mamba (13) proposes a hybrid CNN-SSM architecture to capture both localized fine-grained features and long-range dependencies in images, to solve the biomedical image segmentation task. In this work, we explore how to transfer the success of Mamba to solve the more challenging multimodal learning tasks, which often require modeling of both vision and language modalities and complex reasoning.

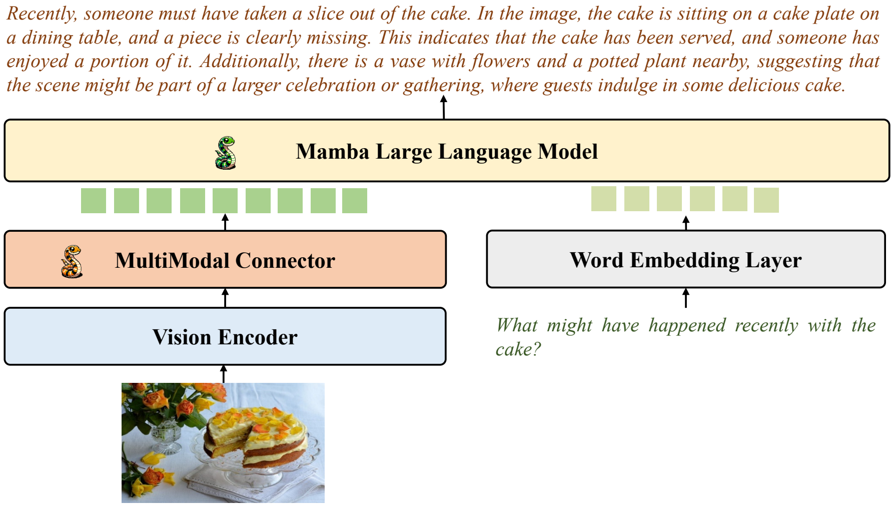

Recently, someone must have taken a slice out of the cake. In the image, the cake is sitting on a cake plate on a dining table, and a piece is clearly missing. This indicates that the cake has been served, and someone has enjoyed a portion of it. Additionally, there is a vase with flowers and a potted plant nearby, suggesting that the scene might be part of a larger celebration or gathering, where guests indulge in some delicious cake.
Mamba Large Language Model
MultiModal Connector
Word Embedding Layer
Vision Encoder
What might have happened recently with the cake?

Figure 1: The architecture of VL-Mamba. It contains a vision encoder, a multimodal connector (MMC), and a language model. We utilize the pre-trained Mamba Large Language Model (Mamba LLM) as its language model, and the pre-trained Vision Transformer model as its vision encoder.

3 Method

In this section, we first introduce the preliminary concepts of state space models. Then, we describe the details of our proposed VL-Mamba, which mainly includes the Vision Encoder, MultiModal Connector, and the Mamba LLM.

3.1 Preliminaries

State space models (SSMs) (33) are commonly considered linear time-invariant systems that map stimulation $x(t) \in \mathbb{R}^L$ to response $y(t) \in \mathbb{R}^M$ through a hidden state $h(t) \in \mathbb{R}^N$. Mathematically, these models are typically formulated as linear ordinary differential equations (ODEs), where the parameters include $\mathbf{A} \in \mathbb{C}^{N \times N}$, $\mathbf{B} \in \mathbb{C}^N$ for a state size N, and the skip connection $\mathbf{D} \in \mathbb{C}^1$. The system dynamics and output equations are given by:

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

Subsequently, the process of discretization is commonly employed to incorporate Eq. 1 practical deep learning algorithms. In this context, $\mathbf{\Delta}$ represents the timescale parameter that is used to convert the continuous parameters $\mathbf{A}, \mathbf{B}$ into discrete parameters, $\overline{\mathbf{A}}, \overline{\mathbf{B}}$. The zero-order hold (ZOH) method is commonly utilized for this discretization, and it is described as follows:

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

Once discretized, Eq. 2 can be reformulated with the step size Δ as:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{k-1} + \overline{\mathbf{B}}x_k, \\ y_t &= \mathbf{C}h_k + \mathbf{D}x_k. \end{aligned} \tag{3}$$

Nevertheless, the formulation in 3 is predicated on a Linear Time Invariance (LTI) system where parameters are invariant despite changes in the input. To address this constraint, the recent work Mamba (33) explored integrating a selective scan technique, in which the matrices $\overline{\mathbf{B}}$, $\mathbf{C}$, and $\mathbf{\Delta}$ are derived from the input data. This change equipped Mamba with the ability to dynamically focus on information from the input sequence, which increased the model's capability.

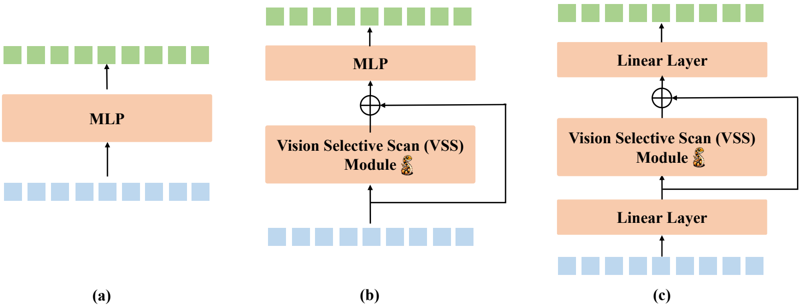

MLP
MLP
Vision Selective Scan (VSS) Module
Linear Layer
Vision Selective Scan (VSS) Module
Linear Layer
(a)
(b)
(c)

Figure 2: Three architectures of MultiModal Connector: (a) MLP; (b) VSS-MLP; (c) VSS-L2.

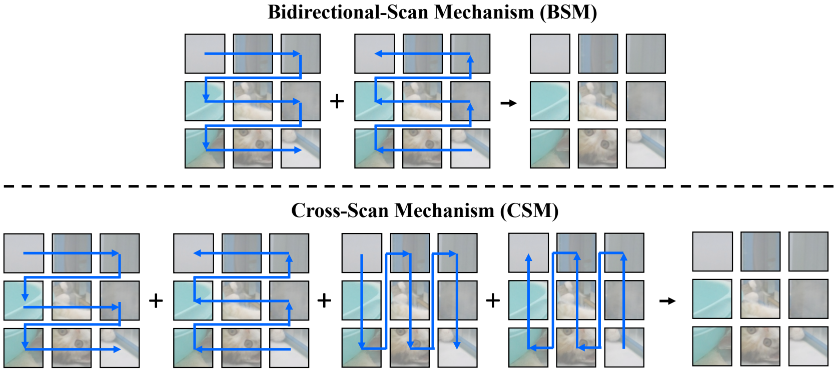

Bidirectional-Scan Mechanism (BSM)
+
Cross-Scan Mechanism (CSM)
+
+
+

Figure 3: Illustration of two different Vision Selective Scan (VSS) Mechanisms: Bidirectional-Scan Mechanism (BSM) (top) and Cross-Scan Mechanism (CSM) (bottom).

3.2 Architecture

As shown in Fig. 1, the architecture of VL-Mamba consists of a pretrained vision encoder, a randomly initialized MultiModal Connector (MMC) which incorporates the 2D vision selective scan mechanism, and a pretrained Mamba Large Language Model (Mamba LLM). Taking an image as input, we first obtain visual features through the visual encoder, then feed the visual sequences into MMC, and then this output vector combined with a tokenized text query is fed into Mamba LLM to generate the corresponding response.

Vision Encoder The vision encoder of VL-Mamba uses the Vision Transformer (ViT) (41) architecture that generates a sequence of patch features from raw images. The vision encoder f_V, takes an image I as input and produces a sequence of the visual patch features V_{img}, as follows:

$$V_{img} = f_V(I). \tag{4}$$

MultiModal Connector (MMC) Since the state space models are designed to process 1D sequential data such as language sequences that have causal relationships, but the visual sequences generated by the vision encoder are non-causal data, 2D vision selective scan mechanisms are proposed to solve computer vision tasks. In this work, we try to apply the 2D vision selective scan mechanisms for multimodal learning by ensembling them in the multimodal connector of VL-Mamba. Specifically, we explore three variants of multimodal connectors:

- **MLP**: a two-layer Multi-Layer Perceptron (MLP), which is depicted in Fig. 2(a).
- **VSS-MLP**: a Vision Selective Scan (VSS) module combined with an MLP. The architecture is shown in Fig. 2(b).
- **VSS-L2**: a VSS module combined with two linear layers, which is depicted in Fig. 2(c).

The VSS module aims to bridge the gap between the 1D sequential processing capabilities inherent in the SSM and the 2D non-causal visual information. Specifically, the VSS module consists of a 2D vision scan mechanism and one mamba layer. In this work, we utilize two 2D scan mechanisms: Bidirectional-Scan Mechanism and Cross-Scan Mechanism, as follows:

- **Bidirectional-Scan Mechanism (BSM)** scans the image patch features in both forward and backward directions, which aims to capture a broader context without increasing computational complexity, as illustrated in the top of Fig. 3.
- **Cross-Scan Mechanism (CSM)** unfolds image patch features into sequences along rows and columns and scans them in four directions (diagonally across the image), as shown in the bottom of Fig. 3.

After the scan process, these sequences are passed through the mamba layer and reshaped back into the original image patch order, and all such features are merged to form a comprehensive representation.

As shown in Fig. 2(b), the input of the multimodal connector is the sequential image patch features V_{img} extracted from the input images via the transformer-based vision encoder. These feature vectors are then passed through a Vision Selective Scan (VSS) module to obtain the visual scanned feature V_{scan}. After the VSS module, the output vectors V_{scan} are combined with the original image patch features V_{img} through a skip connection. The combined vector is then passed into a norm layer and a two-layer Mult-Layer (MLP):

$$\begin{aligned} V_{scan} &= \mathbf{VSS}(V_{img}), \\ V_{out} &= \mathbf{MLP}(\mathbf{Norm}(V_{scan} + V_{img})). \end{aligned} \tag{5}$$

And for the variant MMC in Fig. 2(c), the feed-forward pass progress can be formulated as follows:

$$\begin{aligned} V_{img}^{'} &= \mathbf{Linear}(V_{img}), \\ V_{scan} &= \mathbf{VSS}(\mathbf{GELU}(V_{img}^{'})), \\ V_{out} &= \mathbf{Linear}(\mathbf{Norm}(V_{scan} + V_{img}^{'})). \end{aligned} \tag{6}$$

Mamba LLM We use the pre-trained Mamba Large Language Model (Mamba LLM) (33) f_L as our language model. Given a natural language query Q, we utilize the tokenizer and embedding module f_T to map the text input into the embedding space. Then the visual vector V_{out} and textual T are concatenated and put into the MambaLLM to obtain the response R.

$$R = f_L(V_{out}, f_T(Q)). \tag{7}$$

4 Experiment

4.1 Settings

Implementation details Following (16; 30), the training process contains two stages: vision-and-language alignment pre-training and multimodal instruction tuning. During the pretraining stage, we freeze the vision encoder and Mamba LLM and only keep the multimodal connector updated. The training data is the same as LLaVA. Then we finetune both the multimodal connector and the Mamba LLM in the instruction tuning stage. Our model is trained on 8 NVIDIA Tesla A800 GPUs.

Benchmarks We evaluate our model across a variety of benchmarks, including VQA-v2 (42), GQA (43), ScienceQA-IMG (44), TextVQA (45), POPE (46), MME (47), MMBench (48), MM-Vet (49), MMMU (50), and SEED (51).

4.2 Quantitative Evaluation

As is shown in Table 1, we compare our proposed model VL-Mamba with some SoTA multimodal large language models. Compared with the MobileVLM-3B (24) model with similar scale parameters and the same amount of multimodal training data, our model surpasses the performance on SQA$^{\mathrm{I}}$ (65.4 v.s. 61.2), VQA$^{\mathrm{T}}$ (48.9 v.s. 47.5), and MME (1369.6 v.s. 1288.9), though the Mamba LLM uses much less pretrained tokens (627B) than the backbone MobileLLaMA (1.3T) of MobileVLM. Compared with the LLaVA-Phi (56) model with a SoTA language model Phi-2-2.7B with 1.4T pretrained tokens, our performance shows superior on VQA-v2 (76.6 v.s. 71.4), MME (1369 v.s.

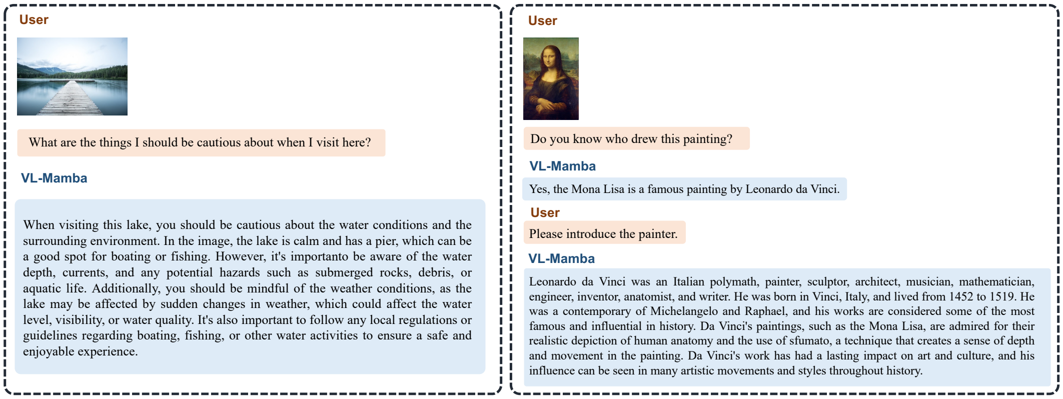
Table 1: **Comparison with SoTA methods.** Benchmark names are abbreviated due to space limits. PT and IT indicate the number of samples in the pretraining and instruction tuning stages, respectively.

Method	LLM	PT	IT	VQA^{v2}	GQA	SQA^{I}	VQA^{T}	POPE	MME	MMB	MM-Vet	MMMU	SEED
MiniGPT-4 (32)	Vicuna-7B	5M	5K	-	32.2	-	-	-	581.7	23.0	-	-	-
BLIP-2 (28)	Vicuna-13B	129M	-	41.0	41.0	61.0	42.5	85.3	1293.8	-	22.4	-	46.4
InstructBLIP (29)	Vicuna-13B	129M	1.2M	-	49.5	63.1	50.7	78.9	1212.8	-	25.6	-	-
Shikra (52)	Vicuna-13B	600K	5.5M	77.4	-	-	-	-	-	58.8	-	-	-
Otter (53)	LLaMA-7B	-	-	-	-	-	-	-	1292.3	48.3	24.6	29.1	-
mPLUG-Owl (23)	LLaMA-7B	2.1M	102K	-	-	-	-	-	967.3	49.4	-	-	-
IDEFICS-80B (54)	LLaMA-65B	353M	1M	60.0	45.2	-	30.9	-	-	54.5	-	-	-
Qwen-VL (55)	Qwen-7B	1.4B	50M	78.8	59.3	67.1	63.8	-	-	38.2	-	-	-
Qwen-VL-Chat (55)	Qwen-7B	1.4B	50M	78.2	57.5	68.2	61.5	-	1487.5	60.6	-	32.9	58.2
LLaVA-1.5 (30)	Vicuna-7B	558K	665K	78.5	62.0	66.8	58.2	85.9	1510.7	64.3	30.5	-	-
LLaVA-1.5 (30)	Vicuna-13B	558K	665K	80.0	63.3	71.6	61.3	85.9	1531.3	67.7	35.4	-	68.2
MobileVLM-3B (24)	MobileLLaMA-2.7B	558K	665K	-	**59.0**	61.2	47.5	84.9	1288.9	59.6	-	-	-
LLaVA-Phi (56)	Phi-2-2.7B	558K	665K	71.4	-	**68.4**	48.6	**85.0**	1335.1	**59.8**	28.9	-	-
TinyGPT-V (57)	Phi-2-2.7B	-	-	38.9	-	-	-	-	-	-	-	-	-
VL-Mamba (Ours)	Mamba LLM-2.8B	558K	665K	**76.6**	56.2	65.4	**48.9**	84.4	**1369.6**	57.0	**32.6**	**30.6**	**60.5**

1335.1), and MM-Vet (32.6 v.s. 28.9). It is worth noting that though our proposed model has fewer parameters and limited training data, it also achieves comparable performance compared with some models with a larger number of parameters. Its performance on the POPE benchmark is similar to LLaVA-1.5 (16), where the LLM parameters are 13B, which is approximately 4.6 times larger than the Mamba LLM. Although the results are not groundbreaking, the primary goal of this work is to explore the potential of SSMs as an alternative to Transformers in multimodal learning tasks. The comparable performance we observed indicates that SSMs can effectively handle the challenges posed by multimodal data, suggesting that further exploration into optimizing and adapting SSMs for such tasks could yield promising outcomes.

4.3 Qualitative Result

We present some examples to see the qualitative results of the VL-Mamba. As shown in Fig. 4, the VL-Mamba could well understand the user's question and respond accurately.



User

What are the things I should be cautious about when I visit here?

VL-Mamba

When visiting this lake, you should be cautious about the water conditions and the surrounding environment. In the image, the lake is calm and has a pier, which can be a good spot for boating or fishing. However, it's importanto be aware of the water depth, currents, and any potential hazards such as submerged rocks, debris, or aquatic life. Additionally, you should be mindful of the weather conditions, as the lake may be affected by sudden changes in weather, which could affect the water level, visibility, or water quality. It's also important to follow any local regulations or guidelines regarding boating, fishing, or other water activities to ensure a safe and enjoyable experience.

User

Do you know who drew this painting?

VL-Mamba

Yes, the Mona Lisa is a famous painting by Leonardo da Vinci.

User

Please introduce the painter.

VL-Mamba

Leonardo da Vinci was an Italian polymath, painter, sculptor, architect, musician, mathematician, engineer, inventor, anatomist, and writer. He was born in Vinci, Italy, and lived from 1452 to 1519. He was a contemporary of Michelangelo and Raphael, and his works are considered some of the most famous and influential in history. Da Vinci's paintings, such as the Mona Lisa, are admired for their realistic depiction of human anatomy and the use of sfumato, a technique that creates a sense of depth and movement in the painting. Da Vinci's work has had a lasting impact on art and culture, and his influence can be seen in many artistic movements and styles throughout history.

Figure 4: Examples of response generated by VL-Mamba.

4.4 Ablation Study and Discussions

The Effect of Variants of Language Model Table 2 shows the ablation experiment of evaluating the effectiveness of different variants of the language model. We conduct experiments on three different variants, Mamba-1.4B which has 1.4B parameters and is trained on Pile (58) with 300B tokens, Mamba-2.8B-Pile with 2.8B parameters and trained on Pile 300B tokens and Mamba-2.8B-Slimpj trained on SlimPajama with 627B tokens. Specifically, we construct the baseline models by using the same variant CLIP-ViT as the vision encoder, Mamba language models as backbone large language models, and vanilla MLP MultiModal Connectors without 2D vision selective scan modules. We can see with the increase of model scale and training tokens, Mamba-2.8B-Slimpj outperforms the other two variants on all benchmarks. Thus, we choose Mamba-2.8B-Slimpj for other experiments.

The Effect of Different Vision Encoders To evaluate the effectiveness of different vision encoders, we conduct an ablation study which is shown in Table 3. We study two different vision encoders, CLIP-ViT-L (31) and SigLIP-SO (59). The baseline models utilize Mamba-2.8B-Slimpj as LLM

Table 2: Ablation study of the variants of the language model.

Method	VQA^{v2}	GQA	SQA^{I}	VQA^{T}	POPE	MME	MMB	MM-Vet
Mamba-1.4B	71.7	49.9	56.1	42.6	84.5	1277.7	46.9	24.0
Mamba-2.8B-Pile	73.6	53.0	60.8	42.7	84.7	1321.3	52.1	28.5
Mamba-2.8B-Slimpj	**74.5**	**54.4**	**63.4**	**44.6**	**84.9**	**1381.8**	**55.8**	**30.6**

and vanilla MLP multimodal connectors. We can see that the CLIP-based model falls behind the SigLIP-based model in most benchmarks except the MME benchmark, where the CLIP-based model surpasses the SigLIP-based model by a large margin. Considering the comprehensive performance, we choose SigLIP-SO as the vision encoder to build the final VL-Mamba.

Table 3: Ablation study of the vision encoder.

Method	VQA^{v2}	GQA	SQA^{I}	VQA^{T}	POPE	MME	MMB	MM-Vet
CLIP-ViT-L (31)	74.5	54.4	63.4	44.6	84.9	**1381.8**	55.8	30.6
SigLIP-SO (59)	**76.7**	**55.4**	**66.3**	**47.5**	**85.2**	1349.4	**56.4**	**30.9**

Ablation on Different MMC Architectures We also explore the impact of different architectures of Multimodal Connector (MMC). We evaluate three different MMC variants: MLP, VSS-MLP, and VSS-L2. As shown in Table 4, by comparing the three architectures, we observe that VSS-L2 shows relatively better performance on most benchmarks, especially on the VQA^{T}, MME, MMB, and MM-Vet. The scores are 48.9, 1369.6, and 32.6 respectively, which proves the effectiveness of the VSS module combined with linear layers. Note that these models utilize SigLIP-SO as the vision encoder, Mamba-2.8B-Slimpj as the language model and Bi-directional selective scan mechanism.

Table 4: Ablation study of the different architectures of MMC.

Method	VQA^{v2}	GQA	SQA^{I}	VQA^{T}	POPE	MME	MMB	MM-Vet
MLP	**76.7**	55.4	**66.3**	47.5	85.2	1349.4	56.4	30.9
VSS-MLP	**76.7**	54.9	65.4	45.6	**85.3**	1335.8	56.4	30.6
VSS-L2	76.6	**56.2**	65.4	**48.9**	84.4	**1369.6**	**57.0**	**32.6**

Ablation on Different Scan Mechanisms We compare two scan mechanisms Bidirectional-Scan Mechanism (BSM) and Cross-Scan Mechanism (CSM) in the MMC module. As shown in Table 5, although BSM and CSM perform similarly in some benchmarks, such as they all score 76.6 in the VQA^{v2}, BSM exhibits superior performance in most benchmarks. Especially on the MMB benchmark, BSM scored 1369.6, 5.6 points higher than CSM, highlighting its strength in processing 2D vision information for multimodal learning tasks. Although CSM has two more scan directions than BSM, the results are similar, and the performance of BSM outperforms CSM on some benchmarks (POPE 5.6↑, MM-Vet 1.5↑). This may be because bidirectional scan methods have already acquired sufficient visual information. Blindly increasing the scan direction will not significantly improve performance, but will cause the potential impact of information redundancy.

Table 5: Ablation study of the scan mechanisms.

Method	VQA^{v2}	GQA	SQA^{I}	VQA^{T}	POPE	MME	MMB	MM-Vet
Bidirectional-Scan Mechanism (BSM)	**76.6**	**56.2**	**65.4**	48.9	84.4	**1369.6**	**57.0**	**32.6**
Cross-Scan Mechanism (CSM)	**76.6**	55.8	64.2	48.8	**85.0**	1364.0	56.3	31.1

5 Limitation

One limitation of our approach is that, compared to transformer-based multimodal models, our VL-Mamba requires more computational resources and has longer training times. In future work, we plan to explore a hybrid Transformer-Mamba architecture, which we believe could leverage the strengths of both models to improve efficiency and maintain competitive performance.

6 Conclusion

In this paper, we present VL-Mamba, the first study to investigate the use of the Mamba state space model for addressing multimodal learning tasks. The VL-Mamba consists of a language model, a vision encoder, and a multimodal connector. To be specific, we utilize the pre-trained Mamba Large Language Model (Mamba LLM) as the language model. Then, we study three architectures of MultiModal Connector (MMC) and introduce a Vision Selective Scan (VSS) module in MMC to bridge the gap between 2D non-causal image information and the inherent causal modeling capabilities of state space models (SSMs). In the VSS module, we propose two 2D scan mechanisms: the Bidirectional Scanning Mechanism (BSM) and Cross Scanning Mechanism (CSM). We conduct extensive experiments on eight multimodal benchmarks and achieve comparable performance with some SoTA MLLMs, and we also conduct ablation studies to evaluate the effectiveness of language variants, different vision encoders, different MMC architectures, and different scan mechanisms. The results demonstrate the effectiveness of our proposed model and prove the potential of the SSMs applied to multimodal learning.

Acknowledgments and Disclosure of Funding

This work was supported by the Centre for Augmented Reasoning, an initiative by the Department of Education, Australian Government.

References

[1] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *CVPR*, pp. 3128–3137, 2014.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *CVPR*, 2015, pp. 3156–3164.

[3] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, “Mattnet: Modular attention network for referring expression comprehension,” in *CVPR*, 2018, pp. 1307–1315.

[4] Y. Qiao, C. Deng, and Q. Wu, “Referring expression comprehension: A survey of methods and datasets,” *IEEE TMM*, vol. 23, pp. 4426–4440, 2020.

[5] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, “Vqa: Visual question answering,” *IJCV*, vol. 123, pp. 4 – 31, 2015.

[6] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, “A-okvqa: A benchmark for visual question answering using world knowledge,” in *ECCV*, 2022.

[7] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.

[8] Y. Yang, Z.-Y. Xing, and L. Zhu, “Vivim: a video vision mamba for medical video object segmentation,” *ArXiv*, vol. abs/2401.14168, 2024.

[9] Z.-Y. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, “Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation,” *ArXiv*, vol. abs/2401.13560, 2024.

[10] J. Ruan and S. Xiang, “Vm-unet: Vision mamba unet for medical image segmentation,” *arXiv preprint arXiv:2402.02491*, 2024.

[11] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, “Vision mamba: Efficient visual representation learning with bidirectional state space model,” *ArXiv*, vol. abs/2401.09417, 2024.

[12] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, “Vmamba: Visual state space model,” *ArXiv*, vol. abs/2401.10166, 2024.

[13] J. Ma, F. Li, and B. Wang, “U-mamba: Enhancing long-range dependency for biomedical image segmentation,” *ArXiv*, vol. abs/2401.04722, 2024.

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” *ArXiv*, vol. abs/2302.13971, 2023.

[15] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,” March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[16] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *ArXiv*, vol. abs/2310.03744, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263672058

[17] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "Opt: Open pre-trained transformer language models," *ArXiv*, vol. abs/2205.01068, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248496292

[18] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. M. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. C. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. García, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Díaz, O. Firat, M. Catasta, J. Wei, K. S. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, pp. 240:1–240:113, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247951931

[19] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[20] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. H. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. R. Florence, "Palm-e: An embodied multimodal language model," in *ICML*, 2023.

[21] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[22] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[23] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi *et al.*, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[24] X. Chu, L. Qiao, X. Lin, S. Xu, Y. Yang, Y. Hu, F. Wei, X. Zhang, B. Zhang, X. Wei, and C. Shen, "Mobilevlm : A fast, strong and open vision language assistant for mobile devices," *ArXiv*, vol. abs/2312.16886, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266573855

[25] M. He, Y. Liu, B. Wu, J. Yuan, Y. Wang, T. Huang, and B. Zhao, "Efficient multimodal learning from data-centric perspective," *ArXiv*, vol. abs/2402.11530, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267751050

[26] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *NeurIPS*, vol. 35, pp. 23 716–23 736, 2022.

[27] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[28] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.

[29] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *NeurIPS*, 2023.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[32] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[33] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[34] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[35] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "Hippo: Recurrent memory with optimal polynomial projections," *Advances in neural information processing systems*, vol. 33, pp. 1474–1487, 2020.

[36] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," in *NeurIPS*, 2021, pp. 572–585.

[37] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *ICLR*. OpenReview.net, 2022.

[38] J. T. H. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," in *ICLR*, 2023.

[39] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," *NeurIPS*, vol. 35, pp. 22 982–22 994, 2022.

[40] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," in *NeurIPS*, 2022.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*. OpenReview.net, 2021.

[42] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *CVPR*, 2017.

[43] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," *CVPR*, pp. 6693–6702, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:152282269

[44] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *ArXiv*, vol. abs/2209.09513, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252383606

[45] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," *CVPR*, pp. 8309–8318, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:85553602

[46] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J. rong Wen, "Evaluating object hallucination in large vision-language models," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258740697

[47] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, and R. Ji, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *ArXiv*, vol. abs/2306.13394, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259243928

[48] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, "Mmbench: Is your multi-modal model an all-around player?" *ArXiv*, vol. abs/2307.06281, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259837088

[49] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, "Mm-vet: Evaluating large multimodal models for integrated capabilities," *ArXiv*, vol. abs/2308.02490, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260611572

[50] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," *ArXiv*, vol. abs/2311.16502, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265466525

[51] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *ArXiv*, vol. abs/2307.16125, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260334888

[52] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.

[53] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, “Otter: A multi-modal model with in-context instruction tuning,” *ArXiv*, vol. abs/2305.03726, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258547300

[54] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. Rush, D. Kiela *et al.*, “Obelics: An open web-scale filtered dataset of interleaved image-text documents,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[55] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” 2023.

[56] Y. Zhu, M. Zhu, N. Liu, Z. Ou, X. Mou, and J. Tang, “Llava-phi: Efficient multi-modal assistant with small language model,” *arXiv preprint arXiv:2401.02330*, 2024.

[57] Z. Yuan, Z. Li, and L. Sun, “Tinygpt-v: Efficient multimodal large language model via small backbones,” *ArXiv*, vol. abs/2312.16862, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266572996

[58] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima *et al.*, “The pile: An 800gb dataset of diverse text for language modeling,” *arXiv preprint arXiv:2101.00027*, 2020.

[59] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” *ICCV*, pp. 11 941–11 952, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257767223