VisualRWKV: Exploring Recurrent Neural Networks for Visual Language Models

Haowen Hou[*] and **Peigen Zeng**[+] and **Fei Ma**[*] and **Fei Richard Yu**[+†]
[*]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[+]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[†]Shool of Information Technology, Carleton University, Canada
{houhaowen, mafei, yufei}@gml.ac.cn [*]

Abstract

Visual Language Models (VLMs) have rapidly progressed with the recent success of large language models. However, there have been few attempts to incorporate efficient linear Recurrent Neural Networks (RNNs) architectures into VLMs. In this study, we introduce VisualRWKV, the first application of a linear RNN model to multimodal learning tasks, leveraging the pre-trained RWKV language model. We propose a data-dependent recurrence and sandwich prompts to enhance our modeling capabilities, along with a 2D image scanning mechanism to enrich the processing of visual sequences. Extensive experiments demonstrate that VisualRWKV achieves competitive performance compared to Transformer-based models like LLaVA-1.5 on various benchmarks. Compared to LLaVA-1.5, VisualRWKV has a speed advantage of 3.98 times and can save 54% of GPU memory when reaching an inference length of 24K tokens. To facilitate further research and analysis, we have made the checkpoints and the associated code publicly accessible at the following GitHub repository: https://github.com/howard-hou/VisualRWKV.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional performance in natural language processing tasks (Touvron et al., 2023b; Brown et al., 2020). Extending LLMs to support visual inputs has garnered significant attention in the research community (OpenAI, 2023). Visual Language Models (VLMs) inherit powerful capabilities from LLMs, such as strong instruction following, zero-shot generalization, and in-context learning (Liu et al., 2023b; Zhu et al., 2024a). By integrating visual and textual information, VLMs not only enhance the understanding of visual content but also provide richer context for language understanding and generation. VLMs hold tremendous potential for solving visual problems and advancing various vision-language tasks.

However, despite the excellent performance of existing LLMs and VLMs, their inherent computational and memory complexity due to the self-attention mechanism in the Transformer architecture results in quadratic growth in computation and memory requirements with the increase in sequence length (Katharopoulos et al., 2020). This leads to high inference costs and limits the deployment and application of Transformer-based VLMs on edge devices.

The Receptance Weighted Key Value (RWKV) model, a novel Recurrent Neural Network (RNN) architecture, presents a promising solution to the bottleneck of long-sequence modeling (Peng et al., 2023a). It surpasses Transformers in large-scale data performance and exhibits linear scalability with sequence length, positioning itself as a promising successor to Transformers in language modeling (Peng et al., 2023b).

Currently, there is a notable gap in research exploring how this efficient architecture can be leveraged for multimodal tasks. In this study, we introduce the VisualRWKV model, marking the first application of a linear RNN model to multimodal learning tasks. Specifically, we utilize the pre-trained RWKV language model as the foundational language model and explore several novel mechanisms applied to VisualRWKV.

VisualRWKV introduces: (1) an innovative data-dependent recurrence to enhance the capabilities and capacity of the RWKV model. (2) a novel sandwich prompt designed to provide richer conditions when processing visual sequences. (3) a new 2D image scanning mechanism to enhance the 2D modeling capabilities of visual sequences.

[*]This work is supported in part by Shenzhen Science and Technology Program under Grant ZDSYS20220527171400002, the National Natural Science Foundation of China (NSFC) under Grants 62406197, 62271324, 62231020 and 62371309. Corresponding author: F. Richard Yu.

Proceedings of the 31st International Conference on Computational Linguistics, pages 10423–10434
January 19–24, 2025.

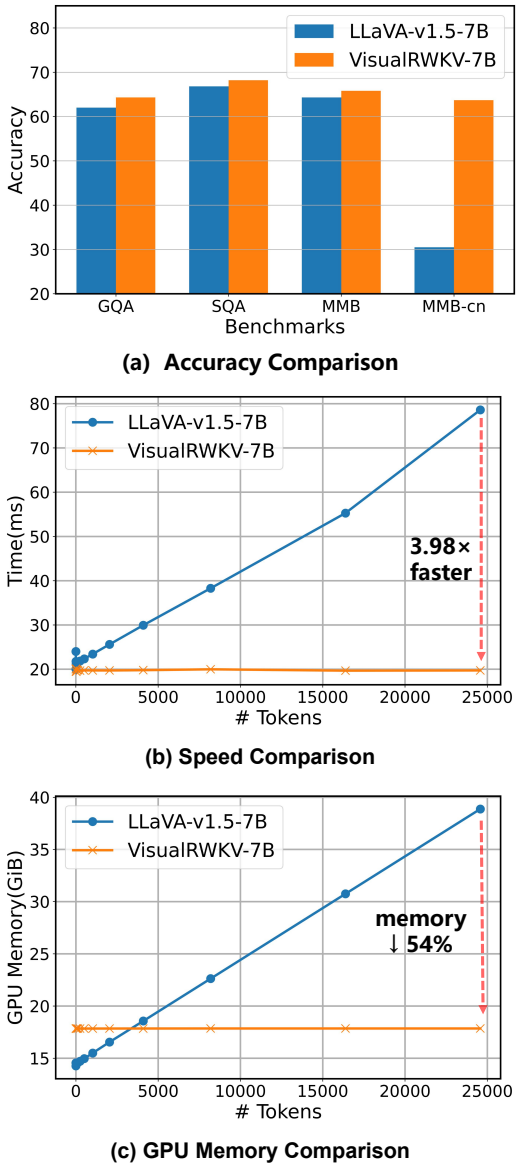

Figure 1: **VisualRWKV** outperforms the SoTA LLaVA-1.5 (Liu et al., 2023a) on 4 tasks (a), with high computational efficiency (b) and low, stable memory usage (c).

Extensive experiments on various multimodal learning benchmarks validate the effectiveness of VisualRWKV, as shown in Figure 1. Compared to other Transformer-based models of similar size, such as LLaVA-1.5 (Liu et al., 2023a), VisualRWKV demonstrates competitive performance, achieving outstanding results on multiple popular benchmarks.

In summary, this study presents the VisualRWKV model, explores the impact of various novel designs on VisualRWKV, introduces the innovative sandwich prompt to enhance representation capabilities, and conducts extensive experiments across diverse multimodal learning benchmarks.

2 Related Works

2.1 Visual Language Models

Following the success of LLMs, recent research has pivoted towards VLMs (Achiam et al., 2023; Team et al., 2023) for enhancing visual understanding and reasoning capabilities. Expanding on various pre-trained LLM architectures, researchers have proposed diverse methodologies for incorporating visual information. Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023c) introduce distinct techniques for modality fusion, integrating visual tokens with frozen large language models through gated attention or query transformers. Building on the effectiveness of instruction tuning, LLaVA (Liu et al., 2023b,a) and MiniGPT-4 (Zhu et al., 2024a; Chen et al., 2023a) utilize visual instruction tuning to align visual input with LLMs, showcasing notable achievements. Recent advancements, such as Kosmos-2 (Peng et al., 2023c) and Shikra (Chen et al., 2023b), further enhance VLMs with grounded visual understanding capabilities. Despite their promising potential for general-purpose visual reasoning and planning tasks, these models are generally expensive and challenging to train and deploy.

2.2 Linear RNN Large Language Model

Recent advancements in LLMs, such as GPT (Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023), LLaMA (Touvron et al., 2023a,b), and PaLM (Anil et al., 2023; Chowdhery et al., 2023), have showcased remarkable prowess across various natural language processing tasks. However, traditional Transformer-based LLMs suffer from quadratic complexity $O(L^2)$ issues in both computation and memory, prompting the emergence of linear RNNs as potential successors.

RNNs model sequential data with temporal dependencies by generating a hidden state h_t at each time step, which is then utilized as input for the subsequent step. Classical RNN variants like LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho et al., 2014) excel in inexpensive inference, operating typically at $O(1)$ time complexity per step relative to sequence length. Nonetheless, their older designs often pose challenges in parallelization across time dimensions during training.

Linear RNNs present themselves as promising successors to the Transformer, offering a more efficient token mixing method. They enable a space complexity of $O(L)$ and an inference complexity

of $O(1)$. Leveraging Parallel Prefix Sum Scan (Harris et al., 2007) for acceleration can further enhance their efficiency. The RWKV (Peng et al., 2023b; Hou and Yu, 2024), a linear RNN-based LLM, has showcased competitive performance compared to GPT models of similar scale. RWKV introduces temporal decay to gradually reduce the influence of past information, implicitly incorporating positional information. Additionally, it integrates a token-shift mechanism facilitating linear interpolation between current and previous inputs. This allows the model to naturally aggregate and regulate information within input channels. Furthermore, RWKV boasts a time complexity of $O(L)$ and an inference complexity of $O(1)$, ensuring consistent inference time per token. As a result, the overall inference duration scales linearly with sequence length. The memory footprint of RWKV remains constant, regardless of sequence length, contributing to its efficiency and scalability.

3 Methods

In this section, we initially introduce the fundamental concepts of the RWKV language model. (Section 3.1). Following that, we elaborate on the transformation of the RWKV language model into our proposed VisualRWKV visual language model (Section 3.2), which mainly includes data-dependent recurrence, sandwich prompting, and image scanning.

3.1 Preliminaries

The RWKV(Peng et al., 2024) backbone is structured using stacked residual blocks, with each block containing a time-mixing and a channel-mixing sub-block. These components embody recurrent structures designed to leverage past information.

Data-independent Token Shift As shown in Figure 3, trainable variable μ_g, μ_r, μ_k, μ_v are used in a linear combination of x_t and x_{t-1}, to achieve a simple mixing, which interpolate between the inputs of the current and previous time-steps. The combination of shifted previous step and current step was linear projected through projection matrix within the block:

$$\alpha_t = (\mu_\alpha \odot x_t + (1 - \mu_\alpha) \odot x_{t-1})W_\alpha \quad (1)$$

where α serves as a notation for the variables r, g, k, and v, given that they are subject to an identical linear combination formula. Please note that the linear combination used here is data independent, meaning the value of μ_α is not dependent on x_t or x_{t-1}.

Data-independent Time Mixing In vanilla RWKV, the time mixing is articulated through the update of the WKV vectors and the WKV operator is input-data independent. The formula of single head WKV operator is given by:

$$wkv_t = \mathrm{diag}(u) \cdot k_t^{\mathrm{T}} \cdot v_t + \sum_{i=1}^{t-1} \mathrm{diag}(w)^{t-1-i} \cdot k_i^{\mathrm{T}} \cdot v_i \quad (2)$$

where w and u are two trainable parameters. The parameter u serves as a term weight for the current token when the model encounters it for the first time. It enables the model to efficiently process the token by focusing more on important tokens and quickly filtering out unimportant ones. Another important parameter is w, which is a channel-wise time decay vector per head. Furthermore, we transform parameter w by $w = \exp(-\exp(w))$. This transformation ensures that all values of w are within the range $(0, 1)$, ensuring that $\mathrm{diag}(w)$ represents a contraction matrix.

The output from the single-head WKV operator undergoes processing by the layer normalization and the SiLU activation. Then, all outputs are concatenated to form the output vector o_t:

$$o_t = concat(SiLU(g_t) \odot \mathrm{LayerNorm}(r_t \cdot wkv_t))W_o \quad (3)$$

where LayerNorm operates on each head separately. For further details and formulas of the models, one can refer to Peng et al. (2024) and Hou and Yu (2024).

3.2 VisualRWKV

Method	Size	VQA	SQA	TQA	GQA
VisualRWKV-Base	1.6B	51.08	41.94	35.19	48.09
+Data-dep Recurrence	1.6B	65.82	46.55	40.26	49.06
+Bidirection +Sandwich	1.6B	64.96	56.72	41.94	48.04
+Better Learning Rate	1.6B	69.42	59.05	43.57	55.23
+Scale up to 3B	3B	71.52	65.34	48.68	59.56
+Scale up to 7B	7B	75.82	68.22	51.01	64.27

Table 1: **Scaling results** on model. We choose to conduct experiments on VQA-v2(VQA), ScienceQA(SQA), TextVQA(TQA) and GQA to examine model's capabilities.

3.2.1 VisualRWKV Baseline

VisualRWKV is a follow-up work to RWKV. RWKV paper (Peng et al., 2024) proposed a simplified version of VisualRWKV that employed data-independent recurrence (Fig. 3), unidirection image scanning (Fig. 4), and image first prompting

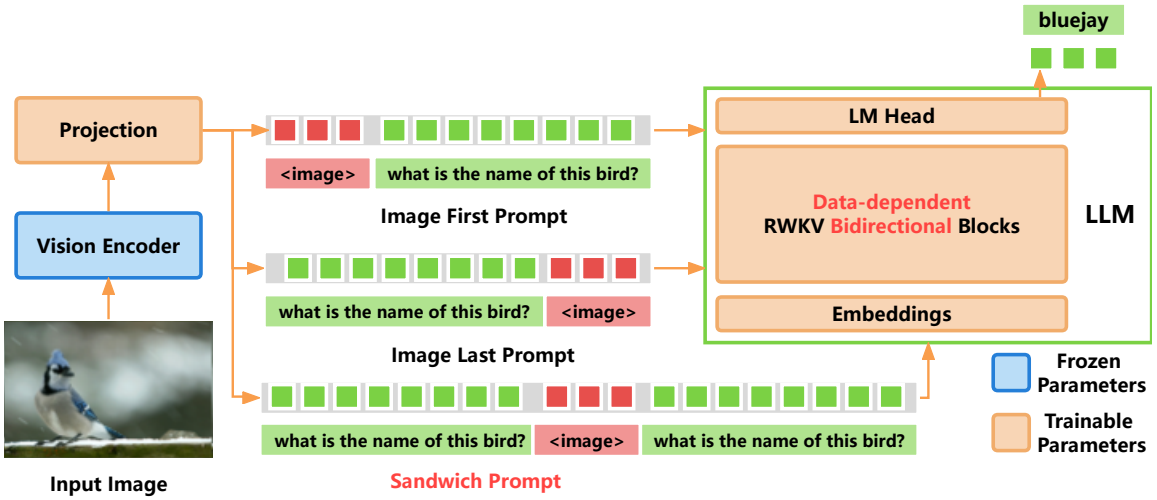

Figure 2: VisualRWKV architecture overview and three prompting method. **Image First Prompt:** place image tokens before instruction tokens; **Image Last Prompt:** place image tokens after instruction tokens; **Sandwich Prompt:** place image tokens in the middle of instruction tokens. Red words indicate the key contributions.

(Fig. 2). We used that version of VisualRWKV as the baseline and starting point for our research, as shown in Table 1. We denote this initial model without any modifications as **VisualRWKV-Base**.

3.2.2 Data-dependent Recurrence

The Data-dependent Recurrence mechanism introduces two key enhancements: the Data-dependent Token Shift and the Data-dependent Time Mixing.

Data-dependent Token Shift First, we define **lo**w-**r**ank **a**daptation (lora) and **d**ata-**d**ependent **l**inear int**erp**olation (ddlerp) as follow:

$$\text{lora}_\alpha(x) = \lambda_\alpha + \tanh(xA_\alpha)B_\alpha \tag{4}$$

$$\text{ddlerp}_\alpha(a,b) = a + (b-a)\odot \text{lora}_\alpha(a + (b-a)\odot \mu_x) \tag{5}$$

Then, the Data-dependent Token Shift is defined as:

$$\alpha_t = \text{ddlerp}_\alpha(x_t, x_{t-1})W_\alpha \tag{6}$$

where α serves as a notation for the variables r, g, k, and v. A_α, B_α, λ_α and W_α are trainable parameters. The data-dependent token shift seeks to broaden the model's capacity. It dynamically allocates the ratio of new to existing data per channel, depends on the input at both current and previous time steps.

Data-dependent Time Mixing The key improvement over data-independent time mixing (Eq. 2) lies in the evolution of the time decay vector from a fixed parameter w to a dynamic one w_t that reacts to the input data x_t at time step t. The dynamic nature of w_t allows the model to adjust more nimbly to diverse input data, unbound by rigid, predefined structures. Equations are as follow:

$$d_t = \text{lora}_d(\text{ddlerp}_d(x_t, x_{t-1})) \tag{7}$$

$$w_t = \exp(-\exp(d_t)) \tag{8}$$

$$wkv_t = \text{diag}(u)\cdot k_t^{\mathrm{T}}\cdot v_t + \sum_{i=1}^{t-1}\text{diag}\left(\bigodot_{j=1}^{i-1} w_j\right)\cdot k_i^{\mathrm{T}}\cdot v_i \tag{9}$$

The LoRA mechanism utilizes vectors learned from data-independent time mixing and enhances them at a low cost with additional offsets modulated by the incoming input. It should be noted that the computation of the new time-varying decay w_t employs a token-shifted value $\text{ddlerp}_d(x_t, x_{t-1})$ as its input, not just the current token x_t. As shown in Table 1, the VisualRWKV equipped with data-dependent recurrence exhibits a significant improvement in performance.

3.2.3 Sandwich Prompt

The motivation for designing the sandwich prompt is as follows: Unlike the attention mechanism in Transformers, RNN models such as RWKV, due to their sequential nature, cannot revisit historical information repeatedly. Instead, they must decide immediately whether to store a token or image token in memory upon encountering it. Therefore, carefully designing tailored prompts is essential for enhancing VisualRWKV's ability to effectively acquire and utilize information. For this purpose, we have specifically designed three types of prompting methods, as shown in Figure 2:

- Image First Prompt: Place image tokens prior to the instruction tokens.

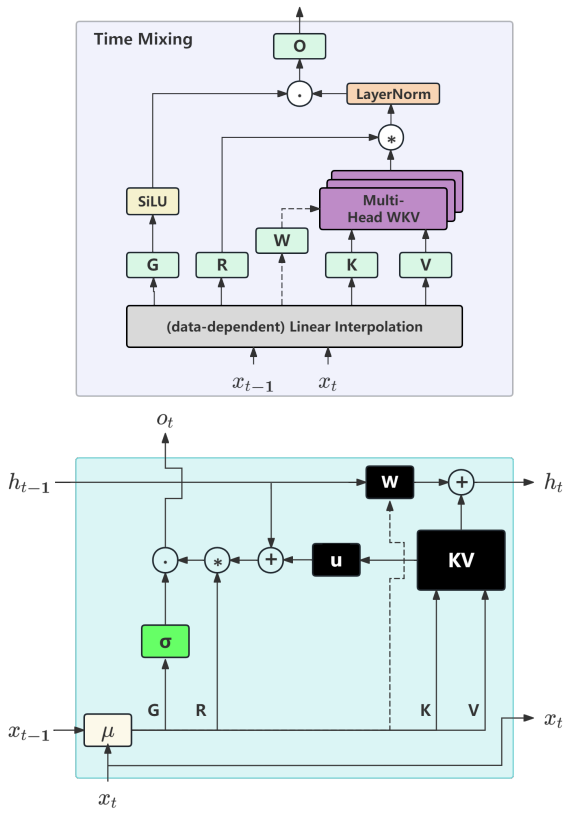

Figure 3: Data-dependent recurrence. **Top:** Semantic diagram of the time-mixing block; **Bottom:** Time-mixing block as an RNN cell. Dashed arrows represent connections in data-dependent recurrence, not present in data-independent recurrence.

- Image Last Prompt: Place image tokens following the instruction tokens.
- Sandwich Prompt: Insert image tokens between the instruction tokens.

The sandwich prompt is designed to provide optimal conditions that assist the model in making these decisions more effectively. Specifically, the first prompt helps the model efficiently extract relevant information from the image, while the second prompt focuses on improving the model's ability to answer questions.

For instance, the Image Last Prompt can cause the model to occasionally forget the question embedded in the prompt, while the Image First Prompt may result in the model processing the image without considering the question, hindering its ability to analyze the image contextually. In contrast, the sandwich prompt resolves these issues and achieves a synergistic effect, enabling the model to perform better than the sum of the individual prompts. The experimental results show that the Sandwich Prompt achieves the best performance, as presented in Table 3.

3.2.4 Image Scanning

The motivation for designing the image scanning techniques is as follows: Language is inherently unidirectional, while images are multidirectional by nature. As a result, unidirectional language models face inherent limitations when processing visual information. By implementing bidirectional or multidirectional image scanning strategies, these challenges can be effectively mitigated.

Vanilla RWKV is designed for 1D sequential data with causal relationships, such as language sequences. However, the visual sequences generated by vision encoders are non-causal. To bridge this gap, we propose a 2D scanning mechanism to improve VisualRWKV's performance on visual tasks. This work integrates the 2D scanning mechanism into RWKV blocks, exploring three variants of multimodal RWKV blocks, which are illustrated in Figure 4:

- Unidirectional Blocks: Only containing the Forward Scanning Block, which is the basic scanning pattern of RWKV and other linear RNN models. This serves as the Base.
- Bidirectional Blocks: Comprising both Forward Scanning and Backward Scanning Blocks, arranged in an alternating fashion.
- Multidirectional Blocks: Including blocks for Forward Scanning, Backward Scanning, Upward Scanning, and Downward Scanning, with the sequence of Forward, Backward, Upward, and Downward arranged in an alternating order.

Our design alternates different scanning directions within layers, which does not introduce additional computational overhead and preserves the efficiency of the architecture. The experimental results (Table 4) have also verified the effectiveness and necessity of such scanning techniques in enhancing the model's ability to handle and understand visual sequences, thereby improving the overall performance of VisualRWKV in visual language processing tasks.

4 Experiments

The following section is dedicated to showcasing the key experiments and outcomes related to Visu-

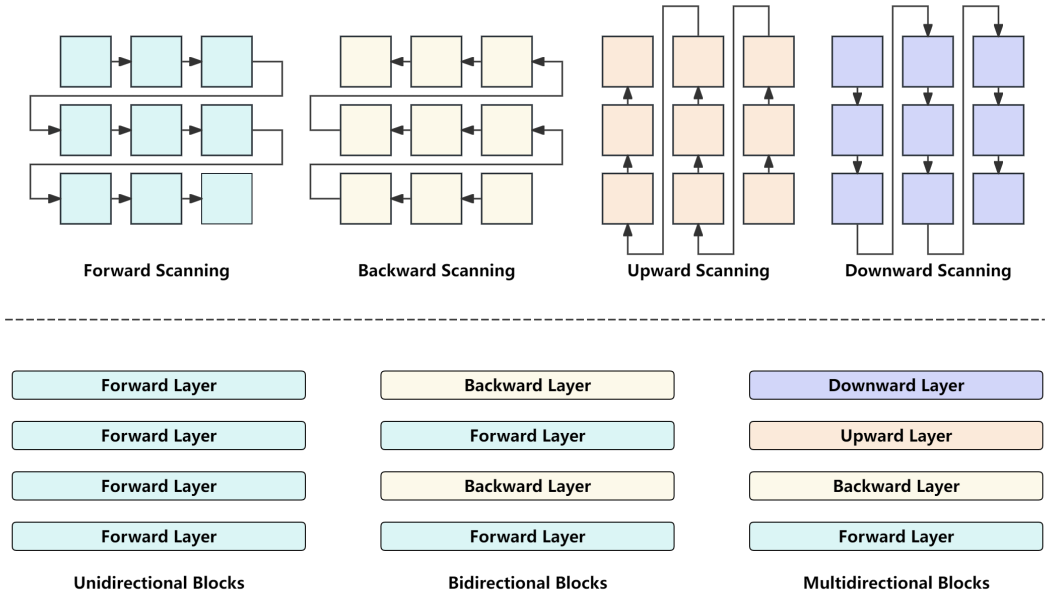

Figure 4: Illustration of 3 different multimodal RWKV Blocks: Unidirectional Blocks (left), Bidirectional Blocks (middle), and Multidirectional Blocks (right). The four scanning modes are also depicted at the top.

alRWKV. All results presented in this section are derived from a single run.

4.1 Experiment Setup

Following Liu et al. (2023a,b), the training process of VisualRWKV consists of two stages: vision-and-language alignment pretraining and visual instruction tuning. In the pretraining stage, the vision encoder and RWKV LLM are frozen, with only the projector being updated. During the visual instruction tuning stage, we finetune both the projector and the RWKV LLM, as shown in Figure 2. Details of training data and hyper-parameters can be found in Appendix A.

4.2 Benchmarks

We evaluated VisualRWKV across 8 benchmark tests tailored to assess the model's performance in academic tasks.

For assessing visual perception capabilities, VQA-v2 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019) presented open-ended short questions. Following the methodology outlined in LLaVA (Li et al., 2023b), we utilized the image subset of ScienceQA (Lu et al., 2022) to gauge the model's zero-shot generalization in answering scientific questions via multiple-choice questions. TextVQA (Singh et al., 2019) focused on visual question answering with rich text content.

Regarding benchmarks tailored for VLMs, various assessments evaluated the model's performance across diverse domains and applications, encompassing different response formats. MME-Perception (Fu et al., 2023) scrutinized the model's visual perception abilities through true/false questions. MMBench (Liu et al., 2023c) assessed the robustness of the model's answers by rigorously shuffling multiple-choice options. MMBench-CN, the Chinese counterpart of MMBench, was employed to evaluate the model's multilingual capabilities. POPE (Li et al., 2023d) assesses the model's hallucination degree on three sampled subsets of COCO (Lin et al., 2014): random, common, and adversarial, reporting the average F1 score across all three splits.

4.3 Quantitative Evaluation

4.3.1 Main Results

Table 2 presents a comparison of our proposed VisualRWKV model with some state-of-the-art (SOTA) multimodal large language models. VisualRWKV achieved the best performance in 3 out of 8 benchmarks and came in second place in SQA benchmark. Compared with LLaVA-1.5 7B, which has similar scale parameters and the same amount of multimodal training data, Our model(VisualRWKV-7B) outperformed it in 4 benchmarks: SQA (68.2 vs. 66.8), GQA (64.3 vs. 62.0), MMB (65.8 vs. 64.3), and MMB-cn (63.7 vs. 30.5). It is noteworthy that VisualRWKV and

LLaVA-1.5 used completely identical training data. Yet, on the MMB-cn Chinese test set, VisualRWKV showed a substantial lead. This may indicate that the RWKV language model has stronger multilingual capabilities. These promising results not only confirm the effectiveness of the VisualRWKV model, but also highlight the significant potential of the Linear RNN model in multimodal learning tasks.

4.3.2 Gain Analysis on Different Benchmarks

VisualRWKV excels in academic benchmarks like VQA, GQA, and SQA, where both the questions and answers are short texts. The model faces no fundamental obstacles in handling such tasks, leading to significant performance improvements. As a result, VisualRWKV achieves results that are comparable to, and even surpass, the Transformer-based LLaVA-1.5 on these benchmarks.

Although VisualRWKV shows notable improvements on the TextVQA (TQA) benchmark, it still lags behind LLaVA-1.5 in this task (51.0 vs. 58.2). TextVQA requires recalling information from images, which is similar to the Multi-Query Associative Recall (MQAR) task (Arora et al., 2023), which is often a limitation for RNN-like architectures. However, our latest work, VisualRWKV-HD/UHD (Li and Hou, 2024), has shown that higher resolution and better-quality image features can significantly alleviate these limitations.

4.4 Ablation Study

4.4.1 Ablation on Data-dependent Recurrence

To verify the effectiveness of data-dependent recurrence described in Section 3.2.2, we conducted a rigorous ablation study, ensuring that the model size, training data, environment, and all hyperparameters were strictly consistent. As depicted in Table 1, the outcomes demonstrate significant enhancements in the data-dependent VisualRWKV across the four monitored benchmarks, affirming that data-dependence is essential for the success of linear RNN-type models in the VLM domain.

4.4.2 Ablation on Prompting Method

As shown in Table 3, among the three prompting approaches, the sandwich prompt outperforms the others, followed by the image-first prompt, with the image-last prompt being the least effective. The effectiveness of the sandwich prompt is attributed to its ability to allow the model to review the instructions before engaging with the image, enabling a more targeted extraction of information and enhancing the conditional aspects of image information retrieval.

However, simply placing the instructions before the image is insufficient. The image-last prompt performs poorly because linear RNN models tend to forget the instructions after processing the image, making it necessary to repeat the instructions for better results. Additionally, our research shows that the sandwich prompt can effectively mitigate information loss even with a reduced number of image tokens, maintaining robust performance. Further experimental results and analyses can be found in Appendix E.

4.4.3 Ablation on Scanning Method

We compared three image scanning mechanisms: Uni-directional scanning (UniDir), Bi-directional scanning (BiDir), and Multi-directional scanning (MultiDir). As shown in Table 4, UniDir performs the worst because it is inherently unsuitable for 2D visual information. BiDir and MultiDir show comparable outcomes across various benchmark assessments, but BiDir outperforms in the majority, highlighting its strength in handling 2D visual information for multimodal learning tasks.

The image scanning techniques are applied during both training and inference, and it is essential to maintain train-test consistency. We have made simple attempts to rearrange the order of layers with different directions, but the performance was not robust. Specific layers have already been specialized to process image information from particular directions.

4.4.4 Ablation on Learning Rate

As shown in Table 1, correct learning rate is crucial for the performance of benchmarks. Table 10 shows a comparison of our model with different learning rate. From the Table, it can be observed that a higher initial learning rate has a significant impact on the model's performance. Our hypothesis is that the substantial divergence in tasks from the textual to the visual domain necessitates a higher learning rate to facilitate the model's adaptation.

It has been observed that there is a substantial discrepancy between the optimal learning rates of VisualRWKV and LLaVA(Liu et al., 2023a), with the optimal initial learning rate for LLaVA-1.5-7B being $2e^{-5}$ and for VisualRWKV-7B being $4e^{-5}$. This caused considerable difficulties in our work

Method	LLM	Res.	PT/IT	VQA	GQA	SQA	TQA	POPE	MME	MMB	MMB-cn
BLIP-2 (Li et al., 2023c)	Vicuna-13B	224	129M/ -	41.0	41.0	61.0	42.5	85.3	1293.8	–	22.4
MiniGPT-4 (Zhu et al., 2024a)	Vicuna-7B	224	5M/5K	-	32.2	-	-	-	581.7	23.0	-
InstructBLIP (Dai et al., 2023)	Vicuna-7B	224	129M/1.2M	–	49.2	60.5	50.1	–	–	36	26.2
InstructBLIP (Dai et al., 2023)	Vicuna-13B	224	129M/1.2M	–	49.5	63.1	50.7	78.9	1212.8	–	25.6
Shikra (Chen et al., 2023b)	Vicuna-13B	224	600K/5.5M	77.4	–	–	–	–	–	58.8	–
Otter (Li et al., 2023a)	LLaMA-7B	224	-	-	-	-	-	-	1292.3	48.3	24.6
mPLUG-Owl (Ye et al., 2023)	LLaMA-7B	224	2.1M/102K	-	-	-	-	-	967.3	49.4	-
IDEFICS-9B (IDEFICS, 2023)	LLaMA-7B	224	353M/1M	50.9	38.4	–	25.9	–	–	48.2	–
IDEFICS-80B (IDEFICS, 2023)	LLaMA-65B	224	353M/1M	60.0	45.2	–	30.9	–	–	54.5	–
Qwen-VL (Bai et al., 2023)	Qwen-7B	448	1.4B/50M	**78.8**	59.3	67.1	**63.8**	–	–	38.2	–
Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B	448	1.4B/50M	78.2	57.5	68.2	61.5	–	1487.5	60.6	–
LLaVA-1.5 (Liu et al., 2023a)	Vicuna-7B	336	558K/665K	78.5	62.0	66.8	58.2	**85.9**	**1510.7**	64.3	30.5
LLaVA-Phi (Zhu et al., 2024b)	Phi2-2.7B	336	558K/665K	71.4	-	**68.4**	48.6	85.0	1335.1	59.8	28.9
MobileVLM-3B (Chu et al., 2023)	LLaMA-2.7B	336	558K/665K	-	59.0	61.2	47.5	84.9	1288.9	59.6	-
VL-Mamba (Qiao et al., 2024)	Mamba-2.8B	224	558K/665K	76.6	56.2	65.4	48.9	84.4	1369.6	57.0	32.6
VisualRWKV-Base (Peng et al., 2024)	RWKV5-1.6B	336	558K/665K	51.1	48.1	41.9	35.2	73.1	-	-	-
VisualRWKV	RWKV6-1.6B	336	558K/665K	69.4	55.2	59.1	43.6	83.2	1204.9	55.8	53.2
VisualRWKV	RWKV6-3B	336	558K/665K	71.5	59.6	65.3	48.7	83.1	1369.2	59.5	56.3
VisualRWKV	RWKV6-7B	336	558K/665K	75.8	**64.3**	68.2	51.0	84.7	1387.8	**65.8**	**63.7**

Table 2: **Comparison with SoTA methods on 8 benchmarks.** Due to space constraints, benchmark names are abbreviated. VQA (Goyal et al., 2017); GQA (Hudson and Manning, 2019); SQA: ScienceQA-IMG (Lu et al., 2022); TQA: TextVQA (Singh et al., 2019); POPE (Li et al., 2023d); MME (Fu et al., 2023); MMB: MMBench (Liu et al., 2023d); MMB-cn: MMBench-CN (Liu et al., 2023d). PT and IT denote the quantity of samples involved in the pre-training and instruction-tuning phases. "Res." stands for "Resolution.

Method	Size	Prompt	VQA	SQA	TQA	GQA
VisualRWKV-Base	7B	First	67.93	**65.59**	47.13	48.52
VisualRWKV-Base	7B	Last	63.07	57.66	48.52	44.19
VisualRWKV-Base	7B	Sandwich	**69.71**	65.20	**50.25**	**50.50**

Table 3: Results for three prompting method.

Method	Size	Scanning	VQA	SQA	TQA	GQA
VisualRWKV-Base	1.6B	UniDir	51.03	41.94	35.19	48.09
VisualRWKV-Base	1.6B	BiDir	65.62	**47.30**	**37.13**	48.60
VisualRWKV-Base	1.6B	MultiDir	**66.04**	44.03	35.84	**49.95**
VisualRWKV	1.6B	BiDir	**69.26**	**57.61**	**43.17**	**54.85**
VisualRWKV	1.6B	MultiDir	69.20	57.31	42.97	54.63

Table 4: Results for three scanning methods.

at the beginning and also confirmed the significant divergence between the RWKV architecture and the Transformer architecture.

4.5 Efficiency Analysis

As shown in Figure 1, we compared the inference speed and GPU memory consumption directly with LLaVA-1.5 of the same parameter size. VisualRWKV has a constant single token inference speed, while the inference speed of a single token in LLaVA-1.5 slows down as more tokens are generated. On the other hand, VisualRWKV has a constant GPU memory consumption, while the memory consumption of LLaVA-1.5 increases linearly. In practice, compared to LLaVA-1.5, VisualRWKV has a speed advantage of 3.98 times and can save 54% of the GPU memory when reaching an inference length of 24576 tokens. Since VisualRWKV retains a fixed state size throughout inference, GPU memory usage remains nearly constant, which is illustrated as a straight line in Figure 1(c).

4.6 Text-only Capability

According to Lin et al. (2024), LLMs face the issue of degraded text capabilities after visual instruction tuning. As shown in Table 5, no degradation of text abilities was observed in VisualRWKV. Conversely, enhancements in performance were noted across various text-only English datasets, which we credit to the integration of a large set of English samples in our fine-tuning dataset. Furthermore, it was observed that VisualRWKV did not face text ability degradation across multiple languages, as shown in Table 5. The capabilities were fundamentally aligned with those of the text-only RWKV. This may be due to the incorporation of the multilingual ShareGPT4. More details about text-only capability can be found in Appendix G.

Besides the results previously stated, we also compared the outcomes of single-stage and two-

Method	Size	LAMBADA ppl	English avg%	MultiLang avg%
RWKV	1.6B	4.63	59.82	59.97
VisualRWKV	1.6B	4.15	61.01	59.83

Table 5: Results for text-only capability: The English score is the average of 10 English benchmarks, while the Multilingual score is the average of 4 Multilingual benchmarks.

stage training approaches; conducted ablation studies on the method of cross-entropy loss reduction; assessed the influence of Weight Decay on the model; and explored a basic hybrid model known as VisualRWKV Hybrid. Due to space limitations, we have included these contents in the Appendix.

5 Conclusions

In this paper, we introduce for the first time VisualRWKV, which explores the construction of a visual language model using the linear RNN model RWKV. VisualRWKV incorporates three innovative designs: data-dependent recurrence to enhance the model's information extraction capabilities, sandwich prompt for better conditioning, and bidirectional scanning for more effective extraction of 2D visual information. We conducted extensive experiments on eight multimodal benchmarks and achieved comparable performance with some of the most advanced VLMs; we also carried out ablation studies to evaluate the effectiveness of data-dependent recurrence, prompting methods, and various scanning mechanisms. The results validate the effectiveness of our proposed model and demonstrate the potential of applying RNNs to VLMs.

Limitations

Despite the encouraging results achieved by VisualRWKV, several limitations must be acknowledged. Firstly, due to the lack of data following such instructions and the limited context length, VisualRWKV is currently unable to process multiple images. Secondly, although VisualRWKV shows good performance on academic datasets, its ability to handle certain tasks, such as TextVQA, may be constrained by the limitations in the recall ability of efficient language models (Arora et al., 2023). These constraints can potentially be mitigated by further architectural improvements. Lastly, to maintain consistency with LLaVA-1.5, this study did not investigate the effects of the choice of vision encoder or the quality of training data on VisualRWKV. In the future, we aim to explore more advanced visual encoders and utilize higher-quality training data to further enhance its performance.

Risks Although VisualRWKV significantly reduces the occurrence of hallucinations, it can still generate hallucinations and occasionally disseminate misinformation. Therefore, its application in critical fields, such as the medical industry, should be approached with great caution.

Acknowledgments

Thanks to Peng Bo, the author of RWKV, for participating in the discussion and providing valuable suggestions for modifications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. 2023. Zoology: Measuring and improving recall in efficient language models. *arXiv preprint arXiv:2312.04927*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed

Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, and Chunhua Shen. 2023. Mobilevlm : A fast, strong and open vision language assistant for mobile devices. *ArXiv*, abs/2312.16886.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Mark Harris, Shubhabrata Sengupta, and John D Owens. 2007. Parallel prefix sum (scan) with cuda. *Graphics Processing Unit Gems*, 3(39):851–876.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Haowen Hou and F. Richard Yu. 2024. Rwkv-ts: Beyond traditional recurrent neural network for time series tasks. *ArXiv*, abs/2401.09093.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

IDEFICS. 2023. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. *ArXiv*, abs/2402.07865.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. 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*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zihang Li and Haowen Hou. 2024. Visualrwkv-hd and uhd: Advancing high-resolution processing for visual language models. *ArXiv*, abs/2410.11665.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. *CVPR*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Ves Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *ArXiv*, abs/2112.10668.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

OpenAI. 2023. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023a. Rwkv: Reinventing rnns for the transformer era. *Preprint*, arXiv:2305.13048.

Bo Peng, Eric Alcaide, Quentin G. Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, G Kranthikiran, Xuming He, Haowen Hou, Przemyslaw Kazienko, Jan Kocoń, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Xiangru Tang, Bolun Wang, Johan Sokrates Wind, Stansilaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui Zhu. 2023b. Rwkv: Reinventing rnns for the transformer era. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemys l aw Kazienko, G Kranthikiran, Jan Koco'n, Bartlomiej Koptyra, Satyapriya Krishna, Ronald McClelland, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Stanislaw Wo'zniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Ruijie Zhu. 2024. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *ArXiv*, abs/2404.05892.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023c. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

E. Ponti, Goran Glavavs, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

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. *ArXiv*, abs/2403.13600.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *Preprint*, arXiv:2106.12066.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024a. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. 2024b. Llava-phi: Efficient multi-modal assistant with small language model. *ArXiv*, abs/2401.02330.