ModRWKV: Transformer Multimodality in Linear Time

Jiale Kang[1*†], **Ziyin Yue**[1*], **Qingyu Yin**[2*],
Jiang Rui[1,3], Weile Li[1], Zening Lu[1], Zhouran Ji[1]
[1]Yuanshi Inc, [2]Zhejiang University, [3]The Hong Kong University of Science and Technology
jiale@rwkvos.com

Abstract

Currently, most multimodal studies are based on large language models (LLMs) with quadratic-complexity Transformer architectures. While linear models like RNNs enjoy low inference costs, their application has been largely limited to the text-only modality. This work explores the capabilities of modern RNN architectures in multimodal contexts. We propose ModRWKV—a decoupled multimodal framework built upon the RWKV7 architecture as its LLM backbone—which achieves multi-source information fusion through dynamically adaptable heterogeneous modality encoders. We designed the multimodal modules in ModRWKV with an extremely lightweight architecture and, through extensive experiments, identified a configuration that achieves an optimal balance between performance and computational efficiency. ModRWKV leverages the pretrained weights of the RWKV7 LLM for initialization, which significantly accelerates multimodal training. Comparative experiments with different pretrained checkpoints further demonstrate that such initialization plays a crucial role in enhancing the model's ability to understand multimodal signals. Supported by extensive experiments, we conclude that modern RNN architectures present a viable alternative to Transformers in the domain of multimodal large language models (MLLMs). Furthermore, we identify the optimal configuration of the ModRWKV architecture through systematic exploration.

https://github.com/JL-er/WorldRWKV

1 Introduction

Linear complexity model (Peng et al., 2025; Gu and Dao, 2024; Yang et al., 2024a, 2025, 2024b) have emerged as an efficient alternative to the attention-based Transformer architecture (Vaswani et al., 2023; Yin et al., 2024) in Large Language Models (LLMs) (Touvron et al., 2023; Achiam et al., 2023). Among various linear models, recurrent neural networks (RNNs) (Peng et al., 2025) have become a competitive approach. Characterized by constant memory usage, RNNs can perform inference at a lower cost compared to the linearly increasing KV cache of Transformers. Recent research has also enabled their parallel training capabilities (Yang et al., 2024a, 2025), facilitated by hardware-aware designs optimized for modern GPU architectures (Dao et al., 2022).

Currently, LLMs are undergoing a paradigm shift—from single-modality processing to cross-modal collaboration (Liu et al., 2023; Fang et al., 2025; Chen et al., 2022; Défossez et al., 2024; Li et al., 2025b,a). By leveraging transfer learning from pre-trained LLM weights, these models achieve cross-modal semantic alignment in tasks such as visual question answering and speech dialogue. However, this practice has primarily been employed within the traditional Transformer architecture. In the context of linear models, few works have expanded their understanding to modalities beyond natural language. This disparity highlights a crucial gap in the current landscape of linear models.

In this paper, we describe ModRWKV. It is the first RNN-based linear model that extends its capabilities to the cross-modal domain. ModRWKV is based on RWKV7, a RNN-based architecture powered by generalized delta rule with vector values gating, in-context learning rates, and relaxed value replacement rule. We hypothesize that the inherent sequential processing capabilities of RNNs, coupled with a carefully designed shared parameter base, can effectively capture both intra-modal and inter-modal dependencies across diverse data types.

We take advantage of the RWKV7 architecture to propose an innovative unified training paradigm

* Equal contributions.
† The corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4104–4116
November 4-9, 2025

for multimodal fusion. MODRWKV adopts a lightweight shared parameter base with a modality-specific encoder framework, where simply switching the front-end encoder enables seamless transfer across multimodal tasks. This approach systematically explores the representation capabilities of RNN architectures within cross-modal semantic spaces, aiming to break the Transformer-dominated research paradigm. It offers new theoretical and practical insights into the deployment of large RNN-based models in the multimodal domain.

Our contributions can be summarized as threefold:

1. Proposed the MODRWKV framework, pioneering a unified multimodal training paradigm based on an RNN architecture. By adopting a plug-and-play design for modality encoders, it significantly enhances cross-modal scalability and integration efficiency.

2. Conducted a comprehensive and systematic evaluation of MODRWKV's full-modality understanding capabilities, establishing a benchmark paradigm for assessing the cross-modal performance of RNN-based architectures.

3. Extensive Ablation experiments validate the optimal multimodal processing design that achieves a desirable balance between performance and computational efficiency.

2 Background

RWKV7: Modern RNN Architecture Simple linear RNNs (Qiao et al., 2024; Gu and Dao, 2024) can be written in the following recurrent form:

$$\boldsymbol{h_t} = \boldsymbol{W}\boldsymbol{h_{t-1}} + \boldsymbol{U}\boldsymbol{x_t}, \tag{1}$$

which enables parallelized training but lacks strong language performance and long-term dependency preservation. RWKV combines the efficiency of linear RNNs (constant memory and time complexity during inference) with powerful modeling capabilities through its time-mixing block. It uses keys $\boldsymbol{k_t}$ and values $\boldsymbol{v_t}$, linearly projected from $\boldsymbol{x_t}$, and updates the state $\boldsymbol{s_t}$ with input-dependent decay $\boldsymbol{w_t}$ and receptance $\boldsymbol{r_t}$:

$$\boldsymbol{s_t} = e^{-\boldsymbol{w_t}} \cdot \boldsymbol{s_{t-1}} + \boldsymbol{k_t}\boldsymbol{v_t^T}, \tag{2}$$

In RWKV7, the state update is enhanced for greater expressiveness with the form:

$$\boldsymbol{s_t} = \boldsymbol{G_t}\boldsymbol{s_{t-1}} + \boldsymbol{a_t}\boldsymbol{k_t}\boldsymbol{v_t^T}, \tag{3}$$

where employed a generalized delta rule with two improvements: (1) **In-context learning rate.** the term $\boldsymbol{a_t}$, a vector-valued learning rate projected as $\boldsymbol{a_t} = \boldsymbol{W_a}\boldsymbol{x_t}$, controls the influence of the new information $\boldsymbol{k_t}\boldsymbol{v_t^T}$ on the state. (2) **Vector value gating.** The dynamic transition matrix $\boldsymbol{G_t} = (\boldsymbol{I} - \boldsymbol{a_t}\boldsymbol{k_t}\boldsymbol{k_t^T})\mathrm{diag}(e^{-e^{\boldsymbol{w_t}}})$ incorporates $\boldsymbol{w_t}$, a vector-valued gating parameter from $\boldsymbol{w_t} = \boldsymbol{W_w}\boldsymbol{x_t}$, enabling channel-specific decay rates. This input-dependent design makes $\boldsymbol{s_t}$ highly adaptive to context.

Multimodal Large Language Models LLMs have traditionally been trained on natural language data and are primarily designed to understand and generate text. These models excel in text-based tasks but are inherently limited to the domain of human language. Recently, many works have begun to explore the potential of large language models beyond their linguistic roots, pushing their capabilities into other modalities. From a modality perspective, MLLMs now handle a variety of data types beyond text, including images (Liu et al., 2024a), audio (Défossez et al., 2024), and video. Structurally, these models adapt by incorporating modality-specific encoders, such as visual transformers for images or audio transformers for sound. Input integration varies between unified tokenization, where all modalities are converted into a single token sequence, and cross-modal attention, where the model attends to features across modalities.

3 Methodology

MODRWKV is the first RNN-based multimodal architecture that integrates the MLLM training paradigm with a linear model, achieving exceptional hardware efficiency. In Section 3.1, we present the encoder selection design of MODRWKV. In Section 3.2, we detail the adapter design of MODRWKV. In Section 3.3, we describe the sequence compression method for efficiently processing diverse multimodal data.

3.1 Multimodal Encoder

Vision Encoder. We evaluated CLIP (Radford et al., 2021) and SigLIP2 (Tschannen et al., 2025) as alternative visual encoders for MODRWKV, applying identical adaptation frameworks to each model independently. Each vision-language encoder processes raw images to generate sequential feature embeddings that are then aligned with the

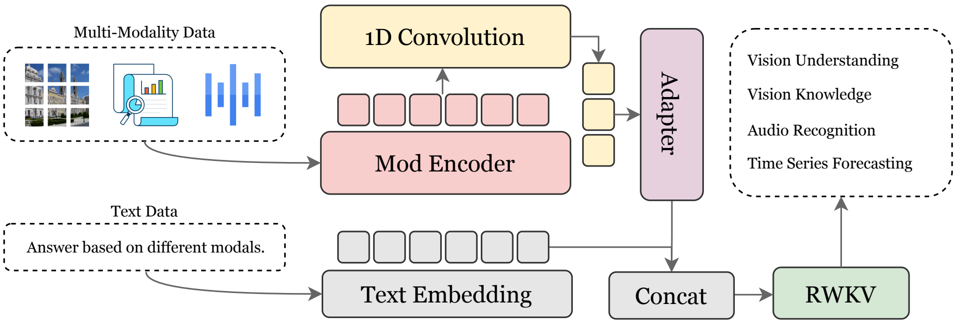

Figure 1: ModRWKV network architecture. Multi-modality data streams undergo initial processing via an encoder, a 1D Convolutional layer, and an adapter. (The 1D Convolutional layer is employed to compress the sequence length of multi-modal inputs, which significantly reduces the computational overhead during training.) Concurrently, text data is transformed through a Text Embedding module. The outputs from the adapter and Text Embedding layers are subsequently concatenated.

RWKV large language model through lightweight adapter layers. Our experiments validated ModRWKV's strong inherent capacity for visual information processing, with this framework demonstrating excellent cross-modal adaptability even without architectural modifications to the base language model.

Audio Encoder. In our study, we employ WavLM (Chen et al., 2022) and Whisper (Radford et al., 2022) as audio encoders for ModRWKV. We select encoder models with sizes ranging from approximately 100M to 400M parameters, specifically choosing WavLM base+, WavLM large, Whisper small, and Whisper medium for evaluation. These encoders process audio sampled at 16,000 Hz and generate feature vectors at a frequency of 50 Hz. For the Whisper encoder, each audio segment is padded to a duration of 30 seconds.

Time Series Encoder. We adopt WaveNet (Van Den Oord et al., 2016) and Timer (Liu et al., 2024b) as alternative temporal encoders for ModRWKV. Timer is initialized with pre-trained weights, with the weights frozen during training, while WaveNet is trained from scratch without pre-trained weights. However, during inference, both encoders are frozen to enable zero-shot evaluation. Each encoder transforms raw time-series data into high-level feature embeddings, which are then aligned with the RWKV blocks via lightweight adapters.

3.2 Adapter Design

We introduce a single-MLP adapter (Liu et al., 2023) for dimension alignment between modalities, reducing the adapter's parameter. This forces the RWKV7 backbone to handle the majority of cross-modal reasoning, providing a rigorous test of RNN-based architectures in multimodal settings:

$$\boldsymbol{h} = \text{Linear}_2(\text{ReLU}(\text{Linear}_1(\boldsymbol{x}))). \quad (4)$$

Table 1: Multimodal Benchmark Evaluation

Benchmark	Description
VQA-v2 (Goyal et al., 2017)	Image Understanding
TextVQA (Singh et al., 2019)	Text-Image Integration
GQA (Hudson and Manning, 2019)	Reasoning
ScienceQA (Lu et al., 2022)	Scientific Reasoning
POPE (Li et al., 2023)	Hallucination
MMMU (Yue et al., 2024)	Reasoning
MMBench (Liu et al., 2024c)	Assessment
LibriSpeech (Panayotov et al., 2015)	Speech Recognition
Aishell-1 (Bu et al., 2017)	Speech Recognition
GIFT-Eval (Aksu et al., 2024)	Time Series
UTSD (Liu et al., 2024b)	Time Series

3.3 Sequence Compression

To address the computational challenges of long sequences in LLMs, we employ 1D convolution to effectively compress multimodal sequences (e.g., image patches, audio spectrograms). This approach significantly reduces processing overhead while maintaining model performance. For an input $\boldsymbol{x} \in \mathbb{R}^{C_{\text{in}} \times L}$, a convolutional kernel $\boldsymbol{W} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k}$, stride $s \geq 1$, padding p, the c-th output channel $\boldsymbol{Y} \in \mathbb{R}^{C_{\text{out}} \times L'}$ is computed as:

$$\boldsymbol{y}_c = \underbrace{\sum_{i=1}^{C_{\text{in}}} \left(\sum_{j=0}^{k-1} \boldsymbol{W}_{c,i,j} \cdot \boldsymbol{x}_{i,s \cdot t+j} \right) + b_c}_{\text{Conv1D}}, \quad (5)$$

where $t = 0, \ldots, L' - 1$ and L' is computed as $L' = \left\lfloor \frac{L+2p-k}{s} \right\rfloor + 1$.

Table 2: **Comparison with SoTA methods on 7 benchmarks.** Benchmark names are abbreviated due to space limits. VQA-v2; GQA; SQA^{I}: ScienceQA-IMG; VQA^{T}: TextVQA; POPE; MMB: MMBench; MMMU. 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	MMB	MMMU
LLaVA-1.5	Vicuna-7B	558K	665K	78.5	62.0	66.8	58.2	86.5	64.3	-
LLaVA-1.6	Vicuna-7B	558K	665K	**81.8**	**64.2**	**72.8**	**65.7**	86.7	67.7	35.8
LLaVA-Phi	Phi-2-2.7B	558K	665K	71.4	-	68.4	48.6	85.0	59.8	-
MobileVLM-3B	MobileLLaMA-2.7B	558K	665K	-	59.0	61.2	47.5	84.9	59.6	-
VL-Mamba	Mamba LLM-2.8B	558K	665K	76.6	56.2	65.4	48.9	84.4	57.0	
+ShareGPT4V	LLaMA2-13B	558K	665K	80.6	63.2	73.1	65.3	84.8	70.8	
ModRWKV	RWKV7 LLM-3B	558K	665K	78.3	60.8	70.9	51.1	**87.1**	66.6	**38.7**

Table 3: Model's WER(%) on Librispeech dataset and CER(%) on Aishell-1 dataset.

Dataset	**Data (h)**	**Encoder**	**Clean WER(%)**	**Other WER(%)**	**Dev CER(%)**	**Test CER(%)**
Librispeech	960	wavlm large	2.43	6.51	-	-
		wavlm base+	3.08	10.38	-	-
		whisper medium	5.33	12.28	-	-
		whisper small	6.24	16.92	-	-
Aishell-1	178	wavlm large	-	-	9.68	10.33
		wavlm base+	-	-	12.40	13.46
		whisper medium	-	-	5.08	5.83
		whisper small	-	-	6.29	6.95

4 Experiments

4.1 Experimental Details

Training Settings (1) *Vision*. Our implementation follows the phased training paradigm of LLaVA (Liu et al., 2023) for both vision and audio understanding. In Phase I, we first freeze the encoder and the RWKV model, training only a linear adapter with a single MLP and layer norm to project multimodal features into the embedding space of the language model. In Phase II, we then unfreeze both the adapter and RWKV parameters, while the encoder remains frozen to preserve pretrained representations. To comprehensively assess the impact of encoder choice and model scale on RWKV7 performance, we performed experiments on four vision languagemarks using three model sizes (0.4B, 1.5B and 3B) for each encoder. Our models are trained on 8×NVIDIA A800 GPUs. Details of training settings can be found at Appendix 12. (2) *Audio*. Training was conducted in two phases: Phase I trained only the audio adapter (LR=1e-4), while Phase II jointly trained the adapter and RWKV (LR decayed from 1e-4 to 5e-5). For LibriSpeech, we ran 1 epoch in each phase; for Aishell-1, 2 epochs in Phase I and 4 in Phase II. The default batch size was 32, reduced to 16 for the Whisper encoder due to GPU constraints, with epochs halved accordingly to match training steps. All experiments used 44×090 GPUs. (3) *Time series*. In the Time series task, We conducted experiments using dual NVIDIA RTX 4090 (24GB) GPUs, training on a 441,725-sample short-duration univariate dataset.

Datasets We consider diverse datasets in vision, audio, and time series (Refer to Table 1). For vision understanding ability, we use LLaVA-595K as training dataset for Phase I, and LLaVa-665k for Phase II. For audio, We train our ModRWKVmodel using two open-source datasets: (1) LibriSpeech (Panayotov et al., 2015), which comprises 960 hours of English reading audio data; and (2) Aishell-1 (Bu et al., 2017), which includes 170 hours of Chinese audio data. For each, we trained our model exclusively on the respective training dataset. In the time series task, we utilized public datasets from GIFT-Eval (Aksu et al., 2024). After thorough sorting and cleaning, we derived a small number of univariate datasets. Additionally, we incorporated UTSD (Liu et al., 2024b) public datasets later in the process.

Table 4: Zero-shot MSE with Adapter Scaling 4× use gift-eval datasets (WaveNet Encoder) (Qiu et al., 2024)

Model	LB-FL	ECL	ETTh1	ETTh2	ETTm1	ETTm2	WTH	Traffic
TimeFM	720-96	**0.119**	0.421	0.326	0.363	0.206	**0.123**	**0.327**
Timer	720-96	0.221	0.414	0.305	0.440	0.203	0.178	0.526
UniTS	720-96	0.175	0.377	0.323	0.761	0.249	0.194	0.481
TTM	720-96	0.170	**0.368**	0.286	0.415	**0.186**	0.152	0.509
MOIRAI	720-96	0.212	0.394	**0.285**	0.516	0.222	0.208	1.359
ROSE	720-96	0.209	0.382	0.298	0.512	0.224	0.200	0.572
ModRWKV(25% gift-eval)	720-96	0.342	0.746	0.633	0.754	0.559	0.797	0.512
ModRWKV(100% gift-eval)	720-96	0.342	0.648	0.453	**0.227**	0.426	0.203	0.342

Benchmarks To rigorously evaluate our model's capabilities across diverse reasoning scenarios, we employed a comprehensive evaluation framework spanning from basic visual recognition to advanced knowledge-intensive tasks. This framework systematically verifies our model's cross-modal competence at various cognitive levels by assessing it on seven multimodal benchmarks: VQA-v2 (Goyal et al., 2017) for fundamental image understanding and question-answering, TextVQA (Singh et al., 2019) to evaluate optical character recognition (OCR) and text-image integration, GQA (Hudson and Manning, 2019) for compositional reasoning and real-world visual understanding, ScienceQA (Lu et al., 2022) to assess scientific multimodal reasoning through multiple-choice questions, POPE (Li et al., 2023) to quantify object hallucination via binary classification tasks, MMMU (Yue et al., 2024) to challenge models with college-level, cross-discipline problems, and MMBench (Liu et al., 2024c), which represents a systematically designed, objective evaluation framework for comprehensive assessment that uses circularEval strategy for assessment stability, ETT (Qiu et al., 2024), which focuses on long-term multivariate time-series forecasting using electricity transformer temperature data, serving as a standard benchmark for evaluating temporal modeling capabilities under various sequence lengths and prediction horizons, WeatherBench (Rasp et al., 2020) to evaluate spatiotemporal forecasting using global atmospheric data as a standard benchmark for data-driven weather prediction, etc.. Additionally, we evaluated our ModRWKVmodel using the corresponding open-source datasets: LibriSpeech (Panayotov et al., 2015), which comprises 960 hours of English reading audio data, and Aishell-1 (Bu et al., 2017), which includes 170 hours of Chinese audio data.

4.2 Qualitative Evaluation

Vision Understanding As summarized in Table 2, ModRWKV demonstrates strong overall performance across eight widely-used multimodal benchmarks, outperforming existing state-of-the-art (SoTA) methods in its parameter range. Compared to VL-Mamba-2.8B, ModRWKV-3B consistently achieves higher scores on all evaluated tasks, reflecting its superior capability in visual question answering, compositional reasoning, and image-conditioned instruction following.

Notably, despite having a significantly smaller language backbone than LLaVA-1.5-7B, ModRWKV achieves competitive or superior results on several benchmarks. It surpasses LLaVA-1.5-7B in ScienceQA-IMG, POPE, and MMBench, while maintaining comparable performance on VQAv2. Furthermore, ModRWKV attains the highest reported score among peers on the MMMU benchmark, highlighting its generalization ability in challenging multi-modal understanding scenarios.

These results collectively suggest that ModRWKV offers a favorable trade-off between performance and model size. Its effectiveness stems not merely from scale, but from architectural efficiency and a well-designed multimodal integration strategy, positioning it as a competitive alternative to larger vision-language models.

Vision Knowledge The following examples in Table 13 showcase the capabilities of the ModRWKV QA chatbot. These examples illustrate how ModRWKV effectively integrates visual information with general knowledge, while also performing basic logical reasoning to address common user queries.

Audio Recognition Table 3 presents the Word Error Rate (WER) for the LibriSpeech test_clean and test_other test sets, as well as the Character Error Rate (CER) for the Aishell-1 development and test sets. For the LibriSpeech dataset, the model

Table 5: Zero-shot MSE on Public Datasets: ECL, ETTh, ETTm, WTH, Traffic (Timer Encoder) (Liu et al., 2024b)

Dataset Size	Adapter Scaling	ECL	ETTh1	ETTh2	ETTm1	ETTm2	WTH	Traffic
Gift-Evel	2×	0.641	0.785	0.882	0.949	0.719	0.633	0.988
Gift-Evel + UTSD	2×	0.516	0.637	0.848	0.891	0.672	0.512	0.683
Gift-Evel + UTSD	4×	**0.453**	**0.629**	**0.547**	0.843	**0.648**	**0.461**	0.641
Gift-Evel + UTSD	8×	0.535	0.629	0.652	**0.828**	0.762	0.566	**0.617**

Table 6: MODRWKV Visual Models with different Encoders and parameters tested on benchmarks.

Vision	Size	VQA^{v2}	VQA^{T}	GQA	SQA^{I}
CLIP	0.4B	62.04	31.72	49.32	51.10
	1.5B	72.31	40.27	54.56	62.77
	3B	73.13	45.56	57.00	70.66
SigLIP2	0.4B	72.04	38.75	55.52	43.32
	1.5B	76.95	44.96	58.88	63.10
	3B	78.30	51.09	60.75	70.93

achieved a WER of 2.43% on the test_clean subset, indicating a high level of precision in recognizing clear speech. On the test_other subset, the model attained a WER of 6.51%, demonstrating reasonable performance in handling more challenging noisy speech samples without data augmentation. For the Aishell-1 dataset, the model achieved CERs of 5.08% on the development set and 5.83% on the test set, using the Whisper medium encoder. These results reflect the model's effectiveness in handling non-English speech recognition tasks with limited training data.

During adapter training, we observed a phenomenon akin to the *capability emergence* described by (Ma et al., 2024). However, the timing of this emergence was inconsistent and heavily influenced by the initialization of the adapter's weights. In some instances, the adapter failed to converge during Phase I.

Time Series Forecasting We conducted comparative experiments on two temporal encoder architectures: Timer and WaveNet. Results (See Table 4) show that although Timer has a larger parameter count (based on pre-trained weights), it consistently underperforms WaveNet on downstream time-series forecasting tasks. We hypothesize that this performance gap arises from WaveNet's use of causal dilated convolutions, which effectively capture long-range temporal dependencies through hierarchically expanding receptive fields. Additionally, unlike Timer's patch-wise embedding, WaveNet adopts a point-wise embedding strategy, allowing it to extract finer-grained temporal features. For training data preparation, we constructed two fine-tuning datasets: a baseline dataset (GIFT-Eval)(Aksu et al., 2024) and an augmented dataset composed of GIFT-Eval and a partially processed subset of UTSD(Liu et al., 2024b). Experiments indicate that models trained on the augmented dataset—containing anomalous samples—achieved superior generalization in zero-shot evaluations across public benchmarks including ECL, ETT, WTH, and Traffic. Notably, this training strategy enables the model to maintain stable predictions even under distribution shifts, demonstrating strong robustness and generalization. Architecture ablation studies further revealed that the scaling factor of the adapter modules plays a significant role in performance. A scaling factor of 4× yielded the best overall results on the validation set (see Table 5), outperforming the 8× and 2× settings by approximately 10.0% and 13.5%, respectively.

Overall, even under constrained conditions—no data augmentation, limited training data, and fewer training steps—the MODRWKV model achieved competitive accuracy on time-series forecasting tasks, providing empirical evidence for its applicability in real-world, complex scenarios.

4.3 Ablation Study

The Effect of Different Vision Encoders In order to evaluate the impact of different vision encoders on the performance of multimodal models, this study designed rigorous comparative experiments. We selected two representative visual encoder architectures for comparison: the contrastive learning-based CLIP and the recently proposed SigLIP2. In the experimental design, we specifically controlled the following variables: the length of the encoded visual feature sequences for both (google/siglip2-base-patch16-384 and openai/clip-vit-large-patch14-336) was set to 577, to eliminate any potential confusion caused by differences in sequence length that could affect the understanding

Table 7: By controlling the kernel and stride of conv1d, control the sequence length of multimodal signals to compare performance differences.

Size	(k,s)	Token	VQA^{v2}	VQA^{T}	GQA	SQA^{I}
1.5B	(0,0)	577	76.95	44.96	58.88	63.10
	(3,2)	288	75.21	45.75	58.28	66.02
	(4,3)	192	74.17	44.27	57.53	65.72
	(5,4)	144	73.21	42.65	57.07	65.29

Table 8: Performance differences under different pretraining weights

Size	Model	VQA^{v2}	VQA^{T}	GQA	SQA^{I}
0.4B	base	72.04	38.75	55.52	43.32
	g1	73.21	41.13	57.34	55.58
1.5B	base	76.95	44.96	58.88	63.10
	g1	77.87	50.91	60.18	64.63

ability of LLMs; cross-validation was conducted on LLMs of different scales (ranging from 0.4B to 3B parameters) to ensure the generalizability of the experimental conclusions.

As shown in Table 6, SigLIP2 encoder consistently outperforms CLIP encoder all evaluated benchmarks, including VQAv2, TextVQA, GQA, and ScienceQA. Notably, the SigLIP2-based model achieves significant improvements in both general and text-based visual question answering tasks, as well as in compositional reasoning. Despite its encoder containing only 90M parameters—approximately 30% of the CLIP encoder's size—SigLIP2 demonstrates superior performance, particularly in tasks requiring fine-grained visual-text alignment and semantic understanding. These results underscore that model effectiveness in multimodal understanding is influenced more by encoder design and pretraining methodology than by parameter scale alone.

Efficiency of Sequence Compression via 1D Convolution It is well known that the efficiency problem in processing long sequences has long been one of the main bottlenecks limiting the performance of LLMs. This challenge is particularly prominent in multimodal tasks, where signals from different modalities often generate a large number of tokens after encoding. For example, in the MODRWKVmodel, a single image encoded through the SigLIP2 encoder generates 577 tokens, and when extended to video sequences, the length increases by an order of magnitude. To address this issue, this section systematically investigates the optimization effects of convolutional dimensionality reduction (Conv1D), aiming to provide new technical insights for sequence compression research.

We conducted empirical research (See Table 7 and visualization in Figure 2) on the MODRWKV-1.5B model architecture using the LLaVA training dataset, and performed comprehensive evaluations across multiple benchmark datasets, including VQAv2, TextVQA, GQA, and ScienceQA. The experimental results show that when the sequence length is compressed by 50%, the model exhibits only a slight decrease in performance (on average) while achieving a 4.6% accuracy improvement on the ScienceQA task. Further research reveals that as the kernel size and stride increase, although the model performance exhibits a gradual decline, the computational efficiency is significantly improved. We tested MODRWKV-1.5B on single 4090 GPU without any acceleration; The results indicate that increasing the compression ratio of token sequences can substantially accelerate inference speed, showing a clear efficiency gain. This highlights an effective strategy for balancing computational efficiency and model performance, offering valuable insights for practical deployment.

G1 reasoning model We experimentally validate the effect of text pretraining weights on the ability of large language models to understand multimodal information by comparing two pretraining weights (base and g1) of the RWKV7-0.4B model. It is important to note that the g1 model is an improved version of the base model, obtained through post-training by introducing a large amount of 'think'-type data. Although both models perform similarly in pure text NLP benchmark tests, as shown in Table 8, fine-tuning with the g1 pretraining weights significantly outperforms the base model across all metrics, with an exceptionally significant improvement observed in the SQA metric (specific improvement is 28%). This empirical result strongly confirms that an appropriate text pretraining strategy can effectively enhance the language model's ability to understand multimodal information, thereby improving its overall performance in downstream tasks.

Time Series Forecasting encoder Compare with Timer and WaveNet In a feedforward neural network (FFN), activation functions such as ReLU introduce sparsity by setting some outputs to zero, which in turn reduces the rank of the output matrix and may impact the model's representational

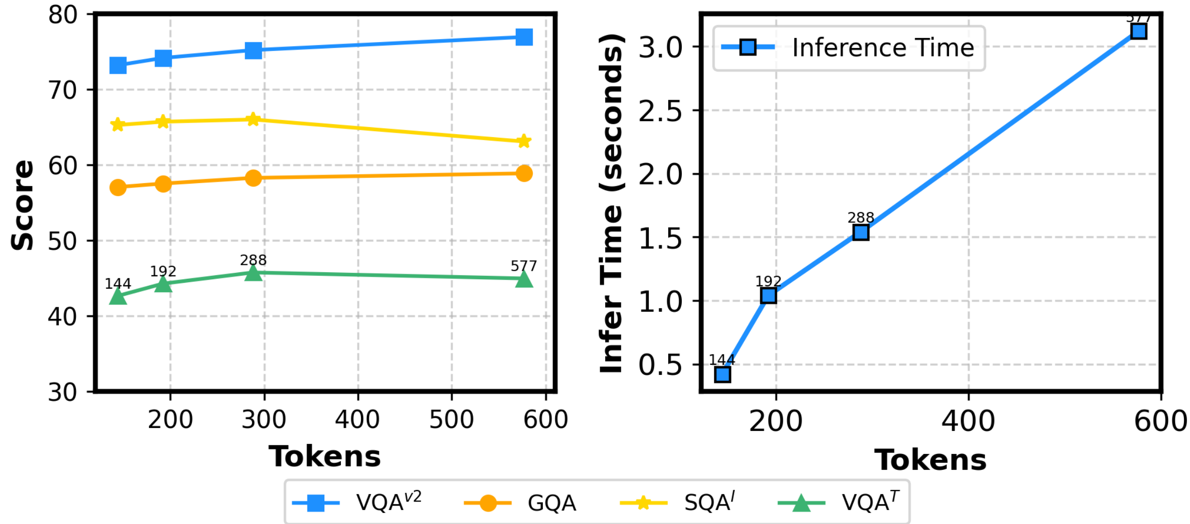

Figure 2: Performance and efficiency of MODRWKV. **Left.** The scaling curve of tokens with the performance score. **Right.** The inference time of MODRWKV with the number of tokens.

capacity. Through both theoretical analysis and empirical experiments using the results in Table 5, we observed that the effect is suboptimal when the hidden layer dimension is set to 2x or 8x the input dimension.

Table 5 presents the zero-shot mean squared error (MSE) performance of different adapter scaling configurations on multiple public datasets, including ECL, ETTh, ETTm, WTH, and Traffic. The results indicate that increasing the adapter scaling factor from 2× to 4× significantly improves performance across most datasets, with the lowest MSE values observed at 4× scaling. Specifically, the Gift-Evel + UTSD model with 4× scaling achieves the best results on ECL (0.453), ETTh1 (0.629), ETTh2 (0.547), ETTm2 (0.648), WTH (0.461), and Traffic (0.641), demonstrating that this configuration effectively enhances model accuracy.

However, further increasing the scaling factor to 8× does not consistently improve performance, with some datasets showing increased error values. This suggests that excessively large hidden layer dimensions may introduce instability or diminish representational efficiency. Based on these findings, we recommend setting the hidden layer dimension to at least four times the input dimension to preserve sufficient rank, thereby enhancing the model's representational power and stability.

$$p = 1 - \frac{\sum_{i=m}^{n} \binom{n}{i}}{2^n}$$

Table 4 presents the zero-shot mean squared error (MSE) results for various models using the WaveNet encoder with adapter scaling 4× on public datasets, leveraging the gift-eval dataset. The models are evaluated on multiple time-series forecasting benchmarks, including ECL, ETTh, ETTm, WTH, and Traffic, with a lookback length of 720 and a forecast length of 96.

From the results, TimeFM achieves the best performance on ECL (0.119), WTH (0.123), and Traffic (0.327), demonstrating strong predictive capabilities on these datasets. TTM performs best on ETTh1 (0.368) and ETTm2 (0.186), while MOIRAI achieves the lowest error on ETTh2 (0.285). Our proposed model, MODRWKV (100% gift-eval), outperforms other models on ETTm1 (0.227), showing its effectiveness in short-term forecasting for this dataset.

Comparing MODRWKV(25% gift-eval) and MODRWKV(100% gift-eval), we observe that increasing the proportion of gift-eval data significantly improves performance across most datasets, particularly on ETTh2 (from 0.633 to 0.453) and ETTm1 (from 0.754 to 0.227). This suggests that leveraging a larger portion of the gift-eval dataset enhances ours model generalization and stability.

Overall, the results highlight the varying strengths of different models across datasets, emphasizing the importance of dataset composition and model architecture in achieving optimal forecasting performance.

5 Empirical Validation and Comparison

5.1 Inference Throughput

We first measured inference throughput on a single NVIDIA 4090 GPU. The results show that Mod-

RWKV achieves significantly higher throughput than both LLaVA and Qwen2.5-VL, with the performance gap widening at larger batch sizes.

A key reason for this advantage is the efficiency of our vision encoder. ModRWKV uses a lightweight SigLIP2 encoder that consistently generates a fixed-length representation of 577 tokens per image. In contrast, the Qwen2.5-VL model employs a much heavier vision encoder that dynamically generates a variable number of tokens based on image resolution, often resulting in a larger token count and thus higher computational cost.

Table 9: Throughput (tokens/s)

Model	Enc.	Params	bs1	bs4	bs16
qwen2.5-vl-3B	qwen-vit	630 M	40	43	43
llava-1.6-7B	clip-L/14	304 M	199	293	421
modrwkv-3B	siglip2-B/16	93 M	433	742	1245

5.2 Scalability with Sequence Length

To directly test the impact of sequence length on performance, we measured the prefill throughput (tokens/second) for text generation using the `llama.cpp` framework on an NVIDIA 4090 GPU. This experiment clearly demonstrates the theoretical complexity differences.

As shown in Table 10, the throughput of the Transformer-based Qwen2.5-3B degrades severely as the sequence length increases, dropping by over 90% from 1k to 64k tokens. This is a direct result of its quadratic time complexity. In stark contrast, the RWKV-7B model maintains a nearly constant throughput across all sequence lengths, empirically confirming its linear scaling and efficiency for long-context processing.

Table 10: Scalability with Sequence Length

Length	qwen2.5-3B (t/s)	rwkv7-2.9B (t/s)
1k	18725.66 ± 58.62	11917.64 ± 18.35
2k	17040.16 ± 25.96	11991.45 ± 3.43
4k	13768.85 ± 14.30	12004.39 ± 25.49
8k	9997.83 ± 3.98	12015.55 ± 10.16
16k	6306.18 ± 0.36	12017.75 ± 18.61
32k	3457.69 ± 0.23	12047.45 ± 13.47
64k	1826.48 ± 0.06	12010.09 ± 1.20

5.3 Training and Evaluation Time

Finally, we compared the evaluation time on the GQA and TextVQA benchmarks. Despite being trained on significantly fewer tokens (2.5B for ModRWKV vs. 4.1T for Qwen2.5-VL), ModRWKV demonstrates competitive or superior evaluation speed. This highlights that our model not only performs well but also achieves its results with greater training and evaluation efficiency.

Table 11: Training and Evaluation Time

Model	Tokens	GQA/min	TVQA/min
ModRWKV-3B	2.5B	60.5 / 24	55.2 / 14
qwen2.5vl-3B	4.1T	60.2 / 45	77.8 / 70
llava-ov-0.5B	5B	52.9 / 30	59.9 / 31
ModRWKV-0.4B	2.5B	56.9 / 9	42.0 / 7

6 Conclusion

In this paper, we propose MODRWKV, a multimodal understanding framework that enables modality switching via interchangeable encoders. Built upon RWKV7, MODRWKV provides a comprehensive analysis and evaluation of the capabilities of modern RNN architectures in the multimodal domain.

7 Limitations

This paper presents a systematic evaluation of the proposed MODRWKV framework across a range of benchmark tasks involving different modalities, demonstrating the feasibility of applying linear-structured models to multi-modal large language models (MLLMs). Nonetheless, this work does not yet explore more complex multi-modal fusion scenarios, such as tri-modal tasks involving speech, vision, and language. Future work will aim to address these richer multi-modal settings.

8 Acknowledgments

We thank Bo Peng and Jiaming Kong for their insightful discussions. We are deeply grateful to RWKVOS for all the support that enabled us to successfully complete all the experiments.

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*.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. 2024. Gift-eval: A benchmark for general time series forecasting model evaluation. *arxiv preprint arxiv:2410.10393*.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models.

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.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. pages 6693–6702.

Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. 2025a. M^2iv: Towards efficient and fine-grained multimodal in-context learning via representation engineering.

Yanshu Li, Tian Yun, Jianjiang Yang, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025b. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. *arXiv preprint arXiv:2505.17098*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. 2023. Evaluating object hallucination in large vision-language models.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning.

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

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024b. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player?

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024. An embarrassingly simple approach for llm with strong asr capacity.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S. Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. 2025. Rwkv-7 "goose" with expressive dynamic state evolution.

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.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. 2024. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *arXiv preprint arXiv:2403.20150*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. Weatherbench: a benchmark data set for data-driven weather forecasting.

Journal of Advances in Modeling Earth Systems, 12(11):e2020MS002203.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. pages 8309–8318.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. 2024a. Gated linear attention transformers with hardware-efficient training.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. 2024b. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. 2025. Parallelizing linear transformers with the delta rule over sequence length.

Qingyu Yin, Xuzheng He, Xiang Zhuang, Yu Zhao, Jianhua Yao, Xiaoyu Shen, and Qiang Zhang. 2024. Stablemask: Refining causal masking in decoder-only transformer. *arXiv preprint arXiv:2402.04779*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

A Example Appendix

Table 12: Hyperparameter Settings for Visual Instruction Tuning of MODRWKV

Hyperparameter	Step1	Step2
bsz	256	128
ctx_len	2048	
lr	1e-3	2e-5
warmup_steps	100	
beta1	0.9	
beta2	0.99	
adam_eps	1e-8	
precision	bf16	
lr_schedule	wsd	

A.1 Examples

Visual Examples	
	
User	How many Cars in this picture?
ModRWKV	There are two cars in the picture, with one being the taxi cab the man is standing on.
	[*Start a new conversation, and clear the history*] 
User	What is the name of this picture?
ModRWKV	The name of this picture is "Starry Night "
User	Who drew it?
ModRWKV	Vincent van Gogh drew this picture.
	[*Start a new conversation, and clear the history*] 
User	Is this real in current life?
ModRWKV	No, this is not a real animal. It is a computer-generated image (CGI) of a dinosaur, specifically a T-Rex, walking through a forest.
User	Depict the environment
ModRWKV	The environment in the image is a forest with trees, rocks, and a dirt path. The CGI dinosaur is walking through this natural setting.
User	Based on the facial expression, what is it's emotion?
ModRWKV	The dinosaur appears to be angry, as it is depicted with its mouth open and teeth showing.
User	Based on its emotion and the environment, what do you think this dinosaur is doing right now?
ModRWKV	Based on its angry expression and the forest setting, it is possible that the dinosaur is either searching for food, defending its territory, or reacting to a perceived threat. The fact that it is walking through the forest suggests that it might be exploring its habitat or looking for prey. However, without more context or information, it is difficult to determine the exact reason for its current behavior.

Table 13: Visual Example.

Audio Examples	
User	What is your name?
MODRWKV	**Hello!** My name is RWKV, and I'm your friendly voice assistant here to help with any questions or tasks you have.
User	Can I drink coffee before I go to bed?
MODRWKV	Yes, you can drink coffee before you go to bed, but it's important to do so in moderation. Drinking coffee too close to bedtime can interfere with your sleep quality. Try to have your last cup of coffee at least 6 hours before bedtime. If you're sensitive to caffeine, you might want to avoid it altogether. Also, consider other relaxing activities before bed to help you wind down
User	Tell me something about machine learning.
MODRWKV	Machine learning is a branch of artificial intelligence that focuses on creating systems that can learn from and make decisions based on data. It involves algorithms that enable computers to identify patterns and make predictions or decisions without being explicitly programmed for each task. This technology is widely used in various applications, such as recommendation systems, speech recognition, and autonomous vehicles. Machine learning models can improve over time as more data is processed, making them more accurate and efficient. It's a powerful tool for automating tasks, enhancing user experiences, and driving innovation across many industries.

Table 14: Interactive Audio Examples Demonstrating Question-Answer Capabilities