MEXA: Towards General Multimodal Reasoning with Dynamic Multi-Expert Aggregation

Shoubin Yu[*] Yue Zhang[*] Ziyang Wang
Jaehong Yoon Mohit Bansal
UNC Chapel Hill

https://github.com/Yui010206/MEXA

Abstract

Combining pre-trained expert models offers substantial potential for scalable multimodal reasoning, but building a unified framework remains challenging due to the increasing diversity of input modalities and task complexity. For instance, medical diagnosis requires precise reasoning over structured clinical tables, while financial forecasting depends on interpreting plot-based data to make informed predictions. To tackle this challenge, we introduce MEXA, a training-free framework that performs modality- and task-aware aggregation of multiple expert models to enable effective multimodal reasoning across diverse and distinct domains. MEXA dynamically selects expert models based on the input modality and the task-specific reasoning demands (i.e., skills). Each expert model, specialized in a modality-task pair, generates interpretable textual reasoning outputs. MEXA then aggregates and reasons over these outputs using a Large Reasoning Model (LRM) to produce the final answer. This modular design allows flexible and transparent multimodal reasoning across diverse domains without additional training overhead. We extensively evaluate our approach on diverse multimodal benchmarks, including Video Reasoning, Audio Reasoning, 3D Understanding, and Medical QA. MEXA consistently delivers performance improvements over strong multimodal baselines, highlighting the effectiveness and broad applicability of our expert-driven selection and aggregation in diverse multimodal reasoning tasks.

1 Introduction

The rapid advancement of multimodal learning has significantly improved AI systems' ability to understand, reason, and interact with the real world, benefiting diverse tasks such as visual question answering (Antol et al., 2015; Hudson and Manning, 2019; Marino et al., 2019; Mathew et al., 2021; Tanaka et al., 2023; Winata et al., 2024; Wang et al., 2025; Zhang et al., 2024b), medical image diagnosis (Lau et al., 2018; He et al., 2020; Liu et al., 2021; Azam et al., 2022; Cai et al., 2023; Bai et al., 2024), and embodied AI (Brohan et al., 2023; Driess et al., 2023; Liu et al., 2024b; Durante et al., 2024; Majumdar et al., 2024; Zhang et al., 2024c). As AI systems become more integrated into complex, real-world applications, the ability to flexibly understand and reason upon heterogeneous multimodal inputs is increasingly essential. For example, a medical diagnostic system may need to interpret CT scans, extract structured information from clinical notes, and reason about temporal patterns in patient histories, while a financial forecasting system must effectively analyze and interpret plot data to make informed predictions about market trends and economic risks.

However, despite recent advances in multimodal learning, the increasing diversity and complexity of multimodal data pose significant challenges for developing a flexible and unified framework that can reason effectively across modalities and generalize across tasks requiring diverse skills. Existing multimodal architectures typically require training separate encoders tailored to each modality, along with designing complex cross-modal alignment mechanisms (Tan and Bansal, 2019; Li et al., 2022, 2023b; Liu et al., 2023a; Hong et al., 2023; Yu et al., 2024). While effective, these models require fine-tuning for each specific task, leading to substantial training overhead and limiting their adaptability to new modalities or tasks. Moreover, previous approaches (Jaegle et al., 2021; Team, 2024) often implicitly fuse multimodal inputs at early stages, restricting transparency and interpretability of the reasoning process. Instead, an ideal multimodal framework should seamlessly process inputs from any modality for any given task, dynamically directing these inputs to the most suitable multimodal expert models. This necessitates a modular design

[*] Equal contribution.

where each expert model is associated explicitly with skills corresponding to particular modality-task, thereby ensuring precise skill matching and enhancing interpretability.

In this paper, we propose **M**ultimodal **Ex**pert **A**ggregator (MEXA), a novel **training-free multi-expert aggregation framework** that dynamically coordinates a pool of specialized experts. MEXA selectively activates and aggregates reasoning outputs from the most relevant expert models for each input instance, guided by both the input modalities (*e.g.*, image, audio, 3D, medical scan) and the required level of reasoning demands (i.e., skills) for that instance. MEXA spans a wide spectrum of tasks, from low-level perceptual reasoning (*e.g.*, object recognition or OCR) to high-level cognitive inference (*e.g.*, temporal event understanding in video or diagnostic interpretation in medical imaging). Unlike existing methods that rely on monolithic, end-to-end fine-tuned multimodal models or fixed, statically defined expert pipelines, MEXA's modular expert coordination enables flexible, interpretable, and scalable multimodal reasoning.

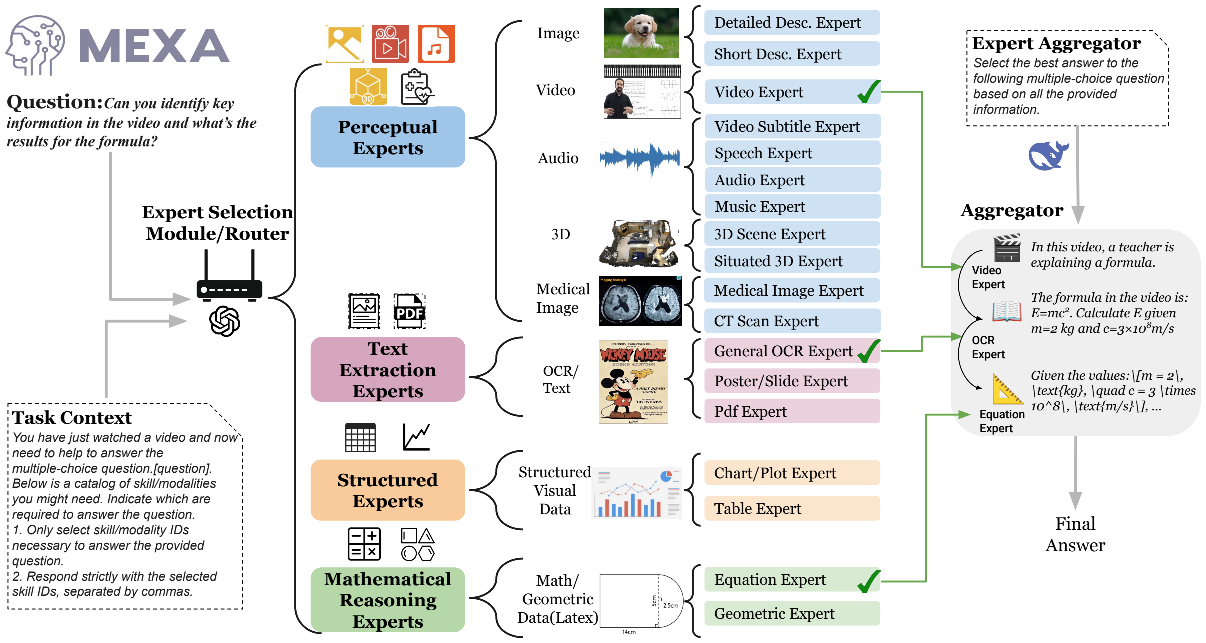
Specifically, we first assemble a diverse pool of expert models, each designed to handle distinct aspects of modalities and tasks, to address mainstream multimodal challenges effectively. Each expert specializes in extracting information unique to its respective skill and encoding it into a unified textual representation. Then, to effectively coordinate these experts, we design an expert selection module that employs a Multimodal Large Language Model (MLLM) as a versatile router within our framework. This router dynamically selects and activates the appropriate experts by analyzing the modality of the input data and identifying the specific skills required based on the input query and task requirements. Finally, we introduce an aggregator that reasons over the outputs of the selected experts using a Large Reasoning Model (LRM), which excels at long-context understanding and producing complex long CoT reasoning during LRM inference. Instead of relying on heuristic merging, the aggregator systematically integrates the complementary information from each expert and infers the final answer based on the full context of the target task.

We perform extensive evaluations of our framework on a diverse set of challenging multimodal benchmarks containing various modalities and requiring diverse skills, including expert knowledge heavy video reasoning (Video-MMMU (Hu et al., 2025a)), audio QA (MMAU (Sakshi et al., 2024)), 3D scene understanding (SQA3D (Ma et al., 2023a)), and medical imaging-based QA (M3D (Bai et al., 2024)). Experimental results show that our proposed method consistently outperforms strong multimodal baseline models across all evaluated benchmarks, with accuracy gains of +5.7% on Video-MMMU, +12.2% on MMAU, +1.7% on SQA3D, and +1.6% on M3D in the corresponding metrics, demonstrating the effectiveness, flexibility, and robustness of our expert-driven multimodal aggregation framework. We further conduct ablation studies on expert-selection router design and LLM aggregator to provide more insights for future work. Our contributions are summarized as follows:

- We present MEXA, a training-free, modality-extensible, and flexible framework that handles general multimodal reasoning via dynamic expert selection and aggregation.
- MEXA achieves strong performance on challenging multimodal benchmarks across diverse modalities, against specialized models as a more general framework.
- We conduct ablation studies and analyses based on our MEXA framework, providing insights into the design choices and effectiveness of each component.

2 Related Work

Mixture of Multiple Expert Models. Recently, the Mixture-of-Experts (MoE) paradigm (Shazeer et al., 2017; Zoph et al., 2022; Chen et al., 2023b; Zhou et al., 2022; Dai et al., 2022; Zhang et al., 2023; Chowdhury et al., 2023; Lee et al., 2024), which integrates multiple specialized parametric *expert modules*, has been widely adopted across a range of machine learning tasks to leverage complementary capabilities. These methods typically employ sparsely-gated mechanisms that activate only a subset of expert modules during each forward pass, improving computational efficiency and scalability. Building on this motivation, recent advances have explored leveraging a mixture of *expert models* (or *agents*) (Li et al., 2024b; Wang et al., 2024; Chen et al., 2025; Li et al., 2025; Cao et al., 2025) that combine independently pre-trained models across diverse knowledge domains, moving beyond the traditional layer-wise sparse activation used in MoE-based approaches. Rather than operating as interchangeable sub-modules within a



Task Context

You have just watched a video and now need to help to answer the multiple-choice question.[question]. Below is a catalog of skill/modalities you might need. Indicate which are required to answer the question.
1. Only select skill/modality IDs necessary to answer the provided question.
2. Respond strictly with the selected skill IDs, separated by commas.

Expert Aggregator

Select the best answer to the following multiple-choice question based on all the provided information.

In this video, a teacher is explaining a formula.

The formula in the video is: $E=mc^2$. Calculate E given m=2 kg and $c=3\times10^8$m/s

Given the values: \[m = 2\, \text{kg}, \quad c = 3 \times 10^8\, \text{m/s}\], ...

Figure 1: Overview of the MEXA Architecture. Given the input task context and question, MEXA first employs an MLLM router (Sec. 3.2.2) to select the appropriate experts based on input modality and required reasoning skills. The aggregator (Sec. 3.2.3) then reasons over the outputs from the selected experts to generate the final answer.

single model, such agent-based (or model-based) mixtures selectively aggregate the outputs of independently trained models through dynamic routing mechanisms to solve tasks. This design allows for more flexible and targeted knowledge utilization, promoting improved generalization across tasks and greater extensibility. However, existing approaches primarily focus on single-modality settings (Chen et al., 2025) or lack support for complex reasoning tasks (Li et al., 2024b; Wang et al., 2024; Li et al., 2025; Cao et al., 2025). In contrast, our method extends this line of work by scaling expert modularity to handle heterogeneous multimodal inputs and reasoning, enabling more versatile and adaptive problem-solving across domains.

Many-Modal Understanding and Reasoning. Real-world environments are increasingly dynamic, requiring the need for AI systems to perceive and process a broader range of modalities beyond unimodal learning. This growing demand has led to advances in models that integrate diverse signals such as images, audio, text, and 3D point clouds to improve learning and generalization. Among these, Vision-Language Models (VLMs) (Huang et al., 2023; Li et al., 2023a; Gong et al., 2023; Chen et al., 2023a) support tasks like speech recognition and audio captioning by fusing acoustic and textual information. Similarly, 2D-3D Joint Models (Li et al., 2020; Hou et al., 2021, 2023; Lei et al., 2024) combine spatial and geometric features for enhanced 3D scene understanding. However, these models are typically limited to fixed modality-task combinations and struggle to generalize or scale to novel modality inputs. This motivates the development of flexible many-modality systems (Zellers et al., 2022; Han et al., 2023; Li et al., 2023b; Girdhar et al., 2023; Liu et al., 2023b; Yu et al., 2024) that adapt to heterogeneous inputs and reasoning demands across domains. Yet, most existing systems are primarily designed to fuse multiple input modalities and learn perception and reasoning implicitly within a unified architecture. While effective for many tasks, these approaches often struggle with scalability and tend to overlook the explicit use of LLMs' advanced reasoning capabilities. In contrast, our modular framework selects and coordinates expert models through an LLM-driven router and aggregator, directly leveraging reasoning abilities that are insufficiently exploited by unified multimodal architectures.

3 MEXA: Dynamic Multi-Modal Expert Aggregation

We first highlight the challenges faced by general-purpose multimodal reasoning systems in Sec. 3.1 and introduce our skill-specialized expert aggregation framework as a solution in Sec. 3.2. Within Sec. 3.2, we first describe the organization of a pool of expert modules and our strategy for extracting expertise-specific text representations from the specialized experts(Sec. 3.2.1). Next, we introduce an expert selection module that dynamically determines which experts to activate based on the input query, target modality, required skill, and reasoning complexity (Sec. 3.2.2). Finally, we present our expert aggregation strategy, which integrates the outputs from the selected expert to generate a final answer (Sec. 3.2.3).

3.1 General-Purpose Multimodal Reasoning

Existing general-purpose multimodal reasoning systems (Han et al., 2023; Liu et al., 2023b; Yu et al., 2024) typically rely on fixed fusion architectures that process raw multimodal inputs within a shared representation space. These systems aim to implicitly learn both multimodal alignment and reasoning capabilities within a single end-to-end model. Formally, such a system can be framed as a question-answering model that takes a natural language question Q_{ques} and a collection of supported modalities, such as 2D images, 3D point clouds, videos, text, and audio, denoted as $\mathcal{M} = \{m_1, m_2, \ldots, m_n\}$, where n is the number of input modalities as input. The input modalities are first encoded and fused into a joint representation space, which is then processed by a monolithic reasoning model trained to reason over this fused representation. The model produces a contextually grounded answer A, defined as:

$$A = f_{\text{fuse}}(f_{\text{enc}}(\mathcal{M}), Q_{\text{ques}}), \tag{1}$$

where$f_{\text{enc}}(\cdot)$ denotes the encoding and fusion of the multimodal inputs $\mathcal{M}$, and $f_{\text{fuse}}(\cdot)$ represents the general-purpose reasoning function applied to the fused multimodal representation to generate the answer. While this approach offers architectural simplicity, it often struggles with scalability, interpretability, and adaptability, especially when dealing with diverse input modalities and varying reasoning demands. This highlights the need for a novel multimodal reasoning framework capable of seamlessly processing inputs across modalities and dynamically activating the appropriate skills for different reasoning tasks.

3.2 Skill-Specialized Mixture of Expert Models

To overcome the limitations of existing general-purpose multimodal models in handling diverse modalities and tasks, we propose MEXA, a flexible and expert-driven multimodal reasoning framework. Instead of processing all modality inputs uniformly through fixed architectures, MEXA dynamically routes queries to a set of specialized expert models, selected based on input modality and task complexity. Formally, our system takes a question Q_{ques} as input and generates an answer A in response. To accomplish this, we first employ an expert selection module that serves as a router (denoted as `Router`), which selects a subset of expert models, denoted as E_s. The outputs from the selected experts are then passed to a reasoning module, functioning as an aggregator (denoted as `Aggregator`), which integrates and reasons over the information extracted by expert models to generate the final answer A:

$$\begin{aligned} A = \texttt{Aggregator}(\{E_s \mid E_i \in \texttt{Router}(\\ Q_{\text{ques}}, \cdots)\}). \end{aligned} \tag{2}$$

In contrast to Equation 1, which implicitly fuses all modalities within a single architecture, our method enables specialization and modular design. It supports dynamic adaptation by selecting relevant experts based on the task, where each expert plays a well-defined role.

3.2.1 Design Principles for the Expert Pool

Our expert pool is designed based on the following key principles.

Design Principle 1: Task-Aware and Modality-Sensitive Reasoning. To ensure that our expert pool is both modular and broadly applicable, we construct it by analyzing the modalities and skills commonly required across diverse multimodal tasks. While our approach is designed to be task-agnostic and extensible, we use the *Video-MMMU* dataset (Hu et al., 2025a) as a representative case study to guide the initial design and evaluation. This benchmark includes videos from a wide range of domains such as medicine, mathematics, and the arts, allowing us to capture a diverse set of modalities and task combinations. For each domain, we systematically identify the core perception and reasoning skills required to answer the associated questions, ensuring that our expert pool is both modular and generalizable across tasks. This design allows our framework to scale beyond any single dataset and adapt to new domains with minimal effort.

In Fig. 1, we illustrate the complete set of expert modules assembled in our framework. Specifically, we categorize these experts into four distinct types: (1) **Perceptual Experts**, specialized in extracting visual information directly from images, videos, audio, 3D point clouds and medical images; (2) **Text Extraction Experts**, which leverage optical character recognition (OCR) to distill textual content from visuals such as slides or embedded text; (3) **Structured Experts**, designed to analyze structured visual data such as charts, tables, and diagrams; and (4) **Mathematical Reasoning Expert**, focused on interpreting and solving questions related to mathematical and geometric equations presented in LaTeX format. Furthermore, for each expert category, we design skill-specialized experts that target distinct reasoning demands across modalities and task types. We focus on aligning each expert's functionality with the specific requirements of the task. For instance, for 2D image perception, we design experts aligned with skills corresponding to different levels of reasoning granularity: a fine-grained image description expert, responsible for generating comprehensive captions covering multiple visual elements and their relationships, and a concise summarization expert, which produces brief, high-level summaries highlighting only the salient visual content. Within the domain of 3D scene understanding, we develop two experts with specialized skills for different task demands: a general 3D scene expert, which provides descriptions capturing the overall spatial layout and major objects in the scene, and a situated 3D expert, designed to generate detailed, viewpoint-grounded descriptions reflecting the specific perspective of an embodied agent positioned within the environment.

Design Principle 2: Unified Textual Representation. We design expert models that convert diverse modality-specific inputs into a shared textual representation, facilitating the integration of heterogeneous multimodal data. To achieve this, we leverage a series of state-of-the-art captioning models combined with modality- and skill-specific prompting strategies. Each expert in our framework functions as a captioner (descriptor), converting modality- and skill-specific information into a natural language format. This textual abstraction further enables interpretable perception and supports reasoning at varying levels of complexity. Formally, given a set of supported input modalities $\mathcal{M}$, the objective is to obtain a unified textual representation x by converting the relevant modality-specific input into natural language. This is accomplished through a set of modality-specific expert functions E_i, and textual output from each expert is denoted as:

$$x_i = E_i(m_i, p_i), \tag{3}$$

where p_i is a carefully designed prompt that reflects the reasoning objective for different modalities and reasoning demands.

3.2.2 Expert Selection Module

After constructing the pool of expert models, the next step is to determine which experts to activate for a given task. We design an expert selection module that leverages a multimodal LLM (MLLM) as a router for different experts. This expert selection process leverages the commonsense reasoning capabilities of the MLLM, allowing it to deeply understand the semantics of the task, the relationships among modalities, and the contextual intent behind the question when deciding which experts to activate. As illustrated in Fig. 1, the router takes task context and question as input. Task context provides high-level guidance on the task type, and helps the router infer the set of skills required to answer the question and explicitly constrains the router by ensuring that only relevant experts are considered. Based on these inputs, the expert selection module activates a subset of expert models most suited to addressing the question.

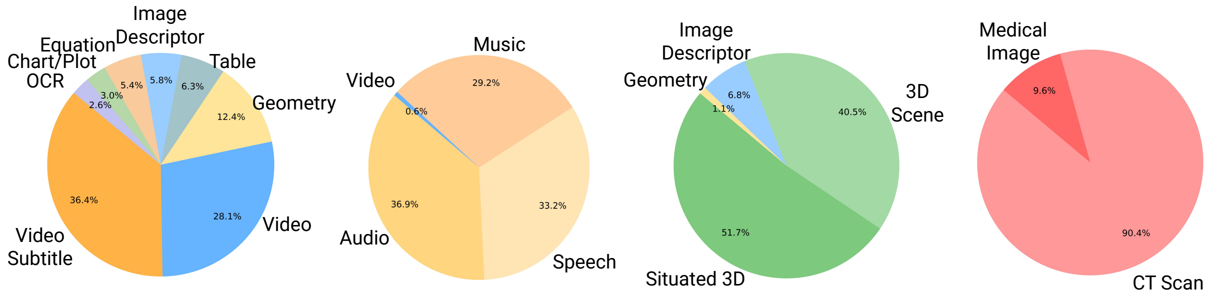
Formally, given question Q_{ques}, a task description T_r, also a set of available modality inputs $\mathcal{M}$, the MLLM router selects a subset of expert models based on needed modality and tasks. We denote the selected experts models as $\{E_s\}_{s\in\mathcal{S}}$, where $\mathcal{S} \subseteq \mathcal{M}$. Once the experts are selected, each expert $\{E_s\}$ receives the specific prompt p_s and transforms the input into a natural language description x_s. The resulting set of expert-generated textual output is represented as :

$$\begin{aligned}\mathcal{T} = \{x_s = & E_s(m_s, p_s) \mid E_s \in \texttt{Router}(\\ & Q_{\text{ques}}, T_r, \mathcal{M}), s \in \mathcal{S} \subseteq \mathcal{M}\}\end{aligned} \tag{4}$$

where $\mathcal{T}$ represents the set of intermediate textual outputs generated by the selected experts, serving as a unified and interpretable representation that is later aggregated to produce the final answer.

3.2.3 Aggregating Expert Information via Reasoning

Once the selected experts have generated modality- and task-specific textual representations, we employ a reasoning module based on LRM, selected



(1) Video MMMU (2) MMAU (3) SQA3D (4) M3D

Figure 2: Expert distributions selected by MEXA across different benchmarks, covering video (Video-MMMU), audio (MMAU), 3D (SQA3D), and medical imaging (M3D).

for its strong long-context reasoning and deep inference capabilities. This LRM-based aggregator systematically integrates the diverse outputs from the selected experts and reasons over them to generate the final answer. Formally, the expert outputs are represented as $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$, where k denotes the number of selected expert. The aggregator inputs the entire set $\mathcal{T}$, along with the task description for the expert, which is denoted as T_a, and generates the final answer, denoted as:

$$A = \texttt{Aggregator}(\mathcal{T}; T_a) \tag{5}$$

By aggregating expert-generated textual outputs and performing reasoning over these unified representations, MEXA enables more accurate, interpretable, and task-aligned answers, demonstrating clear advantages over single end-to-end models that process implicit representations without explicit decomposition or modular coordination.

4 Experimental Results

4.1 Experimental Setup

We use GPT-4o (Hurst et al., 2024) as our multimodal expert selection module to dynamically select the most relevant experts based on the input modalities and task requirements. Subsequently, we employ Deepseek as our aggregator to reason over the expert-generated textual information to obtain the final answer.

To effectively extract textual information from diverse expert modules, we adopt the widely used captioners to generate high-quality textual descriptions across modalities. Specifically, for perceptual experts, we employ different captioning approaches to each modality. For 2D image inputs, we utilize OmniCaptioner-Qwen2-5-7B (Lu et al., 2025) to generate both detailed and concise descriptions. For video modality experts, we adopt NVILA-8B (Liu et al., 2024c) to generate comprehensive video-level captions. For experts specializing in 3D scene understanding, we integrate LEO-Vicuna7B (Huang et al., 2024) to obtain general, top-down scene descriptions and Spartun3D-Vicuna7B (Zhang et al., 2024d) for generating situated captions explicitly grounded in the agent's viewpoint. Regarding audio modality experts, we employ Qwen-2.5-Omni (Xu et al., 2025), which is adept at generating contextualized captions for both speech and music. For all non-perceptual experts, we apply OmniCaptioner-qwen-5-7B (Lu et al., 2025) with prompts designed to emphasize the specific functionality of each expert. We provide detailed prompts for the selection module, aggregator, and all expert models in the Appendix.

4.2 Evaluation Datasets

We validate our framework across various challenging multimodal tasks, including Video Reasoning (Video-MMMU (Hu et al., 2025b)), Audio QA (MMAU (Sakshi et al., 2024)), 3D Situated Reasoning (SQA3D (Ma et al., 2023b)), and Medical QA (M3D (Bai et al., 2024)). We specifically chose these benchmarks because they represent diverse reasoning complexities, modality interactions, and practical application scenarios. Please see more details in the Appendix.

4.3 Quantitative Results

We evaluate MEXA on all datasets under the multiple-choice QA setting, and report performance based on standard accuracy metrics across all experiments.

Video-base Multimodal Reasoning. For the video-based multimodal reasoning capability, we evaluate MXEA on the Video-MMMU benchmark. We compare our methods with three types of baselines: open-source large multimodal models (LMMs) (Li et al., 2024a,c), proprietary LMMs

Method	Overall	Results by Track			Results by Discipline					
		Perception	Comprehension	Adaptation	Art.	Biz.	Sci.	Med.	Hum.	Eng.
Random Choice	14.0	12.0	14.0	16.0	11.1	12.9	12.1	22.5	10.5	13.6
Open-source LMMs										
LLaVA-OneVision-72B (Li et al., 2024a)	48.3	59.7	42.3	43.0	61.9	46.21	40.2	54.3	60.0	44.0
LLaVA-Video-72B (Zhang et al., 2024a)	49.7	59.7	46.0	43.3	69.8	44.7	41.7	58.9	57.1	45.1
Aria (Li et al., 2024c) (8 × 3.5B)	50.8	65.7	46.7	40.0	71.4	47.7	44.7	58.9	62.9	43.7
Proprietary LMMs										
Gemini 1.5 Flash (Team et al., 2024)	49.8	57.3	49.0	43.0	63.5	53.0	43.2	49.6	59.1	45.7
Gemini 1.5 Pro (Team et al., 2024)	53.9	59.0	53.3	49.3	57.1	59.1	49.1	57.4	58.1	50.3
GPT-4o (Hurst et al., 2024)	61.2	66.0	62.0	55.7	69.5	66.9	51.6	64.8	69.5	57.1
Claude-3.5-Sonnet (Anthropic, 2024)	65.8	72.0	69.7	55.7	66.7	75.0	56.1	58.1	75.2	66.1
Human Expert	74.4	84.3	78.7	60.3	81.0	78.8	74.2	70.5	84.8	69.9
MEXA (Ours)	**71.5**	**77.0**	**76.7**	**60.0**	**76.2**	**77.0**	**63.8**	**75.8**	**78.1**	**67.7**

Table 1: Video-MMMU Evaluation Results across three cognitive tracks (Perception, Comprehension, Adaptation) and six disciplines (Art, Business, Science, Medicine, Humanities, Engineering). We highlight the best model performance for each metric.

(Hurst et al., 2024; Anthropic, 2024; Team et al., 2024), and human experts (Hu et al., 2025b). Results show that MEXA significantly outperforms the leading open-source LLM (Li et al., 2024c) by 23.6% in overall accuracy. Moreover, our method outperforms the powerful MLLM like GPT-4o (Hurst et al., 2024) by 6% in overall accuracy. Our method also achieves performance on par with the human expert on most metrics, showing the effectiveness of the proposed framework. Specifically, we see strong results across all six disciplines that validate the effectiveness of the combination of expert modules. For comparison, GPT-4o (Hurst et al., 2024) achieves only 51.6% and 57.1% on science and engineering, performing 12.2% and 10.5% lower than MEXA, respectively. This gap highlights the advantage of explicitly aggregating and reasoning over information from multiple specialized experts, rather than relying solely on a single multimodal LLM.

Audio-based Multimodal Reasoning. Tab. 2 presents the results of the MMAU benchmark, evaluating models on three audio QA types: *Sound*, *Music*, and *Speech*. Our proposed MEXA consistently outperforms all audio large language models, including GAMA, MuLLaMA, and SALAMONN, across all categories. Notably, MEXA achieves a substantial improvement of 4.1%, 5.9%, and 26.3% over the best-performing Audio LLM baselines, demonstrating its strong and generalizable ability to understand diverse non-visual sensory inputs, such as audio. In contrast, existing models struggle particularly with complex music and speech QA tasks. Compared to GPT-4o with generated captions using (Kim et al., 2024; Liu et al., 2024a; Radford et al., 2023), MEXA achieves comparable overall performance while maintaining a more balanced accuracy across all audio types. These results validate the effectiveness of our modular expert-based approach in addressing the diverse challenges of audio understanding.

3D Scene Understanding. Results on the SQA3D benchmark shown in Tab. 3 demonstrate the effectiveness of our method in 3D situated reasoning tasks. By integrating situated captions and general scene descriptions through our routing and aggregation system, MEXA consistently improves accuracy across all question types. Specifically, our approach achieves a 2% accuracy gain over SOTA 3D-based LLMs, which rely on general-purpose encoder-fusion strategies with static unified architectures. This improvement highlights the advantage of our dynamic, modality-aware expert selection and aggregation framework in the 3D domain.

Medical QA. To demonstrate the effectiveness of our approach beyond general domains such as audio and video, we additionally evaluate MEXA on the medical video QA domain using the M3D benchmark. This benchmark includes five distinct medical QA types: Plane classification, Phase recognition, Organ identification, Abnormality detection, and Location estimation. As shown in Tab. 4, MEXA significantly outperforms strong general-purpose MLLM, CREMA, MiniCPM-o,

Method	Sound	Music	Speech	Average
GPT-4o (Hurst et al., 2024) w/ caption	39.3	39.5	58.3	45.7
GAMA-7B (Ghosh et al., 2024)	41.4	32.3	18.9	30.9
MULLaMA-7B (Liu et al., 2024a)	40.8	32.6	22.2	31.9
SALAMONN-13B (Tang et al., 2024)	41.0	34.8	25.5	33.7
MEXA (Ours)	**45.1**	**40.7**	**51.8**	**45.9**

Table 2: MMAU evaluation results across three different audio question types.

Method	What	Is	How	Can	Which	Others	Avg.
CREMA (Yu et al., 2024)	23.2	**52.5**	34.8	41.5	34.1	37.5	37.3
LEO (Huang et al., 2024)	10.2	15.7	11.6	9.4	10.6	16.9	12.4
Spartun3D (Zhang et al., 2024d)	22.3	50.9	33.8	40.9	34.4	34.5	36.1
MEXA (Ours)	**23.4**	51.1	**35.2**	**42.5**	**36.9**	**37.9**	**37.8**

Table 3: SQA3D evaluation results across fine-grained question types.

Method	Plane	Phase	Organ	Abnormality	Location	Avg.
CREMA (Yu et al., 2024)	14.9	26.7	15.9	17.3	13.0	17.2
MiniCPM-o (Yao et al., 2024)	60.7	39.2	35.2	**53.0**	42.0	44.7
GPT-4o	**83.3**	42.7	50.0	41.3	41.3	51.7
MEXA (Ours)	65.0	**48.1**	**60.9**	44.8	**48.0**	**53.3**

Table 4: M3D evaluation results across different kinds of medical QA types.

Router	Aggregator	Video-MMMU	M3D
Qwen2.5-VL (7B)	GPT-4o	57.4	34.7
Qwen2.5-VL (7B)	DeepSeek	70.4	45.8
GPT-4o	GPT-4o	58.9	48.2
GPT-4o	DeepSeek	**71.5**	**53.4**

Table 5: Ablation study on MLLM router and LLM aggregator.

and GPT-4o, which directly process 3D scan images as input, and observe that MEXA consistently achieves superior performance across most categories, improving average accuracy by 36.1%, 8.7% and 1.6%, respectively. These results underscore MEXA's robust capability in medical content understanding, even without extensive domain-specific fine-tuning.

4.4 Ablation Study and Analysis

Ablation on Router and Aggregator. Table 5 presents ablation results for different router and aggregator configurations. We observe that, in addition to the reasoning strength of the model itself, the correct deployment and targeted usage of these models significantly contribute to performance. Specifically, we compare the effectiveness of Qwen2.5-VL and GPT-4o as multimodal expert selection modules, alongside GPT-4o and DeepSeek as aggregators, across two benchmarks: Video-MMMU and M3D. We observe that GPT-4o consistently outperforms Qwen2.5-VL when utilized as the router, indicating its superior capability in multimodal expert selection. Furthermore, DeepSeek demonstrates better performance than GPT-4o in the aggregator role, underscoring its effectiveness in reasoning over the outputs from specialized experts. This superior performance indicates that our aggregator requires strong reasoning abilities, including the effective handling of long-context information and nuanced inference across diverse expert outputs, in which DeepSeek particularly excels.

Expert Distributions. In Fig. 2, we visualize the expert selection distributions from MEXA's expert selection module. For the Video-MMMU benchmark, multiple experts are frequently selected, aligning well with the benchmark's inherent multimodal and multidisciplinary nature, which demands diverse skills for comprehensive video understanding. In the MMAU audio reasoning benchmark, the audio-specific experts (music, audio, and speech) are consistently activated, with their selections evenly distributed, reflecting balanced reliance on all audio modalities. For the 3D reasoning task (SQA3D), the situated 3D and general 3D scene experts are predominantly selected, with occasional inclusion of the image descriptor expert. Finally, for the medical dataset (M3D-VQA), the CT scan expert is primarily selected, complemented occasionally by the medical image expert. These observations confirm that our selection module effectively and adaptively matches experts to the specific modality requirements and reasoning contexts of each task. We provide qualitative examples of Video-MMMU and SQA3D in the Appendix.

5 Conclusion

In this paper, we introduce MEXA, a dynamic multimodal reasoning framework that adaptively leverages a diverse set of expert modules with the skills for specific modalities and tasks. By effectively integrating an expert selection module and aggregation mechanism, our method achieves superior performance across a variety of challenging benchmarks, including Video Reasoning, Audio QA, 3D Situated Reasoning, and Medical QA. Extensive experimental analyses and visualization results demonstrate the clear advantages of our dynamic and modality-aware design, highlighting the significance of selectively combining complementary expert skills. We believe that our framework provides a robust foundation for future research in generalized multimodal reasoning, enabling models to tackle increasingly complex real-world multimodal applications.

Limitations

Our framework achieves competitive reasoning across diverse domains by leveraging modular expert outputs. However, since it relies on frozen pre-trained experts, its reasoning fidelity can be constrained by their inherent capabilities, occasionally leading to incorrect or hallucinated rationales. Furthermore, the quality of the final prediction is affected by the precision and expressiveness of individual experts. As the ecosystem of multimodal and language models evolves, integrating more advanced experts holds promise for improving both accuracy and robustness without altering the overall framework.

6 Acknowledgment

This work was supported by the National Institutes of Health (NIH) under other transactions 1OT2OD038045-01, ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, DARPA ECOLE Program No. HR00112390060, and NSF-AI Engage Institute DRL211263. The views, opinions, and/or findings contained in this article are those of the authors and not of the funding agency.

References

Anthropic. 2024. Claude 3.5 sonnet.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Muhammad Adeel Azam, Khan Bahadar Khan, Sana Salahuddin, Eid Rehman, Sajid Ali Khan, Muhammad Attique Khan, Seifedine Kadry, and Amir H Gandomi. 2022. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in biology and medicine*, 144:105253.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Linqin Cai, Haodu Fang, and Zhiqing Li. 2023. Pre-trained multilevel fuse network based on vision-conditioned reasoning and bilinear attentions for medical image visual question answering. *The Journal of Supercomputing*, 79(12):13696–13723.

Juntai Cao, Xiang Zhang, Raymond Li, Chuyuan Li, Shafiq Joty, and Giuseppe Carenini. 2025. Multi2: Multi-agent test-time scalable framework for multi-document processing. *arXiv preprint arXiv:2502.20592*.

Justin Chih-Yao Chen, Sukwon Yun, Elias Stengel-Eskin, Tianlong Chen, and Mohit Bansal. 2025. Symbolic mixture-of-experts: Adaptive skill-based routing for heterogeneous reasoning. *arXiv preprint arXiv:2503.05641*.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023a. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. 2023b. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837.

Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. 2023. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks. In *International Conference on Machine Learning*, pages 6074–6114. PMLR.

Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Stablemoe: Stable routing strategy for mixture of experts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7085–7095.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, and 1 others. 2023. Palm-e: An embodied multimodal language model.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, and 1 others. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *2305.04790*.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023. Onellm: One framework to align all modalities with language. *arXiv preprint arXiv:2312.03700*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494.

Ji Hou, Xiaoliang Dai, Zijian He, Angela Dai, and Matthias Nießner. 2023. Mask3d: Pre-training 2d vision transformers by learning masked 3d priors. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. 2021. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025a. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025b. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *Preprint*, arXiv:2501.13826.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2024. An embodied generalist agent in 3d world. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20413–20451.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

OpenAI: Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. 2024. Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. 2024. Becotta: nput-dependent online blending of experts for continual test-time adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Weixian Lei, Yixiao Ge, Jianfeng Zhang, Dylan Sun, Kun Yi, Ying Shan, and Mike Zheng Shou. 2024. Vit-lens: Towards omni-modal representations. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 preprint arXiv:2305.03726*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Satvik Chaudhary, Lijie Hu, and Jiayi Shen. 2024b. Smoa: Improving multi-agent large language models with sparse mixture-of-agents. *arXiv preprint arXiv:2411.03284*.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, and 1 others. 2024c. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. 2025. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*.

Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. 2020. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024a. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. 2023b. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024b. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, and 1 others. 2024c. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

Yiting Lu, Jiakang Yuan, Zhen Li, Shitian Zhao, Qi Qin, Xinyue Li, Le Zhuo, Licheng Wen, Dongyang Liu, Yuewen Cao, Xiangchao Yan, Xin Li, Tianshuo Peng, Shufei Zhang, Botian Shi, Tao Chen, Zhibo Chen, Lei Bai, Bo Zhang, and Peng Gao. 2025. Omnicaptioner: One captioner to rule them all. *Preprint*, arXiv:2504.07089.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023a. Sqa3d: Situated question answering in 3d scenes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023b. Sqa3d: Situated question answering in 3d scenes. *Preprint*, arXiv:2210.07474.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, and 1 others. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *Preprint*, arXiv:2410.19168.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI National Conference on Artificial Intelligence (AAAI)*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *arXiv preprint arXiv:2410.12705*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. 2024. Crema: Generalizable and efficient video-language reasoning via multimodal modular fusion. *arXiv preprint arXiv:2402.05889*.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yihua Zhang, Ruisi Cai, Tianlong Chen, Guanhua Zhang, Huan Zhang, Pin-Yu Chen, Shiyu Chang, Zhangyang Wang, and Sijia Liu. 2023. Robust mixture-of-expert training for convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 90–101.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024a. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. 2024b. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pages 399–415. Springer.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. 2024c. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*.

Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. 2024d. Spartun3d: Situated spatial understanding of 3d world in large language models. *arXiv preprint arXiv:2410.03878*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and 1 others. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Appendix

A.1 Prompts for Router and Aggregator

Fig. 6 and Fig. 5 illustrate the prompt designs used in our framework, respectively. Fig. 6 presents the expert selection prompt, which identifies skill-specialized experts for the input task context and question. Fig. 5 shows the aggregator prompt, which integrates the outputs from the selected experts and guides LRM to reason over them and generate the final answer.

To generate captions from different experts that explicitly focus on specialized skills, we design and apply skill-specific prompts for each expert. For instance, to obtain detailed image descriptions, we prompt Omnicaptioner-QWen2.5-7B with *“You are a helpful assistant focused on providing detailed descriptions and background information for images. Analyze the given image and generate a comprehensive caption that includes the visual style, spatial relationships between elements, texture details, descriptions of the main objects, and relevant world knowledge to enhance understanding."* whereas, for concise image summaries, we use the prompt *“You are a helpful assistant focused on creating short captions for images. Analyze the provided image and generate a concise caption that highlights the main subject."* Similarly, for other experts, we provide clear and direct prompts explicitly highlighting the required skill.

A.2 Evaluation Datasets

We validate our framework across various challenging multimodal tasks, including Video Reasoning (Video-MMMU (Hu et al., 2025b)), Audio QA (MMAU (Sakshi et al., 2024)), 3D Situated Reasoning (SQA3D (Ma et al., 2023b)), and Medical QA (M3D (Bai et al., 2024)). We specifically choose these benchmarks because they represent diverse reasoning complexities, modality interactions, and practical application scenarios.

Video-MMMU (Hu et al., 2025b) provides a robust evaluation of models' abilities to integrate multimodal educational content and reason across diverse knowledge domains from educational videos. In our evaluation, we test on the full 900 video reasoning questions.

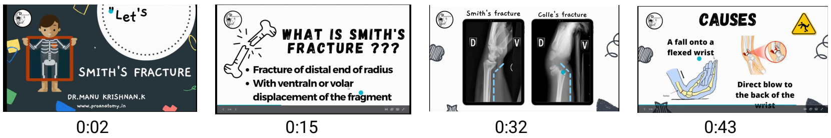
MMAU (Sakshi et al., 2024) is a benchmark designed to evaluate multimodal audio understanding with curated audio clips paired with human-annotated natural language questions spanning speech, environmental sounds, and music. Evaluation is conducted on a 1K validation set with available ground truth annotations.

SQA3D (Ma et al., 2023b) is a benchmark specifically designed for evaluating situated 3D scene understanding. It presents situations such as *"You are standing beside a table"*, requiring models to reason about 3D spatial relationships between the described viewpoint and surrounding objects. In our evaluation, we use the SQA3D test set, which includes around 3K human-annotated question-answer pairs.

M3D-VQA (Bai et al., 2024) is a benchmark designed for expert-level reasoning over medical data, specifically focusing on 3D CT scans. It consists of natural-language question-answer pairs covering five diagnostic dimensions: plane, phase, organ, abnormality, and location. For evaluation, we sample 500 question-answer pairs under the closed-ended QA setting.

A.3 Qualitative Examples

Fig 3 and Fig 4 show two qualitative examples from Video-MMMU and SQA3D, respectively. In the Video-MMMU example, our framework effectively selects the most relevant experts, including the video expert and the medical image expert. The aggregator then filters and prioritizes key information extracted by the medical image expert, allowing the reasoning module to accurately fill in the missing information required to answer the question. In the SQA3D example, both the general 3D scene expert and the situated 3D scene expert are activated. Their outputs are jointly considered by the aggregator to produce a coherent answer.


"Let's
SMITH'S FRACTURE
DR.MANU KRISHNAN.K
WHAT IS SMITH'S FRACTURE ???
Fracture of distal end of radius
With ventraln or volar displacement of the fragment
Smith's fracture
Colle's fracture
D
V
D
V
CAUSES
A fall onto a flexed wrist
Direct blow to the back of the wrist
0:02
0:15
0:32
0:43

```
"Question": "In 0:32, the left is ____, the right is ____.
"Answer Choices":
A: Smith's fracture, Normal X-Ray, Colle's fracture, Smith's fracture
B: Normal X-Ray, Colle's fracture, Smith's fracture, Smith's fracture
C: Normal X-Ray, Smith's fracture, Smith's fracture, Colle's fracture
D: Smith's fracture, Smith's fracture, Colle's fracture, Normal X-Ray
E: Smith's fracture, Colle's fracture, Smith's fracture, Normal X-Ray
F: Smith's fracture, Normal X-Ray, Smith's fracture, Colle's fracture
G: Colle's fracture, Smith's fracture, Normal X-Ray, Smith's fracture
H: Colle's fracture, Normal X-Ray, Smith's fracture, Smith's fracture
I: Colle's fracture, Smith's fracture, Smith's fracture, Normal X-Ray
J: Smith's fracture, Colle's fracture, Normal X-Ray, Smith's fracture
```

Expert Selection

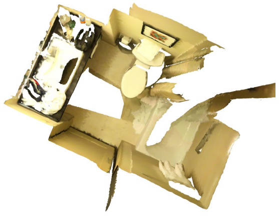
Video Expert: This educational video, presented by Dr. Manu Krishnan from ProAnatomy, introduces and explains Smith's fracture, a specific type of wrist injury. The video begins with a friendly visual of a child holding an X-ray frame over their torso, introducing the topic in an engaging and accessible way. The video defines Smith's fracture as a fracture of the distal end of the radius, accompanied by ventral or volar displacement of the broken fragment.

Medical Image Expert: Image in 0:32 frame presents lateral X-ray views of the wrist comparing two types of distal radius fractures: Smith's fracture and Colles' fracture. In the left panel, the X-ray shows a fracture at the distal end of the radius, with the distal fragment displaced volarly (anteriorly). In the right panel, the Colles' fracture also involves the distal radius, but with dorsal (posterior) displacement of the distal fragment.

Aggregator

Looking at the description provided by the Medical Image Expert: the left panel shows volar (anterior) displacement, and the right panel shows dorsal (posterior) displacement. So the left should be Smith's because of the volar displacement, and the right is Colles' because of the dorsal displacement. In 0:32, the left is Smith's fracture, the right is Colles' fracture. **The answer is C.**

Figure 3: A qualitative example of Video-MMMU.



```
"Situation": Standing in front of the sink
and facing the mirror.
"Question: How many picture in the bathroom?
```

Expert Selection

3D Scene Expert: The scene depicts a bathroom containing a sink and a toilet, a picture hanging on the wall, and a white bathtub.

Situated 3D Scene Expert: In front of me, there is a white sink and a mirror. To the right, there is a white toilet and a picture hanging on the wall. Behind me, there is a white bathtub with a curtain, and to the left, there is a white door.

Aggregator

Both experts mention one picture in their descriptions. The first 3D Scene Expert states: "a picture hanging on the wall." The Situated 3D Scene Expert specifies its location: "to the right, there is a white toilet and a picture hanging on the wall." No conflicting or additional pictures are mentioned. **The answer is one.**

Figure 4: A qualitative example of SQA3D.

Prompts for Aggregator

You are an answerer for a video question answering, audio question answering, 3D situated question answering, or medical visual question answering. Below is information provided by multiple expert modules relevant to solving the question:

- **Expert 1:** {Text Descriptions from Expert 1}
- **Expert 2:** {Text Descriptions from Expert 2}
- **Expert 3:** {Text Descriptions from Expert 3}
- ...

Using the information above, please select the best answer to the question and provide a brief explanation if needed.
Question: `"{}"`

Figure 5: Prompts for Aggregator.

Prompts for Expert Selection

You are an expert multimodal reasoning assistant. Given a multimodal question (e.g., related to video, audio, 3D scenes, medical images, etc.), your task is to select all relevant skills and modalities required to accurately answer the question.
Task Type: `"{}"`
Question: `"{}"`
Options: `{}`
Available Skills and Modalities:
General Visual Perception

- A1. Detailed Image Description
- A2. Medium Image Description
- A3. Short Image Description

Audio Perception

- B1. Video Subtitle Extraction
- B2. Audio Description
- B3. Music Description

3D Visual Understanding

- C1. 3D Scene Description
- C2. 3D Situated Context Description

Medical Visual Understanding

- D1. CT Scan Interpretation
- D2. Medical Image Description

OCR/Text Extraction

- E1. General OCR
- E2. Poster/Slides Caption
- E3. PDF Text Extraction

Structured Visual Data Interpretation

- F1. Chart/Plot Description
- F2. Table Description

Mathematics and Geometry Extraction

- G1. Equation (LaTeX format)
- G2. Mathematics & Geometry (LaTeX format)

Instructions:

1. Only select skill/modality IDs necessary to answer the provided question.
2. Respond strictly with the selected skill IDs, separated by commas.

Selected IDs:

Figure 6: Prompts for expert selection.