Progressive Multimodal Chain-of-Thought Tuning for Vision-Indispensable Reasoning

Anonymous ACL submission

Abstract

Recent advancements in multimodal large language models (MLLMs) have showcased their impressive capabilities in multimodal understanding and generation. Nevertheless, current open-source MLLMs still encounter challenges in complex reasoning and problem solving, especially in vision-indispensable scenarios. In this paper, we present **VILAMR**, an MLLM tailored for vision-indispensable reasoning. To endow VILAMR with powerful reasoning capabilities, we initially construct a multimodal instruction-following dataset, **MCoT-Instruct**, featuring 266K high-quality chain-of-thought responses. Subsequently, we equip VILAMR with a novel connector to selectively integrate different visual features and facilitate alignment between correlated vision and language content. Finally, we fine-tune VILAMR on MCoT-Instruct with a meticulously designed *reasoning progressive-enhancement tuning* scheme, encouraging VILAMR to follow the cognitive process of "*understanding before reasoning*". Experiments on multiple multimodal benchmarks and datasets demonstrate the effectiveness of VILAMR and the contribution of MCoT-Instruct in bolstering MLLM reasoning capabilities.

1 Introduction

Multimodal large language models (MLLMs) (Liu et al., 2024b; Lin et al., 2023; Chen et al., 2023b; Bai et al., 2023b) have recently garnered considerable attention for their powerful capabilities in multimodal understanding and generation. Building on the foundation of open-source large language models (LLMs) such as QWen (Bai et al., 2023a) and Llama (Touvron et al., 2023), MLLMs incorporate visual modality into LLMs and learn how to perform multimodal tasks through instruction tuning (Liu et al., 2023b), showcasing exceptional abilities in various downstream tasks. Despite these advances, current open-source MLLMs still struggle with complex reasoning and problem solving, especially in vision-indispensable scenarios (Chen et al., 2024b). In such scenarios, MLLMs are required to accurately capture the task-related visual content from given images and then elicit the chain-of-thought (CoT) (Wei et al., 2022) reasoning capabilities of LLMs to derive a final response conditioned on the obtained content (*cf*. Figure 1).

In light of this task paradigm, the suboptimal performance of MLLMs in vision-indispensable reasoning can be primarily attributed to two issues: (*i*) *Limited CoT reasoning capability*. While fine-tuning MLLMs on multimodal CoT instruction-following datasets is a feasible approach to empower them with CoT reasoning capabilities, there remains a notable scarcity of high-quality multimodal CoT instruction data in the open-source community. In addition, prevailing instruction tuning approaches (Liu et al., 2023b; Zhang et al., 2023c; Zhao et al., 2023a; Liu et al., 2024a) tend to advocate a uniform development of understanding and reasoning skills in MLLMs, overlooking the hierarchical nature of cognition where reasoning is contingent upon understanding, thereby limiting further improvement of reasoning skills. (*ii*) *Misalignment of correlated vision and language content*. Although existing state-of-the-art MLLMs like LLaVA-NeXT (Liu et al., 2024b) have been proficient in acquiring fine-grained visual content (*e.g.*, Figure 1 (a), (c), and (d)), they frequently fail to align visual details with the corresponding language context, resulting in incorrect visual conditions for reasoning (*e.g.*, Figure 1 (a) and (d)).

In this paper, we strive to improve open-source MLLMs toward vision-indispensable reasoning by tackling the above-identified two issues. To address the first issue, we initially introduce a multimodal instruction-following dataset with high-quality CoT responses (**MCoT-Instruct**), featuring 76K instances for vision-intensive understanding

Figure 1: **Response demonstration of GPT-4V, LLaVA-NeXT-34B, and VILAMR-13B.** Compared to GPT-4V and LLaVA-NeXT, VILAMR consistently observes and understands the given image before reasoning and is more proficient in performing complex multimodal reasoning and problem solving. Blue and red respectively highlight correct and incorrect intermediate reasoning steps or rationales leading to the final response.

and 190K instances for vision-indispensable reasoning. Building on MCoT-Instruct, we then develop a multimodal LLM, dubbed **VILAMR**. To further bolster its reasoning capability, we propose a *reasoning progressive-enhancement tuning* scheme to train VILAMR, encouraging it to follow the cognitive process of "*understanding before reasoning*". To tackle the second issue, we shift our focus to the architecture and pretraining of vision-language connectors. Specifically, we design a novel connector that selectively integrates different features from a hybrid visual encoder via a gate attention mechanism and captures global context using prefix token embeddings. This architectural design ensures a better alignment between visual details and corresponding language context. Furthermore, our connector is pre-trained on a subset of ShareGPT4V (Chen et al., 2023b) containing detailed image captions, allowing for finer alignment between vision and language content. Experiments on six multimodal benchmarks and four datasets demonstrate the effectiveness of VILAMR.

Our contributions are encapsulated as follows: (*i*) We develop VILAMR, a multimodal LLM adept at complex reasoning and problem solving. (*ii*) We propose a reasoning progressive-enhancement tuning scheme to further improve the reasoning capability of VILAMR and a connector to promote the alignment between correlated vision and language content. (*iii*) We introduce a high-quality multimodal CoT instruction-following dataset with 266K instances, aiming to serve as a foundational resource for improving MLLMs toward vision-indispensable reasoning.

2 Related Work

2.1 Multimodal Large Language Models

Bridging Visual Encoder with LLMs. To extend the remarkable capabilities of LLMs to multimodal tasks, MLLMs bridge visual encoders with LLMs via specialized modules (Song et al., 2023), which can be broadly categorized into modality *convertor* and *connector*. The former directly converts visual input into texts using a captioning model (Zhang et al., 2021). In contrast, to align with LLMs, some works (Tsimpoukelli et al., 2021; Driess et al., 2023; Zhang et al., 2023a; Gao et al., 2023b; Liu et al., 2023b; Luo et al., 2023) utilize linear projection layers as *connector* to map visual features into the textual space. Another line of works (Li et al., 2023; Dai et al., 2023; Alayrac et al., 2022; Gong et al., 2023; Ye et al., 2023) introduce cross-attention layers into the *connector* to achieve interaction between different modalities.

Instruction Tuning. After aligning visual encoders with LLMs, the subsequent goal is to enable MLLMs to perform multimodal tasks. The de facto practice is instruction tuning (Liu et al., 2023b), which involves fine-tuning MLLMs on multimodal instruction-following datasets, such as LLaVAR (Zhang et al., 2023c), MiniGPT-4 (Zhu et al., 2023), SVIT (Zhao et al., 2023a), and LRV-Instruction (Liu et al., 2024a). Generally, MLLMs tuned in this manner adeptly handle multimodal tasks by adhering to given instructions and exhibit strong generalization. However, due to their limited reasoning abilities, MLLMs may encounter challenges in complex scenarios.

2.2 Multimodal Chain-of-Thought

Multimodal CoT Reasoning. CoT, referring to a series of *intermediate reasoning steps or rationales* that lead to the final reasoning outcome (Wei et al., 2022), has been extensively utilized to elicit the powerful reasoning capabilities of LLMs (Cheng et al., 2024; Fu et al., 2023; Wang et al., 2023; Diao et al., 2023). Multimodal CoT reasoning aims to leverage CoT prompting (Gao et al., 2024; Mitra et al., 2023; Lu et al., 2023a) or CoT tuning (Wang et al., 2024; Zhang et al., 2023e) to better perform multimodal reasoning tasks, such as decision making (Chen et al., 2023a) and robot planning (Mu et al., 2023). Multimodal CoT prompting is usually employed under zero-shot (Kojima et al., 2022) or few-shot (Zhang et al., 2023d) paradigm to guide large multimodal models like GPT-4V (Achiam et al., 2023) and Gemini (Team et al., 2023) to engage in step-by-step thinking before reaching the final outcomes.

Multimodal CoT Tuning. Multimodal CoT tuning is essentially the instruction tuning of MLLMs using multimodal CoT instruction datasets. Given an image and the corresponding language context, the success of this tuning is closely linked to the quality of free-text CoT responses. The common methods for collecting multimodal CoT data are manual collection (Zellers et al., 2019; Lu et al., 2022; Schwenk et al., 2022) and LLM-assisted generation (Zhao et al., 2023b). Compared to manual collection, the latter can reduce human preference and generate more diverse CoT responses, thereby better instructing MLLMs to elicit CoT reasoning capabilities. However, due to the scarcity of high-quality multimodal CoT instruction data, existing works (Zhang et al., 2023e; Wang et al., 2024) typically fine-tune MLLMs on a limited amount of manually collected data, yielding excellent in-domain performance but poor generalization.

3 MCoT-Instruct

In this paper, we introduce a multimodal instruction dataset, MCoT-Instruct, comprising 266K high-quality CoT responses. As demonstrated in Table 1, MCoT-Instruct distinguishes itself by concentrating on vision-indispensable reasoning and featuring a significant number of complex reasoning instances, setting it apart from other accessible multimodal instruction datasets collected with GPT assistance.

Specifically, MCoT-Instruct is built on existing VQA datasets that provide explanations or rationales for correct answers, such as VCR (Zellers et al., 2019), ScienceQA (Lu et al., 2022), A-OKVQA (Schwenk et al., 2022), GPT-VQA (Zhao et al., 2023b), TabMWP (Lu et al., 2023b), and

Dataset	Image Source	Objective	#TInst.	#RInst.	CoT
LRV-Instruction (Liu et al., 2024a)	VG	Mitigating MM hallucination	400K	-	✗
LLaVAR (Zhang et al., 2023c)	LAION	Text-rich image understanding	20K	-	✗
ShareGPT4V (Chen et al., 2023b)	LAION, COCO, CC, *etc.*	Vision-language alignment	100K	-	✗
LLaVA (Liu et al., 2023b)	COCO	MM comprehension and reasoning	158K	77K	✗
VisCoT (Shao et al., 2024)	Flickr30k, GQA, *etc.*	RoI understanding and reasoning	373K	10K	Box
MCoT-Instruct	COCO, VCR, GeoQA, *etc.*	Vision-indispensable reasoning	266K	190K	Text

Table 1: **Comparison with multimodal instruction data collected with GPT assistance.** #TInst. and #RInst. denote the total number of instances and the number of MViR instances. Image sources are VG (Krishna et al., 2017), LAION (Schuhmann et al., 2021), COCO (Lin et al., 2014), CC (Sharma et al., 2018), Flickr30k (Plummer et al., 2015), GQA (Hudson and Manning, 2019), VCR (Zellers et al., 2019), GeoQA (Chen et al., 2021), *etc.*

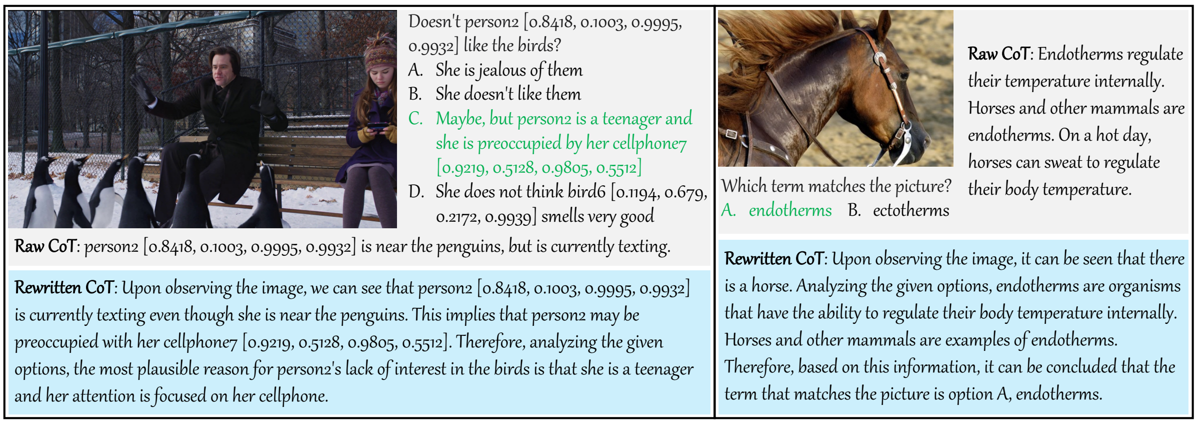


Figure 2: **Comparison of CoT response before and after CoT rewriting.** The rewritten CoT remains faithful to the given context but is more detailed and logically coherent.

GeoQA-T (Gao et al., 2023a). To enhance the quality of CoT responses, we instruct GPT to refine and standardize raw explanations from these datasets through three steps: *CoT Rewriting*, *Quality Verification and Data Filtering*, and *Instance Grouping*. As shown in Figure 2, the improved CoTs still adhere to the provided context but are more detailed and standardized. Ultimately, we obtain 76K instances for vision-intensive understanding (MViU) and 190K instances for vision-indispensable reasoning (MViR). (*cf*. Appendix A for more details.)

4 ViLaMR

To improve open-source MLLMs toward vision-indispensable reasoning, we develop ViLaMR on top of MCoT-Instruct. As illustrated in Figure 3, ViLaMR incorporates a hybrid visual encoder to richly represent image content, a novel vision-language connector to selectively integrate different visual features into LLM-friendly token embeddings, and a LLM to efficiently generate instruction-following CoT responses. The training of ViLaMR includes two consecutive stages: vision-language alignment pretraining and progressive multimodal CoT tuning. To further improve the reasoning capability of ViLaMR, we introduce a reasoning progressive-enhancement tuning scheme in the second training stage, prompting ViLaMR to follow the paradigm of "*understanding before reasoning*".

4.1 Model Architecture

Hybrid Visual Encoder. The premise for models to perform vision-indispensable reasoning is that they can comprehensively understand the input image. To enhance the representation of image contents, ViLaMR considers mixing visual features from different sources and thus combines the pretrained CLIP with ViT (Radford et al., 2021) and ConvNeXt (Woo et al., 2023) as a hybrid visual encoder to extract detailed image appearance features. Specifically, given a 336 $\times$ 336 image, the ViT encoder captures long-range interactions and outputs the features $\boldsymbol{v}_v \in \mathbb{R}^{576\times1024}$ with rich semantic details. Simultaneously, the ConvNeXt encoder with a 384 $\times$ 384 image input encodes neighboring dependencies and outputs the features $\boldsymbol{v}_c \in \mathbb{R}^{576\times1536}$ with rich spatial details.

Vision-Language Connector. To better bridge the hybrid vision encoder with the LLM, we propose **GateMLP**. As shown in Figure 3, the proposed connector initially employs two distinct linear lay-

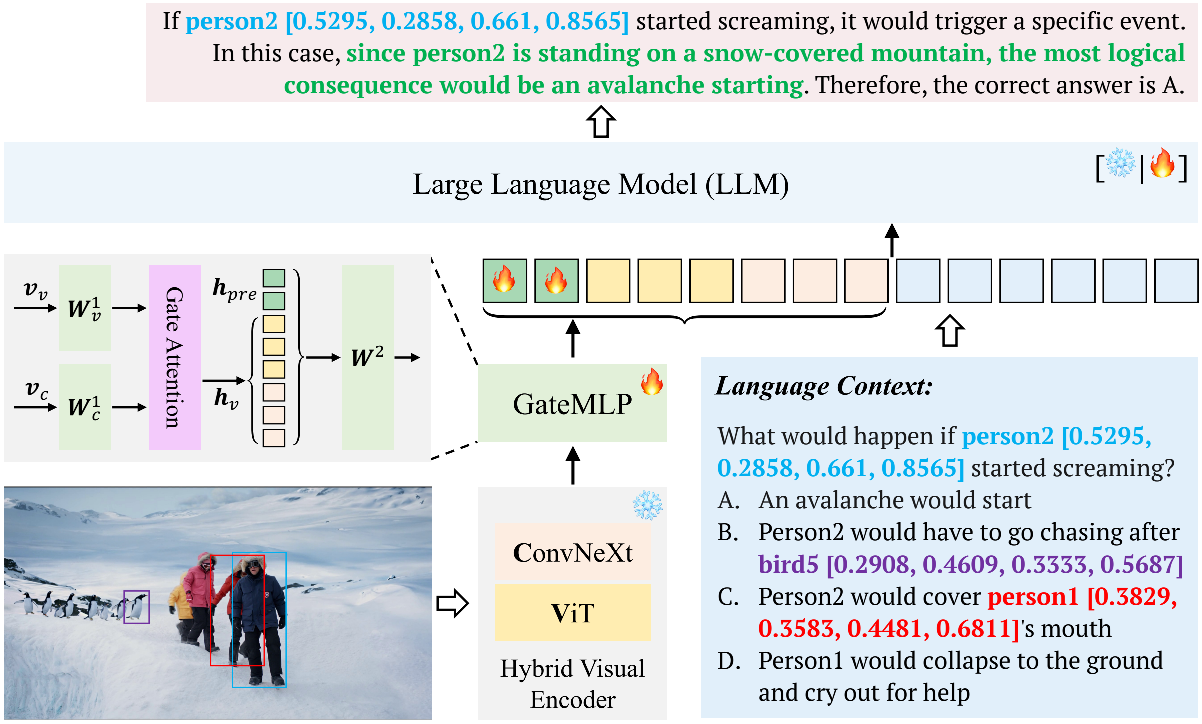


Figure 3: **Overview of ViLaMR**, a multimodal LLM adept at vision-indispensible reasoning and problem solving. Given an image and the corresponding language context, ViLaMR first utilizes a hybrid visual encoder to thoroughly represent the visual input. Subsequently, the proposed GateMLP integrates different visual features into LLM-friendly token embeddings. Finally, LLM generates instruction-following CoT responses conditioned on the combined vision-language embeddings.

ers to map $\{\boldsymbol{v}_v, \boldsymbol{v}_c\}$ into a unified embedding space, resulting in $\{\boldsymbol{h}_v, \boldsymbol{h}_c\} \in \mathbb{R}^{576\times d}$. In order to maintain the current length of visual tokens without significant alteration, $\boldsymbol{h}_v$ and $\boldsymbol{h}_c$ are element-wisely mixed using a gate attention mechanism, *i.e.*,

$$\boldsymbol{\alpha} = \sigma(\boldsymbol{W}_{ga}[\boldsymbol{h}_v;\, \boldsymbol{h}_c] + \boldsymbol{b}_{ga}), \quad (1)$$

$$\boldsymbol{h} = (1 - \boldsymbol{\alpha}) \odot \boldsymbol{h}_v + \boldsymbol{\alpha} \odot \boldsymbol{h}_c, \quad (2)$$

where $\boldsymbol{W}_{ga} \in \mathbb{R}^{d\times 2d}$, $\boldsymbol{b}_{ga} \in \mathbb{R}^{d}$, σ denotes the Sigmoid function, $\odot$ and $[;]$ respectively represent the operations of element-wise matrix multiplication and vector concatenation. Subsequently, we sequence-wisely insert a learnable token embedding $\boldsymbol{h}_{pre} \in \mathbb{R}^{N_{pre}\times d}$ at the beginning of $\boldsymbol{h}$ to facilitate ViLaMR to further capture visual context. In addition, the prefix token embedding somewhat improves the generalization of ViLaMR. Finally, the integrated visual features are transformed into the language embedding space via a linear projection layer. With the proposed GateMLP, ViLaMR can improve visual representations for correlated vision-language alignment.

Large Language Model. CoT reasoning, as one of the typical emergent capabilities of LLMs, is more prominent in relatively large-scale LLMs. Therefore, this work primarily employs the open-source Vicuna-13B (Chiang et al., 2023) as the LLM decoder, which takes the concatenated embeddings of visual and language tokens as input to generate instruction-following CoT responses. In subsequent work, larger LLMs will be incorporated as the LLM decoders for ViLaMR.

4.2 Model Training

We train ViLaMR with a two-stage strategy.

Stage I: Vision-Language Alignment Pretraining. The first stage aims to assist ViLaMR in forming conceptual links between visual and linguistic elements within the embedding space. In this stage, we follow the training setups of LLaVA-1.5 (Liu et al., 2023a), tuning only the weights of GateMLP while keeping the weights of the hybrid visual encoder and LLM fixed. In contrast, to facilitate a comprehensive understanding of details depicted in the images and consider the connector capacity, we instead utilize a subset of ShareGPT4V (Chen et al., 2023b) as the pretraining dataset, which contains 676K high-quality image-text pairs with informative and diverse captions.

Stage II: Progressive Multimodal CoT Tuning. In the second stage, we jointly train the connector and LLM on MCoT-Instruct to bolster the capabil-

Methods	SFT	Connector	CoT	M3U	MMS	RQA	PCA	MMB^{d}	MMB^{t}
▶ **Close-source LMMs**									
Gemini Pro (Team et al., 2023)				47.9	38.6	60.4	51.7	75.2	73.6
Qwen-VL-Max (Bai et al., 2023b)				51.4	49.5	61.3	49.0	78.1	77.6
GPT-4V (Achiam et al., 2023)				56.8	56.0	68.0	68.0	81.4	81.0
▶ **Open-source MLLMs (w/ 13B LLM)**									
VisCoT (Shao et al., 2024)	2M	MLP	✓	-	-	-	-	-	67.5
LLaVA-1.5 (Liu et al., 2023b)	665K	MLP	✗	36.4	34.3	55.3	35.0	69.2	69.2
ShareGPT4V (Chen et al., 2023b)	665K	MLP	✗	36.6	38.3	57.0	-	69.6	69.8
LLaVA-NeXT (Liu et al., 2024b)	760K	MLP	✗	36.2	40.4	57.6	-	70.7	70.0
LLaVA-CCoT (Mitra et al., 2023)	665K	MLP	✓	-	-	-	-	-	70.7
Sphinx-V2 (Lin et al., 2023)	>1M	HybridMLP	✗	-	-	-	-	69.1	71.0
Honeybee (Cha et al., 2023)	>1M	Abstractor	✗	37.3	-	-	-	74.3	74.3
ViLamr w/o RPE	266K	GateMLP	✓	42.9	41.4	59.3	51.7	74.7	74.9
ViLamr	266K	GateMLP	✓	**43.6**	**43.7**	**62.0**	**53.7**	**75.9**	**75.6**

Table 2: **Comparison with accessible MLLMs involving complex multimodal understanding and reasoning** on MMMU val (M3U), MMStar (MMS), RealWorldQA (RQA), PCA-Bench (PCA), MMBench dev (MMB^{d}) and test (MMB^{t}). ▇: mixing of task-oriented public datasets and instruction data, ▇: only instruction data.

ity of ViLamr to follow instructions and perform CoT reasoning. Given that the vision-indispensable reasoning invariably relies on the comprehensive understanding of multimodal inputs, we propose a *reasoning progressive-enhancement* tuning scheme, which round-wisely increases the proportion of reasoning instances from MViR during the supervised fine-tuning process. Formally, in the i-th $(0 < i \leq N_i)$ training round, the composition of visual-intensive understanding instances $\{\mathcal{U}_i\}$ from MViU and visual-indispensable reasoning instances $\{\mathcal{R}_i\}$ from MViR are determined by a sampling ratio β_i. Thus, the total instances $\mathcal{T}_i$ in the i-th training round can be expressed as

$$\mathcal{T}_i = \left[\{\mathcal{U}_{i,m}\}_{m=1}^{(1-\beta_i)\mathcal{N}} ; \{\mathcal{R}_{i,n}\}_{n=1}^{\beta_i \mathcal{N}} \right], \qquad (3)$$

where $\mathcal{N}$ represents the total number of instances in each training round. This progressive tuning scheme encourages ViLamr to adopt the problem solving paradigm of *"understanding before reasoning"*, thereby enhancing its ability to perform complex multimodal reasoning tasks.

5 Experiment

5.1 Experimental Setups

Implementation Details. ViLamr is first pre-trained on the filtered ShareGPT4V-676K subset for a single epoch, using a learning rate of 2e-3 and a batch size of 128. Subsequently, we fine-tune ViLamr on our newly introduced MCoT-Instruct-266K dataset over $N_i = 3$ rounds, with a batch size of 64 and a learning rate of 5e-5. The sequence length N_{pre} of $\boldsymbol{h}_{pre}$ is set to 24. During the second training stage, the total number of instances $\mathcal{N}$ in each round is set to 238K, and the sampling ratio β is set to 0.4, 0.6, and 0.8 for the three rounds, respectively. We adopt AdamW as the optimizer and cosine annealing scheduler as the learning rate scheduler. Both the first and second training stages are implemented on 8 NVIDIA L20 48G GPUs, taking approximately 21h and 22h respectively.

Methods	MLLM	GDR	VCR	QA^{Geo}	SQA^{I}
GIVL	✗	72.0	-	-	-
GPT4RoI	✓	-	78.6	-	-
G-LLaVA	✓	-	-	67.0	-
T-SciQ	✗	-	-	-	**94.7**
ViLamr w/o PRE	✓	84.9	82.3	67.9	84.1
ViLamr	✓	**85.8**	**83.9**	**69.7**	84.7

Table 3: **Results on in-domain datasets**. GIVL (Yin et al., 2023), GPT4RoI (Zhang et al., 2023b), G-LLaVA (Gao et al., 2023a), and T-SciQ (Wang et al., 2024) are the state-of-the-art methods on GD-VCR (GDR), VCR, GeoQA (QA^{Geo}), and SQA^{I}, respectively.

Evaluation Benchmarks and Datasets. We evaluate ViLamr on six public multimodal benchmarks and four in-domain datasets. These multimodal benchmarks include MMMU val (Yue et al., 2023), MMStar (Chen et al., 2024b), RealWorldQA (X.AI, 2024), PCA-Bench (Chen et al., 2024a), MMBench dev and test (Liu et al., 2023d). The in-domain datasets comprise GD-VCR (Yin et al., 2021), VCR val (Zellers et al., 2019), GeoQA test (Chen et al., 2021), and SQA-IMG test (Lu et al., 2022). As some training samples from these in-domain datasets are used to construct MCoT-Instruct, we thus evaluate the in-domain performance of ViLamr on their reserved splits. Each benchmark or

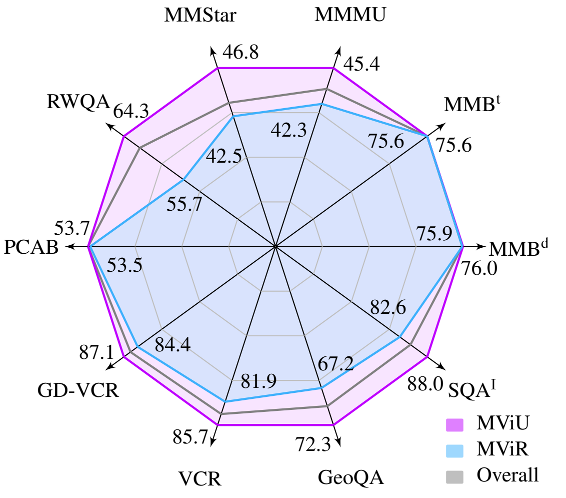


Figure 4: **ViLAMR performance** on vision-intensive understanding (MViU) and vision-indispensable reasoning (MViR) instances.

Methods	**M3U**	**MMS**	**RQA**	**PCA**	**MMB^{d}**	**GDR**	**Avg.**
GateMLP	43.6	43.7	62.0	53.7	75.9	85.8	**60.8**
w/o $\boldsymbol{h}^{pre}$	44.2	42.5	61.3	52.4	74.6	84.3	59.9
w/o GA	42.3	41.7	57.6	53.2	75.1	84.2	59.0

Table 4: **Ablation study on GateMLP**, which improves feature integration via gate attention mechanism (GA) and prefix token embeddings ($\boldsymbol{h}^{pre}$).

dataset, especially MMStar and GeoQA, assesses the complex reasoning capabilities of MLLMs to some extent. These benchmarks and datasets engage the task mode of multi-choice question answering and utilize top-1 accuracy as the evaluation metric. During inference, ViLAMR employs *greedy decoding to generate free-format CoT responses*, from which we parse the option letter of the final outcome for performance evaluation.

5.2 Quantitative Evaluation

Evaluation on Multimodal Benchmark. We first compare ViLAMR with existing state-of-the-art MLLMs using LLMs of the same size in Table 2. Unlike the MLLMs in the table, ViLAMR only utilizes MCoT-Instruct for instruction tuning, without mixing additional task-oriented datasets. Overall, ViLAMR consistently outperforms open-source MLLMs on all benchmarks, and delivers comparable performance to proprietary LMMs such as Gemini Pro and Qwen-VL-Max on RQA, PCA, and MMB, demonstrating the effectiveness of ViLAMR for multimodal understanding and reasoning. Furthermore, we additionally compare "ViLAMR w/o RPE", which fine-tunes ViLAMR using a vanilla instruction tuning method rather than the proposed *reasoning progressive-enhancement* tuning scheme,

Ψ=	**M3U**	**MMS**	**RQA**	**PCA**	**MMB^{d}**	**GDR**	**Avg.**
Ψ1	42.0	41.5	62.9	52.8	76.0	85.2	60.1
Ψ2	43.6	43.7	62.0	53.7	75.9	85.8	**60.8**
Ψ3	43.7	44.2	61.3	53.4	75.6	85.4	60.6

Table 5: **Impact of sampling ratio** Ψ=[β_1, β_2, β_3] on the reasoning progressive-enhancement tuning scheme, where Ψ1=[0.3, 0.5, 0.7], Ψ2=[0.4, 0.6, 0.8], and Ψ3=[0.5, 0.7, 0.9].

meticulously designed to improve the reasoning capability of MLLMs. From Table 2, we can observe that "ViLAMR w/o RPE" continues to outperform all open-source MLLMs on all benchmarks, further validating the effectiveness of ViLAMR.

Evaluation on In-Domain Dataset. Subsequently, we evaluate ViLAMR on in-domain datasets. Considering that the output format of ViLAMR is CoT response, we employ the task mode of Q→AR for GD-VCR and VCR. Results in Table 3 indicate that ViLAMR significantly outperforms existing state-of-the-art methods on GD-VCR, VCR, and GeoQA, except for SQA-IMG. The performance gap with T-SciQ (Wang et al., 2024) (*i.e.*, state-of-the-art on SQA-IMG) can be primarily attributed to their data augmentation strategy, which additionally generates a substantial volume of data analogous to SQA-IMG for training a relatively small-scale vision-language model.

Evaluation on Different Types of Instances. Finally, we conduct a detailed evaluation of ViLAMR on different instance types. Toward this end, we initially utilize the same approach as described in $\mathcal{S}_3$ of Appendix A to categorize these instances from the above benchmarks and datasets. The performance of ViLAMR on MViU and MViR are simply illustrated in Figure 4, from which we can observe that the performance gap between MViR and MViU is not significant. In particular, the performance gap on the PCAB, MMB^{d} and MMB^{t} benchmarks is less than 0.3. These findings indirectly suggest the proficiency of ViLAMR in performing vision-indispensable reasoning tasks.

5.3 Ablation Study

We conduct ablation experiments on M3U, MMS, RQA, PCA, MMB^{d}, and GDR, using their average accuracy as the main criterion to analyze the effectiveness of ViLAMR and MCoT-Instruct.

GateMLP. To efficiently integrate different visual features extracted by the hybrid visual encoder without significantly altering the current length of visual tokens, we propose a vision-language con-

ID	(*i*) VR-OAR	(*ii*) VCR	(*iii*) SciQS	M3U	MMS	RQA	PCA	MMB[d]	GDR	Avg.
#1	✓			39.8	37.3	54.5	47.9	73.5	66.7	53.3
#2	✓	✓		42.7	37.1	58.8	47.3	74.5	84.4	57.4
#3	✓		✓	41.2	39.3	55.4	47.9	75.0	66.7	54.3
#4		✓	✓	43.1	39.7	58.4	48.6	71.0	81.6	57.1
#5	✓	✓	✓	42.9	41.4	59.3	51.7	74.7	84.9	**59.2**

Table 6: **Impact of source data on MCoT-Instruct.** Exp.#5 is equivalent to VILAMR w/o PRE.

nector to combine features in parallel. In terms of architecture, GateMLP first employs a gate attention mechanism for element-wise feature integration, allowing for more precise adjustments to the specific values and distribution patterns of each visual feature. Then, GateMLP uses prefix token embeddings $\boldsymbol{h}^{pre}$ to aid in capturing additional global features. Consequently, as shown in Table 4, we ablate the two critical components within GateMLP to analyze their effect. The average performance degradation observed without $\boldsymbol{h}^{pre}$ ($\downarrow$ 0.9) or gate attention ($\downarrow$ 1.8) demonstrates the effectiveness of our connector design.

Reasoning Progressive-Enhancement Tuning Scheme. This scheme works to further improve the reasoning capability of VILAMR. As demonstrated in Table 2 and Table 3, compared with the vanilla tuning approach (*i.e.*, VILAMR w/o RPE), fine-tuning VILAMR with this progressive scheme delivers considerable performance improvements, demonstrating its effectiveness. As for the scheme itself, its performance markedly depends on the sampling ratio Ψ=[β_1, β_2, β_3], which determines the proportion of reasoning instances in each training round. Thus, we analyze the impact of Ψ on the VILAMR performance by considering the following three sets of values: $\Psi \in \{[0.3, 0.5, 0.7], [0.4, 0.6, 0.8], [0.5, 0.7, 0.9]\}$. Results in Table 5 demonstrate that progressively increasing the number of reasoning instances while consistently maintaining a certain proportion of understanding instances is key to the effectiveness of this scheme.

Contribution of Source Data to MCoT-Instruct. The source datasets for our MCoT-Instruct can be categorized into three groups: (*i*) VR-OAR, which focuses on fine-grained visual reasoning related to the attributes and relations among objects in natural images, including datasets such as A-OKVQA and GPT-VQA; (*ii*) VCR, which emphasizes spatial commonsense reasoning conditioned on a thorough understanding of the visual content within video frames; and (*iii*) SciQS, comprising GeoQA, SQA and TabMWP, which involves solving science problems (*e.g.*, in mathematics, geometry, and physics) using given image content alongside grade-level knowledge and commonsense. We consider VR-OAR as the foundational reference and analyze the impact of incorporating additional types of reasoning data. Notably, to intuitively assess the influence of different data types on MCoT-Instruct, we employ *the vanilla instruction tuning method to fine-tune our model in this ablated experiment*.

Results are presented in Table 6. From the table, we observe that combining all types of source data (Exp.#5) achieves the highest overall average score (59.2) across all benchmarks, demonstrating the indispensability of these three types of reasoning data to MCoT-Instruct. Moreover, the comparative analysis, *i.e.*, Exp.#1 *vs*. Exp.#3 *vs*. Exp.#5, indicates that while the inclusion of SciQS offers advantages within its specific domain and contributes to slight overall improvements, its impact on outcomes in non-science oriented tasks is minimal.

6 Conclusion and Future Work

In this work, we first introduced a multimodal CoT instruction dataset comprising 266K high-quality CoT responses. Building upon this foundation, we developed VILAMR, equipped with a specialized connector to selectively integrate different visual features into LLM-friendly token embeddings. Furthermore, we proposed a reasoning progressive-enhancement tuning scheme to further improve the reasoning ability of VILAMR. Comprehensive experiments across multimodal benchmarks and datasets validated the effectiveness of VILAMR.

In future research, we plan to enrich our MCoT-Instruct with additional reasoning types, scale up VILAMR to larger sizes, and mitigate the factual hallucinations in VILAMR via introducing external knowledge. These forthcoming expansions aim to further improve the capability of VILAMR in complex multimodal reasoning and problem solving, with the aspiration of establishing MCoT-Instruct as a foundational resource to advance the open-source community.

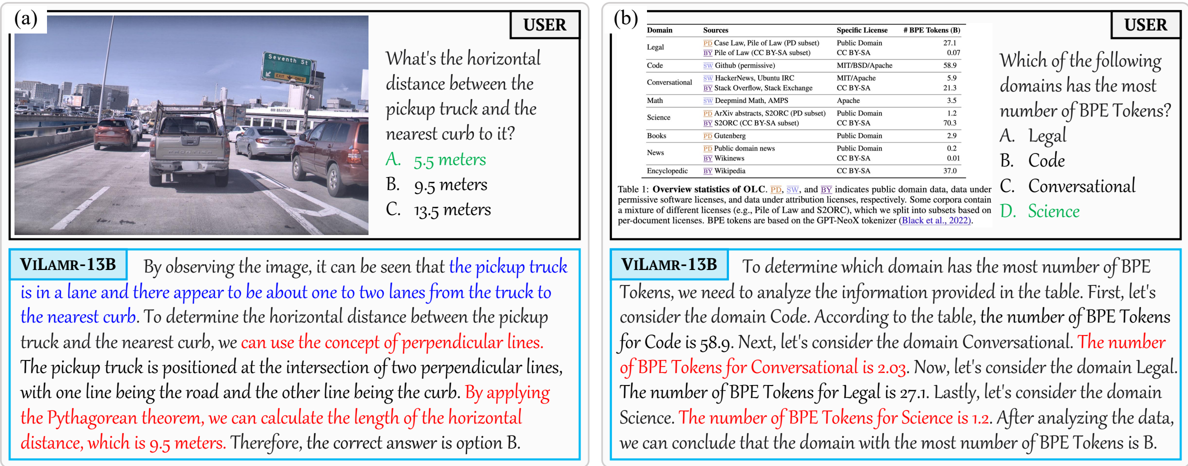


Figure 5: Failure cases caused by a lack of (a) **world commonsense** and (b) **strong OCR capabilities**.

Limitation

Despite the improved reasoning performance after training VILAMR on the introduced MCoT-Instruct dataset, certain limitations remain to be addressed in future work. Through an in-depth analysis of failure cases, we identify that VILAMR still struggles to address two types of reasoning problems: (*i*) *Reasoning requiring world knowledge or commonsense*. As illustrated in Figure 5 (a), VILAMR accurately locates the positions of the pickup truck and the nearest curb, but it fails in reasoning due to a lack of world commonsense (*i.e.*, typical lane widths on urban roads or highways range from about 3.5 to 4.5 meters). (*ii*) *Reasoning conditioned on text-rich image content*. As shown in Figure 5 (b), VILAMR faces challenges in determining which domain has the most number of BPE tokens since it incorrectly recognizes the number of BPE tokens for Conversational and Science domains from the textual table. Therefore, it would be intriguing to mitigate factual hallucination or improve the reasoning capability of MLLMs in text-rich multimodal scenarios.

Ethical Considerations

This work introduced a multimodal CoT instruction dataset, MCoT-Instruct, and developed VILAMR based on this dataset. All source datasets of MCoT-Instruct and the foundational model of VILAMR are open-source and publicly available, without any permission issues or ethical implications. In addition, we will make our dataset and code publicly accessible to facilitate ease of use for researchers and practitioners, thereby promoting transparency and reproducibility in our research.

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*.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint*.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *ACL*, pages 513–523.

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. 2023a. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *arXiv preprint*.

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. 2024a. PCA-Bench: Evaluating multimodal large language models in perception-cognition-action chain. *arXiv preprint*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint*.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. ChainLM: Empowering large language models with improved chain-of-thought prompting. In *LREC-COLING*, pages 2969–2983.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *EMNLP*, pages 8928–8942.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, pages 49250–49267.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. In *ICML*, pages 8469–8488.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023a. G-LLaVA: Solving geometric problem with multi-modal large language model. *arXiv preprint*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023b. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint*.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. *arXiv preprint*.

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. *arXiv preprint*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*, pages 22199–22213.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023. SPHINX: the joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*, pages 34892–34916.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP*, pages 2511–2522.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023a. Chameleon: Plug-and-play compositional reasoning with large language models. In *NeurIPS*, pages 43447–43478.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *ICLR*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. In *NeurIPS*, pages 29615–29627.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2023. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint*.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In *NeurIPS*, pages 25081–25094.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123:74–93.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565.

Shezheng Song, Xiaopeng Li, and Shasha Li. 2023. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint*.

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

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

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *NeurIPS*, pages 200–212.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *AAAI*, pages 19162–19170.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pages 24824–24837.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, pages 16133–16142.

X.AI. 2024. Grok-1.5 vision preview.

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

Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. 2023. GIVL: Improving geographical inclusivity of vision-language models with pre-training methods. In *CVPR*, pages 10951–10961.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *EMNLP*, pages 2115–2129.

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. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint*.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. 2023b. GPT4RoI: Instruction tuning large language model on region-of-interest. *arXiv preprint*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023c. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023d. Automatic chain of thought prompting in large language models. In *ICLR*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023e. Multimodal chain-of-thought reasoning in language models. *arXiv preprint*.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023a. Svit: Scaling up visual instruction tuning. *arXiv preprint*.

Zhiyuan Zhao, Linke Ouyang, Bin Wang, Siyuan Huang, Pan Zhang, Xiao wen Dong, Jiaqi Wang, and Conghui He. 2023b. Mllm-dataengine: An iterative refinement approach for mllm. *arXiv preprint*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint*.

A MCoT-Instruct

A.1 MCoT-Instruct Construction

MCoT-Instruct is constructed using existing VQA datasets that provide explanations or rationales for correct answers. We engage GPT to refine and standardize raw explanations from these datasets to generate high-quality CoT responses. The generation process unfolds in three steps:

$\mathcal{S}_1$ **CoT Rewriting.** CoT is crucial for MLLMs to perform complex reasoning and problem solving, as it not only determines the rationality and controllability of intermediate reasoning processes but also directly affects the accuracy of reasoning outcomes. To improve the diversity and logical consistency of CoTs, as illustrated in Figure 6, we design a specialized prompt to instruct text-only GPT-4 to refine and standardize raw CoTs. These rewritten CoTs will remain faithful and consistent with the given context but become more detailed, logically coherent, and standardized.

$\mathcal{S}_2$ **Quality Verification and Data Filtering.** To further guarantee the quality of rewritten CoTs, we employ GPT to evaluate free-text CoTs across three dimensions: faithfulness, relevance, and completeness. Inspired by the success of LLMs in automatic evaluation (Chiang and Lee, 2023; Liu et al., 2023c), we design a base prompt as shown in Figure 7 to instruct text-only GPT-4 to assign a score (0 - 1) to each rewritten CoT in terms of these three aspects and then average the three scores as an overall score. After that, we filter out these instances with an overall score below 0.6.

$\mathcal{S}_3$ **Instance Grouping.** Following the taxonomy of perception and reasoning capabilities presented in recent works (Zhao et al., 2023b; Liu et al., 2023d), we first direct GPT-3.5 to identify the task type for each selected instance. These instances are subsequently categorized

Data Type	Formatting prompt of CoT response
(*i*) VR-OAR	Provide the rationales that arrive at the correct answer to the question and finally give the option's letter for the correct answer in the format 'ANSWER: X'.
(*ii*) VCR	Perform detailed reasoning based on the context and finally give the option's letter for the correct answer in the format 'ANSWER: X'.
(*iii*) SciQA	Provide the intermediate reasoning steps that lead to the correct answer to the question and finally give the option's letter for the correct answer in the format 'ANSWER: X'.

Table 7: CoT response formatting prompt for different types of source data.

into two groups: *vision-intensive understanding* (MViU) and *vision-indispensable reasoning* (MViR). Specifically, MViU involves a thorough understanding of the input visual content, whereas MViR emphasizes that reasoning must be conditioned on an in-depth understanding of the given visual content, such as the social relation between objects.

A.2 MCoT-Instruct Details

With the above three steps, we ultimately constructed the MCoT-Instruct with 190K MViR instances and 76K MViU instances. Figure 8a and Figure 8b respectively present the source data statistics of MCoT-Instruct and the detailed composition of the source data for MViR and MViU. Additionally, in order to format the CoT response of MLLMs and improve the diversity of task instructions, as depicted in Table 7, we design different task prompts tailored to different source data types within MCoT-Instruct.

B Demonstration

Figures 9 to 11 illustrate the CoT responses of ViLAMR for fine-grained visual understanding and reasoning ((*i*) VR-OAR), spatial commonsense understanding and reasoning ((*ii*) VCR), and science problem solving ((*iii*) SciQA), respectively. In addition to final outcomes, ViLAMR provides detailed intermediate reasoning steps or rationales that lead to the final outcomes, which markedly improves its reliability and interpretability.

System message
You are an AI assistant that can do text rewritten.

Prompt
I want you to act as a Chain-of-Thought (CoT) Rewriter. Given a question with several options and its CoT response (i.e., the intermediate reasoning steps or rationales that lead to the correct answer to the question), your objective is to rewrite the given CoT into a more standardized version.

The rewritten CoT must follow the following rules:
1) Keep the logic of reasoning-then-answering to ensure that the reasoning can be performed step by step.
2) Be faithful enough to ensure that the reasoning can accurately lead to the correct answer.
3) Be clear and concise, without factual errors or repeated content, and no key intermediate reasoning steps are omitted.
4) Do not mention or refer to the given CoT in your responses directly.

You can rewrite the given CoT using the following methods:
1. Improve existing reasoning steps or rationales to make the CoT more coherent and smooth.
2. Add more intermediate reasoning steps or rationales to make the CoT more specific and detailed.

Please make sure you have read and understood these instructions carefully.

Following are two exemplars:
Exemplar 1:

Given question & options:
Is person1 [0.308, 0.1621, 0.7121, 0.9822] the penguin trainer?
A. Yes, person1 is telling all the birds what to do
B. No, person2 [0.8418, 0.1003, 0.9995, 0.9932] doesn't own the penguins
C. No, person1 is a baseball coach
D. No person1 is not at a zoo
Given CoT:
bird4 [0.1462, 0.6635, 0.2965, 0.9813] and bird5 [0.2836, 0.6617, 0.3801, 0.9779], and bird6 [0.1194, 0.679, 0.2172, 0.9939] all are paying attention to person1 [0.308, 0.1621, 0.7121, 0.9822].
Rewritten CoT:
Upon observing the image, it can be seen that person1 [0.308, 0.1621, 0.7121, 0.9822] is receiving attention from bird4 [0.1462, 0.6635, 0.2965, 0.9813], bird5 [0.2836, 0.6617, 0.3801, 0.9779], and bird6 [0.1194, 0.679, 0.2172, 0.9939]. This suggests that person1 has some level of authority or control over the birds. Therefore, analyzing the given options, we can conclude that person1 is likely the penguin trainer who is giving instructions to the birds. Hence, the correct answer to the question is option A.

Exemplar 2:

Given question & options:
Which person is in charge?
A. person3 [0.1231, 0.1601, 0.4296, 0.9848] is the leader
B. person2 [0.5440, 0.1151, 0.7528, 0.9990] is in charge
C. person3 [0.1231, 0.1601, 0.4296, 0.9848] seems to be the one in charge
D. person3 [0.1231, 0.1601, 0.4296, 0.9848] has the higher rank
Given CoT:
person3 [0.1231, 0.1601, 0.4296, 0.9848] is walking ahead of the other two and speaking while person1 [0.3602, 0.0979, 0.5477, 0.9689] looks to be more of a worker.
Rewritten CoT:
The person who is in charge can be determined by observing their behavior and position. In this case, person3 [0.1231, 0.1601, 0.4296, 0.9848] is walking ahead of the other two individuals and is also the one who is speaking, indicating that they have a leadership role. On the other hand, person1 [0.3602, 0.0979, 0.5477, 0.9689] appears to be more of a worker. Based on these observations, it can be concluded that person3 seems to be the one in charge. Hence, the correct answer is B.

Here is the example to be rewritten:
Given Question & Options:
{}
Given CoT:
{}
Rewritten CoT:

Now you can start to rewrite the given CoT.

Figure 6: **Prompt template of CoT rewriting for the VCR dataset.** For other source datasets, please replace the given two exemplars with data-specific examples.

System message

You are a helpful AI assistant that can evaluate the quality of free-text chain-of-thought (CoT) responses generated by a multimodal large language models (MLLM).

Prompt

You will be provided with the input context to the MLLM (i.e., an image description, a question, and several options for the question), along with the corresponding CoT response generated by the MLLM. Your task is to evaluate the free-text CoT responses and give a final overall score (0 - 1) based on the following three perspectives:

- **Faithfulness** (0 - 1): it refers to how accurately the CoT response reflect the actual reasoning process of the MLLM. A faithful CoT response is one that genuinely represents the factors and logic the MLLM used to arrive at its answer. For example, if the MLLM generates an answer based on certain key points in the given context, a faithful CoT response would accurately describe how it picked those points and how they led to the answer. The focus of faithfulness is on the transparency and truthfulness of the explanation.
- **Relevance** (0 - 1): it measures how the CoT response aligns with and supports the answer generated by the MLLM. A consistent CoT response should logically justify the answer, demonstrating a clear and direct connection between the CoT response and the inferred answer. That is, a consistent CoT response should not only be aligned with the answer but also provide sufficient and convincing reasons for why the answer is valid.
- **Completeness** (0 - 1): it evaluates whether the CoT response provided by the MLLM encompasses all essential information and reasoning necessary to understand the MLLM's answer reasoning process. A complete CoT response should cover all critical aspects and steps of the MLLM's reasoning without omitting key details.

Evaluation Steps:

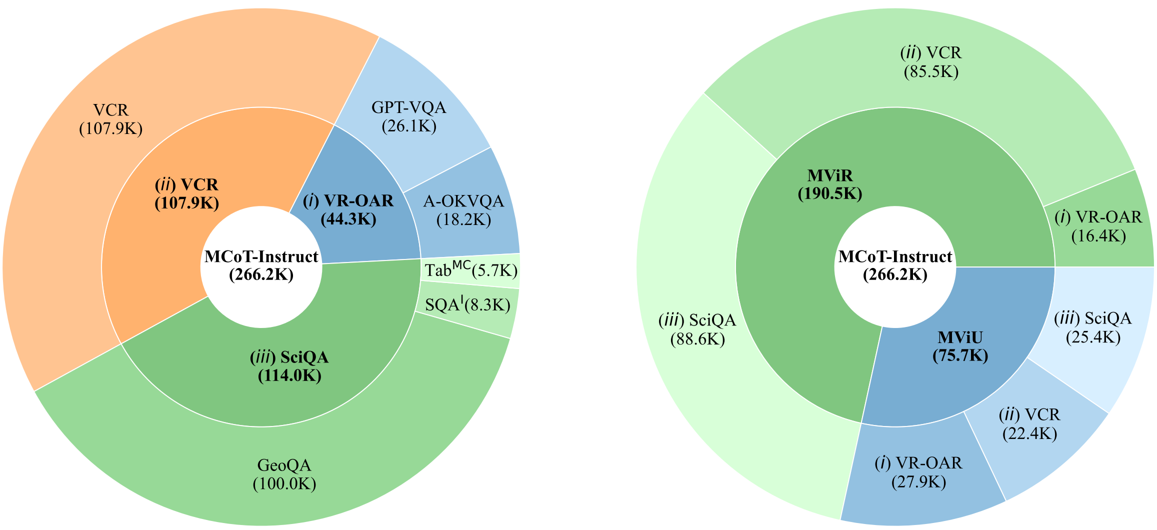
1. Understand and analyze the provided image description, question, and options.
2. Read the MLLM's response and systematically assess the CoT response from the three perspectives of Faithfulness, Relevance, and Completeness.
3. Assign a final overall score (0 - 1) by averaging Faithfulness, Relevance, and Completeness.

Please make sure you read and understand these instructions carefully.

The sample to be scored:
Image Description:
{}
Question & Options:
{}
CoT Response:
{}

Evaluation Form:
Answer by starting with "Scoring:" and then give the explanation of the score by "Explanation:"
- Overall:

Figure 7: **Prompt template for GPT-4 assisted CoT response evaluation.**



(a) Statistics on different types of source data

(b) Statistics on MViR and MViU

Figure 8: **MCoT-Instruct statistics.**

Tags: *action prediction, MViR*

USER

Can we make a U-turn?
A. Yes
B. No

VILAMR-13B To determine if we can make a U-turn, we need to consider the current situation. In this case, we are at an intersection, and there is a sign indicating that U-turns are not allowed. Therefore, the answer to the question is option B, No.

Tags: *object localization, MViU*

USER

Is the bicyclist in this image facing?
A. Toward us
B. Away from us

VILAMR-13B By observing the image, it can be seen that the bicyclist is facing away from the viewer. The bicyclist is located towards the right side of the image and is positioned in such a way that their back is facing the viewer. Therefore, the correct answer to the question is B.

Tags: *future prediction, MViR*

USER

Is it legal to start making a right turn now?
A. Yes
B. No

VILAMR-13B To determine if it is legal to start making a right turn now, we need to consider the current traffic situation. In this case, the traffic light is displaying a red light, indicating that vehicles should stop and not proceed. Therefore, it is not legal to start making a right turn now.

Tags: *spatial relationship, MViU*

USER

Where is the dog relative to the stuffed animal?
A. The dog is to the right of the stuffed animal
B. The dog is to the left of the stuffed animal
C. The dog is a stuffed animal

VILAMR-13B Upon observing the image, the dog is positioned to the left of the stuffed animal. The dog is a real animal, not a stuffed animal, and it is not located to the right of the stuffed animal. Therefore, the answer to the question is option B.

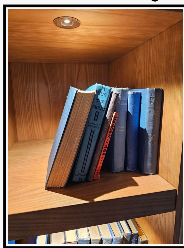
Tags: *attribute comparison, MViU*

USER

Which knife is larger?
A. Both knives are the same size
B. The knife on the left is larger
C. The knife on the right is larger

VILAMR-13B By observing the image, it can be seen that the knife on the left occupies a larger area of the image compared to the knife on the right. Therefore, it can be inferred that the knife on the left is larger. Hence, the correct answer is option B.

Tags: *attribute recognition, MViU*



USER

Which book is facing the wrong way?
A. The first book on the left is facing the wrong way
B. The red book is facing the wrong way
C. All of the books pictured are facing the same direction

VILAMR-13B Upon observing the image, it can be seen that the first book on the left is facing the wrong way compared to the other books. The other books are all facing the same direction, which is not the case with the first book on the left. Therefore, the correct answer is option A.

Figure 9: **VILAMR response demonstration for fine-grained visual understanding and reasoning** (*i.e.*, (*i*) VR-OAR) on RealWorldQA (X.AI, 2024).

Tags: social relation, South Asia, MViR

USER

Why are person2 [0.6549, 0.047, 0.9989, 0.9984], person3 [0.0695, 0.0, 0.3425, 0.9829], and person4 [0.471, 0.086, 0.6813, 0.9689] qualified for this party?

A. Because they are rich people in Indian society
B. They are part of an army
C. They are students in a private school
D. person2, person3, and person4 are here for a special event and that is her costume

VILAMR-13B By observing the image, we can identify that the reason why person2, person3, and person4 are qualified for this party is because they are rich people in Indian society. This can be inferred from the fact that they are dressed in expensive clothing and are drinking champagne. Additionally, person2 is wearing a gold ring, which further suggests their wealth. Therefore, the answer to the question is option A.

Tags: image scene, West, MViU

USER

Where are person1 [0.0914, 0.0677, 0.3482, 0.8531] and person2 [0.4387, 0.0073, 0.9793, 0.9959]?

A. They are in person1 and person2's house
B. They're at a wedding
C. They are in class
D. They are at a bar

VILAMR-13B By observing the image, it can be seen that person1 and person2 are currently in a location where there are other individuals present. Specifically, person3 [0.623, 0.0014, 0.749, 0.3766], person4 [0.7514, 0.0014, 0.8663, 0.3766], person5 [0.8663, 0.0014, 0.9992, 0.3766], and person6 [0.9015, 0.0014, 0.9992, 0.3766] are also present in this location. Additionally, person1 and person2 are seated at a table where there are drinks, including wineglass7 [0.0002, 0.7712, 0.0396, 0.9986], wineglass8 [0.0396, 0.7712, 0.089, 0.9986], and wineglass9 [0.089, 0.7712, 0.1384, 0.9986]. Based on this information, it can be inferred that person1 and person2 are at a bar. Therefore, the answer to the question is option D.

Tags: image scene, East Asia, MViR

USER

What is person2 [0.0155, 0.113, 0.8708, 0.994] going to do?

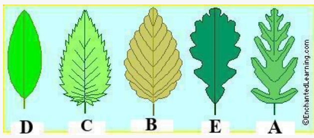
A. person2 is going to film his experiences while visiting the country
B. person2 will turn and say something to person2
C. person2 is seeking a job
D. Get the autograph of an important person

VILAMR-13B By observing the image, we can find that a crowded scene where people appear to be holding flyers or papers up, which often suggests an event like a job fair, a protest, or some form of public gathering where information is distributed. Additionally, a recruitment display board is visible on the left side of the image, confirming that person2 is likely participating in job-seeking activities. Therefore, the answer to the question is option C.

Figure 10: **VILAMR response demonstration for spatial commonsense understanding and reasoning** (*i.e.*, (*ii*) VCR) on GD-VCR (Yin et al., 2021).

Tags: biology problem, MViR

USER



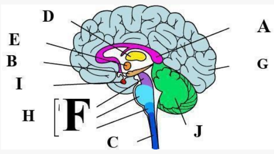
Which is lobed leaf?

A. E
B. A
C. C
D. B

ViLaMR-13B A lobed leaf is a type of leaf that has multiple sections or lobes. Among the leaf shapes in the image, Leaf A and Leaf E display such characteristics with distinct divisions along their margins that extend well towards the center. But, Leaf E shows more pronounced and deeper indentations, characteristic of a lobed structure. Therefore, the correct is option A.

Tags: biology problem, MViU

USER



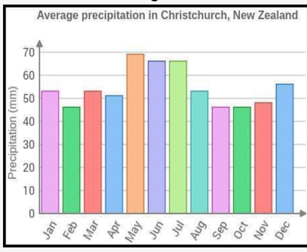
The letter A points at what part of the brain?

A. The corpus callosum
B. The cerebral cortex
C. The spinal cord
D. The thalamus

ViLaMR-13B The letter A in the diagram is pointing towards a specific part of the brain. This part is known as the corpus callosum. The corpus callosum is a large bundle of nerve fibers that connects the two hemispheres of the brain. It allows for communication between the two hemispheres. Therefore, the correct answer is A.

Tags: attribute comparison, MViU

USER



Which month is the wettest on average in Christchurch?

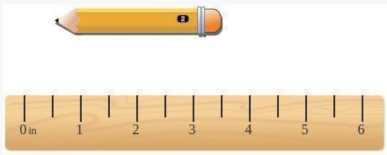
A. August
B. April
C. May
D. March

ViLaMR-13B Upon observing the image, we can find a graph that displays the average amount of precipitation each month. Taller bars on the graph indicate higher average precipitation. By examining the data, we can conclude that the month with the highest average precipitation is May. Therefore, the correct answer is C.

Tags: numeric calculation, MViU

USER

Move the ruler to measure the length of the pencil to the nearest inch. The pencil is about () inches long.

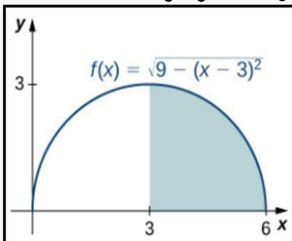


A. 0.5
B. 1
C. 2
D. 3

ViLaMR-13B To determine the length of the pencil in inches, we can use the ruler as a measuring tool. By comparing the markings on the ruler with the length of the pencil, we can estimate its size. In this case, the pencil is approximately 3 inches long. Therefore, the answer to the question is option D.

Tags: geometry problem, MViR

USER



What is the size of the shaded area under the curve?

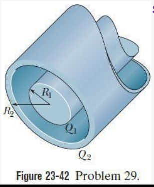
A. 7.07
B. 6.97
C. 7.14
D. 9.00

ViLaMR-13B The size of the shaded area under the curve can be determined using the formula for the area of a sector. The area of the sector is equal to $(1/2) \times \pi \times 3^2$, which simplifies to $9/2\times\pi$. Since the shaded area is equal to half of the sector area, the shaded area is equal to $1/2 \times 9/2\times\pi$, resulting in $9/4\times\pi$. Therefore, the size of the shaded area is A, 7.07.

Tags: physics problem, MViR

USER

Figure 23-42 is a section of a conducting rod of radius R_1=1.30mm and length L=11.00m inside a thin-walled coaxial conducting cylindrical shell of radius R_2=10.0R_1 and the (same) length L.



The net charge on the rod is Q_1=+3.40×10-12C; that on the shell is Q_2=-2.00Q_1. What is the magnitude E of the electric field at radial distance r=2.00R_2?

A. -0.21 B. 0.21 C. 21 D. 0

ViLaMR-13B We can calculate the electric field using the formula $E = Q/\epsilon_0 2\pi rL$, where the total charge $Q=Q_1+Q_2=-3.4\times10^{-12}$C, the vacuum permittivity $\epsilon_0=8.854\times10^{-12}$F/m, r=2.00$R_2$=2.6×$10^{-2}$m, and L=11.00m. Substituting the given values, we have $E = =-3.4\times10^{-12}C /(8.854\times10^{-12}F/m\times2\pi\times2.6\times10^{-2}m\times11m)$. Simplifying this expression, we get E≈0.213V/m. Therefore, the answer is option B.

Figure 11: **ViLaMR response demonstration for science problem solving** (*i.e.*, (*iii*) SciQA) on MMStar (Chen et al., 2024b).