MARS: Multimodal Adaptive Reasoning Model for Avoiding Overthinking

Tan Yue[1], **Qiong Wu**[1], **Dongyan Zhao**[1,2*]

[1]Wangxuan Institute of Computer Technology, Peking University
[2]State Key Laboratory of General Artificial Intelligence
yuetan@pku.edu.cn, wu_qiong@stu.pku.edu.cn, zhaodongyan@pku.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have shown advanced performance in vision-language tasks. However, existing multimodal reasoning models often suffer from excessive reasoning steps, leading to high computational costs and inefficiency. In this paper, we propose the Multimodal Adaptive Reasoning Model (MARS), which enables adaptive adjustment of the reasoning strategy based on question difficulty. Specifically, MARS adopts a three-stage training framework based on our constructed training dataset (MART): 1) CoT Masking Learning to enhance reasoning logicality by predicting masked reasoning steps. 2) Adaptive Reasoning Instruction Learning to train the model to skip or keep reasoning steps according to difficulty levels. 3) CoT Lightweight Reinforcement Learning with the Information Bottleneck Principle based GRPO algorithm to reduce CoT length while maintaining performance and generalizability. Results on both in-domain and out-of-domain datasets show that MARS significantly reduces the CoT length (90.2% decrease) while improving accuracy (0.54%), outperforming existing SOTA open-source and proprietary MLLMs.

Introduction

Multimodal Large Language Models (MLLMs) demonstrate strong perception and understanding capabilities in multimodal understanding (Xia et al. 2024) and visual question answering (VQA) (Zheng et al. 2024) by fusing multimodal information (Li et al. 2023; Fan et al. 2024). Current MLLMs can be categorized into non-thinking models (Chiang et al. 2023; Wu et al. 2024; Bai et al. 2025) and thinking models (Team et al. 2025; Anthropic 2025). Non-thinking models usually have strong instruction-following capabilities and generate responses efficiently, making them suitable for a wide range of daily question answering (Q&A) tasks (Touvron et al. 2023). However, these models show significant performance deficiencies when dealing with complex, multi-step reasoning tasks (Lu et al. 2023; Wang et al. 2024a). Thinking models, on the other hand, construct long Chain-of-Thought (CoT) to derive answers step by step, and show better accuracy and stability in tasks that require complex logical reasoning. However, the process of generating CoT is often time-consuming and requires high computational cost (Sui et al. 2025; Yang et al. 2025). To address the limitation, some studies use selective decoding methods to determine early stopping by confidence (Bajpai et al. 2023; Wei et al. 2025), while other works compress the CoT length by analyzing the contribution of CoT tokens to the final answer (Xia et al. 2025; Wang et al. 2025a).

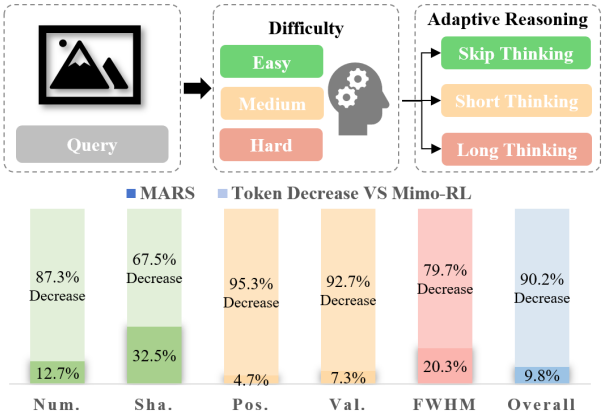

Figure 1: Output length of Mimo-RL (baseline) and MARS across question difficulty (easy: green, medium: yellow, hard: red, overall: blue). Details are shown in Table 4.

However, 1) most of works focus on unimodal reasoning tasks (e.g., math). 2) Existing studies also ignore the question difficulty rating for adaptive reasoning. 3) The lack of high-quality training data for multimodal reasoning leads to significant performance decreases due to compression of the CoT. Accordingly, we aim to make MLLMs capable of adaptive thinking, intelligently adjusting reasoning strategy according to the difficulty level of the question, leading to efficient and accurate answer generation. For easy questions, the model does not need a lengthy CoT and should skip the reasoning steps to provide the answer directly. For medium-difficulty questions, reasoning steps are necessary to improve accuracy but should be kept concise. For hard questions, we enhance reasoning logic and penalize repeated ineffective steps to avoid overthinking.

We propose the **M**ultimodal **A**daptive **Rea**soning Model (MARS). As shown in Fig. 1 and 2, MARS enables adaptive reasoning based on question difficulty, and significantly

*Corresponding author

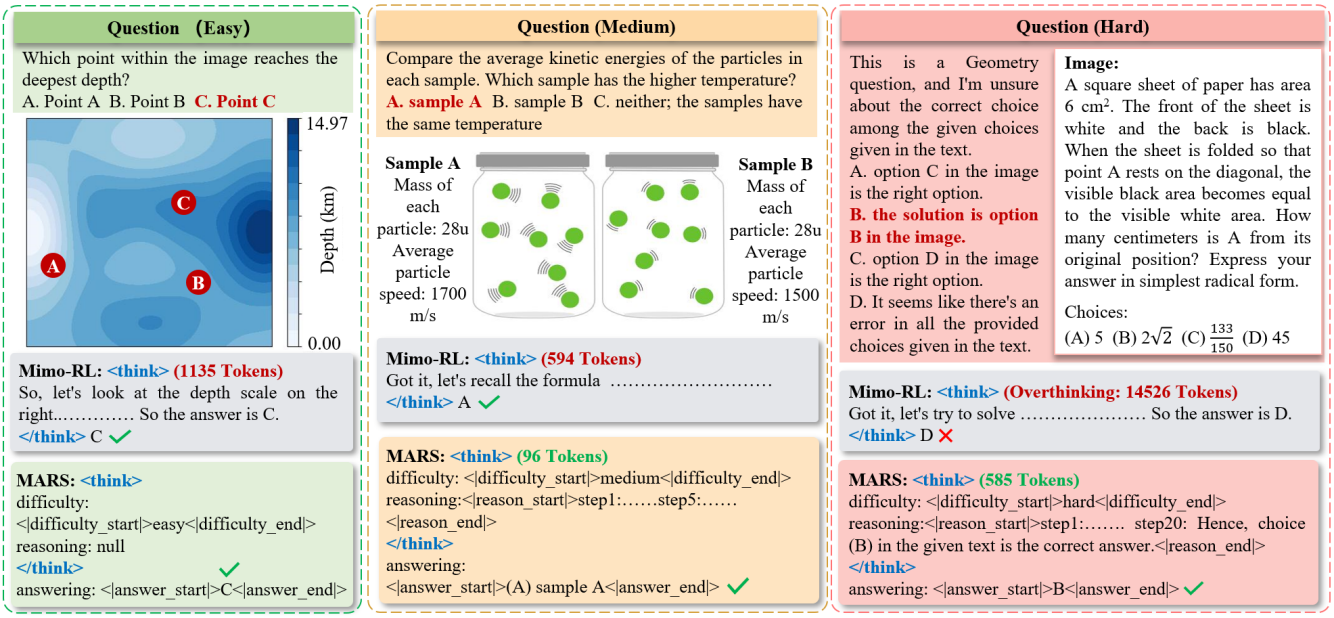

Figure 2: Comparison of the MARS and Mimo-RL models in multimodal reasoning Q&A tasks of different difficulty levels.

reduces the CoT length compared to the baseline Mimo-RL (90.2% overall decrease), while maintaining good performance (MARS-42.31% VS Mimo-RL-41.77%). Specifically, we design a three-stage training framework to achieve the adaptive reasoning. **CoT Masking Learning:** We randomly mask reasoning steps in the CoT and train the model to reconstruct the missing parts from contextual logic, enhancing its understanding of the reasoning process and mitigating overthinking caused by internal logical errors. **Adaptive Reasoning Instruction Learning:** We instruct the model to determine the difficulty level of the question (easy/medium/hard) before reasoning. For easy questions, the model skips the reasoning step and answers directly; for medium and hard questions, model remains the reasoning step and generates CoT before giving the answer. **CoT Lightweight Reinforcement Learning:** we propose IB-GRPO by combining the Information Bottleneck (IB) Principle with the Group Relative Policy Optimization (GRPO) algorithm. This approach maintains performance while encouraging the model to reduce CoT length based on question difficulty through a hierarchical length penalty reward, achieving a balance between accuracy and efficiency.

Our contributions are summarized as follows: 1) We propose a three-stage training framework of "CoT Masking Learning–Adaptive Reasoning Instruction Learning–CoT Lightweight Reinforcement Learning", which effectively achieves the adaptive adjustment of reasoning strategy. 2) We construct a Multimodal Adaptive Reasoning Training Dataset (MART-Dataset), with more than 38,000 human annotated high quality samples from 18 domains and 105 sub-domains, for three-stage training. 3) We propose the IB-GRPO algorithm, and design a hierarchical length penalty reward based on the question difficulty to effectively achieve accurate and low-redundancy CoT generation.

Related Works

Thinking and Non-Thinking MLLMs

"Thinking" MLLMs generate intermediate reasoning processes (e.g., CoT) in multimodal tasks, while "non-thinking" MLLMs generate final answers directly from multimodal inputs (Yue et al. 2025c; Wang et al. 2025b; Zhang et al. 2025).

Specifically, non-thinking MLLM focuses on cross-modal alignment and instruction following by inputting visual features into a language model through a projection layer, and then fine-tuning the model with multimodal data to enable direct answer generation (Dai et al. 2023; Yue et al. 2023; Zhu et al. 2024). This end-to-end training process is simple and efficient but lacks intermediate validation, and thus tends to be incapable of multi-step logic questions (Wang et al. 2022; Zhao et al. 2023; Yue et al. 2024).

Thinking MLLMs (Xu et al. 2024; Kil et al. 2024), on the other hand, add the intermediate CoT during training and inference. MLLMs that employ explicit "thinking" outperform direct-response models on complex reasoning tasks. The CoT enables the model to utilize multimodal information step-by-step and avoid omissions (Cheng et al. 2025), resulting in better performance on tasks such as scientific Q&A (Wang et al. 2024c; Yue et al. 2025a). However, it is always accompanied by longer response and higher computational cost (Qiao et al. 2025; Yue et al. 2025b).

Overthinking Optimization

With the adoption of CoT in large language models, researchers have observed that the generated reasoning processes are often excessively long and contain redundant steps (Bajpai et al. 2023; Sui et al. 2025). Although appropriate step-by-step reasoning can enhance the performance of complex tasks (Kang et al. 2025), excessively long CoTs

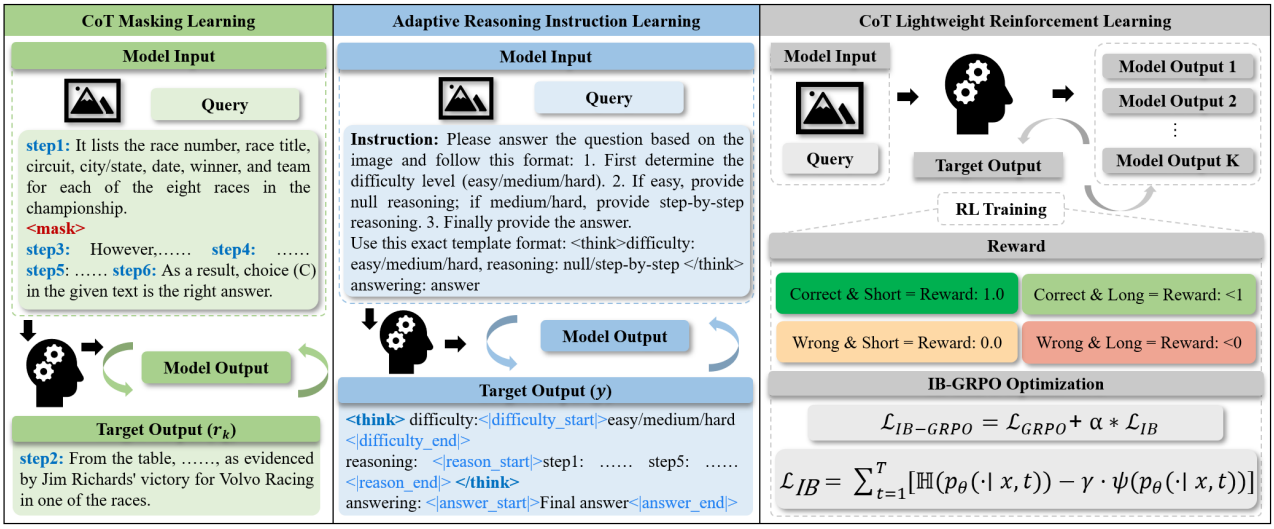

Figure 3: Our proposed three-stage training framework.

often introduce redundant steps and repetitive content (Xue et al. 2023; Yu, Ma, and Wang 2025), thereby increasing computational cost and potentially leading to error accumulation, which is termed as "Overthinking".

To address the challenge of overthinking, researchers have proposed various optimization strategies. Google proposes the CALM method (Schuster et al. 2022) to design a confidence-based early stopping mechanism. The AoT-O3 method (Sel et al. 2025) plans the solution in Algorithm-of-Thought style and penalizes unnecessary steps through a reward function. The TokenSkip method (Xia et al. 2025) compresses the CoT length by analyzing the contribution of tokens in a long CoT to the final answer and skipping those tokens and sentences with low contribution.

Although recent works have made progress, they ignore hierarchical thinking based on the question difficulty. Many easy questions in daily Q&A scenarios do not require complex reasoning. In addition, for extremely difficult questions, even generating a lengthy CoT may still lead to incorrect answers, resulting in an inefficient and fruitless "pseudo-thinking" process. Therefore, it is crucial to enable multimodal reasoning models to adaptively adjust their reasoning strategy and improve their logical consistency in CoT generation.

Methodology

To address the overthinking issue in current multimodal reasoning models, we propose the Multimodal Adaptive Reasoning Model (MARS). As shown in Fig. 3, the model adaptively adjusts the depth of reasoning according to the difficulty of questions by the three-stage training framework, which improves the reasoning efficiency and reduces the computational cost while ensuring the accuracy.

CoT Masking Learning

In the first stage, we train the model to complete reasoning chains by randomly masking parts of the reasoning steps in the training data, aiming to enhance its ability to understand and reconstruct the underlying logical structure of the reasoning chains. The target of this stage is to enhance the fundamental ability of the model to generate logically coherent CoTs, while enhancing reasoning robustness and reducing overthinking caused by errors in the CoT.

Given a CoT $R = \{r_1, r_2, \ldots, r_n\}$ composed of n reasoning steps, we randomly sample one step r_k as the target step to be predicted, and construct an incomplete chain R^k. The model is then tasked to predict the missing step r_k based on the provided question q, answer a, image I, and the incomplete CoT R^k. The training objective is to minimize the negative log-likelihood (NLL) loss of the predicted reasoning step. Formally, the objective function is:

$$\mathcal{L}_{\text{mask}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, q, a, R^k, I; \theta), \qquad (1)$$

where $r_k = \{y_1, y_2, ..., y_T\}$ is the tokenized ground-truth reasoning step, θ denotes the model parameters.

Adaptive Reasoning Instruction Learning

In the second stage, we design a training approach with task instruction that prompt the model to evaluate the difficulty level of a question *[easy / medium / hard]* before generating the CoT and answer. The model then determines whether to perform CoT based on the predicted difficulty level: easy questions are answered directly, while medium and hard questions are answered by generating a CoT first. This stage enables the model to classify question difficulty and selectively apply CoT, allowing for dynamic adjustment of its reasoning process.

Given a question q, an image I, and the answer a, the training instance also includes a labeled difficulty level $d \in \{\texttt{easy}, \texttt{medium}, \texttt{hard}\}$ and the CoT $R = \{r_1, r_2, ..., r_n\}$. Based on these inputs, the model is trained to generate a

structured output in the specified format (shown in Fig. 3). If the difficulty level is `easy`, the reasoning part R is set to *null*. Otherwise, the model must generate a step-by-step CoT.

x denote the full input consisting of the question, image, CoT, labeled difficulty level, and instruction. y denote the expected output (answer). We optimize the model using the standard language modeling objective:

$$\mathcal{L}_{\text{ada}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x; \theta). \tag{2}$$

Here, T is the length of the target output y, y_t is the t-th token in the output sequence, $y_{<t}$ denotes all previously generated tokens before time t, and θ represents the model parameters.

CoT Lightweight Reinforcement Learning

In the third stage, on the basis of the model's adaptive reasoning ability, we further optimize its ability to control the length of the CoT through reinforcement learning. We introduce the Information Bottleneck Principle to improve the GRPO algorithm, and train the model to retain key information and compress redundant content. In addition, we design a hierarchical length penalty reward that constrains the CoT length based on question difficulty, encouraging the model to reduce redundant reasoning while ensuring the efficiency of the CoT.

Reward Computation The reward r is computed by combining two components: the relaxed answer evaluation and the length penalty based on question difficulty.

We evaluate the correctness of $\hat{a}$ against the ground-truth answer a using the *relaxed accuracy criterion* (Methani et al. 2020) to calculate the accuracy (*acc [0,1]*). If both answers are numerical values (including percentages), we allow a relative error tolerance δ (e.g., 5%).

Next, we define the length penalty as a ReLU-style linear function of output length L:

$$\text{Penalty}(L) = \lambda \cdot \max(0, L - T). \tag{3}$$

Here, L is the token count of output y, λ is a difficulty-dependent penalty coefficient, and T is a threshold beyond which overthinking is penalized. These values are set based on difficulty level $d \in \{\texttt{easy}, \texttt{medium}, \texttt{hard}\}$:

$$(\lambda, T) = \begin{cases} (\lambda_e, T_e) & \text{if } d = \texttt{easy} \\ (\lambda_m, T_m) & \text{if } d = \texttt{medium} \\ (\lambda_h, T_h) & \text{if } d = \texttt{hard} \end{cases}$$

Finally, the reward is:

$$r = \text{acc} - \lambda \cdot \max(0, L - T). \tag{4}$$

Information Bottleneck (IB) Based GRPO Optimization We propose the IB-GRPO algorithm, which combines the Information Bottleneck (IB) Principle with the GRPO algorithm to encourage the model to generate concise and accurate CoT. The IB Principle seeks an intermediate representation Z that retains essential task-relevant information while discarding irrelevant details from input X. This is achieved by minimizing $I(Z;X)$ while preserving $I(Z;Y)$.

$$\min_{p(Z|X)} \quad I(Z;X) - \beta \cdot I(Z;Y). \tag{5}$$

We approximate $I(Z;X)$ via the token-level output entropy and a confidence-based regularization term that encourages peaked distributions. Specifically, for each output token y_t, we define:

$$\begin{aligned} \mathcal{L}_{\text{IB}} &= \sum_{t=1}^{T} \left[\mathbb{H}\left(p_\theta(\cdot \mid x, t)\right) - \gamma \cdot \psi\left(p_\theta(\cdot \mid x, t)\right) \right] \\ &= \sum_{t=1}^{T} \Bigg[-\sum_{j} p_\theta(y_t = j \mid x) \log p_\theta(y_t = j \mid x) \\ &\qquad - \gamma \cdot \max_j p_\theta(y_t = j \mid x) \Bigg], \end{aligned} \tag{6}$$

where $\mathbb{H}(\cdot)$ denotes the entropy, $\psi(\cdot)$ denotes the maximum confidence, and γ is the confidence margin coefficient. $p_\theta(y_t = j \mid x)$ denotes the predicted probability of token j at position t given input x. This encourages the model to reduce distributional entropy while increasing confidence in dominant predictions.

To optimize with reinforcement signals, we follow the GRPO algorithm. Let θ denote current model parameters and θ_{old} the frozen baseline. For each input x, K responses $\{y_1, \ldots, y_K\}$ are generated (temperature=0.6), with associated rewards $\{r_1, \ldots, r_K\}$. The normalized advantage is computed as:

$$A_k = \frac{r_k - \mu_r}{\sigma_r + \epsilon_{\text{std}}}, \quad \sigma_r = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(r_k - \mu_r)^2}, \tag{7}$$

where $\mu_r = \frac{1}{K}\sum_{k=1}^{K} r_k$, $\epsilon_{\text{std}} = 10^{-8}$. The GRPO objective is:

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{K}\sum_{k=1}^{K} \min\left(\rho_k A_k,\ \text{clip}(\rho_k, 1-\epsilon_{\text{c}}, 1+\epsilon_{\text{c}}) A_k\right), \tag{8}$$

where $\rho_k = \exp\left(\log p_\theta(y_k \mid x) - \log p_{\theta_{\text{old}}}(y_k \mid x)\right)$ denotes the likelihood ratio between current and baseline policies. $\epsilon_{\text{c}} = 0.2$. Finally, the IB-GRPO optimization objective:

$$\mathcal{L}_{\text{IB-GRPO}} = \mathcal{L}_{\text{GRPO}} + \alpha \cdot \mathcal{L}_{\text{IB}}, \tag{9}$$

where α is the coefficient controlling the strength of information compression.

Compared to the standard GRPO algorithm, our IB-GRPO introduces a structured inductive bias inspired by the Information Bottleneck Principle. By penalizing dispersed representations and redundant outputs, the model is encouraged to generate shorter and coherent reasoning chains. Unlike the standard GRPO, which often produces lengthy CoT, IB-GRPO achieves comparable accuracy with significantly more concise CoT.

MART-Dataset

Training for multimodal adaptive reasoning model relies on a large amount of high-quality data. Therefore, we construct a Multimodal Adaptive Reasoning Training Dataset (MART-Dataset) for multimodal adaptive reasoning task training. We sample data from 21 publicly available multimodal Q&A datasets covering a total of 18 main domains and 105 subdomains. The dataset contains 38,148 samples and 42,947 CoT steps (shown in Fig. 4 and Table 1). We divide the MART-Dataset into three subsets, which are used for the three-stage training. All of the total 36,019 samples used for training stages 2 and 3 are classified into three difficulty levels for each question by human annotators with following criteria. Easy: questions that can be answered directly by extracting a small amount of explicit information. Medium: questions that require comparison, combination, or simple computation of multiple elements in the image. Hard: questions that need to be answered through lots of steps of logical reasoning and integration of specialized knowledge. Each difficulty label is annotated by two annotators, with disagreements resolved by a extra annotator. Final labels are determined by majority vote.

Stage 1: CoT Masking Learning. The goal of this training stage is to avoid the problem of "wrong answers and overthinking due to flawed CoT steps" by training the model to complete the masked CoT steps and improve its logical consistency and causal reasoning ability. We select samples with a complete CoT from publicly available multimodal reasoning Q&A datasets. After first filtering, we collect over 10,000 raw samples. Then, to ensure that the questions are challenging, we use five MLLMs to answer the collect Q&A samples and set the following filtering strategy: A sample is retained if the corresponding question is incorrectly answered by more than three MLLMs, indicating that it is challenging. With the above two-stage filtering strategy, we finally select 2,129 high-quality samples, containing 24,984 CoT steps. Each sample contains image, question, answer, and CoT.

Stage 2: Adaptive Reasoning Instruction Learning. To achieve the adaptive reasoning capability, we construct a subset for the instruction fine-tuning in the second stage, aiming to instruct the model to adjust its depth of thinking according to the question difficulty, thus achieving a balance between efficiency and performance. The subset contains 5,193 multimodal Q&A samples, each with human annotated difficulty level labels *[easy, medium, hard]* and the complete CoT, covering 17,963 high-quality CoT steps. Unlike the first stage which focuses on the CoT complementation, the training objectives in this stage include the following two main aspects: 1) Question difficulty perception ability: the trained model has the ability to classify the difficulty level of the input question, which provides the basis for the subsequent reasoning strategy selection. 2) Adaptive reasoning ability following instructions: the trained model outputs according to the template, including difficulty level, reasoning steps, and final answer. For easy questions, the model should skip the reasoning steps and answer directly; for medium or hard questions, it should generate a CoT before giving the answer (shown in Fig. 3).

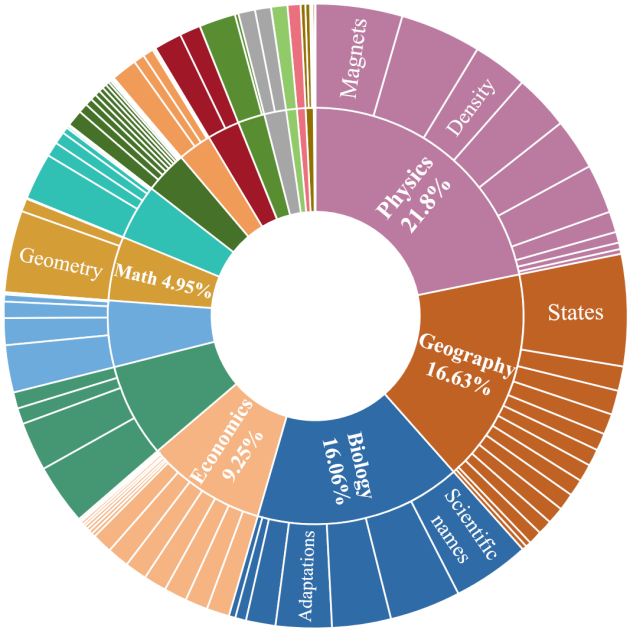

Figure 4: Statistics of domains for the MART-Dataset.

Training Stage	Sample	CoT Step
Stage 1	2129	24984
Stage 2	5193	17963
Stage 3	30826	-
Total	38148	42947

Table 1: Statistics of three-stage sample numbers.

Stage 3: CoT Lightweight Reinforcement Learning. The third training stage aims to further compress redundant CoT while maintaining performance, enhancing both reasoning efficiency and generalization. To support this training stage, we construct a subset of over 30,000 multimodal Q&A samples. All samples have human-annotated difficulty level labels. This subset is used to train the model by the proposed hierarchical CoT length penalty reward and the IB-GRPO algorithm, making the model to generate more concise CoT according to different question difficulty.

Experimental Setup

Baselines: To comprehensively evaluate the performance of our proposed MARS model, we conduct comparisons with 19 proprietary and open-source MLLMs, covering the parameter scale from 1B to 1000B+. The proprietary models include the GPT series (GPT-4o, GPT-4o-mini) (Achiam et al. 2023), Gemini series (e.g., Gemini-2.0-Flash-Thinking) (Team et al. 2024), and the Claude series (e.g., Claude 3.7) (Anthropic 2025). For the open-source MLLMs, we select Qwen series (Bai et al. 2025), LLaVA-1.5 (Liu et al. 2024), ChartVLM (Xia et al. 2024), Kimi-VL-A3B-Thinking (Team et al. 2025), MM-CoT (Zhang et al. 2024), MMR (Wei et al. 2024), MC-CoT (Tan et al. 2024),

	O	T	LLM-Acc	R-Acc
7B+ MLLMs				
Claude-3-Haiku	✗	✗	17.81	14.33
GPT-4o-Mini	✗	✗	27.46	25.17
Gemini-Pro-Vision	✗	✗	32.83	32.14
Gemini-2.0-Flash-T	✗	✓	30.05	33.53
Claude-3-7-Sonnet-T	✗	✓	33.63	38.81
GPT-4o	✗	✗	33.63	41.79
3B-7B MLLMs				
ChartVLM	✓	✗	13.83	8.96
Qwen-VL	✓	✗	21.98	22.98
Kimi-VL-A3B-Instruct	✓	✗	24.68	25.10
Kimi-VL-A3B-Thinking	✓	✓	26.56	27.78
Mimo-VL-7B-SFT	✓	✗	33.11	37.96
Mimo-VL-7B-RL	✓	✓	33.29	41.77
MARS (7B)(Ours)	✓	✓	**34.21**	**42.31**

Table 2: Results of the out-of-domain evaluation on Scichart. O=Open source. T=Thinking model.

	Parameter	Acc
0-1B MLLMs		
MM-CoT-base	0.2B	44.85
MM-CoT-large	0.7B	48.73
MMR	0.8B	50.64
MC-CoT-base	0.2B	53.51
MC-CoT-large	0.7B	57.69
7B-17B MLLMs		
Chameleon	7B	32.30
Qwen2-VL	7B	46.00
LLaMA-Adaper	7B	54.89
LLaVA-V1.5	7B	56.74
CogVLM	17B	58.25
LLaVA-V1.5	13B	59.50
MARS (Ours)	7B	**62.55**

Table 3: Results of the in-domain evaluation on M3CoT.

Chameleon (Team 2024), CogVLM (Wang et al. 2024b), and Mimo-VL (Xiaomi 2025).

Datasets: Since most of the well-known datasets included training subsets have been sampled for the diversity of our MART-Dataset, we conduct in-domain and out-of-domain evaluation for fair comparisons. We select the sampled M3CoT dataset for the in-domain evaluation. We compare our MARS with other models that have also been trained on M3CoT training set. To test generalizability and robustness, we use the SciChart dataset for the out-of-domain evaluation. This dataset has not been trained for any models and reflects real-world scientific chart-based Q&A scenarios. The evaluation is conducted under the "zero-shot" setting without any additional fine-tuning or adaptation.

Settings: We use the Mimo-VL-7B-RL model as the backbone model. AdamW is the optimizer with a learning rate of 2×10^{-5}. $\lambda_e = 0.03, T_e = 32, \lambda_m = 0.001, T_m = 512, \lambda_h = 0.0005, T_h = 1024; \alpha = 0.5, \gamma = 0.01$. For generation, we use sampling with Top-p $= 0.95$. For proprietary models, we use the official API. For open-source models, we use the official GitHub or Huggingface code and reproduce it using the default parameters in the paper. We conduct experiments using eight GPUs (H20-96 GB), and report the average results over five runs.

Evaluation: Following Bai et al. (2023) and Xia et al. (2024), we use *Relaxed-Accuracy (R-Acc)* (Methani et al. 2020) and *LLM-Accuracy (LLM-Acc)* (Xia et al. 2024) for SciChart dataset. *Accuracy (Acc)* is used for M3CoT dataset.

Results

Out-of-Domain (SciChart) Performance As shown in Table 2, on the SciChart dataset, we find that the thinking capability leads to a significant improvement. The Kimi-VL-A3B-Thinking model improve the R-Acc by 2.68% (from 25.10% to 27.78%) compared to its non-thinking version. The Mimo-VL-RL model further improved the R-Acc from 37.96% to 41.77% with the enhanced thinking mechanism, on the basis of the excellent performance achieved. This trend suggests the importance of training models to think before answering to improve accuracy in complex tasks. Our proposed MARS model effectively compresses the CoT length while enabling multimodal adaptive reasoning. Results indicate that MARS maintains strong performance and achieves state-of-the-art results, with an R-Acc of 42.31%. This surpasses the base model Mimo-VL-7B-RL (41.77%), as well as all open-source models and leading proprietary MLLMs, including GPT-4o (41.79%) and Claude-3-7-Sonnet-Thinking (38.81%).

In-Domain (M3CoT) Performance Since we sample data from the M3CoT training data to construct our MART-Dataset, to ensure a fair comparison, we compare MARS with models that are also trained on M3CoT training set. Table 3 shows the performance of current mainstream models. In the small model group (0-1B), the best performing model is MC-CoT-large with an accuracy of 57.69%. In the large model group (7B-17B), several models based on LLaVA and CogVLM perform well. LLaVA-V1.5-13B achieves the best accuracy of 59.50%. Our MARS model achieves 62.55% accuracy using 7B parameters, significantly outperforming compared models with even larger number of parameters.

Adaptive Reasoning Results For further analysis of performance in different types of questions, we perform breakdown analysis on five types of questions in the SciChart dataset, including Peak Number, Shape, Peak Position, Peak Value, and Full Width at Half Maximum (FWHM). We count the length of the model response for each type of question (token number) and accuracy (R-Acc) performance, as shown in Table 4. From the overall results, MARS significantly reduces the tokens required for reasoning while maintaining accuracy, with the overall R-Acc increasing from 41.77% to 42.31%, while the number of reasoning tokens decreases considerably from 980.09 to 95.81 (dropped by 90.22%). In terms of question types, it is observed that MARS has excellent adaptive reasoning ability in different difficulty questions: For the easy type questions such

SciChart	Peak Number		Shape		Peak Position		Peak Value		FWHM		Overall	
	R-Acc	Avg. T.	R-Acc	Avg. T.	R-Acc	Avg. T.	R-Acc	Avg. T.	R-Acc	Avg. T.	R-Acc	Avg. T.
Kimi-VL	38.02	871.23	49.34	509.69	21.21	954.61	17.86	1029.50	9.23	1400.41	27.78	939.33
Mimo-RL	56.25	424.08	50.22	159.87	**48.99**	2302.03	35.18	1216.52	**16.95**	902.47	41.77	980.09
Δ (vs Mimo)	↑1.87	↓87.27%	↑1.94	↓67.47%	↓0.43	↓95.33%	↑0.71	↓92.72%	↓1.56	↓79.67%	↑0.54	↓90.22%
MARS (Ours)	**58.12**	**53.98**	**52.16**	**52.01**	48.56	**107.47**	**35.89**	**88.54**	15.39	**183.45**	**42.31**	**95.81**

Table 4: Breakdown results on different difficulty-level five question types. Avg. T.=average tokens. Kimi-VL=Kimi-VL-A3B-Thinking (16B/A3B), Mimo-RL=Mimo-VL-RL (7B).

as Peak Number and Shape, MARS increases the R-Acc by 1.87% and 1.94% respectively, while the number of tokens decreases by 87.27% and 67.47% respectively, indicating that the model effectively "skips unnecessary thinking" and directly gives the correct answer. For medium-difficulty questions (Peak Value and Peak Position), MARS maintains strong performance stability. The performance of Peak Value is especially outstanding, as the R-Acc increases by 0.71% even though the number of tokens decreases by 92.72%. On the hard FWHM questions, the R-Acc of MARS decreases slightly (1.56%), but the number of tokens decreases by 79.67%, which indicates that the model can still maintain a good reasoning ability in the face of complex task, even though it is limited by the reasoning length.

Analysis of the Max Thinking Length Recent studies (Team et al. 2025) point out that *max-thinking-length* significantly affects reasoning performance, we set different maximum thinking length to evaluate the performance of three thinking models, as shown in Fig. 5. From the overall trend, KIMI-VL-Thinking (16B/A3B) and Mimo-VL-RL (7B) models show a significant decrease in R-Acc under the reasoning length limitation. KIMI-VL-Thinking model shows a decrease from 27.78% to 22.72% when the maximum think length is reduced from 30k to 0.5k, and only 21.34% at the extreme limitation (0.1k). Mimo-VL-7B-RL shows a relatively smaller decrease under the same limitations, but it still decreases from 41.77% to 33.43% and the R-Acc drop reaches 8.34%. In contrast, MARS shows significantly adaptability. Even when the maximum thinking length is reduced from 30k to 0.5k, the accuracy decreases by only 1.35% (from 42.31% to 40.96%), and the leading performance is maintained across all length limitations.

Ablation Study To verify the effectiveness of each training stage and the proposed IB-GRPO algorithm, we conduct ablation study (shown in Table 5). Specifically, we remove or combine four parts in our model: CoT Masking Learning (Mask), Adaptive Reasoning Instruction Learning (Ins.), CoT Lightweight Reinforcement Learning (RL), and IB-GRPO algorithm (IB-GRPO). Using only Stage 2 and Stage 3, although the reasoning length is reduced (258.33 tokens), the accuracy decreases to 38.23%, indicating that the lack of training on the logical structure of the CoT reduces the model's reasoning ability. In the case of the raw GRPO algorithm without IB-GRPO, the accuracy of the model after the complete three-stage training is 41.68% with an average output length of 256.68. After introducing our proposed IB-GRPO algorithm, the model achieves a remarkable compression effect by further decreasing the average tokens from 256.68 to 95.81 while achieving the accuracy improvement (42.31%). This verifies that the Information Bottleneck Principle and the hierarchical length penalty reward we introduced are crucial for adaptive reasoning optimization.

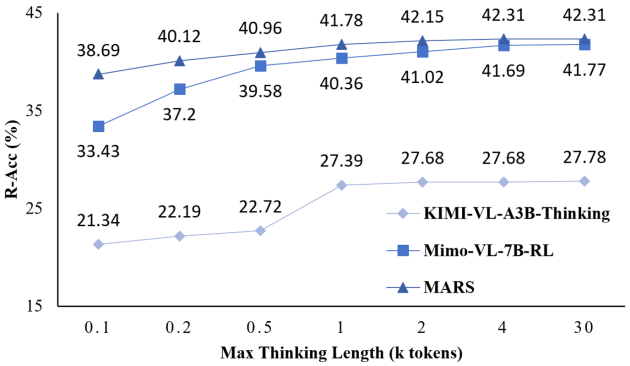

Figure 5: Analysis of the max thinking length.

Mask	Ins.	RL	IB-GRPO	R-Acc	Avg. T.
✗	✗	✗	✗	41.77	980.09
✗	✓	✓	✗	38.23	258.33
✓	✓	✗	✗	40.68	218.65
✓	✗	✓	✗	41.15	686.17
✓	✓	✓	✗	41.68	256.68
✓	✓	✓	✓	**42.31**	**95.81**

Table 5: Ablation study. Avg. T.=average tokens.

Conclusion

In this paper, we propose MARS, a Multimodal Adaptive Reasoning Model that adaptively adjusts its reasoning strategy based on question difficulty. We introduce a three-stage training framework using the improved IB-GRPO algorithm. The MART-Dataset with over 38k high-quality human-annotated samples is constructed to support adaptive reasoning training. Experimental results demonstrate strong generalization on both in-domain and out-of-domain datasets. This work highlights the importance of adaptive reasoning strategies in improving the efficiency and flexibility of multimodal reasoning models.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC, 62576016 and 62506014).

References

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

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-05-20.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bajpai, D. J.; Trivedi, V. K.; Yadav, S. L.; and Hanawal, M. K. 2023. Splitee: Early exit in deep neural networks with split computing. In *Proceedings of the Third International Conference on AI-ML Systems*, 1–9.

Cheng, K.; YanTao, L.; Xu, F.; Zhang, J.; Zhou, H.; and Liu, Y. 2025. Vision-Language Models Can Self-Improve Reasoning via Reflection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8876–8892.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.

Fan, C.; Lin, J.; Mao, R.; and Cambria, E. 2024. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*, 106: 102306.

Kang, Y.; Sun, X.; Chen, L.; and Zou, W. 2025. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24312–24320.

Kil, J.; Tavazoee, F.; Kang, D.; and Kim, J.-K. 2024. II-MMR: Identifying and Improving Multi-modal Multi-hop Reasoning in Visual Question Answering. In *ACL (Findings)*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Methani, N.; Ganguly, P.; Khapra, M. M.; and Kumar, P. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1527–1536.

Qiao, Z.; Deng, Y.; Zeng, J.; Wang, D.; Wei, L.; Meng, F.; Zhou, J.; Ren, J.; and Zhang, Y. 2025. ConCISE: Confidence-guided Compression in Step-by-step Efficient Reasoning. *arXiv preprint arXiv:2505.04881*.

Schuster, T.; Fisch, A.; Gupta, J.; Dehghani, M.; Bahri, D.; Tran, V.; Tay, Y.; and Metzler, D. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35: 17456–17472.

Sel, B.; Huang, L.; Ramakrishnan, N.; Jia, R.; and Jin, M. 2025. LLMs Can Reason Faster Only If We Let Them. In *Forty-second International Conference on Machine Learning*.

Sui, Y.; Chuang, Y.-N.; Wang, G.; Zhang, J.; Zhang, T.; Yuan, J.; Liu, H.; Wen, A.; Zhong, S.; Chen, H.; et al. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Tan, C.; Wei, J.; Gao, Z.; Sun, L.; Li, S.; Guo, R.; Yu, B.; and Li, S. Z. 2024. Boosting the Power of Small Multimodal Reasoning Models to Match Larger Models with Self-consistency Training. In *European Conference on Computer Vision*, 305–322.

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

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, H.; Yue, T.; Ye, X.; He, Z.; Li, B.; and Li, Y. 2022. Revisit finetuning strategy for few-shot learning to transfer the emdeddings. In *The Eleventh International Conference on Learning Representations*.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37: 95095–95169.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2024b. CogVLM: visual expert for pretrained language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 121475–121499.

Wang, Y.; Shen, L.; Yao, H.; Huang, T.; Liu, R.; Tan, N.; Huang, J.; Zhang, K.; and Tao, D. 2025a. R1-Compress: Long Chain-of-Thought Compression via Chunk Compression and Search. *arXiv preprint arXiv:2505.16838*.

Wang, Y.; Wu, S.; Zhang, Y.; Yan, S.; Liu, Z.; Luo, J.; and Fei, H. 2025b. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Wang, Z.; Xia, M.; He, L.; Chen, H.; Liu, Y.; Zhu, R.; Liang, K.; Wu, X.; Liu, H.; Malladi, S.; et al. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37: 113569–113697.

Wei, J.; Tan, C.; Gao, Z.; Sun, L.; Li, S.; Yu, B.; Guo, R.; and Li, S. Z. 2024. Enhancing human-like multimodal reasoning: a new challenging dataset and comprehensive framework. *Neural Computing and Applications*, 36(33): 20849–20861.

Wei, Z.; Chen, W.-L.; Zhu, X.; and Meng, Y. 2025. AdaDecode: Accelerating LLM Decoding with Adaptive Layer Parallelism. *arXiv preprint arXiv:2506.03700*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Xia, H.; Leong, C. T.; Wang, W.; Li, Y.; and Li, W. 2025. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*.

Xia, R.; Zhang, B.; Ye, H.; Yan, X.; Liu, Q.; Zhou, H.; Chen, Z.; Dou, M.; Shi, B.; Yan, J.; et al. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Xiaomi, L.-C. 2025. MiMo-VL Technical Report. arXiv:2506.03569.

Xu, G.; Jin, P.; Li, H.; Song, Y.; Sun, L.; and Yuan, L. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Xue, M.; Liu, D.; Lei, W.; Ren, X.; Yang, B.; Xie, J.; Zhang, Y.; Peng, D.; and Lv, J. 2023. Dynamic voting for efficient reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3085–3104.

Yang, C.; Si, Q.; Duan, Y.; Zhu, Z.; Zhu, C.; Li, Q.; Lin, Z.; Cao, L.; and Wang, W. 2025. Dynamic Early Exit in Reasoning Models. *arXiv preprint arXiv:2504.15895*.

Yu, R.; Ma, X.; and Wang, X. 2025. Dimple: Discrete diffusion multimodal large language model with parallel decoding. *arXiv preprint arXiv:2505.16990*.

Yue, T.; Mao, R.; Shi, X.; Zhan, S.; Yang, Z.; and Zhao, D. 2025a. QAEval: Mixture of Evaluators for Question-Answering Task Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14717–14730.

Yue, T.; Mao, R.; Song, Z.; Hu, Z.; and Zhao, D. 2025b. F2TEval: Human-Aligned Multi-Dimensional Evaluation for Figure-to-Text Task. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 3932–3948.

Yue, T.; Mao, R.; Wang, H.; Hu, Z.; and Cambria, E. 2023. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion*, 100: 101921.

Yue, T.; Shi, X.; Mao, R.; Hu, Z.; and Cambria, E. 2024. SarcNet: a multilingual multimodal sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14325–14335.

Yue, T.; Shi, X.; Mao, R.; Song, Z.; Hu, Z.; and Zhao, D. 2025c. AnaFig: A Human-Aligned Dataset for Scientific Figure Analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 12837–12843.

Zhang, S.; Zhang, J.; Song, W.; Yue, T.; and Zhu, L. 2025. ARISE: Explainable Multi-modal Aggressive Driving Detection via Driver State and Environment Perception. *IEEE Intelligent Systems*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2024. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research*, 2024.

Zhao, J.; Ye, X.; Yue, T.; and Li, Y. 2023. CLDM: convolutional layer dropout module. *Machine Vision and Applications*, 34(4): 63.

Zheng, H.; Wang, S.; Thomas, C.; and Huang, L. 2024. Advancing Chart Question Answering with Robust Chart Component Recognition. *arXiv preprint arXiv:2407.21038*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *ICLR*.