This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Corvid: Improving Multimodal Large Language Models Towards Chain-of-Thought Reasoning

Jingjing Jiang[1,2] Chao Ma[1,*] Xurui Song[2] Hanwang Zhang[2] Jun Luo[2]
[1] Shanghai Jiao Tong University [2] Nanyang Technological University
jingjingjiang2017@gmail.com, chaoma@sjtu.edu.cn, {song0257,hanwangzhang,junluo}@ntu.edu.sg

Abstract

Recent advancements in multimodal large language models (MLLMs) have demonstrated exceptional performance in multimodal perception and understanding. However, leading open-source MLLMs exhibit significant limitations in complex and structured reasoning, particularly in tasks requiring deep reasoning for decision-making and problem-solving. In this work, we present Corvid, an MLLM with enhanced chain-of-thought (CoT) reasoning capabilities. Architecturally, Corvid incorporates a hybrid vision encoder for informative visual representation and a meticulously designed connector (GateMixer) to facilitate cross-modal alignment. To enhance Corvid's CoT reasoning capabilities, we introduce MCoT-Instruct-287K, a high-quality multimodal CoT instruction-following dataset, refined and standardized from diverse public reasoning sources. Leveraging this dataset, we fine-tune Corvid with a two-stage CoT-formatted training approach to progressively enhance its step-by-step reasoning abilities. Furthermore, we propose an effective inference-time scaling strategy that enables Corvid to mitigate over-reasoning and under-reasoning through self-verification. Extensive experiments demonstrate that Corvid outperforms existing o1-like MLLMs and state-of-the-art MLLMs with similar parameter scales, with notable strengths in mathematical reasoning and science problem-solving. Project page: https://mm-vl.github.io/corvid.

1. Introduction

Multimodal large language models (MLLMs) [6, 14, 59, 69, 111, 118] have demonstrated exceptional capabilities in perception and understanding by integrating visual modalities into powerful large language models (LLMs) [5, 95, 99]. However, as illustrated in Figure 1, leading MLLMs, such as Ovis2 [69] and Qwen2.5-VL [6], still exhibit suboptimal performance on complex tasks requiring *deep thinking and extrapolation* for effective problem-solving. Pioneering o1-like MLLMs [71, 97, 109, 110] solve such sophisticated tasks through enhancing chain-of-thought (CoT) reasoning [106] capability, which explicitly unveils logical thought processes and generates step-by-step rationales before arriving at outcomes. Despite these advancements, MLLMs still encounter challenges in complex and structured reasoning.

In this work, we aim to further enhance MLLMs for CoT reasoning by addressing three critical challenges. ***First**, there remains a significant shortage of high-quality multimodal*

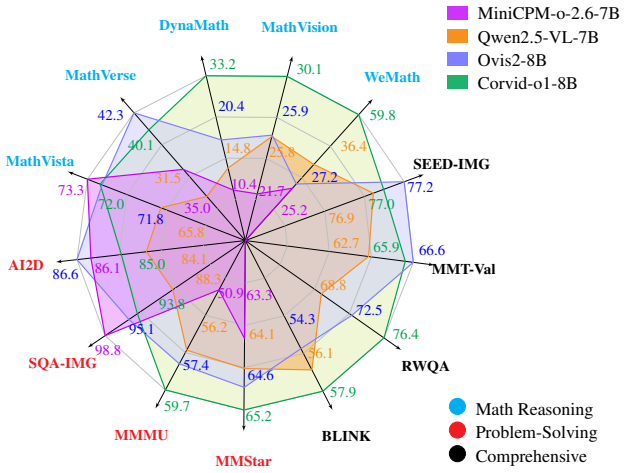


Figure 1. **Comparison with leading open-source MLLMs** with on-par parameter scales. Corvid-o1-8B showcases superiority in mathematical reasoning and problem-solving.

MLLMs	MMStar	MMB	MMVet	MathV	AI2D	Hallusion	Avg.
Baseline	49.8	65.8	57.6	48.6	77.3	40.3	56.9
LLaVA-o1	58.1	75.6	61.7	56.1	78.8	48.2	63.1
LlamaV-o1	59.5	79.9	65.4	54.4	81.2	**63.5**	67.3
Corvid-o1†	62.2	80.3	**67.0**	61.5	81.2	58.7	**68.5**
Mulberry-o1-7B	61.3	75.3	43.9	57.5	79.0	54.1	62.8
Corvid-o1-8B	**65.2**	**82.9**	45.9	**72.0**	**85.0**	54.0	**67.5**

Table 1. **Compared with o1-Like MLLMs**, Corvid demonstrates superior overall performance across multiple benchmarks. Here, Corvid-o1†, LLaVA-o1, and LlamaV-o1 utilize the same baseline MLLM, Llama-3.2-11B-Vision-Instruct.

*Corresponding author.

CoT data. Recent empirical studies [121] have demonstrated that MLLMs trained using direct responses cannot perform step-by-step reasoning, highlighting the urgent need for multimodal CoT data. However, manually created CoTs are typically brief, while AI-generated examples tend to be noisy, rendering them unsuitable for effective CoT-formatted training. ***Second**, MLLMs frequently reason using flawed visual evidence due to insufficient representation and misalignment.* To achieve complex cross-modal reasoning, MLLMs must accurately capture visual information and efficiently transform it into language embedding space, providing the condition and rationale for LLM reasoning. Therefore, optimizing visual representation and cross-modal alignment in MLLMs is essential for further enhancing their reasoning capabilities. ***Third**, MLLMs are prone to over-reasoning and under-reasoning during inference.* Current o1-like MLLMs execute deep reasoning uniformly across all testing instances, regardless of task complexity. In practice, complex reasoning is not always necessary for correct outcomes, particularly for straightforward tasks where inference directly may yield more accurate answers than CoT reasoning. This limitation primarily stems from the context loss and hallucination of MLLMs during long-chain generation.

To address these challenges, we develop an MLLM with advanced CoT reasoning capabilities, referred to as **Corvid**. Architecturally, Corvid integrates two visual foundation models as a hybrid vision encoder for sufficient visual representation and incorporates a novel connector, **GateMixer**, to facilitate modality interaction and alignment through a gate mechanism with selective attention. To enhance Corvid's CoT reasoning capabilities, we first refine and standardize multiple multimodal reasoning datasets spanning diverse reasoning types and domains, resulting in a high-quality multimodal CoT instruction-following dataset, **MCoT-Instruct-287K**. Building upon this foundation, we meticulously curate three specialized datasets to support Corvid's training and optimization. Subsequently, Corvid undergoes a two-stage CoT-formatted training process following its alignment pre-training, to progressively develop step-by-step reasoning while ensuring comprehensive visual understanding. Finally, we propose scaling inference-time computation to mitigate the issues of over-reasoning and under-reasoning. Specifically, we introduce an effective **self-verification** strategy that enables the model to determine whether to perform CoT reasoning conditioned on task complexity.

Corvid is implemented based on the open-source Llama3-8B-Instruct [62] and evaluated across various benchmarks involving mathematical reasoning, science problem-solving, and comprehensive capabilities. As illustrated in Figure 1 and Tab. 1, Corvid outperforms leading open-source MLLMs with on-par parameter scales on 7 out of 13 benchmarks and showcases substantial performance advantages over existing o1-like MLLMs across multiple benchmarks. Furthermore, we conduct systematic ablation studies and analyses to validate the effectiveness and contributions of individual Corvid components. Our main contributions are as follows:

- We develop Corvid, an MLLM with advanced CoT reasoning capabilities. It incorporates a hybrid vision encoder for sufficient visual representation and a novel connector (GateMixer) for enhanced cross-modal alignment.
- We propose an effective inference-time scaling strategy, enabling MLLMs to alleviate over-reasoning and under-reasoning through self-verification.
- We introduce MCoT-Instruct-287K, a high-quality multimodal CoT dataset covering diverse reasoning types.
- Extensive experiments demonstrate the superiority of Corvid against leading open-source MLLMs of comparable parameter sizes and existing o1-like MLLMs.

2. Related Work

Modality Alignment in MLLMs. MLLMs [6, 15, 59, 69, 111] aim to extend the exceptional capabilities of LLMs to complex tasks involving multiple modalities beyond text. To achieve this, MLLMs require specialized modules (*i.e.*, *connectors*) to integrate visual foundation models with LLMs. The most common MLP-based connector [24, 29, 57, 70, 100, 124] utilizes simple linear projection layers to map visual features into the textual embedding space, which is lightweight but short of interaction. In contrast, another line of work [3, 20, 32, 51, 112] incorporates cross-attention layers into the connector to facilitate interaction between modalities. Once structural connections are established, the next objective is to enable LLMs to understand visual concepts by learning from cross-modal calibrated data. A feasible practice is *alignment pre-training*, which involves training the connector on large-scale image-text pairs, such as ShareGPT4V [11] and PixMo [22]. Similarly, our GateMixer, with gate attention, is trained on multi-grained calibrated data with enhanced cross-modal alignment.

CoT Reasoning with MLLMs. Chain-of-thought refers to a series of intermediate reasoning steps or rationales for deriving the final outcome [106] and has been extensively demonstrated to elicit the powerful reasoning capabilities of LLMs [17, 23, 28, 104]. Multimodal CoT reasoning extends this concept by leveraging *CoT prompting* [30, 68, 81] and *CoT-formatted tuning* [71, 97, 109] to perform more complex multimodal tasks, such as mathematical reasoning [120] and robot planning [82]. CoT prompting is typically applied under zero-shot [46] or few-shot [123] paradigms, where large multimodal models like GPT-4o [85] and Claude-3.5-Sonnet [4] engage in step-by-step thinking before reaching an outcome. On the other hand, CoT-formatted tuning is the *visual instruction tuning* of MLLMs using multimodal CoT instruction-following datasets, where performance hinges on the quality of CoT responses. Existing multimodal CoT data primarily originate from manual creation [66, 89, 115] and

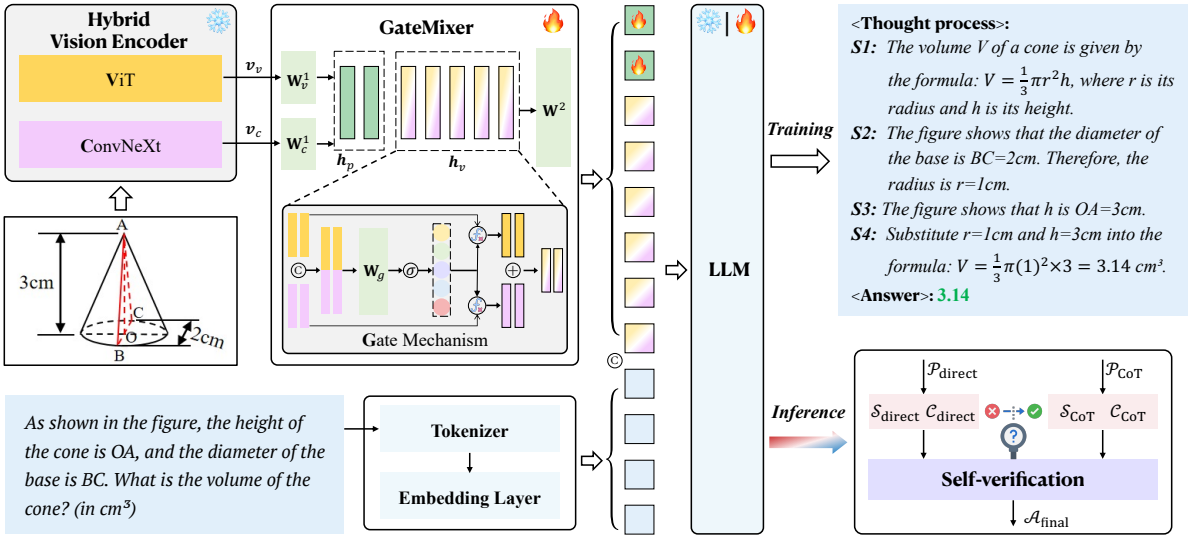


Figure 2. **Model Overview.** Corvid incorporates a hybrid vision encoder to effectively represent visual content and a specially designed connector (GateMixer) to enhance alignment with the LLM. During inference, Corvid implements a self-verification strategy to mitigate over-reasoning and under-reasoning, ensuring more balanced and accurate responses.

AI-assisted generation [109, 120, 124]. Manually-created CoTs are usually accurate but brief, while generated CoTs tend to be more detailed but may contain errors and duplications. In this work, we separately refine and standardize manually-created and AI-generated reasoning datasets with GPT assistance to obtain high-quality CoT data.

Inference-time Scaling in MLLMs. Inference-time scaling [93] seeks to further enhance model performance by enabling iterative reasoning, allowing models to generate multiple responses, engage in self-reflection, and refine their answers through exploratory problem-solving strategies. Recent studies, such as DeepSeek-R1 [21] and OpenAI-o1 [38], have demonstrated the effectiveness of inference-time scaling techniques [33, 36, 103, 107] for depth reasoning in LLM. However, the application of inference-time scaling for MLLM reasoning remains relatively unexplored. Recent approaches have attempted to exploit this technique to improve multimodal reasoning. For instance, LLaVA-o1 [109] and LlamaV-o1 [97] utilize stage-level and sentence-level beam search, respectively, to generate multiple reasoning paths and select the optimal one. Similarly, Mulberry-o1 [110] introduces a collective Monte Carlo tree search algorithm for reasoning path selection. In contrast, we propose an inference-time self-verification strategy to select the final answer without relying on any external verifier.

3. Methodology

3.1. Model Architecture

As illustrated in Figure 2, Corvid comprises the following three crucial modules:

Hybrid Vision Encoder. Visual understanding is fundamental for MLLMs to perform cross-modal reasoning. To maximize this capability in Corvid, we strategically integrate the pretrained SigLIP ViT-SO400M [116] and OpenCLIP ConvNeXt-XXL [61] as a hybrid vision encoder, guided by recent investigations on optimal vision encoders for MLLMs [90, 98, 105]. These studies empirically demonstrate that SigLIP and ConvNeXt-XXL represent the state-of-the-art ViT-based and CNN-based encoders across various multimodal benchmarks. Specifically, the ViT encoder processes input images at a resolution of 384×384 to extract semantically rich features $\boldsymbol{v}_v \in \mathbb{R}^{729\times1152}$, while the ConvNeXt encoder operates the same image at the same resolution to generate multi-stage aggregated features $\boldsymbol{v}_c \in \mathbb{R}^{729\times5760}$ that preserve spatial details.

LLM. Corvid primarily employs the open-source Llama3-8B [62] as its LLM decoder, which encodes concatenated embeddings of visual and language tokens to generate instruction-following CoT responses.

GateMixer. To better align the hybrid vision encoder with LLMs, we design a novel connector to transform $\boldsymbol{v}_v$ and $\boldsymbol{v}_c$ into LLM-friendly token embeddings. As shown in Figure 2, the connector initially employs two distinct linear layers $\mathbf{W}_v^1$ and $\mathbf{W}_c^1$ to map $\{\boldsymbol{v}_v, \boldsymbol{v}_c\}$ into a unified embedding space, resulting in $\{\boldsymbol{h}_v, \boldsymbol{h}_c\} \in \mathbb{R}^{729\times d}$. To maintain the current length of visual tokens without significant alteration, we take insight from the input gate mechanism in LSTM [35], and thus propose to element-wisely mix $\boldsymbol{h}_v$ and $\boldsymbol{h}_c$ using a gate mechanism with selective attention, *i.e.*,

$$\boldsymbol{\alpha} = \sigma(\mathbf{W}_g[\boldsymbol{h}_v;\, \boldsymbol{h}_c] + \boldsymbol{b}_g), \tag{1}$$

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

where $\mathbf{W}_g \in \mathbb{R}^{d\times2d}$, $\boldsymbol{b}_g \in \mathbb{R}^d$, σ denotes the Sigmoid

Name	Data Type (Proportion)	Source Datasets
MGA-1M	Coarse-grained (32.1%)	LLaVA-Pretrain [58] (321K)
	Fine-grained (32.1%)	ALLaVA-4V [10] (195K), Docci [84] (15K), ShareGPT-4o [19] (49K), ShareGPT4V [11] (52K), VG [47] (10K)
	Chart, Math, OCR (35.8%)	ChartCap [43] (30K), MAVIS-Cap [120] (306K), TextCaps [91] (22K)
Corvid-1M	CoT Reasoning (20.2%)	**MCoT-Instruct** (124K), MAVIS-Instruct [120] (81K)
	Direct Reasoning (31.3%)	AI2D [44] (4K), CLEVR [40] (70K), CLEVR-Math [55] (85K), HatefulMemes [45] (8K), VSR [56] (2K), IconQA [65] (27K), Inter-GPS [64] (1K), RAVEN [117] (21K), TallyQA [1] (99K), TQA [2] (1K)
	Chart, Figure, Table (29.7%)	ChartQA [73] (18K), DVQA [41] (49K), FigureQA [42] (49K), HiTab [18] (2K), MapQA [9] (37K), PlotQA [77] (49K), SQA [37] (9K), TAT-QA [127] (2K), WikiSQL [126] (49K), WTQ [87] (38K)
	OCR, Doc (13.8%)	IAM [72] (6K), OCRVQA [79] (80K), InfoGraphicVQA [75] (2K), TextVQA [92] (22K), Visualmrc [94] (3K), DocVQA [74] (10K), ST-VQA [8] (17K)
	Language-only (5.0%)	CamelMath [50] (12K), Dolly [78] (15K), Orca-Math [80] (13K), OpenHermes-2.5 [96] (12K)
o1-320K	CoT Reasoning (100%)	**MCoT-Instruct** (163K), MAVIS-Instruct [120] (137K), CamelMath [50] (20K)

Table 2. **Summary of training data. MCoT-Instruct** is our high-quality multimodal CoT instruction-following dataset, refined and standardized from diverse publicly available reasoning datasets (see Supplementary Material for complete details). Numbers in parentheses indicate the number of conversation instances used for each dataset.

function, $\odot$ and $[;]$ respectively represent the operations of element-wise matrix multiplication and vector concatenation. Subsequently, we sequence-wisely insert learnable prefix token embeddings $\boldsymbol{h}_p \in \mathbb{R}^{N_p \times d}$ at the beginning of $\boldsymbol{h}$ to facilitate context capture and generalization. Finally, the integrated visual features are projected into the language embedding space via a linear projection layer $\mathbf{W}^2$, resulting in $\boldsymbol{h}^0_{img} \in \mathbb{R}^{(N_p+729)\times d}$. With the proposed GateMixer, Corvid can enhance visual representations for correlated vision-language alignment.

3.2. Training Data Curation

Recent studies [76, 98] have demonstrated the pivotal role of data quality and scale in improving MLLM performance. As a result, we introduce three meticulously curated datasets to support the training and optimization of MLLMs.

MAG-1M. Beyond the alignment module, cross-modal calibrated data is essential for MLLMs to establish cross-modal semantic links. To achieve multi-grained alignment between vision and language components, we select nine publicly available image-text pair datasets, as listed in Tab. 2, for alignment pretraining. Here, coarse-grained data supports image-level alignment, fine-grained data facilitates object-level alignment, and the chart, math, and OCR datasets contribute to data diversity.

Corvid-1M. We curate a 1M instruction-following dataset encompassing a diverse range of data types to facilitate supervised fine-tuning. As summarized in Tab. 2, this dataset includes approximately 51.5% reasoning data, with CoT-formatted reasoning data comprising 20.2%, ensuring that MLLMs effectively learn to reason. To enable accurate extraction of visual evidence for reasoning, we augment the dataset using specialized data with various image types, including charts, figures, tables, OCR, and documents, as MLLMs frequently extract critical information from such images for complex reasoning in science, mathematics, and logic. Finally, we incorporate a small portion of language-only instruction-following data to mitigate catastrophic forgetting during model training.

o1-320K. To further strengthen the o1-like reasoning capabilities of MLLMs, we curate a high-quality instruction dataset in a pure CoT format. As shown in Tab. 2, this dataset comprises two multimodal CoT datasets (MCoT-Instruct, MAVIS-Instruct) and one language-only CoT dataset, totaling 320K samples. Here, MCoT-Instruct is our newly refined CoT dataset, featuring standardized and detailed reasoning steps as well as accurate final reasoning outcomes.

3.3. Training Procedure

We train our model in the following three consecutive stages:

Stage 1: Multi-Grained Alignment Pre-training. In this stage, we freeze all parameters of the hybrid vision encoder and LLM, training only the GateMixer on MGA-1M to enable Corvid to establish conceptual links between visual and linguistic elements within the embedding space. In addition to the text generation loss, we introduce a contrastive regularization $\mathcal{L}_{\text{CReg}}$ to further promote semantic correlation between image and text representations, *i.e.*,

$$\mathcal{L}_{\text{CReg}} = -\frac{1}{2b}\sum_{i=1}^{b}\left[\log\frac{\mathcal{S}_{ii}}{\sum_{j=1}^{b}\mathcal{S}_{ji}} + \log\frac{\mathcal{S}_{ii}}{\sum_{j=1}^{b}\mathcal{S}_{ij}}\right], \quad (3)$$

where b denotes the batch size, $\mathcal{S}$ is the similarity measure between the average-pooled image and text representations

(*i.e.*, $\bar{\boldsymbol{h}}_{img} \in \mathbb{R}^{b\times d}$ and $\bar{\boldsymbol{h}}_{txt} \in \mathbb{R}^{b\times d}$) after LLM. Throughout this work, we employ cosine similarity as the measure.

Stage 2: CoT-Enhanced Supervised Fine-tuning. In the second stage, we jointly train the proposed GateMixer and the LLM on Corvid-1M. Upon completion of this stage, we obtained **Corvid-base**, the foundational model capable of following instructions and performing CoT reasoning.

Stage 3: Pure-CoT Instruction Tuning. After enabling Corvid with cross-modal perception abilities and basic CoT reasoning capabilities in the preceding two stages, this stage aims to further strengthen its capability for systematic and structured reasoning. To this end, we fine-tune Corvid-base on o1-320K, resulting in **Corvid-o1**, the final model tailored for complex multimodal reasoning.

3.4. Model Inference

To mitigate *over-reasoning on easy samples* and *under-reasoning on hard ones* during inference, we propose a **self-verification** strategy for handling tasks of varying difficulty. As summarized in Algorithm 1, this strategy first prompts the model to generate both a direct response, $\mathcal{R}_{\text{direct}}$, and a CoT response, $\mathcal{R}_{\text{CoT}}$. It then utilizes a weighted average of the cross-modal representation similarity, $\mathcal{S}$, and the model confidence, $\mathcal{C}$, to determine the final answer. Specifically, $\mathcal{S} \in [0, 1]$ is the cosine similarity between image and text representations after the LLM, while $\mathcal{C} \in (0, 1]$ is the normalized perplexity of the model and formally defined as

$$\mathcal{C} = \exp\left(\frac{\sum_{t=1}^{T} \log P(w_t|w_1, w_2, \ldots, w_{t-1})}{T}\right), \quad (4)$$

where T denotes the number of tokens in the generated text and $P(w_t|w_1, w_2, \ldots, w_{t-1})$ is the conditional probability of generating the t-th token.

Algorithm 1 Inference-Time Self-Verification

Input: Image: $\mathcal{I}$; Question: $\mathcal{Q}$; Task Prompts: $\{\mathcal{P}_{\text{direct}}, \mathcal{P}_{\text{CoT}}\}$; Averaging Weight: α.
Output: Answer: $\mathcal{A}$.

 # Computing Similarity and Confidence during inference
$\mathcal{R}_{\text{direct}}, \mathcal{S}_{\text{direct}}, \mathcal{C}_{\text{direct}} \leftarrow Corvid(\mathcal{I}, \mathcal{Q}, \mathcal{P}_{\text{direct}})$;
$\mathcal{R}_{\text{CoT}}, \mathcal{S}_{\text{CoT}}, \mathcal{C}_{\text{CoT}} \leftarrow Corvid(\mathcal{I}, \mathcal{Q}, \mathcal{P}_{\text{CoT}})$;
 # Determining the final answer $\mathcal{A}$
$\mathcal{A}_{\text{direct}} \leftarrow$ Extracting an answer from $\mathcal{R}_{\text{direct}}$;
$\mathcal{A}_{\text{CoT}} \leftarrow$ Extracting an answer from $\mathcal{R}_{\text{CoT}}$;
if $\mathcal{A}_{\text{CoT}} = \mathcal{A}_{\text{direct}}$ **then**
  return $\mathcal{A}_{\text{CoT}}$
else
  $\mathcal{SC}_{\text{direct}} \leftarrow (1-\alpha)\mathcal{S}_{\text{direct}} + \alpha\mathcal{C}_{\text{direct}}$;
  $\mathcal{SC}_{\text{CoT}} \leftarrow (1-\alpha)\mathcal{S}_{\text{CoT}} + \alpha\mathcal{C}_{\text{CoT}}$;
  if $\mathcal{SC}_{\text{CoT}} >= \mathcal{SC}_{\text{direct}}$ **then**
    return $\mathcal{A}_{\text{CoT}}$
  else
    return $\mathcal{A}_{\text{direct}}$

4. Experiment

4.1. Experimental Setups

Benchmarks. In this work, we primarily evaluate the reasoning capabilities of Corvid across 9 multimodal reasoning benchmarks, including MMStar [12], MMMU [114], SQA-IMG [66], AI2D [44], MathVista [67], MathVerse [119], We-Math [88], MathVision [101], and DynaMath [128]. These benchmarks encompass tasks involving science problem solving and mathematical reasoning. For a comprehensive evaluation of general multimodal abilities, we employ 4 representative benchmarks: SEED-IMG [49], MMT-Val [113], RWQA [108], and BLINK [27]. We report the official performance metric (*i.e.*, accuracy) computed using an open-source evaluation toolkit of MLLMs (*i.e.*, VLMEvalKit [25]) for all the aforementioned benchmarks.

Implementation Details. We set the sequence length N_p in GateMixer to 24. In training phrases, the maximum sequence length of LLMs is set to 2,048. During inference, the weight factor α in Algorithm 1 is set to 0.7, and the maximum length of generated tokens is limited to 1,024. For Corvid-base, we set the $\texttt{temperature}$ of LLM generation to 1.0, whereas for Corvid-o1, we set the $\texttt{temperature}$ and $\texttt{top}_\text{p}$ to 0.4 and 0.9, respectively. All training and inference are conducted on 8 NVIDIA A800 (80G) GPUs. Detailed training setups and costs are provided in the Supplementary Material.

4.2. Main Results

Multimodal Reasoning Benchmarks. In Tab. 3, we compare the proposed Corvid with open-source MLLMs with on-par parameter scales and customized MLLMs using the same base LLM. For problem-solving tasks, Corvid-o1-8B outperforms other custom MLLMs on all benchmarks. Notably, Corvid-o1-8B achieves scores of 65.2 and 59.7 on MMStar and MMMU respectively, surpassing prior best-performing models. For mathematical reasoning, Corvid-o1-8B also performs well across all benchmarks and achieves the highest scores on WeMath (59.8), MathVision (30.1), and DynaMath (33.2). These results confirm Corvid-o1's advanced multimodal reasoning, suggesting that improving the CoT reasoning capability of MLLMs is a feasible solution to handle challenging multimodal tasks.

Comprehensive Benchmarks. Tab. 4 presents a comparison across comprehensive benchmarks, revealing that our models outperform other Llama3-8B-Instruct-based counterparts by a large margin and achieve competitive performance compared to leading MLLMs. In particular, on RWQA, Corvid-base-8B and Corvid-o1-8B surpass the current state-of-the-art by 4.5 and 3.9 points, respectively. On the challenging vision-centric benchmark BLINK, Corvid-o1-8B attains a top score of 57.9, outperforming the runner-up by

MLLMs	*Problem Solving*				*Mathematical Reasoning*				
	MMStar	MMMU	SQA-IMG	AI2D	MathVista	MathVerse	WeMath	MathVision	DynaMath
LLaVA-v1.5-7B [58]	33.1	35.7	69.2	55.5	25.5	4.3	7.0	11.4	1.4
Janus-Pro-7B [13]	46.5	41.6	83.2	68.1	42.5	15.9	9.7	14.7	4.0
Molmo-7B-D [22]	54.4	48.7	92.2	79.6	48.7	4.2	-	16.2	12.6
GLM-4v-9B [31]	54.8	46.9	96.7	71.2	52.2	15.9	11.8	15.0	8.6
MiniCPM-V-2.6-7B [111]	57.5	49.8	96.7	82.1	60.8	17.6	-	18.4	9.8
URSA-8B [71]	-	-	-	-	58.8	31.0	32.8	28.7	13.2
InternVL2.5-4B-MPO [102]	-	-	-	-	64.1	26.0	-	22.5	10.0
VITA-v1.5-7B [26]	60.2	52.6	95.8	79.2	66.2	23.4	19.4	19.5	9.6
POINTS1.5-7B [60]	61.1	53.8	95.0	81.4	66.4	26.6	24.6	22.0	14.2
Ovis2-4B [69]	61.6	49.0	94.0	85.7	69.6	38.5	16.9	21.5	18.0
InternVL2.5-8B [14]	63.2	56.2	-	84.6	64.5	22.8	23.5	17.0	9.4
MiniCPM-o-2.6-7B [111]	63.3	50.9	**98.8**	86.1	**73.3**	35.0	25.2	21.7	10.4
Qwen2.5-VL-7B [6]	64.1	56.2	88.3	84.1	65.8	31.5	36.4	25.8	14.8
Ovis2-8B [69]	64.6	57.4	95.1	**86.6**	71.8	**42.3**	27.2	25.9	20.4
▼ *base LLM: Llama3-8B-Instruct*									
VILA1.5-8B [54]	39.7	37.4	73.2	58.8	37.4	-	-	-	-
Mantis-8B [39]	41.3	41.1	75.5	60.4	32.7	-	-	-	-
Slime-8B [122]	43.5	38.8	78.0	68.5	41.8	22.9	-	-	-
LLaVA-NeXT-8B [59]	43.9	43.1	73.1	72.8	37.7	-	-	-	-
Idefics3-8B [48]	55.0	46.6	91.3	76.5	58.7	-	-	-	-
MiniCPM-V-2.5-8B [111]	51.8	45.8	89.2	78.4	54.5	-	-	-	-
Bunny-8B [34]	45.4	43.4	79.1	69.4	35.2	-	-	-	-
Ovis1.5-8B [69]	57.3	48.3	88.8	82.5	63.0	-	-	-	-
Cambrian-8B [98]	-	42.7	80.4	73.0	49.0	-	-	-	-
Eagle-X5-8B [90]	-	43.8	84.3	76.1	52.7	-	-	-	-
Corvid-base-8B	62.4	57.4	93.2	82.8	64.8	34.8	54.0	26.8	24.5
Corvid-o1-8B	**65.2**	**59.7**	93.8	85.0	72.0	40.1	**59.8**	**30.1**	**33.2**

Table 3. **Comparison with state-of-the-art MLLMs on multimodal reasoning benchmarks.** The best and runner-up results in each section are highlighted in **bold** and underlined, respectively. The performance of compared MLLMs is mainly derived from VLMEvalKit.

1.4 points. Overall, these results demonstrate the robust general multimodal abilities of Corvid.

4.3. Ablation Studies

Effectiveness of GateMixer. As shown in Tab. 5, we first ablate two critical components – the gate attention mechanism ($\mathcal{M}_{\text{GA}}$) and prefix token embeddings ($\boldsymbol{h}_p$) – along with a training objective ($\mathcal{L}_{\text{CReg}}$) of GateMixer to validate its effectiveness. Without $\mathcal{M}_{\text{GA}}$ and $\mathcal{L}_{\text{CReg}}$ alignment training, Corvid's performance declined across all nine benchmarks. Similarly, removing $\boldsymbol{h}_p$ resulted in performance degradation on seven benchmarks. Additionally, we compare the performance of GateMixer with two MLP-based typical architectures, FC_GELU_FC and 2×FC_GELU_FC, to further assess its effectiveness. The results in Tab. 5 show that GateMixer outperforms these MLP-based connectors by 1.6% and 2.3% on average, respectively, highlighting the superiority of our connector. Overall, these findings confirm the effectiveness of GateMixer configurations.

Effect of High-Quality CoT Data. We argue that high-quality CoT-formatted data is essential for improving existing MLLMs towards CoT reasoning. To further investigate this, we compare our final SFT dataset (Corvid-1M) with two alternative datasets: *(i)* $\mathcal{D}_{\text{direct}}$-1M, deleting CoTs and keeping only final answers in Corvid-1M. *(ii)* $\mathcal{D}_{\text{raw rationale}}$-1M, replacing our high-quality CoTs with the corresponding raw rationales from MCoT-Instruct's source datasets. The performance of models fine-tuned on these datasets is presented in Tab. 6, showing that Corvid-1M outperforms *(i)* and *(ii)* across all benchmarks, underscoring the importance of de-

MLLMs	SEED-IMG	MMT-Val	RWQA	BLINK
Monkey-Chat-7B [52]	68.9	53.3	52.4	47.1
DeepSeek-VL-7B [63]	70.1	53.5	54.2	40.9
Molmo-7B-D [22]	74.1	56.8	68.2	46.1
VITA-v1.5-7B [26]	74.1	59.5	66.9	45.0
POINTS1.5-7B [60]	75.1	61.8	67.5	44.0
MiniCPM-V-2.6-7B [111]	74.0	60.8	65.0	55.2
Qwen2.5-VL-7B [6]	76.9	62.7	68.8	56.1
InternVL2.5-8B-MPO [102]	76.8	62.5	68.8	56.6
Ovis2-8B [69]	**77.2**	**66.6**	72.5	54.3
▼ *base LLM: Llama3-8B-Instruct*				
VILA1.5-8B [54]	65.0	48.7	43.4	39.5
Slime-8B [122]	69.8	50.2	58.0	38.8
Mantis-8B [39]	71.2	54.3	59.5	50.1
LLaVA-NeXT-8B [59]	72.5	53.1	58.4	43.5
Bunny-8B [34]	73.5	54.8	60.4	41.6
Eagle-X5-7B [90]	73.6	52.6	63.8	22.4
Ovis1.5-8B [69]	75.4	60.7	64.2	39.8
Idefics3-8B [48]	73.8	58.4	62.6	50.3
Corvid-base-8B	76.4	62.5	**77.0**	55.7
Corvid-o1-8B	77.0	65.9	76.4	**57.9**

Table 4. **Results on comprehensive multimodal benchmarks.**

Connector	**Average**	MMStar	MMMU	SQA-IMG	AI2D	MathVista	MathVerse	WeMath	MathVision	DynaMath
FC_GELU_FC	54.0	61.8	54.9	92.5	80.7	62.1	32.9	49.9	24.2	**26.5**
2×FC_GELU_FC	53.7	61.6	54.6	92.2	82.1	63.2	32.6	51.2	23.6	22.3
GateMixer	**55.6**	**62.4**	57.4	**93.2**	**82.8**	**64.8**	**34.8**	54.0	**26.8**	24.5
w/o $\mathcal{L}_{\text{CReg}}$	55.0	60.1	57.0	91.7	82.0	64.3	32.9	57.1	26.0	23.9
w/o $\boldsymbol{h}_p$	54.8	60.9	**58.8**	92.9	82.5	58.1	32.6	**60.0**	24.1	23.6
w/o GA	54.1	58.1	57.1	89.9	82.1	64.1	29.8	56.5	24.7	24.1

Table 5. **Ablation study on GateMixer.** GateMixer improves feature integration via gate attention mechanism ($\mathcal{M}_{\text{GA}}$) and prefix token embeddings ($\boldsymbol{h}_p$). $\mathcal{L}_{\text{CReg}}$ facilitates cross-modal alignment during pretraining. 2×FC_GELU_FC: two separate FC layers followed by a GELU activation and a FC layer. FC_GELU_FC: one FC layer followed by a GELU activation and a FC layer.

SFT Data	**Average**	MMStar	MMMU	SQA-IMG	AI2D	MathVista	MathVerse	WeMath	MathVision	DynaMath
LLaVA-665K	44.1	51.0	49.5	89.6	66.7	43.8	26.7	35.7	16.7	16.8
Cambrian-1M†	51.9	57.7	60.7	88.8	77.0	57.0	31.1	46.8	25.0	22.7
Ovis-1M†	53.8	61.3	**57.7**	93.0	81.6	61.3	32.9	49.9	24.1	21.9
Corvid-1M	**55.6**	**62.4**	57.4	**93.2**	**82.8**	**64.8**	**34.8**	**54.0**	**26.8**	**24.5**
$\mathcal{D}_{\text{raw rationale}}$-1M	51.2	55.0	55.1	90.7	81.9	54.5	28.8	49.0	23.9	21.9
$\mathcal{D}_{\text{direct}}$-1M	45.7	49.8	44.6	85.7	79.1	50.9	24.3	44.4	13.6	18.5

Table 6. **Effect of high-quality CoT data.** Cambrian-1M† and Ovis-1M† are sampled from Cambrian-10M [98] and Ovis-dataset [69].

MLLMs	Strategy	**Average**	MMStar	MMMU	SQA-IMG	AI2D	MathVista	MathVerse	WeMath	MathVision	DynaMath
Corvid-base	INFER$_{\text{<direct>}}$	44.4	54.1	45.9	87.3	78.3	44.3	22.8	31.6	15.4	20.3
	INFER$_{\text{<CoT>}}$	49.7	54.3	52.8	89.4	67.7	57.6	32.5	46.7	22.4	23.9
	INFER$_{\text{<SV>}}$	**55.6**	**62.4**	**57.4**	**93.2**	**82.8**	**64.8**	**34.8**	**54.0**	**26.8**	**24.5**
Corvid-o1	INFER$_{\text{<direct>}}$	48.9	54.5	47.6	86.8	79.5	49.2	32.2	41.8	24.1	24.7
	INFER$_{\text{<CoT>}}$	56.2	61.1	55.7	91.6	77.5	67.2	36.5	57.1	27.6	31.9
	INFER$_{\text{<SV>}}$	**59.9**	**65.2**	**59.7**	**93.8**	**85.0**	**72.0**	**40.1**	**59.8**	**30.1**	**33.2**

Table 7. **Effectiveness of the proposed self-verification strategy.** Our inference strategy (INFER$_{\text{<SV>}}$) showcases significant performance improvements over direct inference (INFER$_{\text{<direct>}}$) and inference with CoT reasoning (INFER$_{\text{<CoT>}}$).

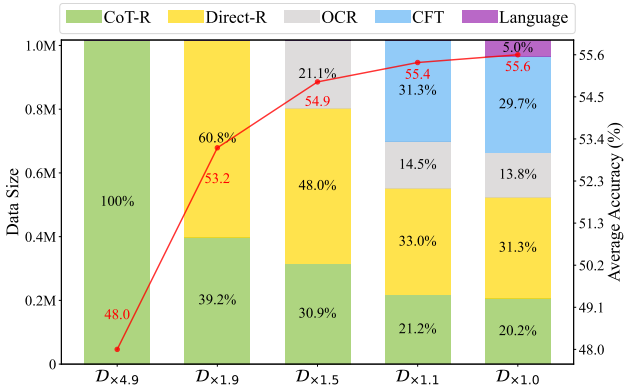


Figure 3. **Effect of SFT data diversity.** The number in the subscript of $\mathcal{D}$ indicates the data repetition rate. $\mathcal{D}_{\times 1.0}$ i.e. Covid-1M.

tailed and standardized CoT data. Additionally, we compare Corvid-1M with three commonly used SFT datasets in Tab. 6, showing that Corvid-1M surpasses them by 11.5%, 8.3%, and 6.7% on average, respectively. These results further emphasize the significance of high-quality CoT-formatted data in improving CoT reasoning capabilities.

Effect of SFT Data Diversity. In Figure 3, we further explore the impact of SFT data diversity on model performance with a controlled data size of 1M. As the dataset transitions from $\mathcal{D}_{\times 4.9}$ (with only CoT reasoning data) to Corvid-1M, the average accuracy steadily increases from 48.0 to 55.6, indicating that incorporating different diverse data types (*e.g.*, direct reasoning and OCR data) facilitates CoT reasoning performance. These results highlight the importance of diverse and well-structured SFT data in improving model reasoning capabilities under a fixed data volume.

Effectiveness of Self-Verification Strategy (INFER$_{\text{<SV>}}$). This strategy is a inference-time scaling technique designed to mitigate *over-reasoning* and *under-reasoning* issues. To evaluate its effectiveness, we compare it against two alternative settings: INFER$_{\text{<CoT>}}$, which formats testing samples as hard instances and performs CoT reasoning before answering, and INFER$_{\text{<direct>}}$, which treats testing samples as easy instances and performs direct inference. The results in Tab. 7 demonstrate that across both Corvid-base and Corvid-o1, INFER$_{\text{<SV>}}$ consistently outperforms the two alternatives. In Corvid-base, it achieves an average score of 55.6, surpassing INFER$_{\text{<direct>}}$ (44.4) by 11.2 points and INFER$_{\text{<CoT>}}$ (49.7) by 5.9 points. A similar improvement is observed in

SigLIP	ConvNeXt	**Average**	MMStar	MMMU	SQA-IMG	AI2D	MathVista	MathVerse	WeMath	MathVision	DynaMath
✓		54.7	**63.2**	**58.9**	**94.2**	81.6	55.5	34.1	**55.0**	25.0	**24.7**
	✓	52.3	59.1	55.6	90.4	80.4	55.4	29.7	50.6	25.4	24.2
✓	✓	**55.6**	62.4	57.4	93.2	**82.8**	**64.8**	**34.8**	54.0	**26.8**	24.5

Table 8. **Effect of the hybrid vision encoder** that combines SigLIP and ConvNeXt for more informative visual representation.

MLLMs	*Similarity-based Metrics*					*LLM-based Offline Evaluator*					GPT-Scorer
	BLUE-1	BLUE-4	ROUGE-L	METEOR	BERTScore	Fact-C	C-Nat.	C-Coh.	C-Eng.	C-Gnd.	Overall
Corvid-base	0.362	0.162	0.445	0.453	0.763	0.792	0.829	0.998	7.137	0.990	0.940
Corvid-o1	0.541	0.396	0.616	0.664	0.894	0.847	0.894	0.999	9.491	0.975	0.990

Table 9. **Automatic quality evaluation of CoT responses on SQA-IMG** with human-provided CoTs as references. We use the following abbreviations: Fact-C for factual consistency, and C-Nat./Coh./Eng./Gnd. for naturalness, coherence, engagingness, and groundedness in conversation. Except for C-Eng. $\in [0, +\infty)$, all other metrics range from 0 to 1.

Corvid-o1, where INFER$_{\langle SV \rangle}$ exceeds them by 11.0 and 3.7 points, respectively. These performance gains showcase the effectiveness of our self-verification strategy.

4.4. Analysis and Discussion

Is Hybrid Vision Encoder Necessary? To investigate this, we conduct the ablation study presented in Tab. 8. The results indicate that leveraging both encoders yields the highest average performance (55.6) across multiple benchmarks, outperforming the use of SigLIP (54.7) or ConvNeXt (52.3) alone. Notably, the hybrid encoder achieves the best scores in AI2D (82.8), MathVista (64.8), MathVerse (34.8), and MathVision (26.8), demonstrating enhanced capabilities in both scientific and mathematical reasoning. These findings confirm the necessity of enriching visual representations using a hybrid vision encoder.

CoT Quality Evaluation. We utilize three types of evaluation methods to assess the quality of Corvid responses: *(i)* similarity-based metrics (BLEU-1&4 [86], ROUGE-L [53], METEOR [7], and BERTScore [83]) to measure the coverage and similarity between generated and reference CoTs. *(ii)* an LLM-based offline evaluator, UNIEVAL [125], to evaluate the factual consistency, naturalness, coherence, engagingness, and groundedness of generated CoTs. *(iii)* a GPT-based online evaluator to assess CoTs across faithfulness, relevance, and completeness, providing an overall score. Results in Tab. 9 indicate that while Corvid's CoT responses exhibit low coverage compared to human-provided ones, they demonstrates high overall performance, particularly excelling in naturalness, coherence, and engagingness, thereby indirectly validating their quality.

Case Study on Corvid's Limitations. To analyze Corvid's reasoning limitations, we conduct an in-depth analysis of failure cases and observe that it struggles with specific types of reasoning problems, particularly those *requiring world knowledge or commonsense*. For example, as illustrated in Figure 4, while Corvid accurately locates the positions

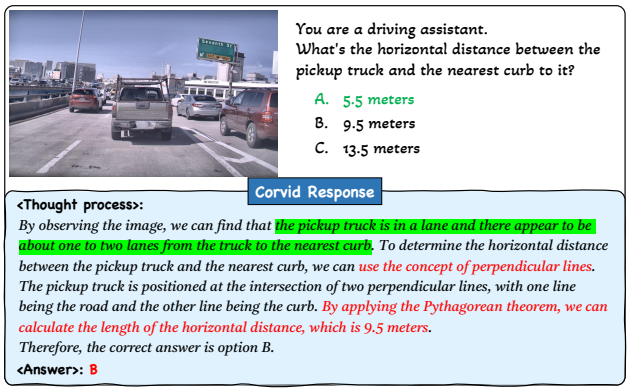


Figure 4. **A failure case** where Corvid fails to reason effectively due to a lack of world commonsense.

of the pickup truck and the nearest curb, it fails to reason correctly 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). Therefore, more techniques, such as group relative policy optimization [21] and retrieval-augmented generation [16], need to be explored to further improve the reasoning capabilities of MLLMs.

5. Conclusion

In this work, we introduce Corvid as the primary exploration into CoT reasoning for MLLMs. Enhanced by critical components such as GateMixer, high-quality CoT data, and an inference-time self-verification strategy, Corvid surpasses existing o1-like MLLMs and state-of-the-art MLLMs with comparable parameter scales across various reasoning and comprehensive benchmarks, showcasing its advanced CoT reasoning capabilities. Despite its impressive performance, several challenges persist that merit future investigation, such as knowledge-intensive reasoning (as discussed in the Limitations section) and efficient scaling techniques. We believe that ongoing research into o1-like capability will further enhance MLLMs to address more sophisticated tasks.

Acknowledgements. This work was supported in part by NSFC (62406189, 62322113, 62376156), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *AAAI*, pages 8076–8084, 2019. 4

[2] Hessa Abdulrahman Alawwad, Areej Alhothali, Usman Naseem, Ali Alkhathlan, and Amani Jamal. Enhancing textbook question answering task with large language models and retrieval augmented generation. *arXiv:2402.05128*, 2024. 4

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, pages 23716–23736, 2022. 2

[4] Anthropic. Grok-1.5 vision preview. https://www.anthropic.com/claude, 2024. 2

[5] 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. Qwen technical report. *arXiv:2309.16609*, 2023. 1

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv:2502.13923*, 2025. 1, 2, 6

[7] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, pages 65–72, 2005. 8

[8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 4

[9] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv:2211.08545*, 2022. 4

[10] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv:2402.11684*, 2024. 4

[11] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*, 2023. 2, 4

[12] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv:2403.20330*, 2024. 5

[13] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv:2501.17811*, 2025. 6

[14] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv:2412.05271*, 2024. 1, 6

[15] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 2

[16] Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv:2407.21439*, 2024. 8

[17] Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. ChainLM: Empowering large language models with improved chain-of-thought prompting. In *COLING*, pages 2969–2983, 2024. 2

[18] Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. Hitab: A hierarchical table dataset for question answering and natural language generation. *arXiv:2108.06712*, 2021. 4

[19] Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o. https://sharegpt4o.github.io, 2024. 4

[20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, pages 49250–49267, 2023. 2

[21] Daya Guo DeepSeek-AI, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*, 2025. 3, 8

[22] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv:2409.17146*, 2024. 2, 6

[23] Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang.

Active prompting with chain-of-thought for large language models. *arXiv:2302.12246*, 2023. 2

[24] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *ICML*, pages 8469–8488, 2023. 2

[25] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv:2407.11691*, 2024. 5

[26] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv:2501.01957*, 2025. 6

[27] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv:2404.12390*, 2024. 5

[28] Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv:2305.17306*, 2023. 2

[29] 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. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*, 2023. 2

[30] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. *arXiv:2404.16033*, 2024. 2

[31] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv:2406.12793*, 2024. 6

[32] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790*, 2023. 2

[33] Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. Glore: When, where, and how to improve llm reasoning via global and local refinements. *arXiv:2402.10963*, 2024. 3

[34] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv:2402.11530*, 2024. 6

[35] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[36] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv:2210.11610*, 2022. 3

[37] Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. Search-based neural structured learning for sequential question answering. In *ACL*, pages 1821–1831, 2017. 4

[38] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv:2412.16720*, 2024. 3

[39] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv:2405.01483*, 2024. 6

[40] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017. 4

[41] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018. 4

[42] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv:1710.07300*, 2017. 4

[43] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In *ACL*, pages 4005–4023, 2022. 4

[44] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016. 4, 5

[45] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, pages 2611–2624, 2020. 4

[46] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, pages 22199–22213, 2022. 2

[47] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 4

[48] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv:2408.12637*, 2024. 6

[49] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*, 2023. 5

[50] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. In *NeurIPS*, pages 51991–52008, 2023. 4

[51] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen

image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 2

[52] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pages 26763–26773, 2024. 6

[53] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 8

[54] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024. 6

[55] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv:2208.05358*, 2022. 4

[56] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv:2205.00363*, 2022. 4

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023. 2

[58] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 4, 6

[59] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next, 2024. 1, 2, 6

[60] Yuan Liu, Le Tian, Xiao Zhou, Xinyu Gao, Kavio Yu, Yang Yu, and Jie Zhou. Points1.5: Building a vision-language model towards real world applications. *arXiv:2412.08443*, 2024. 6

[61] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 3

[62] Meta Llama Team. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 2, 3

[63] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv:2403.05525*, 2024. 6

[64] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, pages 6774–6786, 2021. 4

[65] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv:2110.13214*, 2021. 4

[66] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521, 2022. 2, 5

[67] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*, 2023. 5

[68] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *NeurIPS*, pages 43447–43478, 2023. 2

[69] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024. 1, 2, 6, 7

[70] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. In *NeurIPS*, pages 29615–29627, 2023. 2

[71] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv:2501.04686*, 2025. 1, 2, 6

[72] U-V Marti. The iam-database: an english sentence database for offline handwriting recognition. *IJDAR*, 5:39–46, 2002. 4

[73] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, pages 2263–2279, 2022. 4

[74] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 4

[75] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 4

[76] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv:2403.09611*, 2024. 4

[77] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *ICCV*, pages 1527–1536, 2020. 4

[78] C Mike, H Matt, M Ankit, X Jianwei, W Jun, S Sam, G Ali, W Patrick, Z Matei, and X Reynold. Free dolly: Introducing the world's first truly open instruction-tuned llm. https://simonwillison.net/2023/Apr/13/dolly, 2023. 4

[79] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952, 2019. 4

[80] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *arXiv:2402.14830*, 2024. 4

[81] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *CVPR*, pages 14420–14431, 2024. 2

[82] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In *NeurIPS*, pages 25081–25094, 2023. 2

[83] Reimers Nils. sentence embeddings using siamese bert-networks. In *EMNLP*, pages 3982–3992, 2020. 8

[84] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. *arXiv:2404.19753*, 2024. 4

[85] OpenAI. Hello gpt-4o. https://gpt4o.ai, 2024. 2

[86] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 8

[87] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv:1508.00305*, 2015. 4

[88] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv:2407.01284*, 2024. 5

[89] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162, 2022. 2

[90] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv:2408.15998*, 2024. 3, 6

[91] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020. 4

[92] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 4

[93] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*, 2024. 3

[94] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, pages 13878–13888, 2021. 4

[95] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv:2408.00118*, 2024. 1

[96] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants. https : / / huggingface . co / datasets / teknium/OpenHermes-2.5, 2023. 4

[97] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv:2501.06186*, 2025. 1, 2, 3

[98] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv:2406.16860*, 2024. 3, 4, 6, 7

[99] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 1

[100] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, pages 200–212, 2021. 2

[101] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv:2402.14804*, 2024. 5

[102] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv:2411.10442*, 2024. 6

[103] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv:2203.11171*, 2022. 3

[104] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 2

[105] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *ECCV*, pages 408–424, 2024. 3

[106] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pages 24824–24837, 2022. 1, 2

[107] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. *arXiv:2212.09561*, 2022. 3

[108] X.AI. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, 2024. 5

[109] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv:2411.10440*, 2024. 1, 2, 3

[110] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering

mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv:2412.18319*, 2024. 1, 3

[111] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv:2408.01800*, 2024. 1, 2, 6

[112] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178*, 2023. 2

[113] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv:2404.16006*, 2024. 5

[114] 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. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024. 5

[115] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019. 2

[116] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 3

[117] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*, pages 5317–5327, 2019. 4

[118] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1.5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv:2409.20566*, 2024. 1

[119] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multimodal llm truly see the diagrams in visual math problems? *arXiv:2403.14624*, 2024. 5

[120] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv:2407.08739*, 2024. 2, 3, 4

[121] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv:2410.16198*, 2024. 2

[122] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv:2406.08487*, 2024. 6

[123] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *ICLR*, 2023. 2

[124] Zhiyuan Zhao, Linke Ouyang, Bin Wang, Siyuan Huang, Pan Zhang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Mllm-dataengine: An iterative refinement approach for mllm. *arXiv:2308.13566*, 2023. 2, 3

[125] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *EMNLP*, pages 2023–2038, 2022. 8

[126] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. arxiv 2017. *arXiv:1709.00103*, 2017. 4

[127] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv:2105.07624*, 2021. 4

[128] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv:2411.00836*, 2024. 5