TTRV: Test-Time Reinforcement Learning for Vision Language Models – Supplementary Material –

Akshit Singh[1] **Shyam Marjit**[2] **Wei Lin**[3] **Paul Gavrikov**[1]
Serena Yeung-Levy[4] **Hilde Kuehne**[5,6] **Rogerio Feris**[6] **Sivan Doveh**[4]
James Glass[7] **M. Jehanzeb Mirza**[7]

[1]Independent Researcher [2]IISc Bangalore [3]JKU Linz [4]Stanford
[5]Tübingen AI Center [6]MIT-IBM Watson AI Lab [7]MIT CSAIL

In this supplement, we present additional experiments and explanations that provide further insight and clarity beyond the main manuscript. Section 1 lists additional implementation details and evaluation protocols. Section 2 presents a detailed overview of the datasets used in our study. In Section 3, we describe the prompts utilized in our experiments. Then in Section 4, we present additional ablation studies that highlight further aspects of our method and offer deeper insights into its effectiveness. Finally, Section 5 includes comprehensive pseudocode to facilitate reproducibility and to help readers gain a clearer understanding of the implementation details.

All experiments were conducted on a machine equipped with 4× NVIDIA A100 and 4× NVIDIA A6000 GPUs.

1. Additional Experimental Settings

Implementation Details. We apply TTRV independently on each benchmark and report the results in Tables 1 & 2 (main manuscript). For optimization, we adopt the AdamW optimizer with a cosine learning rate schedule, setting the peak learning rate to 5×10^{-7}. During rollout, we generate 32 candidate responses with a temperature of 1.0 for all experiments. The reward hyperparameter α is fixed at 0.75 for all datasets. We cap the maximum prompt length at 7524 tokens and the maximum response length at 1024 tokens. We generally report results using 20 samples in the main table. These samples are randomly sampled from the test data. In the appendix, we also provide a comparison between 20- and 500-sample adaptation, and in the ablation study, we further evaluate the extreme case 1-sample adaptation, where the model adapts to a single example before being evaluated on the full dataset.

Evaluation Protocol: For evaluation, we use greedy decoding (temperature $= 0$) across all datasets, covering both recognition and VQA tasks. We convert the object recognition task to a four-way multiple-choice questioning task, following Gavrikov et al. [6]. For the VQA tasks, we employ the official dataset prompts and append the same multiple-choice instruction to standardize responses. Two exceptions are made: for Capture, we use free-form answers as recommended by Pothiraj et al. [14], and for MME, we convert yes/no questions into a multiple-choice format. Performance is measured by accuracy against the ground truth for recognition and VQA tasks, while for Capture we report $1 -$ symmetric mean percentage error [14]. For all the zero-shot results we do not employ any chain-of-thought prompting [19], because that evaluation setting is more fair with the setting employed in our work.

2. Dataset Description

To comprehensively evaluate our method, we curated a diverse set of recognition and VQA benchmarks that span multiple task-specific challenges. Table 1 provides detailed statistics of the datasets used in our experiments, including both the original test sizes and the number of images retained after preprocessing.

We employed several widely used recognition datasets that test the robustness and generalization capability of models across distribution shifts. Specifically, we included ImageNet [5], ImageNet-V2 [15], and ImageNet-A [8] to capture generic object recognition in both standard and adversarial settings. In addition, ImageNet-Sketch [16] and ImageNet-R [7] were incorporated to examine robustness under edge-based and texture-based distortions, respectively. To further assess fine-grained and material recognition, we used Food101 [2] and DTD [4], which emphasize category-level detail and texture variation.

To test higher-level reasoning, we included a wide range of VQA datasets spanning mathematical ability, general un-

Dataset	Used Test Size	Original Test Size	Focus
ImageNet-A [8]	7,467	7,500	Generic
ImageNet-V2 [15]	9,772	10,000	Generic
ImageNet [5]	49,032	50,000	Generic
ImageNet-Sketch [16]	35,350	50,000	Edges
ImageNet-R [7]	28,506	30,000	Texture
DTD [4]	5640	5,640	Edges, Texture
Food101 [2]	25,250	25,250	Fine-grained
Resisc45 [3]	4,500	4,500	Satellite Imagery
Mathverse (mcq) [21]	1631	2180	Mathematical Ability
Mathvista [12]	490	1000	Mathematical Ability
Seed [10]	3,881	13,991	General Understanding
MME [20]	1576	2,370	General Understanding
RealworldQA [1]	765	765	Realworld Understanding
Capture [14]	817	962	Counterfactual Understanding
CRPE [17]	7575	7575	Compositionality and Halluncination
AI2D [9]	2704	3090	Grpah and Chart Understanding

Table 1. **Statistics of Recognition and VQA datasets** used in TTRV. We drop images with resolution higher than 1000×1000. Hence, we report both i) the original number of test images and ii) the used number of test images (*i.e.*, those below the 1000×1000 threshold).

derstanding, and compositional reasoning. Mathematical reasoning was evaluated using Mathverse [21] and MathVista [12], while Seed [10] and MME [20] were selected for general multimodal understanding. RealWorldQA [1] was used to benchmark models against real-world scenarios, where all images were first standardized to a maximum resolution of 1000×1000 for consistency across experiments. We also included Capture [14] to probe counterfactual reasoning, CRPE [17] to evaluate compositionality and hallucination resistance, and AI2D [9] to study performance on diagram, graph, and chart-based understanding tasks.

For computational resons, we filtered out images exceeding a resolution of 1000×1000 pixels across all datasets, retaining only those within this threshold. The reported "used test size" in Table 1 reflects this preprocessing step. In particular, for the RealWorldQA dataset, where image dimensions were highly inconsistent, we explicitly resized all images to 1000×1000 resolution to ensure compatibility with our evaluation pipeline.

Overall, the curated dataset collection provides a broad coverage of recognition, reasoning, and real-world understanding challenges, allowing us to rigorously evaluate the generalization capability of our proposed approach.

3. TTRV Prompt Details

In this section, we provide the prompts used in our experiments. For each dataset, we present a representative example of the prompt employed in our study. While the specific prompts may vary depending on the nature of the question, particularly in VQA tasks, we provide a general outline illustrating the structure and format of the prompts used across different datasets.

- **ImageNet:**

```
<image> \n Look at the given image and
 ↪  identify what it shows.
Choose the correct answer from the options
 ↪  below and respond
with only the corresponding option letter
 ↪  (A, B, C, or D).
Do not include any explanation or extra
 ↪  text.
\n Options:\nA. neck brace\nB. shopping
 ↪  cart\nC. guillotine
\nD. garbage truck.
```

- **ImageNet-V2:**

```
<image> \n Look at the given image and
 ↪  identify what it shows.
Choose the correct answer from the options
 ↪  below and respond with
only the corresponding option letter (A,
 ↪  B, C, or D). Do not
include any explanation or extra text. \n
 ↪  Options:\nA. cuirass
\nB. dial telephone, dial phone\nC.
 ↪  beaver\nD. desk.
```

- **ImageNet-R:**

```
<image> \n Look at the given image and
 ↪  identify what it shows.
Choose the correct answer from the options
 ↪  below and respond with
only the corresponding option letter (A,
 ↪  B, C, or D). Do not
```

```
include any explanation or extra text. \n
↪  Options:\nA. skunk
\nB. panda\nC. german_shepherd_dog\nD.
↪  orangutan.
```

- **ImageNet-S:**

```
<image> \n Look at the given image and
↪  identify what it shows.
Choose the correct answer from the options
↪  below and respond with
only the corresponding option letter (A,
↪  B, C, or D). Do not
include any explanation or extra text. \n
↪  Options:\nA. lab coat
\nB. cheetah\nC. ptarmigan\nD. canoe.
```

- **ImageNet-A:**

```
<image> \n Look at the given image and
↪  identify what it shows.
Choose the correct answer from the options
↪  below and respond with
only the corresponding option letter (A,
↪  B, C, or D). Do not
include any explanation or extra text. \n
↪  Options:\nA. feather boa
\nB. garter snake\nC. soap dispenser\nD.
↪  tank.
```

- **Food101:**

```
<image> \n Look at the given image and
↪  identify what it shows.
Choose the correct answer from the options
↪  below and respond with
only the corresponding option letter (A,
↪  B, C, or D). Do not
include any explanation or extra text. \n
↪  Options:\nA. Greek
salad\nB. Red velvet cake\nC. Bibimbap\nD.
↪  Pork chop.
```

- **DTD:**

```
<image> \n Look at the given image and
↪  identify what texture it
shows. Choose the correct answer from the
↪  options below and
respond with only the corresponding option
↪  letter (A, B, C, or D).
Do not include any explanation or extra
↪  text. \n Options:\nA.
banded\nB. crosshatched\nC. freckled\nD.
↪  marbled.
```

- **Resisc45:**

```
<image> \n Look at the given image and
↪  identify what it shows.
Choose the correct answer from the options
↪  below and respond with
only the corresponding option letter (A,
↪  B, C, or D). Do not
include any explanation or extra text. \n
↪  Options:\nA. industrial
area\nB. sea ice\nC. circular farmland\nD.
↪  golf course.
```

- **Mathverse:**

```
<image> \nPlease directly answer the
↪  question and provide the
correct option letter, e.g., A, B, C,
↪  D.\nDo not include any
explanation or extra text.\nQuestion:
↪  Emile is observing a wind
turbine. The vertical distance between the
↪  ground and the tip of
one of the turbine's blades, in meters, is
↪  modeled by $H(t)$
where $t$ is the time in seconds. What is
↪  the meaning of the
highlighted segment?\nChoices:\nA:The
↪  turbine's center is 35
meters above the ground.\nB:The turbine
↪  completes a single
cycle in 35 seconds.\nC:The length of the
↪  blade is 35 meters.
\nD:The turbine has 35 blades.
```

- **Mathvista:**

```
<image> \nHint: Please answer the question
↪  and provide the correct
option letter, e.g., A, B, C, D, at the
↪  end.\nDo not include any
explanation or extra text.\nQuestion: In
↪  the figure above,
triangle ABC is inscribed in the circle
↪  with center O and diameter
AC. If AB = AO, what is the degree measure
↪  of angle ABO \nChoices:
\n(A) 15*\\degree\n(B) 30*\\degree\n(C)
↪  45*\\degree\n
(D) 60*\\degree\n(E) 90*\\degree.
```

- **SEED:**

```
<image> \nWhat is the main color of the
↪  dress worn by the woman
with the density value of [0.6536, 0.5600,
↪  0.7431, 0.7743]?
\nChoose the correct answer from the
↪  options below and respond
with only the corresponding option letter
↪  (A, B, C, or D). Do not
include any explanation or extra
↪  text.\nOptions:\nA. Red\nB.
```

```
None of the above\nC. Brown\nD. Tan
```

- **MME:**

```
<image> \n Is this photo taken in a place
 ↪  of home office? Please
answer with yes or no. Please respond with
 ↪  only the corresponding
option letter (A or B). Do not include any
 ↪  explanation or extra
text. \n Options:\nA. No \nB. Yes.
```

- **RealWorldQA:**

```
<image> \nHow many lanes are there on the
 ↪  left?\n\nOptions are:
\nA. 4\nB. 3\nC. 2\nD. 5\n\nPlease answer
 ↪  directly with only the
letter of the correct option and nothing
 ↪  else.",
```

- **Capture:**

```
<image>\nCount the exact number of
 ↪  sunglasses in the image. Assume
the pattern of sunglasses continues behind
 ↪  any black box. Provide
the total number of sunglasses as if the
 ↪  black box were not
there.\nPlease reason step by step, and
 ↪  put your final
answer within \\boxed{}.
```

- **CRPE:**

```
<image> \nWhat is the person in front
 ↪  of?\nA. The person is in
front of the person.\nB. The person is in
 ↪  front of the tree.\nC.
The person is in front of the mirror.\nD.
 ↪  The person is in front
of the shelf.\nAnswer with the option's
 ↪  letter from the given
choices directly and do not include any
 ↪  explanation or extra text.
```

- **AI2D:**

```
<image> \nWhat is the line that divides
 ↪  the two different plates?
\nChoose the correct answer from the
 ↪  options below and respond
with only the corresponding option letter
 ↪  (A, B, C, or D). Do not
include any explanation or extra
 ↪  text.\n\nOptions:\nA. ground\nB.
fault line\nC. dirt\nD. earthquake.
```

4. Additional Experiments

4.1. Latency versus Number of Samples

To further substantiate our claims, we conducted experiments aimed at analyzing the adaptability of the model under varying conditions. The first experiment investigates how the model's performance changes when it is allowed to adapt using different numbers of training samples. Specifically, we compare the outcomes when the model is adapted with only 20 samples versus when it is adapted with 500 samples. As shown in Table 2, the results demonstrate a consistent improvement in performance as the number of adaptation samples increases. This suggests that providing the model with a richer set of examples enables it to better align with the target task, thereby yielding higher accuracy and robustness. However, this improvement does not come without trade-offs. Increasing the number of samples also leads to a higher computational burden, both in terms of memory consumption and processing time. This is particularly important in real-world applications where inference latency is a critical factor. To quantify this trade-off, we conducted an additional experiment measuring the latency associated with adaptation on different sample sizes. The results in Table 3, reveal that while larger adaptation sets enhance task performance, they simultaneously increase the time required for inference, thereby highlighting an inherent balance between accuracy and efficiency.

All experiments were conducted using the vLLM inference engine, one of the fastest and most recent frameworks for large language model inference. Despite its efficiency, optimized inference remains an active research area, and ongoing improvements in frameworks such as vLLM are expected to further reduce latency. Moreover, the reported times are dependent on the underlying hardware. Access to more powerful GPUs would likely accelerate both inference and adaptation, thereby reducing the overall time required for these tasks.

4.2. Robustness

To evaluate the robustness of our method, we conducted experiments to measure the variance in its performance. The results are summarized in Table 4. As shown, our method, when evaluated using greedy decoding, demonstrates strong robustness and is only subject to minor variations attributable to hardware and software factors.

4.3. Further Cross-Data Generalization Examples

In addition to the results shown in Figure 3 (main manuscript), we present challenging cross-dataset evaluation results in Table 5. Our method consistently improves performance across all transfer settings, demonstrating its effectiveness not only in within-domain accuracy but also in transferring knowledge across diverse domains. For in-

	Food101	DTD	Resisc45	ImageNet	ImageNetv2	ImageNetR	ImageNetS	ImageNetA
InternVL3-2B	67.19	37.24	72.28	56.00	67.43	66.01	62.19	67.92
TTRV 20 samples	95.60	89.73	90.06	98.31	98.25	96.89	94.74	96.31
TTRV 500 samples	96.20	89.99	93.67	98.89	98.95	97.85	95.54	96.38
InternVL3-8B	78.32	59.11	83.62	79.47	62.58	59.32	54.48	57.03
TTRV 20 samples	87.13	77.92	89.19	91.43	85.27	63.51	53.79	81.43
TTRV 500 samples	97.20	89.37	93.82	99.31	97.24	96.88	95.03	96.86

Table 2. **Number of Samples for Adaptation.** Top-1 Accuracy (%) obtained by sampling varying data points from the test data.

	Latency (avg $\pm$ std)	Overhead vs. Normal Inference	% Increase
Normal Inference	25.5 ± 4.5 s	–	–
Adaptation:			
1 sample	2.75 ± 0.43 m	$\approx +2.7$ m	$\approx 547\%$
20 samples	3.77 ± 0.63 m	$\approx +3.8$ m	$\approx 786\%$
500 samples	1 hr 38 m $\pm$16 m	$\approx +1$ hr 38 m	$\approx 23{,}000\%$

Table 3. **Computation Overhead.** Inference and adaptation latency through TTRV. Seconds: s, Minutes: m, Hours: h.

	DTD	Imagenet-A	Imagenet-V2	AI2D	Mathverse
InternVL2.5-4B	46.78$\pm$ 0.03	90.58 $\pm$ 0.05	83.01 $\pm$ 0.03	51.73 $\pm$ 0.01	52.67 $\pm$ 0.55
w/ TTRV	81.87$\pm$ 0.80	96.09 $\pm$ 0.01	96.77 $\pm$ 0.06	64.75 $\pm$ 2.34	53.59 $\pm$ 0.45

Table 4. **Variance of Results.** Results obtained by employing TTRV across 5 independent runs.

	Mathvista $\rightarrow$ Mathverse	Food $\rightarrow$ Mathvista	DTD $\rightarrow$ Seed	IN-V2 $\rightarrow$ IN-R	IN-V2 $\rightarrow$ IN-A	IN-A $\rightarrow$ IN-V2
InternVL2.5-4B	51.69	65.49	56.25	79.53	90.67	83.07
w/ TTRV	52.00	67.14	59.07	95.42	96.30	96.14
Δ	+0.31	+2.52	+2.02	+15.89	+5.63	+13.07

Table 5. **Cross-dataset generalization**. Performance on different dataset combinations, where X $\rightarrow$ Y denotes training on dataset X and testing on dataset Y. "IN" in the table refers to ImageNet.

	Mathverse	Mathvista	Seed	AI2D	MME	RealWorldQA	CRPE
MM-Eureka-7B	60.69	78.92	78.02	81.60	87.55	68.25	73.80
w/ TTRV	61.44	80.94	78.66	82.00	88.28	68.81	74.41
Δ	+0.75	+2.02	+0.64	+0.40	+0.73	+0.56	+0.61
ThinkLite-VL-7B	64.41	78.87	77.82	82.10	87.11	70.00	73.00
w/ TTRV	64.48	80.43	78.63	83.40	87.63	71.40	73.71
Δ	+0.07	+1.56	+0.81	+1.30	+0.52	+1.40	+0.71
VisionReasoner-7B	62.30	78.71	77.10	82.60	86.46	70.33	72.31
w/ TTRV	62.88	80.45	77.83	83.44	86.85	71.11	74.21
Δ	+0.58	+1.74	+0.73	+0.84	+0.39	+0.78	+1.90

Table 6. **Generalization to Model Families.** We provide results for MM-Eureka, ThinkLite-VL and VisionReasoner.

stance, training on ImageNet-V2 leads to significant gains when tested on ImageNet-R (+15.89%) and ImageNet-A (+5.63%). Similarly, training on ImageNet-A improves performance on ImageNet-V2 by +13.07%. We also observe

positive transfer in mathematical reasoning tasks, with a gain of +0.31% when training on MathVista and evaluating on MathVerse. Notably, even when models are trained on visual recognition datasets and evaluated on VQA benchmarks, such as training on Food and testing on MathVista, we achieve a performance improvement of +2.52%. These results indicate that our approach enhances visual understanding in a way that generalizes well across heterogeneous tasks and domains.

4.4. Additional Models

Apart from the results presented on Qwen in Table 7 (main manuscript) and InternVL in the main table. In Table 6 we also report additional results on three new models: MM-Eureka [13] and ThinkLite-VL [18] and VisionReasoner [11]. These results demonstrate consistent improvements and further reinforce our claim that the proposed method is model-agnostic.

5. Pseudocode

The following pseudocode illustrates the main steps of our method, including rollout generation, reward computation, advantage estimation, and policy update.

Pseudocode: Test-Time Reinforcement Learning for Vision Language Models

```
def inference_time_grpo(model, test_sample, N=32, alpha=0.75, lr=0.001):
   ""
   Args:
      model: decoder-based VLM with parameters theta
      test_sample: (x) consisting of image + text prompt
      N: number of rollouts per test sample
      alpha: weight for entropy regularization
      lr: learning rate for policy update

   Returns:
      updated_model: model with adapted parameters
   ""

   # 1. Generate rollouts
   responses = [model.sample(test_sample) for _ in range(N)]
   unique_responses = set(responses)

   # 2. Empirical probabilities
   freq = {y: responses.count(y) for y in unique_responses}
   probs = {y: freq[y]/N for y in unique_responses}

   # 3. Compute rewards
   r1 = {y: probs[y] for y in unique_responses} # Frequency-based reward
   H = -sum(p * log(p) for p in probs.values())
   r2 = -H # diversity control reward
   R = {y: r1[y] + alpha * r2 for y in unique_responses} # total reward

   # 4. Convert rewards -> relative advantages
   mean_R = sum(R[y] for y in responses) / len(responses)
   std_R = (sum((R[y] - mean_R)**2 for y in responses) / len(responses))
       **0.5
   if std_R < 1e-8: # avoid divide by zero
      A = {y: 0.0 for y in responses}
   else:
      A = {y: (R[y] - mean_R) / std_R for y in responses}

   # 5. Policy update (GRPO)
   grad_estimate = 0
   for y in responses:
      logprob = model.log_prob(test_sample, y)
      grad_estimate += A[y] * grad(logprob, model.params)

   grad_estimate /= N

   for param in model.params:
      param += lr * grad_estimate[param]

   return model
```

References

[1] X AI. Grok-1.5 Vision Preview, 2024. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – Mining Discriminative Components with Random Forests. In *Proc. ECCV*, 2014. 1, 2

[3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. In *Proc. IEEE*, 2017. 2

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *Proc. CVPR*, 2014. 1, 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 1, 2

[6] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Bianca Lamm, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Are Vision Language Models Texture or Shape Biased and Can We Steer Them? *arXiv preprint arXiv:2403.09193*, 2024. 1

[7] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proc. ICCV*, 2021. 1, 2

[8] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *Proc. CVPR*, 2021. 1, 2

[9] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram Is Worth A Dozen Images. In *Proc. ECCV*, 2016. 2

[10] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2

[11] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025. 6

[12] 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. In *Proc. ICLR*, 2024. 2

[13] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 6

[14] Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. CAPTURe: Evaluating Spatial Reasoning in Vision Language Models via Occluded Object Counting. *arXiv preprint arXiv:2504.15485*, 2025. 1, 2

[15] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proc. ICML*, 2019. 1, 2

[16] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning Robust Global Representations by Penalizing Local Predictive Power. In *NeurIPS*, 2019. 1, 2

[17] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The All-Seeing Project V2: Towards General Relation Comprehension of the Open World. In *Proc. ECCV*, 2024. 2

[18] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025. 6

[19] 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*, 2022. 1

[20] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*, 2023. 2

[21] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? In *Proc. ECCV*, 2024. 2