Multimodal Reasoning through Reinforcement Learning

Jakhongir Saydaliev
EPFL

Abstract

We study three paradigmsn of multimodal chain-of-thought (CoT) reasoning: *Multimodal-to-Multimodal*, *Text-to-Multimodal*, and *Multimodal-to-Text*. Using Group Relative Policy Optimization (GRPO), our models learn reasoning strategies without annotated chains, guided by task-specific rewards. Experiments on each paradigm-specific VQA datasets reveal that image generation during reasoning often hurts performance, while *Multimodal-to-Text* with visual grounding improves results (e.g., +1.34 F1 on A-OKVQA). We also scale our approach for *Multimodal-to-Text* to 120K samples across 10 datasets, offering some practical insights into multimodal reasoning.

1 Introduction

Chain-of-Thought (CoT) prompting has revolutionized the reasoning capabilities of Large Language Models (LLMs) by enabling them to generate explicit intermediate reasoning steps (Wei et al., 2023). This breakthrough has led to significant improvements in mathematical reasoning, logical deduction, and complex problem-solving tasks. As multimodal models have emerged, researchers have naturally sought to extend CoT reasoning to tasks involving both visual and textual information.

Current approaches to multimodal reasoning primarily follow established patterns from text-only CoT, where models generate textual reasoning chains even when processing visual inputs. However, this raises fundamental questions if visual and textual modalities can both be used during the reasoning process

In this work, we study 3 paradigms for multimodal chain-of-thought reasoning:

Multimodal-to-Multimodal Reasoning: In this paradigm, models receive multimodal inputs (image + text) and generate reasoning chains that interleave both textual and visual thoughts before producing a final textual answer. This approach mirrors human cognition, where visual mental imagery often accompanies verbal reasoning processes.

Text-to-Multimodal Reasoning: Here, models start with purely textual inputs but generate multimodal reasoning chains that include both textual explanations and visual representations. This paradigm is particularly relevant for scenarios where generating mental images can help in problem-solving.

Multimodal-to-Text Reasoning: This paradigm takes multimodal inputs but constrains the reasoning process to textual chains, along with with visual grounding elements (i.e. bounding boxes) that reference specific regions in the input images.

A critical challenge in training models for these reasoning paradigms lies in the scarcity of datasets with ground-truth multimodal reasoning chains. Traditional supervised fine-tuning approaches require extensive annotations of intermediate reasoning steps, which are expensive to collect and often subjective in nature. To address this limitation, we leverage Group Relative Policy Optimization (GRPO) (Shao et al., 2024b), a reinforcement learning method that has recently shown remarkable success in training reasoning models (DeepSeek-AI et al., 2025).

GRPO enables models to learn reasoning strategies through reward-based optimization rather than supervised imitation, requiring only input-output pairs without annotated reasoning chains. We design specific reward functions for each reasoning paradigm, incorporating measures of accuracy, format compliance, visual-textual alignment, and visual grounding quality.

In our experiments we find that generating images during reasoning (Multimodal-to-Multimodal and Text-to-Multimodal paradigms) often degrades the performance compared to purely textual reasoning, while Multimodal-to-Text reasoning with visual grounding achieves the most consistent im-

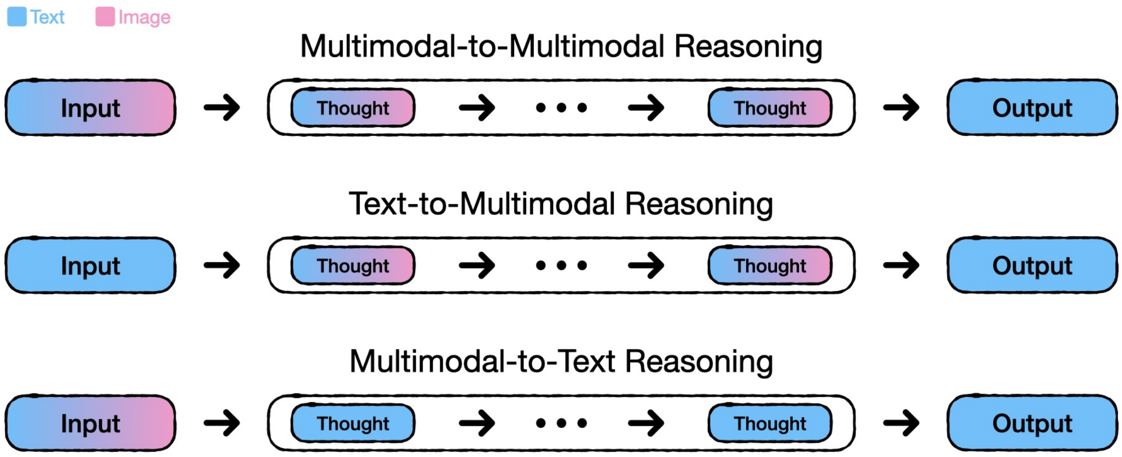


Figure 1: 3 variations of multimodal reasoning paradigms.

provements, particularly when bounding boxes are incorporated into textual reasoning chains to ground the model's attention on specific visual regions.

2 Related Work

Multimodal Chain-of-Thought Reasoning Chain-of-Thought (CoT) prompting (Wei et al., 2023) has significantly improved the reasoning capabilities of large language models (LLMs). To extend CoT to multimodal models, recent research has proposed a variety of approaches. Some methods adopt a two-stage process, where visual information is first transformed and grounded into textual representations (Zhang et al., 2024), structured graphs such as scene graphs (Mitra et al., 2024) or knowledge graphs (Mondal et al., 2024), or bounding boxes (Lei et al., 2024), before initiating the reasoning process. Other works leverage the generative capabilities of multimodal models and directly fine-tune them for Multimodal Chain-of-Thought reasoning (Li et al., 2025; Wu et al., 2024). In this work, we explore different paradigms of Multimodal CoT using Reinforcement Learning (RL), moving beyond reliance on supervised instruction datasets.

GRPO for Vision-Language Models Following the recent success of DeepSeek R1 (DeepSeek-AI et al., 2025), numerous studies have investigated the application of the Group Relative Policy Optimization (GRPO) method (Shao et al., 2024b) to Vision-Language Models (VLMs). For instance, Liao et al. (2025) applies GRPO to enhance visual-spatial reasoning through various prompting strategies, while Zhou et al. (2025) reports a similar "Aha Moment" in a 2B-parameter VLM. In the video domain, Feng et al. (2025) introduces a temporal-aware GRPO variant for video-based reasoning. Furthermore, Shen et al. (2025) demonstrates that GRPO achieves better generalization on out-of-distribution (OOD) datasets, whereas supervised fine-tuning performs better on in-domain data for certain tasks. In contrast, our work focuses on applying GRPO to the Visual Question Answering (VQA) task, incorporating bounding-boxes directly within the reasoning chain of VLMs.

3 Multimodal Chain-of-Thought Reasoning Framework

Humans often create mental imagery to aid decision-making. Instead of relying solely on verbal thought proxies, a multimodal chain-of-thought (CoT) approach enables models to reason by interleaving both visual and textual thoughts during intermediate reasoning steps. This multimodal reasoning process enhances the model's expressiveness and mirrors human cognition more naturally.

We formally define the reasoning process as follows. Let P_θ be a pre-trained Multimodal Large Language Model (MLLM) with parameters θ, x an input (either text, image, or both), and z, v the generated sequences of textual and visual thoughts, respectively, any y the final output. At each reasoning step i, the model generates intermediate thoughts that may include a textual component $\hat{z}_i$ and a visual component $\hat{v}_i$. The next step is conditioned on all previous thoughts and visualizations, as defined in Equations 1 and 2:

$$\hat{v}_i \sim P_\theta(v_i \mid \hat{z}_1, \hat{v}_1, \ldots, \hat{v}_{i-1}, \hat{z}_i) \quad (1)$$

$$\hat{z}_{i+1} \sim P_\theta(z_{i+1} \mid x, \hat{z}_1, \hat{v}_1, \ldots, \hat{z}_i, \hat{v}_i) \quad (2)$$

This training strategy enables the model to align interleaved reasoning traces with corresponding visualizations, enhancing its ability to solve complex multimodal tasks. Depending on the modality of the input, reasoning chain, and output, this framework supports the following multiple reasoning types as illustrated in Figure 1.

Multimodal-to-Multimodal Reasoning In this setting, both the input and reasoning steps are multimodal (image + text), while the final output is textual, i.e. *input*: $x = (x_{\text{text}}, x_{\text{image}})$; *thoughts*: $\hat{z}_i$, $\hat{v}_i$; *output*: $y = y_{\text{text}}$.

Text-to-Multimodal Reasoning Here, the model receives a Textual input, generates multimodal intermediate thoughts and concludes with a textual answer, i.e. *input*: $x = x_{\text{text}}$; *thoughts*: $\hat{z}_i$, $\hat{v}_i$; *output*: $y = y_{\text{text}}$.

Multimodal-to-Text Reasoning In this case, the input is multimodal (image + text), but the reasoning chain and final output remain purely textual, i.e. *input*: $x = (x_{\text{text}}, x_{\text{image}})$; *thoughts*: $\hat{z}_i$; *output*: $y = y_{\text{text}}$.

4 Training Methodology

In this section we focus on Autoregressive MLLMs for both training and inference and describe the training design choices for each reasoning paradigm.

4.1 Multimodal-to-Multimodal Reasoning

We follow the architecture of Chameleon (Team, 2025), which leverages a unified Transformer to process both image and text tokens. The architecture integrates two tokenizers: an image tokenizer based on ERO21 (Esser et al., 2021) and a text tokenizer, which convert images and text into discrete token sequences, respectively. The image tokenizer uses a discrete codebook to encode input images into a sequence of image tokens, while the text tokenizer maps textual data into corresponding token sequences. These token sequences are concatenated and processed by a causal transformer model. We fine-tune the causal Transformer model using the next-token prediction objective, while the image tokenizer and text tokenizer are kept frozen throughout the process.

4.2 Text-to-Multimodal Reasoning

Similar to 4.1, we use an MLLM with unified Transformer for both image and text tokens. However, instead of directly fine-tuning it on a reasoning dataset, we use a recent reinforcement learning method Group Relative Policy Optimization (GRPO) (Shao et al., 2024b), that is, given an input (x), and the final output (y), we train the model to learn the intermediate reasoning steps (z, v) by itself, through appropriate reward functions. The use of GRPO over supervised fine-tuning (SFT) has the following 2 advantages:

- It can be trained on a dataset without the ground-truth reasoning chains
- GRPO has been shown to generalize better to out-of-domain datasets over SFT (Shen et al., 2025)

We train the base model using GRPO with following rule-based reward functions:
Formatting: Checks if the reasoning and the final response are formatted within <think> ...<\think> and <answer>...<\answer> tags. 1 if the format is followed, otherwise 0.
Accuracy: 1 if the final response is correct, otherwise 0.
Image-Text Alignment: Cosine similarity between the embeddings of the scene description and a generated image (if any). Rescaled to a range between 0 and 1.
 # Images: Number of images the model generates as part of its reasoning trace. 0 for no image generation, 1 for a single image, and exponentially decreases as the number of images increases.

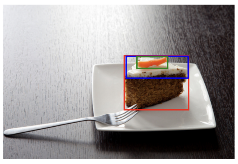
4.3 Multimodal-to-Text Reasoning

Following the same reason as in 4.2, we train a Vision Language Model (VLM) using GRPO. For this task, our goal is to train a VLM to generate reasoning chains interleaved with bounding boxes, which, we hypothesize, will improve visual question answering task performance (Figure 2). As illustrated in Figure 3, our training pipeline consists of 2 stages:

- **SFT-warmup:** We train the base model for to generate bounding boxes as part of its reasoning chain.

Question:
What kind of cake has been served?

A.cinnamon
B.red velvet
C.chocolate
D.carrot



Model's response:
</think>The cake [333.94, 140.46, 178.03, 149.82] and icing [338.24, 141.41, 172.79, 61.11] of the cake gives it away for what type of cake it is, and the carrot [369.06, 142.95, 84.71, 32.57] confirms that it is a carrot cake.</think> <answer>**D**</answer>

Figure 2: An example of Multimodal-to-Text Reasoning with interleaved bounding-boxes.

- **GRPO:** starting from the SFT-warmup model, we apply GRPO for using 3 reward functions.

During the GRPO training, we apply the following 3 reward functions:
Accuracy: 1 if the final response is correct, otherwise 0.
Format: Checks if the reasoning and the final response are formatted within `<think>`...`<\think>` and `<answer>`...`<\answer>` tags. 1 if the format is followed, otherwise 0.
IoU Score: Average intersection over union (IoU) score of the generated and ground-truth bounding-boxes.

5 Experimental Setup

5.1 Multimodal-to-Multimodal Reasoning

Data For Multimodal-to-Multimodal Reasoning training we use PuzzleVQA (Chia et al., 2024), a collection of puzzles based on abstract patterns. This is a dataset of synthetically generated puzzle questions based on fundamental concepts, including colors, numbers, sizes, and shapes. PuzzleVQA provides a code to generate question, reasoning and answer for 18 different puzzles. We use their code to generate up to 10000 examples for each puzzle type, and use 5 of them as our training data, and the rest as out-of-domain evaluation dataset. More information about PuzzleVQA dataset construction, along with examples are provided in Appendix A.

Model We use Anole7B (Chern et al., 2024) model as the backbone in this task. Anole is tuned on Chameleon (Team, 2025) and can generate interleaved text and image, making it well-suited for Multimodal-to-Multimodal Reasoning. We only tune part of the model's parameters with LoRA (Hu et al., 2021) in an instruction tuning manner for 20 epochs, where only the loss from the predictions is optimized. In addition to Multimodal Reasoning, we also fine-tune with Textual Reasoning only as our baseline, and compute the accuracy.

5.2 Text-to-Multimodal Reasoning

Data For Text-to-Multimodal Reasoning training we use ReSQ (Mirzaee and Kordjamshidi, 2022). ReSQ is a human-generated dataset of 1000 samples, where each question contains a description of a scene, followed by a question on spatial relationship of objects in the scene. For example:

Question: A red car is parking in front of a grey house with brown window frames and plants on the balcony. Are the plants in front of the car?
Answer: No

Model We use SEED-LLaMA-8B (Ge et al., 2023) as our base model. This is an autoregressive MMLM pretrained on image-text interleaved datasets such as MMC4[1], OBELISC[2], and further fine-tuned on CoMM[3], which contains 4 images on average per example. The advantage of this model is that it represents an image using only 32 tokens, while having a context length of 2048 tokens. It allows to us to use more number of images while reasoning.

Baselines We compare our method to the following baselines: A direct prompting of the baseline model with CoT, and a model trained with GPRO using Textual reasoning (z) with *accuracy* and *format* rewards. We also use Dall-E 3[4] and GPT-4o[5] as a strong baseline. For a subset of train set, we generate 3 images using Dall-E 3 and pass each image to GPT-4o and sample 3 outputs for each image, ending up with 9 samples per question. We then run evaluation in Textual (z) and Multimodal (z, v) manners.

Metrics During evaluation, we generate 5 samples for each problem and compute the average pass@1 and pass@3 metrics as in Equation 3.

$$\text{Pass@}k = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \tag{3}$$

[1]https://github.com/allenai/mmc4
[2]https://github.com/huggingface/OBELICS
[3]https://github.com/HKUST-LongGroup/CoMM?tab=readme-ov-file
[4]https://openai.com/index/dall-e-3/
[5]https://openai.com/index/hello-gpt-4o/

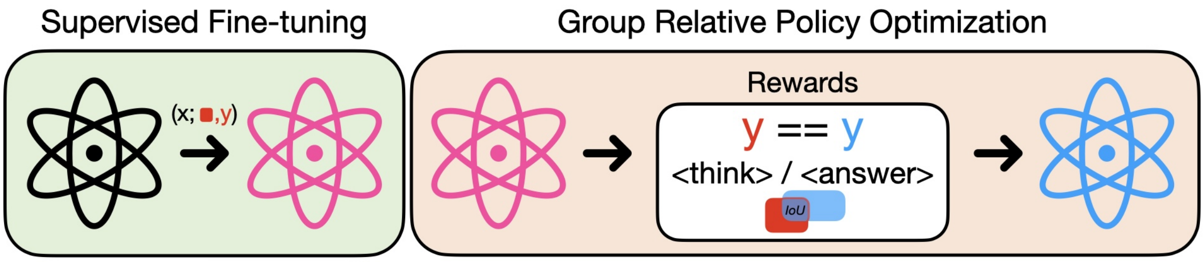


Figure 3: Overview of the method used for Multimodal-to-Text Reasoning training.

where n is the total number of attempts (5 in our case), and c is the number of correct solutions.

5.3 Multimodal-to-Text Reasoning

Data For Multimodal-to-Text Reasoning we use DrivingVQA (Corbière et al., 2025) and A-OKVQA (Schwenk et al., 2022). DrivingVQA and A-OKVQA are visual question answering dataset of ~3000 and ~17000 examples respectively. The choice of using these datasets is based on the availability of ground-truth reasoning traces with interleaved bounding-boxes, made available by (Corbière et al., 2025). The reasoning chains are necessary for the initial SFT-warmup step.

Model We use Qwen2.5-VL-7B (Bai et al., 2025) as our base model, as its pre-training data also contains bounding-box formats for grounding. We train the base model using LoRA in 2 stages as described in 4.3.

Metrics We report the F1 score as the main metrics. Additionally, during our analysis, we observed that sometimes the reasoning of the model contradicts its final answer. We quantified this with an LLM-as-judge protocol using the OpenAI model GPT-4.1[6]. We provide the original question, ground-truth answer, generated reasoning chain, and final answer to the judge, and ask it to assess whether the reasoning supports the answer or contradicts it. Finally we compute an alignment score as the ratio of questions where the final answer aligns with the reasoning chain over the number of all question in the test set. For DrivingVQA, we also report the Exam Score following Corbière et al. (2025).

Puzzle Type	Domain	Multimodal	Textual
Rectangle-Height-Color	ID	0.90	0.98
Polygon-Sides-Number	ID	0.69	0.78
Grid-Number-Color	ID	0.61	0.69
Color-Number-Hexagon	ID	0.26	0.66
Triangle	ID	0.41	0.61
Color-Size-Circle	OOD	0.08	0.35
Shape-Size-Hexagon	OOD	0.11	0.29
Color-Hexagon	OOD	0.34	0.29
Shape-Size-Grid	OOD	0.15	0.29
Polygon-Sides-Color	OOD	0.29	0.28
Shape-Morph	OOD	0.14	0.25
Color-Grid	OOD	0.47	0.25
Venn	OOD	0.00	0.24
Size-Cycle	OOD	0.04	0.23
Shape-Reflect	OOD	0.09	0.21
Size-Grid	OOD	0.02	0.16
Color-Overlap-Squares	OOD	0.12	0.08
Circle-Size-Number	OOD	0.00	0.00
Grid-Number	OOD	0.00	0.00
Rectangle-Height-Number	OOD	0.00	0.00
Mean	-	0.24	**0.33**

Table 1: Accuracy scores of Anole7B model trained on PuzzleVQA using Multimodal and Textual-only reasonings. The Puzzle Types included in the train set are set to in-domain (ID), and those that are not included are set to out-of-domain (OOD).

6 Results & Discussion

6.1 Multimodal-to-Multimodal Reasoning

The experiment results are given in Table 1. First, we can see that when trained with Multimodal or Textual reasoning, the model performs rather poorly for the out-of-domain datasets. We can also see that across almost all puzzle types, both for in-domain and out-of-domain, the model trained with Textual reasoning outperforms the model with Multimodal reasoning. It shows that the generating interleaved images is rather degrading the model's performance.

6.2 Text-to-Multimodal Reasoning

We report the results in Table 2. We observe that including image-based rewards (Image-Text Alignment, # Images) during GRPO training is degrading the performance. Additionally, training with Multimodal reasoning almost always underperforms

[6]https://openai.com/index/gpt-4-1/

Model	Training	Modality	Rewards				Metrics	
			Accuracy	Format	Image-Text Align.	# Images	Pass@1	Pass@3
SEED-LLaMA-CoT	✗	textual	–	–	–	–	0.079	0.213
SEED-LLaMA-GRPO	✓	textual	✓	✓	✗	✗	0.450	0.745
SEED-LLAMA-GRPO	✓	multimodal	✓	✓	✓	✓	0.329	0.659
SEED-LLAMA-GRPO	✓	multimodal	✓	✓	✓	✗	0.387	0.721
SEED-LLAMA-GRPO	✓	multimodal	✓	✓	✗	✗	0.411	0.760
GPT-4o + DALL-E	✗	textual	–	–	–	–	0.761	0.822
GPT-4o + DALL-E	✗	multimodal	–	–	–	–	0.695	0.808

Table 2: Results of training with Text-to-Multimodal reasoning on ReSQ dataset. Whether a model is trained or not is given under the Training column, and the modalities used is given under the Modality column.

the Textual reasoning. It suggests that including images during reasoning is degrading the model's performance, which aligns with the observation we have from 6.1. It is also further highlithed by the Textual and Multimodal reasonings comparison of Dall-E 3 and GPT-4o, where Textual reasoning outperforms the Multimodal counterpart.

6.3 Multimodal-to-Text Reasoning

As reported in Table 3, the GRPO method outperforms the SFT baseline on both datasets.

Interestingly, however, after the GRPO stage the alignment score reduces by ~7% for the DrivingVQA dataset. To evaluate it further we did a human evaluation of 50 samples from both DrivingVQA and A-OKVQA datasets, and computed the Human agreement score as a ratio of number of examples where the LLM-Judge agrees with the Human over 50 samples. The results are given in Table 4. As we can see there is a low agreement when the Judge predicts the reasoning as misalignment. This suggests that the Alignment scores from Table 3 should be taken with a grain of salt.

Bounding-boxes To check if the use of interleaved bounding-boxes during reasoning helps with the performance, we carried out the following experiment. We fine-tuned Qwen2.5-VL-7B on DrivingVQA dataset with and without bounding-boxes, and ran evaluation on the held-out evaluation set. We found that including the bounding boxes during reasoning increased the performance from 63.55% to 66.09% of F1 score.

Reward Functions We also did an experiment on the effect of reward functions. We trained the first stage with A-OKVQA dataset, followed by a second stage with DrivingVQA dataset with different combination of the 3 reward functions. The results are given in Table 5. We have the following observations:

- Incorporating bounding-boxes based reward (IoU) especially helps for the stage-2 dataset.
- The effect of the Format reward is negligible, and we think it is because during the SFT-warmup stage, the model already learns to output its response following the format of <think> and <answer> tags.

7 Scaling up Multimodal-to-Text Reasoning

Motivated by the results in 6.3, we scaled up the training pipeline to 10 visual question answering datasets covering ~120K examples.

Train Datasets Following Shao et al. (2024a), we use datasets from the following 5 domains: **Fine-Grained Understanding** (Birds-200-2011 (Wah)), **Relation Reasoning** (GQA (Hudson and Manning, 2019), VSR (Liu et al., 2023)), **Text/Doc** (TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021b), DUDE (Landeghem et al., 2023), TextCaps (Sidorov et al., 2020), SROIE (Huang et al., 2019)), **General VQA** (Visual7W (Zhu et al., 2016)), **Charts** (InfographicsVQA (Mathew et al., 2021a)), making up ~120K samples in total. We refer to these datasets as *VisCOT*.

Evaluation Datasets We evaluate our models on the following 6 benchmarks: VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), POPE (Li et al., 2023), ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), VizWiz (Gurari et al., 2018).

Following the same pipeline as in 4.3, we train the following models starting from Qwen2.5-VL-3B as our base model.

1. **SFT-no-reason** the base model trained on A-OKVQA and VisCOT datasets for input-output pairs with no reasoning chain.

2. **SFT-warmup** the base model trained on A-OKVQA dataset with reasoning chains that contains bounding-boxes.

Method	DrivingVQA			A-OKVQA		
	F1	Exam Score	Alignment	F1	Exam Score	Alignment
SFT-warmup	51.86 $_{0.88}$	47.26 $_{0.76}$	**97.26**	86.78 $_{0.05}$	-	94.93
GRPO	**53.6** $_{\mathbf{0.93}}$	**49.58** $_{\mathbf{0.92}}$	90.3	**88.12** $_{\mathbf{0.23}}$	-	**95.2**

Table 3: The results of training Qwen2.5-VL-7B on A-OKVQA dataset with SFT and GRPO methods for Multimodal-to-Text Reasoning with interleaved bounding-boxes. Note that A-OKVQA is an in-domain evaluation dataset, while DrivingVQA is out-of-domain.

Alignment Class	DrivingVQA	A-OKVQA
Aligned	92%	100%
Misaligned	68%	40%

Table 4: Human-Judge Agreement scores for 50 samples from DrivingVQA and A-OKVQA for reasoning-answer alignment.

Rewards			DrivingVQA	A-OKVQA
Accuracy	Format	IoU		
✓	✓	✗	57.89	**88.56**
✓	✗	✓	**61.31**	88.3
✓	✓	✓	**61.31**	88.3

Table 5: The F1 score of training Qwen2.5-VL-7B on A-OKVQA for SFT and DrivingVQA for GRPO stages for Multimodal-to-Text Reasoning with interleaved bounding-boxes with different reward functions.

3. **GRPO** starting from SFT-warmup, we train with GRPO on VisCOT dataset using all 3 reward functions.

We report the exact-match accuracy scores in Figure 4. Our observations are as follows:

- The performance gain saturates at a certain point, after which more data does not help with the performance
- GRPO training is always outperforming the SFT-warmup model (0%)

8 Conclusion

In this work, we explored three paradigms of multimodal chain-of-thought reasoning: Multimodal-to-Multimodal, Text-to-Multimodal, and Multimodal-to-Text. Each paradigm involves different ways of combining text and image during the reasoning process. To train these models, we used Group Relative Policy Optimization (GRPO), a reinforcement learning method that trains models using only input-output pairs, without needing annotated reasoning chains.

We developed custom reward functions for each paradigm, focusing on answer accuracy, correct formatting, alignment between text and image, and the quality of visual grounding. Our experiments show that generating image during reasoning (as in Multimodal-to-Multimodal and Text-to-Multimodal) often leads to worse performance. In contrast, the Multimodal-to-Text approach, especially when bounding boxes are used to guide the reasoning, gives more consistent improvements. This suggests that grounded reasoning could be an effective method to improve the visual question answering ability of the models.

9 Future work

The future work mainly focuses on Multimodal-to-Text Reasoning, as we observed encouraging results in this setting.

Dataset As shown in Figure 4, we notice that beyond a certain point, increasing the amount of training data does not lead to further performance improvements. This suggests that focusing on data quality may be more beneficial than simply increasing quantity. We also observed that the model sometimes performs a right but short reasoning and arrives at incorrect conclusions too quickly. One possible reason is that the SFT-warmup stage, which uses around 17K examples, may lead the model to become too deterministic, limiting its ability to learn diverse reasoning patterns during the GRPO stage. As a next step, we can select a smaller, cleaner subset of the A-OKVQA dataset – possibly filtered with GPT – for the SFT-warmup. Additionally, Fan et al. (2025) applies GRPO using only 20 examples over 12 hours of training (with each example seen approximately 1280 times by my estimate), suggesting that applying GRPO on smaller datasets over more epochs could be a promising direction to explore.

Training Method In addition to relying only on rule-based sparse rewards during GRPO, Cui et al. (2025) incorporates an implicit token-level reward signal during training updates. Applying a similar approach to GRPO for VLMs could be interesting,

Figure 4: GRPO training progress of Qwen2.5-VL-3B on VisCOT dataset. Evaluations are done on the 6 given datasets.

as it would provide denser and potentially more informative reward signals, enabling the model to learn more efficiently.

Inference During inference, after generating bounding boxes, feeds the corresponding image patches back into the model, allowing it to extract additional information from the image. This is another inference-time strategy we could experiment with to potentially improve performance.

Evaluation The evaluation results shown in Figure 4 are based on exact matching with the ground truth, which comes with its limitations. It may be more appropriate to use BLEU scores or LLM-as-a-Judge methods for a more accurate assessment of model outputs. Additionally, for open-ended questions, we should explore the use of BLEU-based metircs during training to use as a reward function instead of relying solely on the exact match (as in the Accuracy Reward).

References

Technical report.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *Preprint*, arXiv:2407.06135.

Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. 2024. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *Preprint*, arXiv:2403.13315.

Charles Corbière, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. 2025. Retrieval-based interleaved visual chain-of-thought in real-world driving scenarios. *Preprint*, arXiv:2501.04671.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, and 4 others. 2025. Process reinforcement through implicit rewards. *Preprint*, arXiv:2502.01456.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming transformers for high-resolution image synthesis. *Preprint*, arXiv:2012.09841.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025. Grit: Teaching mllms to think with images. *Preprint*, arXiv:2505.15879.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *Preprint*, arXiv:2503.21776.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. Making llama see and draw with seed tokenizer. *Preprint*, arXiv:2310.01218.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *CVPR*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanisławek. 2023. Document understanding dataset and evaluation (dude). *Preprint*, arXiv:2305.08455.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2024. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *Preprint*, arXiv:2402.12058.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. *Preprint*, arXiv:2501.07542.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. 2025. Improved visual-spatial reasoning via r1-zero-like training. *Preprint*, arXiv:2504.00883.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Preprint*, arXiv:2205.00363.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. Infographicvqa. *Preprint*, arXiv:2104.12756.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. *Preprint*, arXiv:2311.17076.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kamcot: Knowledge augmented multimodal chain-of-thoughts reasoning. *Preprint*, arXiv:2401.12863.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Preprint*, arXiv:2403.16999.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *Preprint*, arXiv:2504.07615.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. *Preprint*, arXiv:2003.12462.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

Chameleon Team. 2025. Chameleon: Mixed-modal early-fusion foundation models. *Preprint*, arXiv:2405.09818.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. *Preprint*, arXiv:2404.03622.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model. *Preprint*, arXiv:2503.05132.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. *Preprint*, arXiv:1511.03416.

A PuzzleVQA

PuzzleVQA contains 18 different puzzle types, 10 of which are illustrated in Figure 5. For each puzzle type, we generate up to 1,000 examples in the format as shown in Figure 6.

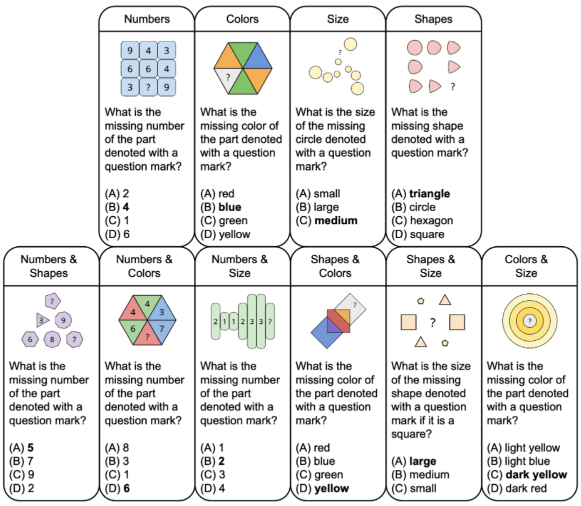
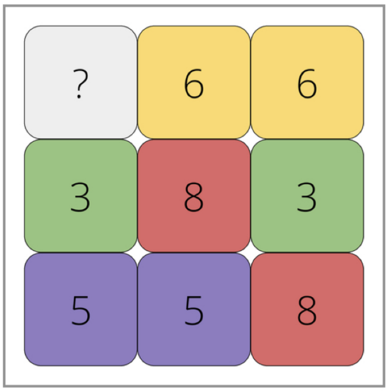


Figure 5: Examples of 10 puzzle types from PuzzleVQA. Adopted from Chia et al. (2024)

.



 # Question:
<image>What is the missing color if the part denoted with the question mark has the number 6?

 # Options:
(A) purple
(B) red
(C) yellow
(D) green

 # Pattern:
There is a 3x3 colored grid of numbers. The first row has number-color pair [(6, '?'), (6, 'yellow'), (6, 'yellow')], the second row is [(3, 'green'), (8, 'red'), (3, 'green')], and the third and final row is [(5, 'purple'), (5, 'purple'), (8, 'red')]. We observe that the grid cells with number 8 is red in color, the grid cells with number 5 is purple in color, the grid cells with number 3 is green in color, and the grid cells with number 6 is yellow in color. Thus, the pattern is that the grid cell with the same number will have the same color.

 # Option A
Replacing '?' with purple: <image>
Reasoning: Comparing the sum of each row, it does not fit the observed pattern. purple is unlikely to be the correct answer.

 # Option B
Replacing '?' with red: <image>
Reasoning: Comparing the sum of each row, it does not fit the observed pattern. red is unlikely to be the correct answer.

 # Option C
Replacing '?' with yellow: <image>
Reasoning: Comparing the sum of each row, it matches well the pattern, that is each row adds up to the same value. yellow is a strong candidate for the correct answer.

 # Option D
Replacing '?' with green: <image>
Reasoning: Comparing the sum of each row, it does not fit the observed pattern. green is unlikely to be the correct answer.

 # Final Answer:
Based on the reasoning process, the best fit for the missing number is yellow. The pattern that the numbers in each row adding up to the same value holds consistently when we replace '?' with yellow. Therefore, among (A) (B) (C) (D), the answer is: (C).

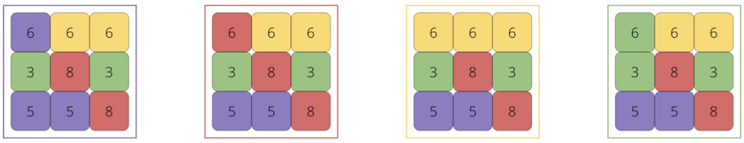


Figure 6: An example of PuzzleVQA train set. The example is from Grid-Number-Color puzzle type.