

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 VISUAL REASONING VIA PERCEPTUAL EXTENSION AND IN-CONTEXT LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

The reasoning ability with visual information has recently gained significant attention in the field of large vision-language models (LVLMs). Existing R1-like reasoning LVLMs are usually finetuned from a base LVLM on a large-scale vision-language dataset, incorporating reinforcement learning (RL) with rewards from verifiable answers. However, such reasoning LVLMs usually requires high-quality multimodal long-chain datasets for supervised finetuning in the cold start stage, and time-consuming multiple response sampling in the RL stage. Therefore, we seek to explore an efficient approach to achieve visual reasoning. To do so, we first investigate the interaction between visual and textual tokens in LVLMs, and find that although the post-trained reasoning LVLM improves the cross-modal interaction, but only at deep layers and for long responses, this improvement is negligible for short responses. Based on these observations and insights, we propose to separate the perception and reasoning process, to avoid the LVLM from generating long responses, so that the LVLM maintains cross-modal interaction ability, and do the reasoning by the LLM, which is not required to integrate cross-modal information. To this end, we leverage the existing reasoning large language models (LLMs) with a VLM extension, to synthesize visual and textual information in advance and then perform the reasoning by the LLM, without any finetuning. Furthermore, to make full use of the training samples, we use a matching mechanism to find the relevant reasoning process and incorporate them by in-context learning. We evaluate our method on the common visual reasoning benchmarks. The results show that, without extra training samples, our method achieves performance comparable to the existing post-trained reasoning LVLMs, and outperforms them with in-context learning.

1 INTRODUCTION

Reasoning with visual information has become a core capability of large vision-language models (LVLMs) to achieve human-level intelligence (Li et al., 2024; Chen et al., 2023; Bai et al., 2023; Xu et al., 2024; Huang et al., 2025). It enables the model to solve complicated questions that require both visual understanding and multi-step reasoning. Moreover, such reasoning ability also improves the interpretability and trustworthiness of these models. Building on the recent success of reinforcement learning with verifiable rewards in large language model (LLM) reasoning (Guo et al., 2025), existing R1-like reasoning LVLMs are usually finetuned from a base LVLM on a large-scale vision-language dataset (Huang et al., 2025; Chen et al., 2025). However, the training process of reasoning LVLMs usually requires high-quality multimodal long-chain datasets for supervised finetuning in the cold start stage, and time-consuming multi-response sampling in the reinforcement learning stage. This training process requires both a high-quality dataset and massive computational resources. To address these challenges, recent work has investigated more efficient approaches. One line of research leverages existing reasoning LLMs while training only a lightweight projector to align visual and textual tokens (Peng et al., 2025). Although this strategy avoids sampling long-chain responses, it still requires a large number of training samples to train the projector. Given powerful reasoning LLMs such as DeepSeek-R1 (Guo et al., 2025) at hand, we ask: **is it possible to achieve comparable visual reasoning performance without expensive finetuning?**

To this end, we examine the difference between base LVLMs and reasoning LVLMs in terms of vision-text information interaction within attention blocks, aiming to uncover insights for more ef-

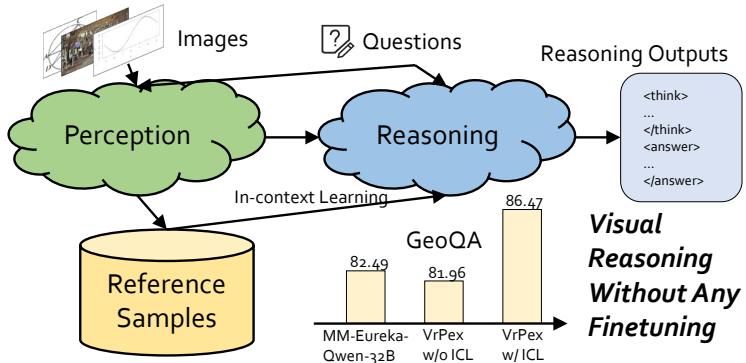


Figure 1: Achieving visual reasoning via a perceptual extension and enhancing it by in-context learning. Our method separates perception and reasoning and does not need any finetuning, while having comparable performance with R1-like reasoning VLMs.

Efficient design. Our analysis shows that although reasoning LVLMs does improve the vision-text information interaction when generating long responses. But when generating short responses, this improvement is negligible. Additionally, we find that this improvement occurs only in the deep layers. In the early layers, reasoning LVLMs still exhibit inadequate integration of visual and textual information. We attribute this limitation to the inherent disparity between visual and textual tokens in their semantic abstraction levels: at the early layers, the LVLM prioritizes processing visual tokens, which reduces its capacity to fuse cross-modal information when producing long responses. These observations suggest that LVLMs may need to avoid generating long responses if we want to maintain the same level of cross-modal interaction. Motivated by these findings, we propose to **decouple perception and reasoning**. In this design, the LVLM focuses on perceiving and integrating visual information without the burden of producing long responses, while the LLM performs the reasoning independently. This separation enables the LVLM to more effectively inject visual content into the output, while the reasoning LLM (unconstrained by cross-modal fusion) can handle the complex reasoning process.

Based on these observations and insights, we propose a method to integrate visual information into the reasoning process without requiring any finetuning. Specifically, our approach leverages a powerful reasoning LLM in combination with a VLM extension. The VLM extension first synthesizes visual and textual information, and the reasoning is subsequently carried out by the LLM. Such combination of the reasoning LLM and the VLM extension can be more efficient compared to a single LVLM, since the VLM extension is no longer required to generate long responses where the visual information is difficult to integrate properly. We refer to this framework as **Visual Reasoning via Perceptual Extension (VrPex)**. In addition, we design a matching mechanism to retrieve relevant reasoning trajectories from training samples and incorporate them through in-context learning with the reasoning LLM. This mechanism allows VrPex to exploit available training data while further improving performance on specific benchmarks.

We conduct extensive experiments to evaluate the effectiveness of our proposed method in terms of reasoning performance. The results show that even without access to training samples, VrPex achieves performance comparable to existing reasoning LVLMs. Moreover, when combined with in-context learning, it further outperforms these models. We observe that incorporating relevant training samples into the input context can further enhance performance on targeted tasks, highlighting the extensibility and flexibility of our approach.

2 RELATED WORKS

2.1 R1-LIKE REASONING LVLMs

Recent advancements of large language models in incentivizing reasoning ability with reinforcement learning (Guo et al., 2025) inspire the researches on R1-like reasoning large vision-language models (Meng et al., 2025; Chen et al., 2025; Huang et al., 2025). These visual reasoning models are usually finetuned with reinforcement learning algorithms (e.g., PPO, DPO, and GRPO) using rewards from verifiable answers. For example, DPO is used in RLHF-V (Yu et al., 2024), LLaVA-Reasoner (Zhang et al., 2024b), and Insight-V (Dong et al., 2025). They construct large-scale preference datasets and directly apply DPO for training. MMPR (Wang et al., 2024) further introduces

108 a quality loss from a Binary Classifier and a generation loss from supervised finetuning (SFT), in
 109 addition to the standard DPO preference loss, thereby strengthening reasoning ability. Meanwhile,
 110 GRPO, which has proven effective in DeepSeek-R1, has become a widely adopted RL strategy for
 111 reasoning LVLMs. Representative works, including MM-Eureka (Meng et al., 2025), Vt-R1 (Zhou
 112 et al., 2025), LLM-R1 (Yingzhe et al., 2025), and R1-V (Chen et al., 2025), apply GRPO to multi-
 113 modal reasoning tasks such as mathematical geometry, achieving promising results.

115 2.2 VISUAL IN-CONTEXT LEARNING FOR LVLMs

117 In-context learning is a paradigm that allows language models to learn new tasks given only a few
 118 examples in the form of demonstration (Brown et al., 2020). At inference, in-context learning re-
 119 search has primarily focuses on three aspects: demonstration organization (Zhao et al., 2021; Lu
 120 et al., 2021), demonstration selection (Liu et al., 2021; Tanwar et al., 2023; Qin et al., 2023), and
 121 demonstration reformatting (Kim et al., 2022; Liu et al., 2023a; Yang et al., 2023). For LVLMs,
 122 in-context learning is extended to use visual information as demonstrations (Sun et al., 2023; Liu
 123 et al., 2023b). Considering the cross-modal challenge in LVLMs, VICL (Zhou et al., 2024) uses
 124 intent-oriented image summary and demonstration composition to address such challenge.

125 In our paper, we focus on visual reasoning and avoid potential challenge in cross-modal interaction
 126 between vision and text by separating perception from the reasoning process. Also, we use utilize
 127 the thought process of the similar ones in the training samples as visual demonstrations to improve
 128 the reasoning in specific areas, which differs from any of the prior works.

129 3 METHODOLOGY

131 3.1 CROSS-MODAL INFORMATION INTERACTION IN REASONING LVLM

133 To better understand what makes reasoning LVLMs perform better than base LVLMs and give in-
 134 sights to efficient construction for visual reasoning system, we investigate the interaction between
 135 visual and textual information in LVLMs.

137 Therefore, we look into the attention weights of the base LVLM and the reasoning LVLM trained
 138 by RL. Suppose the queries and keys in the attention block at layer l to be $\mathbf{Q}^{(l)}, \mathbf{K}^{(l)}$. The attention
 139 map is

$$140 \mathbf{A}^{(l)} = \text{softmax} \left(\mathbf{Q}^{(l)} \mathbf{K}^{(l)\top} \right). \quad (1)$$

141 It can be further decomposed into three parts according to the queries and keys: system prompt
 142 tokens ($\mathbf{K}_{\text{sys}}^{(l)}$), visual tokens ($\mathbf{K}_{\text{vis}}^{(l)}$), and text tokens ($\mathbf{Q}_{\text{text}}^{(l)}, \mathbf{K}_{\text{text}}^{(l)}$). The attention map can be
 143 expressed as a partitioned matrix given the above partition, which is

$$145 \mathbf{A}^{(l)} = \left[\mathbf{A}_{\text{text,sys}}^{(l)}, \mathbf{A}_{\text{text,vis}}^{(l)}, \mathbf{A}_{\text{text,text}}^{(l)} \right]. \quad (2)$$

146 Specifically, we care about the *proportion of attention weights* that the text tokens attend to the visual
 147 tokens (Text-Vision, $\mathbf{A}_{\text{text,vis}}^{(l)}$) and the text tokens (Text-Text, $\mathbf{A}_{\text{text,text}}^{(l)}$), since they reveal how the
 148 information flows before generating the output text.

150 We use prompts to encourage the model generate long chain reasoning process, then compare the
 151 proportion of the attention weights between the base and reasoning models. We use the training
 152 samples of GeoQA (Chen et al., 2021) as the inputs. We take Qwen2.5-VL-7B-Instruct (Bai et al.,
 153 2025) as the base LVLM, and MM-Eureka-Qwen-7B (Meng et al., 2025) as the reasoning LVLM.

154 Firstly, we investigate the attention proportion by the response length. For this experiment, we take
 155 the average attention proportion across the layers, which is,

$$157 158 p_{\text{text,vis}} = \frac{1}{L} \sum_{l=1}^L \sum_{i=1}^N \mathbf{A}_{\text{text,vis},i}^{(l)}, \quad (3)$$

159 where N is the total number of visual tokens. The Text-Text proportion is similarly calculated.

160 As we can see in Figure 2a, 2b. We plot the attention proportion against the response length in
 161 logarithmic scale. And the dashed lines are the linear fitting of the samples. From the absolute value

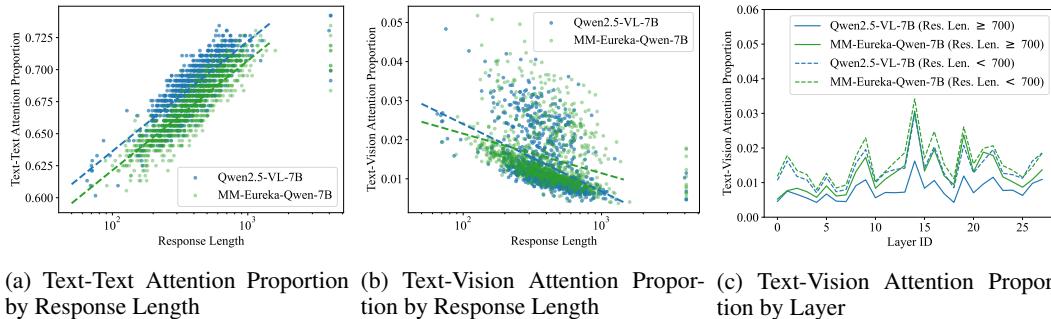


Figure 2: Investigations on attention proportion of LVLMs. (a)(b) Text-Text and Text-Vision attention proportions are roughly proportional and inverse proportional to the response length. The reasoning LVLM is able to generate longer responses while keeping the attention proportion unchanged. (c) For the reasoning LVLM, the gap between long and other responses diminishes as the layer goes deeper. **However, for short responses, the reasoning LVLM only achieves the same level of cross-modal interaction as the base LVLM.**

of the proportion, Text-Vision takes a much smaller attention proportion than Text-Text, for both base and reasoning models. As the response length increases, the Text-Text takes more attention proportion, and the Text-Vision takes less accordingly. This is reasonable because the longer the response, the more text tokens can be attended to, thus the Text-Text attention proportion rises. More importantly, as for the difference between two models, from Figure 2a, the reasoning model (green) generally generates longer responses, while keeping the same Text-Text attention proportion. From Figure 2b, the Text-Vision attention proportion of the reasoning model diminishes more slowly as the response length increases, compared to the base model (blue). We thus posit that the success of the reasoning LVLMs is related to the increased Text-Vision attention proportion for long responses, which helps the model to integrate the visual information better.

Furthermore, we investigate the attention proportion by layer to find more clues, as shown in Figure 2c. In the figure, we divide the responses into long (≥ 700 tokens) and short (< 700 tokens) responses and compare the Text-Vision attention proportion by layer. We can see that longer responses usually take a smaller Text-Vision attention proportion for both models, which is consistent with our prior observation. Interestingly, for the reasoning model, the gap between long and short responses reduces as the layer goes deeper, which clearly contrasts to the base model, where the gap remains relatively unchanged. However, the increased attention proportion for long responses at deeper layers for the reasoning model is merely the same as that for short responses. Therefore, we come to another conclusion that although the Text-Vision attention proportion increases for the reasoning LVLM, it mainly occurs at deep layers and for long responses. For short responses, the reasoning LVLM only achieve the same level of cross-modal interaction as the base LVLM. Moreover, considering the difference between visual and text tokens in terms of the abstraction level, such an inadequacy of information integration is likely the fundamental deficiency for the LVLMs to further achieve better performance when generating long responses.

So far, we have found the merits and demerits of the reasoning LVLMs in terms of visual information integration. Although the reasoning LVLM improves the integration of visual information at deeper layers for long responses, it only achieves the same level of integration for short responses. To construct an efficient visual reasoning system, there is no need for the LVLMs to generate long responses if we want to keep the same level of cross-modal interaction. Therefore, we can avoid the VLM from generating long responses, and make it only do the perception to preserve the cross-modal interaction without further post-training. Then to perform the reasoning, we have to incorporate a reasoning LLM to further integrate the information. As such integration is only in the text domain, we do not worry about the cross-modal interaction when generating long responses. It leads us to the separation of the perception and reasoning process.

3.2 PERCEPTUAL EXTENSION FOR REASONING LLM

To keep the interaction between visual and textual information and avoid the deficiency of the LVLMs in generating long responses, we separate the perception from the reasoning process. Then, we use the existing reasoning LLM to generate long responses for reasoning. Specifically, we sepa-

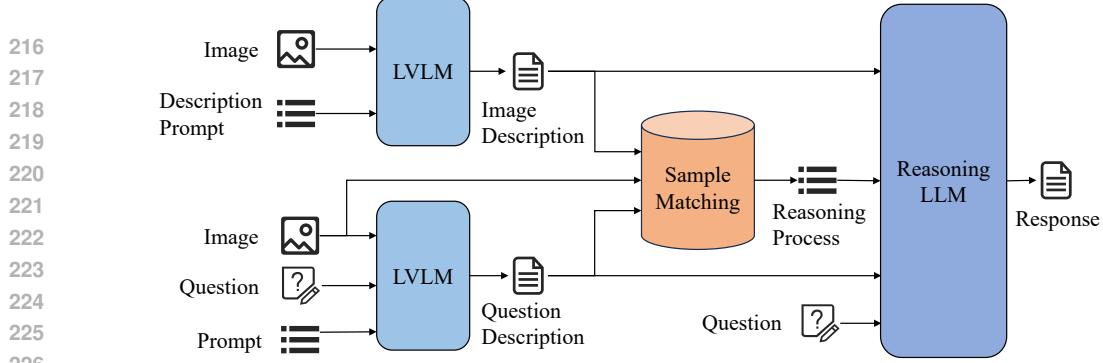


Figure 3: The overview of VrPex with in-context learning. It includes the separation of perception and reasoning, which avoids the inefficient vision-text interaction in the LVLM when generating long responses. It also includes the in-context learning of reasoning process, for performance improvement on specific benchmarks.

rate the perception and the reasoning as follows:

$$\mathbf{o}_{\text{desc}} = \text{LVLM}(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{text}}), \quad (\text{Perception}) \quad (4)$$

$$\mathbf{o}_{\text{ans}} = \text{LLM}(\mathbf{o}_{\text{desc}}, \mathbf{x}_{\text{text}}), \quad (\text{Reasoning}) \quad (5)$$

where we denote the input sample as $(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{text}})$, the description text generated by the LVLM as \mathbf{o}_{desc} , and the output responses as \mathbf{o}_{ans} . Such separation not only avoids the LVLM to generate long responses which the model is probably not good at, but also avoids finetuning on the LVLM which require massive computational resources.

Next, we have to determine what should the description text \mathbf{o}_{desc} be. Naively, the description text can be simply the generated caption of the image. However, such caption is usually general, and shows irrelation to the input question. Thus, the reasoning LLM is rather difficult to extract relative information from the caption. Therefore, we take a step further to make the LVLM generate targeted description related to the input question. We use carefully designed prompts to make the LVLM perform two generations, one with only the image, the other with both the image and the question, to get the image description and question description. Formally,

$$\mathbf{o}_{\text{img_desc}} = \text{LVLM}(\mathbf{p}_{\text{img_desc}}, \mathbf{x}_{\text{img}}), \quad (6)$$

$$\mathbf{o}_{\text{ques_desc}} = \text{LVLM}(\mathbf{p}_{\text{ques_desc}}, \mathbf{x}_{\text{img}}, \mathbf{x}_{\text{ques}}), \quad (7)$$

where \mathbf{p} denotes the corresponding prompt.

With the perception results above, the reasoning LLM needs to perform the final reasoning given the related information about the question and the image.

$$\mathbf{o}_{\text{ans}} = \text{LLM}(\mathbf{p}_{\text{reasoning}}, \mathbf{o}_{\text{img_desc}}, \mathbf{o}_{\text{ques_desc}}, \mathbf{x}_{\text{ques}}). \quad (8)$$

We use existing reasoning LLMs which is trained to generate long responses for the final answers. The visual illustration of VrPex is in Figure 3.

3.3 IN-CONTEXT LEARNING OF REASONING PROCESS

So far, we only use the existing LVLMs and reasoning LLMs to construct a system to incorporate visual information in reasoning. In this section, we consider improving the performance further by incorporating the training samples via in-context learning. Therefore, we design a matching mechanism to find the relevant training sample according to the input sample. In the previous section, we use the LVLM to generate the image description and question description for the input sample. We can also use the generated descriptions as keys to find the relevant training samples.

Firstly, we preprocess the training samples for better sample matching and reasoning process output. The raw training samples are usually in the form of image-question-answer triplets. We use GPT-4o (OpenAI, 2025) to generate the structured reasoning process for each training sample, including the image description, question descriptions, and possible reasoning process according to the given answer. The reasoning process does not contain any specific numbers, only contains the general

270 reasoning steps, in order to exclude the distraction of question-specific information in the training
271 samples.

272 Then, we leverage a pre-trained image-text encoder to encode the images and texts in the training
273 samples into embeddings. Suppose the training samples to retrieve are represented by
274 $(\mathbf{z}_{\text{img}}^{(i)}, \mathbf{z}_{\text{img_desc}}^{(i)}, \mathbf{z}_{\text{ques_desc}}^{(i)}, \mathbf{z}_{\text{reasoning}}^{(i)})_i^n$, the encoding can be represented as
275

$$276 \mathbf{e}_{\text{img}}^{(i)} = \text{Enc}_{\text{img}}\left(\mathbf{z}_{\text{img}}^{(i)}\right), \mathbf{e}_{\text{img_desc}}^{(i)} = \text{Enc}_{\text{text}}\left(\mathbf{z}_{\text{img_desc}}^{(i)}\right), \mathbf{e}_{\text{ques_desc}}^{(i)} = \text{Enc}_{\text{text}}\left(\mathbf{z}_{\text{ques_desc}}^{(i)}\right). \quad (9)$$

276 The embeddings of the training samples can be stored as a database for efficient retrieval at inference.
277 For a specific input sample, we also encode the image and generated descriptions into the
278 embeddings as $(\mathbf{e}_{\text{img}}, \mathbf{e}_{\text{img_desc}}, \mathbf{e}_{\text{ques_desc}})$.

279 We calculate the similarities between the embeddings of the input sample and the training samples
280 in terms of each key. We use the cosine similarity to measure the similarity between the embeddings,
281 which is

$$282 s_{\text{img}}^{(i)} = \cos\left(\mathbf{e}_{\text{img}}, \mathbf{e}_{\text{img}}^{(i)}\right), \quad (10)$$

283 and it is similar for other keys ($s_{\text{img_desc}}^{(i)}$ and $s_{\text{ques_desc}}^{(i)}$).

284 To select the most relevant training samples, we keep the Pareto front of the three similarity
285 measures. Therefore, we obtain a set of the most similar training samples. We then use this set as the
286 demonstrations of in-context learning for the reasoning LLM. The demonstration contains only the
287 reasoning process generated in the sample preprocessing. The number of used demonstrations can
288 be controlled by a hyperparameter n_d . Given n_d , we select the top- n_d samples according to the total
289 similarity measures. The in-context learning can be represented as
290

$$291 \mathcal{D} = \text{Pareto}\left(\left(s_{\text{img}}^{(i)}, s_{\text{img_desc}}^{(i)}, s_{\text{ques_desc}}^{(i)}\right), n_d\right), \quad (11)$$

$$292 \mathbf{o}_{\text{ans}} = \text{LLM}\left(\frac{\mathbf{p}_{\text{reasoning}}, \mathbf{o}_{\text{img_desc}},}{\mathbf{o}_{\text{ques_desc}}, \mathbf{z}_{\text{reasoning}}, \mathbf{x}_{\text{ques}}}\right), i \in \mathcal{D}. \quad (12)$$

293 There are cases where the input question and image are not similar to any training samples. This
294 could happen when the type of the question is not covered by the training samples. In such cases,
295 we use a rejection threshold λ to filter out the training samples that are not similar enough. If
296 the maximum average similarity is below the threshold, we do not use any training samples as
297 demonstrations.

304 4 EXPERIMENTS

305 4.1 EXPERIMENTAL SETUP

306 **Benchmarks.** We evaluate our methods on the common visual reasoning benchmarks, including
307 MathVista mini (Lu et al., 2024), MathVerse mini (Zhang et al., 2024a), GeoQA (Chen et al., 2021),
308 and MMK12 (Meng et al., 2025). MathVista and MathVerse are comprehensive math visual reasoning
309 datasets, incorporating various math problems from different areas. GeoQA is more focused on
310 geometry problems, and MMK12 is a math visual reasoning dataset for K-12 students. Most of the
311 questions in the datasets are multiple-choice questions. Mastering such datasets requires the model
312 to be capable of both perception and reasoning.

313 **Baselines.** We select 4 R1-like reasoning vision-language models as the baselines. They are
314 R1-Onenision-7B, MM-Eureka-Qwen-7B/32B, and Skywork-R1V-38B. R1-Onenision (Yang et al.,
315 2025) and MM-Eureka (Meng et al., 2025) series are both finetuned on curated CoT datasets, then
316 uses reinforcement learning to improve the reasoning ability. Skywork-R1V (Peng et al., 2025) only
317 trains the projector to align the visual and text domains before the reasoning LLM.

318 **Implementation Details.** It is flexible for our method to use different VLMs and reasoning
319 LLMs. In this paper, we test our method using Qwen2.5-VL-3B/7B-Instruct (Bai et al., 2025)
320 and DeepSeek-R1-Distill-Qwen-7B/14B (Guo et al., 2025) as the VLM and reasoning LLM, re-
321 spectively. Such combination forms two parameter scales: 10B (3B + 7B) and 21B (7B + 14B).

Table 1: Results on visual reasoning benchmarks. The results are reported in accuracy (%). The best results are highlighted in **bold**. The second bests are underscored.

Reasoning VLMs		MathVista	MathVerse	GeoQA	MMK12
R1-Onevision-7B	Yang et al. (2025)	61.3	45.94	71.75	37.50
MM-Eureka-Qwen-7B	Meng et al. (2025)	73.2	50.25	80.37	60.80
MM-Eureka-Qwen-32B	Meng et al. (2025)	74.7	<u>54.19</u>	<u>82.49</u>	68.75
Skywork-R1V-38B	Peng et al. (2025)	70.3	46.78	73.87	48.40
Perception VLM		Reasoning LLM			
w/o ICL	Qwen2.5-VL-3B-Instruct	DeepSeek-R1-Distill-Qwen-7B	71.2	48.50	73.47
	Qwen2.5-VL-7B-Instruct	DeepSeek-R1-Distill-Qwen-14B	73.4	54.03	81.96
w/ ICL	Qwen2.5-VL-3B-Instruct	DeepSeek-R1-Distill-Qwen-7B	73.3	51.17	81.96
	Qwen2.5-VL-7B-Instruct	DeepSeek-R1-Distill-Qwen-14B	<u>74.1</u>	55.99	86.47
					65.25

Query Question



Question: As shown in the figure, AB parallel CD , straight line EF intersects AB at point E , intersects CD at point F . EF bisects angle BEF , and it intersects CD at point F , angle $1 = 50^\circ$, then angle 2 is equal to \square
 Choices: A: 50° B: 60° C: 65° D: 90°

Perception VLM image description:
 The image shows two parallel lines, $\setminus(AB)$ and $\setminus(CD)$, intersected by a transversal line $\setminus(EF)$. The intersection points of the transversal with the parallel lines form several angles.
 Line Segments and Points: \square $\setminus(AB)$ $\setminus(CD)$ $\setminus(EF)$ \square E F \square 1 2 6
 In summary, ...

Perception VLM question description:
 To solve this problem, we need to use the properties of parallel lines and angles formed by a transversal intersecting them.
 1. \square Identify the given information: \square 2. \square Use the property of parallel lines: \square ...

Matched Sample



Question: As shown in the figure, AB is parallel to CD , and line EF intersects AB and CD at points E and F , respectively. The bisector of angle AEF intersects CD at point F . If angle $EF6 = 64^\circ$, then the measure of angle AEF is \square $\setminus(AB)$. 32° B: 58° C: 64° D: 128°

Reasoning Process of the Matched Sample:
 {Step1}: Identify the relationships between angles formed by parallel lines and a transversal (corresponding angles, alternate interior angles).
 {Step2}: Analyze the effect of the angle bisector on the given angles.
 {Step3}: Express unknown angle in terms of variables based on the bisector's property.
 {Step4}: Apply the angle sum property of a triangle to establish an equation involving known and unknown angles.
 {Step5}: Solve the equation to find the desired angle measure. }

Figure 4: Case study of the proposed method. As shown, in the perception, the image description extracts all of the geometry elements in the image, and the question description gives targeted descriptions of the question. In the sample matching, it finds the sample using the same theorem, with similar reasoning process.

The specific prompts for perception and reasoning is in the supplementary material. For the default setting of in-context learning, we only use the training samples from GeoQA and set the number of demonstrations n_d to 1. On MMK12, we use a subset of the training samples from MMK12 to perform the sample matching. The image-text encoder we use is jina-clip-v2 (Koukounas et al., 2024), which encodes the input to a unified embedding. As stated in the Methodology section, we use GPT-4o to generate the reasoning process for each training sample, and apply a rejection threshold λ to ensure the relevance of the training samples when testing on benchmarks in various areas.

4.2 BENCHMARK RESULTS

The benchmark results are shown in Table 1. The answers generated by each model is extracted and checked by `math_verify`, which may differ from the results reported in other benchmarks, where they use LLMs to verify the answers. Anyway, it is fair for all of the results in the table. We can see that on most of the benchmarks, VrPex achieves comparable performance with the reasoning LVLMs. On MathVista and MathVerse, VrPex achieves comparable performance with MM-Eureka. On GeoQA, VrPex with ICL outperforms MM-Eureka by a considerable margin. On MMK12, since the samples for matching is sampled from the training set of MMK12, the ICL plays a more significant role. However, MM-Eureka is trained on MMK12, the performance gap is larger compared to other benchmarks. It is notable that VrPex does not require any finetuning, which is easy to construct and deploy. Therefore, our method offers an easier way to achieve visual reasoning with extreme little cost.

We can see that with the in-context learning, VrPex even outperforms some of the reasoning LVLMs in specific benchmarks. Because the samples used for in-context learning are from the training set of GeoQA, which is a geometry reasoning dataset, VrPex has a great performance gain on GeoQA. This shows that the in-context learning can be used to improve the performance on specific benchmarks. This offers more extensibility and flexibility for constructing the visual reasoning solution. In Additional Analysis, we will further investigate the performance gain of in-context learning on other benchmarks in terms of category-wise accuracy.

Furthermore, we explore the flexibility and extensibility of VrPex by using different size of perception VLMs and reasoning LLMs in other scenarios. We perform experiments on perception-intensive

Table 2: Ablation study. We verify the effectiveness of our method by upgrading the method step by step.

Perception VLM	Reasoning LLM	MathVista	GeoQA
✓	✗	55.6	50.27
✗	✓	48.4	68.57
img. desc. only	✓	67.1	70.82
ques. desc. only	✓	69.3	71.88
img. + ques. desc.	✓	71.2	73.47

Table 3: Accuracies of different number of demonstrations for ICL. More number of demonstrations does not come with better performance in general.

Number of Demos	1	2	3	4
MathVista	73.4	73.2	73.3	72.9
GeoQA	81.96	81.83	82.10	81.83

reasoning benchmarks like M3CoT Chen et al. (2024) and RealworldQA X.AI (2024). The details are stated in the Appendix C.

4.3 CASE STUDY

We illustrate the effectiveness of our method by showing some cases of GeoQA in Figure 4. And leave some detailed cases in the supplementary material.

Perception VLM The outputs of the perception VLM are shown in the left part of Figure 4. For the image description, the VLM extracts all of the geometry elements in the image in detail, such as the points, parallel lines, angles, and triangles. For the question description, the VLM gives targeted descriptions of the question, such as the theorems or properties used in the question.

Sample Matching The result of the sample matching is shown in the right part of Figure 4. By the three similarity measures used in the matching, it finds the sample using the same theorem, for example, the theorems of the parallel lines and bisect angles. Therefore, the reasoning process of the matched sample can be effectively used as the demonstration for in-context learning.

4.4 ABLATION STUDY

To further verify the effectiveness of the proposed method, we perform ablations on the GeoQA dataset. We upgrade the method step by step from only using the VLM to answer the question to using both image and question description. The in-context learning is not applied in the ablations. We use the 10B combination of our method in this experiment. The results are shown in 2.

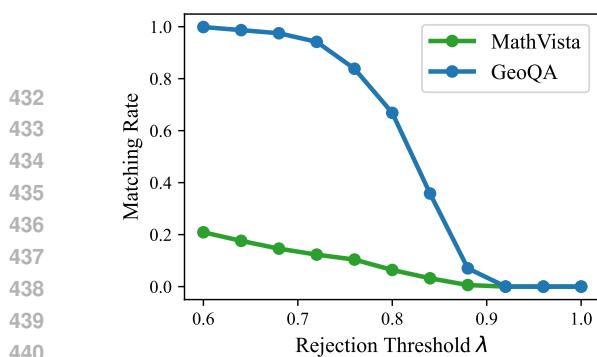
We first checked the performance of only using the VLM to answer the question and only using the text information (row 1 and 2). From these results, we see that reasoning LLM perform much better on GeoQA than MathVista, so we conclude that MathVista requires more perception than GeoQA. It can further inspire us to use better perception VLM on MathVista to handle perception-intensive benchmarks like MathVista. Then we investigate the performance with only the image description, to show the effectiveness of the question description. The results shows that using question description can improve the performance significantly. Also, we see that using question description only is better than using image description only, but the best performance is achieved by using both of them.

4.5 CATEGORY-WISE IMPROVEMENT AFTER ICL

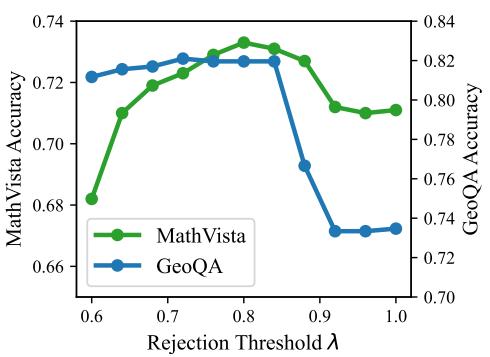
To further investigate the effect of the in-context learning, we look deeper into the category-wise accuracy before and after in-context learning. We take MathVista as an example, showing the results in Table 4 for model scale of 10B. We can conclude that, by using the training set of GeoQA for in-context learning, the accuracy on test samples in geometry-related sources generally increases. It shows that we can specifically improve the capability of the proposed visual reasoning system by using training samples in specific areas with corresponding reasoning process, while not affecting the performance in other areas too much.

Table 4: Category-wise accuracy and relative improvement on Mathvista.

Sources (Partial)	w/o ICL	w/ ICL	Increment
GeoQA+ (6.2%)	77.42	82.26	+4.84
Geometry3K (6.2%)	67.74	72.58	+4.84
UniGeo (6.2%)	85.48	88.71	+3.23
Super-CLEVER (6.2%)	56.45	54.84	-1.61
IQTest (3.7%)	45.95	43.24	-2.70
FigureQA (6.2%)	64.52	61.29	-3.23
Overall (100%)	71.2	73.3	+2.1



432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
Figure 5: Matching rate is decreasing as the re-
jection threshold increases.
Figure 6: Benchmark accuracies is affected by
the rejection threshold.



444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
491
492
493
494
495
496
497
498
499
500
Figure 5: Matching rate is decreasing as the re-
jection threshold increases.
Figure 6: Benchmark accuracies is affected by
the rejection threshold.

4.6 HYPERPARAMETER ANALYSIS

In this section, we analysis the effect of two hyperparameters in our proposed method.

Rejection Threshold for Sample Matching. The rejection threshold λ controls the similarity level that the sample matching accepts the sample as demonstration. We investigate how the rejection threshold affect the accuracy on the benchmarks, the results are shown in Figure 5 and 6.

Matching rate is the proportion of test samples that succeeds matching with at least one sample as the demonstration. We can acknowledge how many test samples actually use the demonstrations with ICL. As we can see from Figure 5, in MathVista, there are many test samples have low similarities (below 0.6) with the matching samples, since we only use the geometry samples for matching and there are other test samples in different areas in MathVista. However, in GeoQA, the similarities are mostly distributed between 0.7 to 0.9. There are much more samples succeed matching with the demonstration as the rejection threshold decreases.

The relation between the rejection threshold and the accuracy on benchmarks are shown in Figure 6. The accuracies generally rises and then drops as the rejection threshold increases. We can choose the best threshold at around 0.8.

Number of Demonstrations for ICL. For the in-context learning, we can take multiple samples from the pareto front of three similarity measures as the demonstrations. We perform experiments with different number of demonstrations n_d on MathVista and GeoQA. The results are shown in Table 3. We can see from the results that the number of demonstrations does not affect the performance too much. It can be explained by the similar reasoning processes of the retrieved samples.

5 CONCLUSIONS

In this paper, we explore an efficient approach to achieve visual reasoning without any finetuning. We first investigate the differences between base VLMs and existing R1-like reasoning VLMs in terms of the interaction between visual and textual tokens. Our study reveals that while the reasoning LVLM exhibits stronger cross-modal interaction than the base model, this improvement mainly occurs in deeper layers and when generating long responses. For short responses, the reasoning LVLM achieves a level of interaction comparable to the base VLM. Motivated by these findings, to avoid finetuning and deficiency of the base VLM on generating long responses, we separate the perception and reasoning process in visual reasoning. We thus propose Visual Reasoning via Perceptual Extension (VrPex), which uses a VLM to generate both image and question description, followed by a reasoning LLM to perform the reasoning. Furthermore, to make full use of the training samples, we incorporate in-context learning in the reasoning LLM by matching relevant reasoning processes from training samples. The experiments show that VrPex can achieve performance comparable to R1-like reasoning VLMs, and with in-context learning it outperforms them, while avoiding the costly finetuning process.

Limitations: Although the proposed approach can achieve visual reasoning without any finetuning, the perception capability relies on the perception VLM, and the reasoning capability relies on the reasoning LLM. To get better perception and reasoning capability, it needs to use larger models and more test-time computation. Although the ICL can offer some flexibility and further improvement for the system, it requires to preprocess the training samples first.

486 REFERENCES
487

488 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang
489 Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities.
490 *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

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

496 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
497 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
498 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

499

500 Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin.
501 Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning.
502 *arXiv preprint arXiv:2105.14517*, 2021.

503 Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman
504 Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large
505 language model as a unified interface for vision-language multi-task learning. *arXiv preprint*
506 *arXiv:2310.09478*, 2023.

507

508 Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization
509 ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. Accessed: 2025-02-02.

510

511 Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M³cot: A novel
512 benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proc. of ACL*, 2024.

513

514 Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei
515 Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In
516 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9062–9072, 2025.

517 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
518 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
519 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

520

521 Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and
522 Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models.
523 *arXiv preprint arXiv:2503.06749*, 2025.

524 Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee.
525 Self-generated in-context learning: Leveraging auto-regressive language models as a demonstra-
526 tion generator. *arXiv preprint arXiv:2206.08082*, 2022.

527

528 Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami,
529 Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. jina-clip-
530 v2: Multilingual multimodal embeddings for text and images, 2024. URL <https://arxiv.org/abs/2412.08802>.

531

532 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan
533 Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint*
534 *arXiv:2408.03326*, 2024.

535

536 Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What
537 makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

538

539 Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning
more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*,
2023a.

540 Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao
 541 Dong. Unifying image processing as visual prompting question answering. *arXiv preprint*
 542 *arXiv:2310.10513*, 2023b.

543

544 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-
 545 Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of
 546 foundation models in visual contexts. In *International Conference on Learning Representations*
 547 (*ICLR*), 2024.

548 Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered
 549 prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint*
 550 *arXiv:2104.08786*, 2021.

551

552 Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng
 553 Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng
 554 Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-
 555 based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

556 OpenAI. GPT-4o (chatgpt-4o-latest). Proprietary AI model, March 2025. URL <https://www.ibm.com/cn-zh/think/topics/gpt-4o>. Multimodal AI system with text, audio, and
 557 image processing capabilities. Release date: March 2025.

558

559 Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao,
 560 Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou.
 561 Skywork r1v: Pioneering multimodal reasoning with chain-of-thought, 2025. URL <https://arxiv.org/abs/2504.05599>.

562

563 Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning
 564 with iterative demonstration selection. *arXiv preprint arXiv:2310.09881*, 2023.

565

566 Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors
 567 for improving visual in-context learning. *arXiv preprint arXiv:2304.04748*, 2023.

568

569 Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual llms
 570 are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*,
 571 2023.

572

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

576

577 X.AI. Grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v>, 2024.

578

579 Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language
 580 models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

581

582 Jinghan Yang, Shuming Ma, and Furu Wei. Auto-ic1: In-context learning without human supervi-
 583 sion. *arXiv preprint arXiv:2311.09263*, 2023.

584

585 Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng
 586 Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing general-
 587 ized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*,
 588 2025.

589

590 Peng Yingzhe, Zhang Gongrui, Zhang Miaosen, You Zhiyuan, Liu Jie, Zhu Qipeng, Yang Kai,
 591 Xu Xingzhong, Geng Xin, and Yang Xu Lmm. r1: Empowering 3b lmms with strong reasoning
 592 abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

593

594 Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu,
 595 Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment
 596 from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on*
 597 *Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

594 Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou,
595 Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the
596 diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024a.
597

598 Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang,
599 Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning.
600 *arXiv preprint arXiv:2410.16198*, 2024b.

601 Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving
602 few-shot performance of language models. In *International conference on machine learning*, pp.
603 12697–12706. PMLR, 2021.

604 Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-
605 “zero’s” aha moment” in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*,
606 2025.

607 Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large
608 vision-language models. *arXiv preprint arXiv:2402.11574*, 2024.

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648 A PROMPTS FOR PERCEPTION AND REASONING
649650 Here we provide the prompt template for the perception VLM and the reasoning LLM.
651652 Image description prompt $p_{\text{img.desc}}$:653 *You are an image descriptor. When the user gives you an image, you need to describe the image in
654 precise description, revealing the relations of all the elements in the image.*655 Question description prompt $p_{\text{ques.desc}}$:656 *You are a question solving assistant. When the user gives you an image, along with a question, you
657 need to output an effective prompt for the next large language model to get the correct answer. The
658 prompt must describe the image in precise description, revealing the relations of all the elements in
659 the image, and point out the possible concepts relating to the question.*660 Reasoning prompt template $p_{\text{reasoning}}$:661 *Here is the image description: $\{o_{\text{img.desc}}$ } and question description: $\{o_{\text{ques.desc}}$ } Think as carefully
662 and methodically about the problem as you need to. Referring to the following demonstrations and
663 hints: $\{z_{\text{reasoning}}$ } Give a true answer to the following question: $\{x_{\text{ques}}$ }.*664 B DETAILED CASE STUDIES
665666 Here we show some cases of VrPex with ICL in detail. The samples are from GeoQA, and the
667 model combination is the 10B one. The results are shown in Figure 7, Figure 8, and Figure 9. We
668 can frequently observe the reasoning model corrects the answer given by the perception VLM, and
669 outputs the correct answer. Even though the image descriptions are not always precise, it can help
670 the matching process to find the best matching sample. And thanks to the three similarity measures,
671 we can easily find the most related samples in the training set with the same reasoning process.
672673 As we can see, the perception VLM and the reasoning model work together to get the final answer.
674 In VrPex, the reasoning model rarely get confused by no sufficient information provided, if there is
675 enough information in the question or the descriptions. However, for those questions that require
676 stronger perception capability, we can use larger VLM, or train better perception VLM. On the other
677 hand, to enhance the reasoning capability, we can use or train better reasoning LLM, both ways
678 avoid training reasoning VLMs.
679680 C PERFORMANCE ON PERCEPTION-INTENSIVE REASONING BENCHMARKS
681682 In Section 4.2, we test VrPex on several visual reasoning benchmarks. However, these benchmarks
683 are relatively biased to text reasoning, and do not require strong perception capability. In this section,
684 we test VrPex on some perception-intensive reasoning benchmarks, including M3CoT Chen et al.
685 (2024), RealworldQA X.AI (2024). These benchmarks contains reasoning questions in real world
686 scenarios, requiring the model to frequently refer to the image to get the correct answer. To deal with
687 these benchmarks, we use a larger perception VLM in VrPex, which is Qwen2.5-VL-32B-Instruct,
688 and use DeepSeek-R1-Distill-Qwen-7B as the reasoning LLM. The results are shown in Table 5. As
689 We can see, VrPex achieves comparable visual reasoning capability in visual-intensive scenarios,
690 which shows the extensibility of VrPex, that we can easily improve the perception capability by
691 using larger and better VLMs.
692693 Table 5: Results on perception-intensive reasoning benchmarks. The results are reported in accuracy
694 (%). VrPex also achieves competitive performance on these benchmarks with a larger VLM.
695

Reasoning VLMs	M3CoT	RealworldQA
R1-Onevision-7B Yang et al. (2025)	46.89	37.38
MM-Eureka-Qwen-7B Meng et al. (2025)	36.97	44.05
MM-Eureka-Qwen-32B Meng et al. (2025)	38.57	46.14
Skywork-R1V-38B Peng et al. (2025)	34.21	35.03
Perception VLM	Reasoning LLM	
Qwen2.5-VL-32B-Instruct	DeepSeek-R1-Distill-Qwen-7B	49.87
		47.58

702

703

Query Question

Question: As shown in the figure, the diameter CD of circle O crosses the midpoint G of chord EF , angle $DCF = 20$, then angle EOD is equal to ()
Choices: A: 10° B: 20° C: 40° D: 80° **Answer:** C

Perception VLM image description:

The image shows a circle labeled as circle O . On the circumference of the circle, there is a point E . A line segment EF is drawn parallel to the diameter CD and intersects the circle at points F and G . The line segment OE connects point O to point E . Point G lies on the line segment EF .

Perception VLM question description:

To find the measure of angle EOD , we can use the properties of circles and triangles. Since CD is the diameter of circle O , angle CDE is a right angle (90 degrees). Also, since G is the midpoint of chord EF , triangle CEF is isosceles with $CE = CF$. Therefore, angle $ECF = \angle CFE = 20$ degrees. Since angle CDE is a right angle, angle $EDF = 70$ degrees. Finally, angle $EOD = 2 * \angle EDF = 140$ degrees. Therefore, the answer is option D.

Matched Sample**Matched Sample**

Question: As shown in the figure, CD is the diameter of circle O , and the chord DE passing through point D is parallel to radius OA . If the degree measure of angle A is 50°, then the degree measure of angle EOD is ()
Choices: A. 50° B. 40° C. 30° D. 25°

Reasoning Process of the Matched Sample:

{Step1}: Use properties of parallel lines and identify equal angles created by transversals', 'Step2': 'Apply circle theorems relating central angles and arcs', 'Step3': 'Calculate the measure of arcs based on known angle measures', 'Step4': 'Relate central angles to inscribed angles and their relationships within isosceles triangles', 'Step5': 'Solve for unknown angles using sum of angles properties in triangles and circle geometry'}

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

Query Question

Question: As shown in the figure, in triangle ABC , angle $C = 36$, rotate triangle ABC anticlockwise around point A by 60 to get triangle AED , AD and BC intersect at point F , then the degree of angle AFC is ()
Choices: A: 84° B: 80° C: 60° D: 90° **Answer:** A

Perception VLM image description:

The image shows a triangle $\triangle ABC$ with vertices A , B , and C . Point $\langle D \rangle$ is located on side $\langle AC \rangle$ such that $\langle AD = DC \rangle$, making $\langle D \rangle$ the midpoint of $\langle AC \rangle$. Point $\langle E \rangle$ is located on side $\langle AB \rangle$ such that $\langle AE = EB \rangle$, making $\langle E \rangle$ the midpoint of $\langle AB \rangle$. Point $\langle F \rangle$ is located on side $\langle BC \rangle$ such that $\langle BF = FC \rangle$, making $\langle F \rangle$ the midpoint of $\langle BC \rangle$. The lines $\langle DE \rangle$ and $\langle EF \rangle$ intersect at point $\langle G \rangle$.

Perception VLM question description:

To find the measure of angle AFC , we can use the properties of rotation and the given angles.
1. **Understanding Rotation**: When triangle ABC is rotated anticlockwise around point A by 60 degrees, each vertex moves 60 degrees clockwise relative to its original position. This means that angle BAC remains unchanged because it is the angle formed by the rotation center (point A) and the points of rotation (B and C).
2. **Angle Relationships**:
- Since angle C is 36 degrees, and triangle ABC is rotated around A by 60 degrees, angle DAE will be equal to angle C plus the rotation angle, which is $60 + 36 = 96$ degrees.
3. **Finding Angle AFC **:
- In triangle AEF , angle AEF is the sum of angle EAD and angle DAB , which is $96 + 60 = 156$ degrees.
- Since angle AEF is 156 degrees, and triangle AEF is a straight line, angle AFC must be $180 - 156 = 24$ degrees.
- Therefore, angle AFC is the sum of angle AFC and angle C , which is $24 + 36 = 60$ degrees.
Answer: C

Matched Sample

Question: As shown in the figure, rotating triangle $\triangle ABC$ around point C clockwise by 35° results in triangle $\triangle A'BC$ at point D . If $\angle A'DC = 90$, then the measure of angle $\angle A$ is ()
Choices: A. 35° B. 65° C. 55° D. 25°

Reasoning Process of the Matched Sample:

{step1}: 'Identify the effect of the rotation on the angles and sides of the triangle.', 'step2': 'Use the given right angle (e.g., $\angle A'DC$) to formulate relationships between angles in the resulting triangle.', 'step3': 'Relate the angles in the rotated triangle back to the original triangle using properties of congruence and rotations', 'step4': 'Apply geometric theorems or properties (e.g., sum of angles in a triangle or complementary angles) to solve for the unknown angle.', 'step5': 'Verify that the solution satisfies all given conditions and relationships.'

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

Query Question

Question: As shown in the figure, in triangle ABC , points D and E are the midpoints of sides AB and AC respectively. If $DE = 1.5$, the length of BC is ()
Choices: A: 3 B: 4 C: 2 D: 1 **Answer:** A

Perception VLM image description:

The image shows a triangle ABC with vertices A , B , and C . Line DE is parallel to line BC , intersecting line AB at point D and line AC at point E .

Perception VLM question description:

To find length of BC , we can use the Midsegment Theorem. According to this theorem, the segment connecting the midpoints of two sides of a triangle is parallel to the third side and half its length. Given that DE is the midsegment of triangle ABC , we have: $DE = 1/2 * BC$. We know that $DE = 1.5$, so we can substitute this value into the equation: $1.5 = 1/2 * BC$. Now, we can solve for BC by multiplying both sides of the equation by 2: $BC = 1.5 * 2 = 3$. Therefore, the length of BC is 3.
Answer: A

Reasoning Model**Responses from the Reasoning Model:**

Okay, let's try to figure out the length of BC in this triangle problem. So, we have triangle ABC , and points D and E are the midpoints of sides AB and AC , respectively. That means AD is half of AB , and AE is half of AC . I remember something called the Midsegment Theorem, which says that the segment connecting the midpoints of two sides of a triangle is parallel to the third side and half its length. In this case, DE is the segment connecting the midpoints D and E , so DE should be parallel to BC and half as long as BC .

The problem tells us that DE is 1.5 units long. Since DE is half of BC , I can set up an equation where DE equals half of BC . So, $1.5 = 1/2 * BC$. To find BC , I just need to multiply both sides of the equation by 2.

Let me do that calculation: $1.5 * 2 = 3$. So, BC must be 3 units long. Looking at the answer choices, option A is 3, which matches our calculation.

Well, I think everything is correct, so I'll double-check. Points D and E are indeed midpoints, so DE should definitely be half of BC . And since DE is given as 1.5, doubling that gives BC as 3. Yeah, that seems right. I don't think I made any mistakes here.

The length of BC is determined using the Midsegment Theorem, which states that the segment connecting the midpoints of two sides of a triangle is parallel to the third side and half its length. Given $DE = 1.5$, we find BC by doubling DE , resulting in $BC = 3$. Answer: A

Matched Sample

Question: In triangle ABC , point D is on side AB , $BD = 2AD$, $DE \parallel BC$ intersects AC at point E . If segment $DE = 10$, then the length of segment BC is ()
Choices: A. 15 B. 20 C. 30 D. 40

Reasoning Process of the Matched Sample:

{solving1}: 'Identify the problem involves using similarity of triangles and ratios to determine segment lengths.'
{solving2}: 'Given the problem involves using the Midsegment Theorem (Thales' Theorem), the relationship between corresponding sides of similar triangles is established. The key steps include identifying parallel lines, setting up proportions based on given segment lengths, and solving for the unknown length.'
{step1}: 'Identify Similar Triangles': 'Use the fact that two triangles are similar if they have corresponding angles equal (due to parallel lines).'
{step2}: 'Express Unknowns in Terms of Given Quantities': 'Use the given ratio or segment relationships to express all relevant segments in terms of one variable.'
{step3}: 'Set Up Proportions': 'Use the similarity property to equate the ratios of corresponding sides.'
{step4}: 'Solve for the Unknown': 'Solve the resulting proportion equation to find the desired segment length.'

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

Figure 9: Detailed case 3. The matched sample is highly related to the input one, with the element of similar triangles.