BBA: Bi-Modal Behavioral Alignment for Reasoning with Large Vision-Language Models

Xueliang Zhao[♠*] **Xinting Huang**[★] **Tingchen Fu**[★] **Qintong Li**[♠]
Shansan Gong[♠] **Lemao Liu**[★] **Wei Bi**[★] **Lingpeng Kong**[♠]
[♠]The University of Hong Kong
[★]Tencent AI Lab
xlzhao22@connect.hku.hk

Abstract

Multimodal reasoning stands as a pivotal capability for large vision-language models (LVLMs). The integration with Domain-Specific Languages (DSL), offering precise visual representations, equips these models with the opportunity to execute more accurate reasoning in complex and professional domains. However, the vanilla Chain-of-Thought (CoT) prompting method faces challenges in effectively leveraging the unique strengths of visual and DSL representations, primarily due to their differing reasoning mechanisms. Additionally, it often falls short in addressing critical steps in multi-step reasoning tasks. To mitigate these challenges, we introduce the Bi-Modal Behavioral Alignment (BBA) prompting method, designed to maximize the potential of DSL in augmenting complex multi-modal reasoning tasks. This method initiates by guiding LVLMs to create separate reasoning chains for visual and DSL representations. Subsequently, it aligns these chains by addressing any inconsistencies, thus achieving a cohesive integration of behaviors from different modalities. Our experiments demonstrate that BBA substantially improves the performance of GPT-4V(ision) on geometry problem solving ($28.34\% \rightarrow 34.22\%$), chess positional advantage prediction ($42.08\% \rightarrow 46.99\%$) and molecular property prediction ($77.47\% \rightarrow 83.52\%$).

1 Introduction

The use of domain-specific language (DSL) (Bowman and Hammerlindl, 2008; Edwards, 1994; Weininger, 1988) aims to incorporate multimodal information by providing a precise and unequivocal alternative form using text.[1] Its application has significantly improved the multimodal reasoning capability, yielding notable improvements in intricate contexts, especially within specialized domains such as symbolic reinforcement learning (McGrath et al., 2022; Zahavy et al., 2023; Ruoss et al., 2024) and diverse scientific fields (Winter et al., 2022).

Multimodal reasoning is a fundamental capability for large vision-language models (LVLMs) (OpenAI, 2023; Yang et al., 2023b), crucial for many of their applications. Despite the considerable progress made by LVLMs in multimodal tasks (Lu et al., 2023; Hu et al., 2023), effectively utilizing them for complex multimodal reasoning, particularly in conjunction with DSLs, remains underexplored. The most direct approach is to feed the LVLMs with both visual data (e.g., images) and its corresponding DSL representation along with the textual queries. They are then guided through the Chain-of-Thought (CoT) (Wei et al., 2023) prompting to process step-by-step reasoning. However, a significant issue with this approach is that the reasoning processes derived from different modalities are often inconsistent, or even conflicting. This inconsistency limits the ability of LVLMs to effectively integrate the strengths of visual and DSL representations (§2.1). Moreover, these models encounter difficulties in executing multi-step reasoning (Wu et al., 2023a; Liu and Chen, 2023), which hampers their effectiveness in addressing critical steps within complex problems (§2.2).

To address these challenges, we propose a Bi-Modal Behavioral Alignment (BBA) prompting method that adeptly integrates DSL into complex multimodal reasoning tasks. BBA begins by prompting LVLMs to generate distinct reasoning chains from both visual and DSL representations, and then aligns these chains by resolving inconsistencies, thereby harmonizing the behaviors elicited from various modalities. BBA offers two primary advantages. Firstly, it adopts a "late fusion" strategy (Ghanem et al., 2018; Owens and Efros, 2018),

[*] This work was done during an internship at Tencent AI Lab.

[1] Figure 2 illustrates an example of a DSL tailored for the geometry domain. Further instances of DSLs can be found in Appendix D.

Findings of the Association for Computational Linguistics ACL 2024, pages 7255–7279

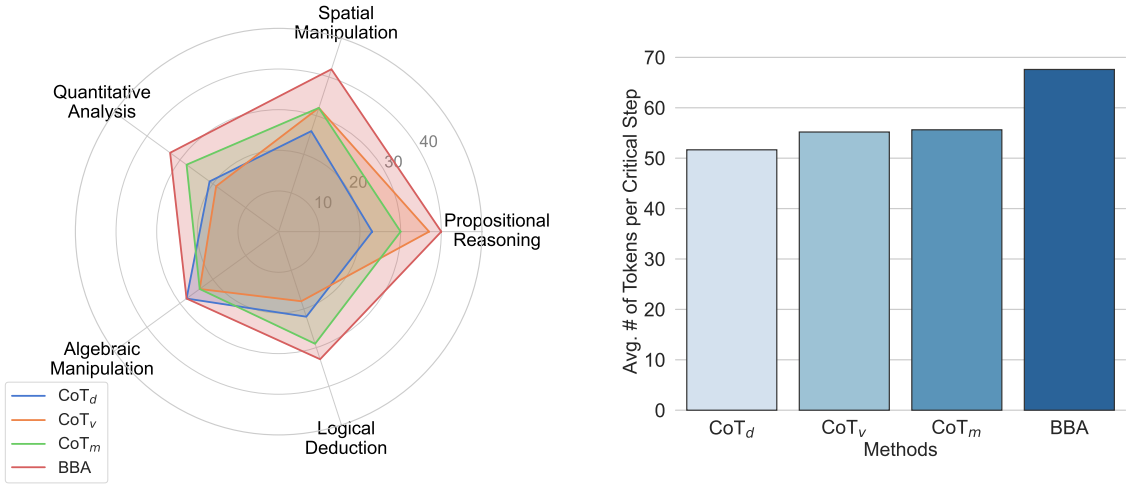


Figure 1: Comparative analyses of different methods in problem-solving and critical step detailing. **Left**: Problem-solving rates across diverse problem types, where CoT$_d$ and CoT$_v$ refer to Chain-of-Thought prompting with DSL and image inputs, respectively, and CoT$_m$ represents the approach combining both inputs. **Right**: Average number of tokens per critical step across different methods.

effectively maintaining the inherent strengths of both the direct vision input and the DSL representation. Secondly, BBA turns the inconsistency across modalities into a beneficial signal that aids in identifying critical steps within reasoning processes. By revealing where the reasoning chains differ, it efficiently allocates more intermediate tokens to these critical steps by resolving the inconsistencies found.

We evaluate BBA on three multimodal reasoning tasks: geometry problem-solving, chess positional advantage prediction, and molecular property prediction. In these diverse applications, BBA demonstrated notable relative improvements, with respective performance improvements of 14.26%, 10.25%, and 6.30%.

2 Pilot Study

In this study, we compare three variants of CoT prompting within domains where DSL is available. These variations include: (1) CoT$_v$, which utilizes only images for grounding responses to queries; (2) CoT$_d$, which relies exclusively on DSL representations for grounding; and (3) CoT$_m$, which integrates both images and DSL representations. We focus on a selection of mathematical geometry problems from the MATH benchmark (Hendrycks et al., 2021), comprising a total of 187 problems that incorporate image inputs. We then explore the difficulties associated with performing multi-modal reasoning using both images and DSL representations, through an empirical examination of distinct success rates across various problem types and the allocation of tokens for critical reasoning steps.

2.1 Performance on Fine-grained Types

Our analysis begins with an assessment of the performance of different models on fine-grained problem types. To this end, we categorically divide the geometry problems based on the primary skills required for resolution, resulting in five categories: (1) Spatial Manipulation, (2) Propositional Reasoning, (3) Logical Deduction, (4) Algebraic Manipulation, and (5) Quantitative Analysis. Additional details on the categorization annotation can be found in Appendix A.1. We proceed to calculate and compare the problem-solving rates for each category.

Figure 1 offers a visual comparison of the models' performances across these categories. It is evident that CoT$_v$ and CoT$_d$ exhibit significantly different levels of effectiveness across these problem types. Specifically, CoT$_v$ shows superior performance in tasks involving spatial manipulation and propositional reasoning, while CoT$_d$ excels in logical deduction, algebraic manipulation, and quantitative analysis. This variation in performance can be attributed to the different reasoning mechanisms enabled by each modality. DSL representations provide detailed information (e.g., precise coordinates) that support logic-oriented operations. On the other hand, images provide intuitive visual

cues that are more conducive to spatial reasoning tasks. Despite the concurrent use of images and DSL representations, CoT$_m$ does not demonstrate uniform improvements across all problem types, indicating the challenge of aligning reasoning mechanisms across modalities. In §4, we elaborate on BBA, which initiates by independently deriving reasoning chains from images and DSL representations, and then aligning these chains by resolving any inconsistencies between them. Unlike CoT$_m$, BBA effectively capitalizes on the strengths of both modalities, achieving comprehensive improvements across all identified problem categories.

2.2 Token Allocation for Critical Steps

In light of recent theoretical advances (Feng et al., 2023; Merrill and Sabharwal, 2023) indicating the effective allocation of intermediate tokens as pivotal for unlocking the expressive power of models in sequential reasoning tasks, we delve into the allocation of intermediate tokens for addressing critical steps in problem-solving. A critical step in solving mathematical problems is defined as the point at which an essential insight, decision, or application of a method is crucial for obtaining the correct solution, typically involving a significant conceptual leap, strategic theorem application, or key calculation that influences the subsequent problem-solving process. For each problem, we identify all critical steps, categorizing each step in the generated solution as either corresponding to one of the identified critical steps or not, and then sum the tokens for steps within a generated solution that are associated with the same critical step. Details on the annotation of critical steps are provided in Appendix A.2.

Figure 1 demonstrates that merely combining images and DSL representations in inputs is insufficient for effectively allocating more tokens to critical steps, thus reducing the expressive power of LLMs and leading to inferior overall performance (as discussed in §5.4). We hypothesize that this limitation arises from the current inefficiencies of LLMs in exploring the solution space for complex problems (Yang et al., 2023b), resulting in their struggle to accurately identify critical steps. As will be discussed in §4.2, BBA is more effective in discerning and addressing critical steps by uncovering and reconciling discrepancies among reasoning chains derived from different modalities.

3 Preliminaries

3.1 Problem Formulation

This study focuses on multi-modal reasoning tasks, specifically where the visual modality is represented as an image, coupled with a DSL that accurately depicts the image. Our objective is to predict an answer to a given question q, associated with an image v and a DSL representation d, adhering to specific task requirements (e.g., solving mathematical problems).

The emergence of LVLMs has streamlined this process. Owing to extensive pre-training on trillions of tokens, these models can accurately interpret various instructions and execute the corresponding tasks. In this paradigm, the model parameters are denoted by $\boldsymbol{\theta}$, and the answer $\hat{a}$ is generated as $\hat{a} = \arg\max_a p(a \mid q, v, d; \boldsymbol{\theta})$, where the inputs are reformulated into well-crafted prompts using specific templates, designed to elicit the desired response from the LVLMs.

3.2 Chain-of-Thought Prompting

Recently, chain-of-thought prompting has gained recognition as an effective technique for enhancing the reasoning capabilities of language models (Wei et al., 2023). This method decomposes the original task into two distinct phases: rationale generation and answer prediction. In the rationale generation phase, a rationale $\hat{r}$ is derived as $\hat{r} = \arg\max_r p(r \mid q, v, d; \boldsymbol{\theta})$, leveraging a query augmented with an instruction designed to initiate stepwise analytical thinking (Kojima et al., 2022)). Subsequently, the answer is often deduced directly from the rationale, utilizing heuristic string-matching methods for precise identification.

4 Method

This work aims to tackle two primary challenges in multi-modal reasoning: (1) the integration of the inherent strengths of both visual and DSL representations, and (2) the identification and resolution of critical steps within these tasks. To address these challenges, we introduce the BBA prompting method, an innovative approach that seeks to unleash the power of DSL in enhancing complex multi-modal reasoning tasks. Figure 2 offers an overview of our proposed methodology. BBA initiates by employing LVLMs to generate reasoning chains separately from visual and DSL inputs. Subsequently, these chains proceed through an alignment phase, wherein inconsistencies are identified

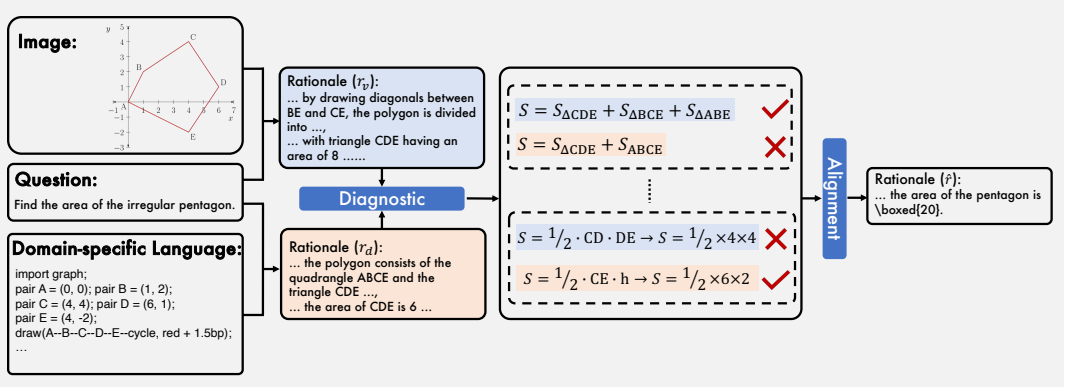


Figure 2: An instantiation of the proposed BBA method.

and reconciled, ensuring the harmonization of behaviors derived from each modality.

Road Map. The rest of this section is structured as follows: We begin by detailing the process of eliciting reasoning chains from both vision and DSL representations in §4.1. This is followed by an elaboration on diagnosing and rectifying inconsistencies between these reasoning chains and the methods of aligning behaviors from different modalities in §4.2. Lastly, in §4.3, we detail how BBA effectively identifies and addresses critical steps in the reasoning process.

4.1 Bi-Modal Behavior Eliciting

The objective of this phase is to effectively harness the unique strengths of vision and DSL representations in answering a given question. Unlike vanilla CoT prompting, which intermingles the reasoning processes of these two modalities, BBA seeks to elicit reasoning chains from each modality independently. This approach allows the vision-based reasoning chain to deliver more credible steps in intuitive and spatial reasoning, while the DSL-based reasoning chain provides steps with greater reliability in precise computation. The formal definition of this process is as follows:

$$\begin{aligned} r_{\mathrm{v}} &= \arg\max_{r} p(r \mid q, v; \boldsymbol{\theta}) \\ r_{\mathrm{d}} &= \arg\max_{r} p(r \mid q, d; \boldsymbol{\theta}). \end{aligned} \tag{1}$$

where r_{v} and r_{d} represent the reasoning chains derived from the vision and DSL representations, respectively.

4.2 Behavior Alignment

This phase is centered on aligning the reasoning chains from different modalities to capitalize on the best of both worlds in multi-modal reasoning. We initiate this process with diagnostic checks to uncover inconsistencies between the chains, including variances in intermediate steps and the final answers. Following this, an aligned reasoning chain is created by addressing the discrepancies identified in the diagnostics. When different methods produce conflicting results, it often indicates an error in at least one approach. The divergence point then becomes a crucial indicator of where deeper understanding or more meticulous application of principles is necessary. The model is subsequently instructed to thoroughly examine the derivations from both modalities and ascertain accurate conclusions. The diagnostic results are formally obtained as follows:

$$r_{\mathrm{inc}} = \arg\max_{r} p(r \mid r_{\mathrm{v}}, r_{\mathrm{d}}; \boldsymbol{\theta}), \tag{2}$$

where r_{inc} denotes the rationale for inconsistencies identified during the diagnostic process. Next, the formation of the aligned reasoning chain is defined as:

$$\hat{r} = \arg\max_{r} p(r \mid r_{\mathrm{v}}, r_{\mathrm{d}}, r_{\mathrm{inc}}; \boldsymbol{\theta}) \tag{3}$$

where the final rationale $\hat{r}$ includes the definitive answer a within special tokens.

4.3 Discussion

The strengths of BBA can be mainly attributed to its capability to address critical steps in multi-step reasoning problems. BBA excels in addressing critical steps primarily due to two reasons: (1) the critical step is more easily identified by contrasting different solutions, revealing their divergences; and (2) learning from these differences allows for a more efficient allocation of intermediate tokens to these critical steps. Drawing from cognitive learning principles observed in humans, it is a plausible

extrapolation that identifying and rectifying disparities between various methods fosters a deeper comprehension of essential aspects of a problem (Munzar et al., 2021). Furthermore, encountering and acknowledging mistakes enhances the reasoning process, paralleling human problem-solving strategies. This not only deepens the understanding but also facilitates the allocation of additional reasoning tokens, thereby amplifying the model's capacity to resolve critical steps (Feng et al., 2023; Merrill and Sabharwal, 2023).

5 Experiments

5.1 Datasets and Evaluation

We assess the efficacy of BBA across three multimodal reasoning tasks spanning distinct domains: geometry problem-solving, chess positional advantage prediction, and molecular property prediction.

Geometry Problem-Solving. This task involves predicting a free-form solution to a given geometry problem. We utilize the geometry subset of the MATH benchmark (Hendrycks et al., 2021) for this task, selecting only those problems that include Asymptote code (Bowman and Hammerlindl, 2008), a domain-specific language (DSL) used for depicting geometric figures. This process resulted in a dataset of 187 problems, which we refer to as **G-MATH**. The official evaluation script from the MATH benchmark is employed to compute accuracy by comparing the predicted answers with the correct answers.

Chess Positional Advantage Prediction. The objective in chess positional advantage prediction is to classify a given chessboard state as being advantageous for White, advantageous for Black, or balanced. This task evaluates the model's capacity to correlate with the actual value of a chessboard state, determined by chess engines after extensive analysis. For evaluation, we compiled a dataset of 183 game snippets, applying Stockfish 15 at a search depth of 18 to assess the winning probability for the white pieces. We classified the winning probabilities into three intervals: 0–33% indicating an advantage for Black, 34–66% denoting a balanced state, and 67–100% suggesting an advantage for White. We refer to this dataset as **ChessAdv**, employing Forsyth-Edwards Notation (FEN) (Edwards, 1994) as the DSL for this domain. Classification accuracy serves as the evaluation metric.

Molecular Property Prediction. Molecular property prediction focuses on determining whether a molecule exhibits a certain property based on its molecular graph. The **MUTAG** benchmark dataset (Debnath et al., 1991) is used for this purpose, comprising 188 chemical compounds categorized into two classes based on their mutagenic effects on a bacterium. The Simplified Molecular-Input Line-Entry System (SMILES) (Weininger, 1988) is utilized as the DSL in this domain, with classification accuracy as the metric for evaluation.

5.2 Baselines

For comparative evaluation, we adopt the following baselines:

DSL or Visual-Only Methods. (1) **CoT**$_v$: Implements chain-of-thought prompting (Wei et al., 2023), omitting DSL representations and relying solely on images; (2) **CoT**$_d$: Utilizes chain-of-thought prompting, excluding images to focus exclusively on DSL representations; (3) **Plan-and-Solve**: Formulates a plan to segment the overall task into manageable subtasks for sequential execution (Wang et al., 2023a); and (4) **Least-to-Most**: Breaks complex problems into simpler, sequential subproblems, leveraging solutions of preceding subproblems to facilitate solving subsequent ones (Zhou et al., 2022).

Integrated DSL and Visual Methods. (1) **CoT**$_m$: Employs chain-of-thought prompting using a combination of both DSL representations and images; (2) **CCoT**: enhances compositional reasoning by integrating visual and DSL inputs, substituting the scene graph with DSL for fair comparison (Mitra et al., 2023); (3) **DDCoT**: Introduces negative-space prompting and multimodal reasoning by dividing cognitive tasks between reasoning and recognition, enhancing reasoning with visual recognition capabilities (Zheng et al., 2023).

All baseline methods, alongside BBA, are implemented on **GPT-4V(ision)** (OpenAI, 2023), utilizing the `gpt-4-vision-preview` version to ensure a fair and consistent comparison.

5.3 Implementation Details

For geometry problem-solving and chess positional advantage prediction, we employ zero-shot prompting. In the case of molecular property prediction, we augment the instruction with four <SMILES, category> pairs, given the challenge this specialized task presents to the GPT-4V(ision). It is cru-

Methods	With DSL	With Figure	G-MATH	ChessAdv	MUTAG	Avg.
CoT$_v$ (Wei et al., 2023)	✗	✓	23.53	40.98	75.82	46.56
CoT$_d$ (Wei et al., 2023)	✓	✗	23.12	38.80	76.92	46.01
Plan-and-Solve (Wang et al., 2023a)	✓	✗	25.67	42.62	78.57	48.73
Least-to-Most (Zhou et al., 2022)	✓	✗	25.13	38.25	73.63	45.47
CoT$_m$ (Wei et al., 2023)	✓	✓	28.34	42.08	77.47	49.09
CCoT (Mitra et al., 2023)	✓	✓	26.74	39.34	68.68	44.75
DDCoT (Zheng et al., 2023)	✓	✓	29.95	37.70	73.08	46.74
BBA (Ours)	✓	✓	**34.22**	**46.99**	**83.52**	**54.71**

Table 1: Evaluation results for geometry problem-solving (**G-MATH**), chess positional advantage prediction (**ChessAdv**), and molecular property prediction (**MUTAG**), including average performance. Numbers in bold denote the best performance.

Methods	With DSL	With Figure	G-MATH	ChessAdv	MUTAG	Avg.
BBA (Ours)	✓	✓	**34.22**	**46.99**	**83.52**	**54.71**
-diagnostic	✓	✓	32.09	41.53	78.57	50.54
-visual	✓	✗	28.34	37.70	61.54	42.39
-dsl	✗	✓	27.27	36.07	75.82	46.20

Table 2: Ablation study results with best performances highlighted in bold.

cial to note that these SMILES representations are excluded from the test cases to prevent data leakage. Detailed instructions for these tasks can be found in Appendix B. To interact with the `gpt-4-vision-preview`, the *temperature* and *top_p* are set to 0 and 1, respectively, to ensure deterministic outputs, while the *max_tokens* parameter is capped at 2048.

5.4 Main Results

The results of our experiments, presented in Table 1, reveal several key observations: (1) BBA surpasses all compared baseline methods, achieving relative improvements of 14.26%, 10.25%, and 6.30% in geometry problem-solving, chess positional advantage prediction, and molecular property prediction, respectively. This superior performance can be attributed to BBA's adeptness at leveraging the combined strengths of both visual and DSL representations, along with its capacity to pinpoint and address critical steps; (2) The integration of DSL and visual information proves advantageous for multi-modal reasoning tasks. Our results demonstrate that CoT$_m$ achieves the second-best average performance, notably excelling in geometry problem-solving. This task benefits markedly from the complementary insights provided by DSL and visual inputs, indicating the value of integrating these modalities; and (3) The process of effectively merging DSL representations with visual data poses a significant challenge, as evidenced by the subpar performance of CCoT.

6 Analysis

6.1 Ablation Study

This ablation study evaluates four variants of our model across three datasets, as shown in Table 2. These variants comprise the full method and three variants: one without the diagnostic check ("-diagnostic"), where the reasoning process is solely based on divergent reasoning chains from different modalities without any verification; one lacking image inputs ("-visual"), where the model's assessment of reasoning chains relies exclusively on the DSL representation and its intrinsic knowledge; and one excluding DSL inputs ("-dsl"), where the evaluation of reasoning chains depends solely on visual information and the model's inherent understanding.

The results demonstrate that our full method outperforms all variants on the datasets, indicating the crucial role of combining DSL and visual inputs alongside diagnostic checks for identifying discrepancies and enhancing problem-solving in critical steps. Notably, the exclusion of visual inputs results in the most significant performance drop, highlighting the vital contribution of images to the efficacy of multi-modal reasoning tasks.

	Level 1	Level 2	Level 3	Level 4	Level 5	Avg.
BBA (Ours)	**71.43**	**53.13**	**44.12**	16.98	**17.02**	**34.22**
CoT_m	61.90	37.50	29.41	**24.53**	10.64	28.34
CoT_v	52.38	37.50	26.47	13.21	10.64	23.53
CoT_d	47.62	50.00	29.41	7.69	6.38	23.12

Table 3: Evaluation results on the geometry problem-solving task. Numbers in bold indicate the best performance.

	Level 1	Level 2	Level 3	Avg.
BBA (Ours)	**57.41**	**43.21**	**41.67**	**46.99**
CoT_m	51.85	37.04	39.58	42.08
CoT_v	48.15	38.27	37.50	40.98
CoT_d	46.30	33.33	39.58	38.80

Table 4: Evaluation results on the chess positional advantage prediction task. Numbers in bold indicate the best performance.

6.2 Analysis on Different Complexities

This experiment delves into how BBA performs under varying problem complexities, comparing it with three variants of chain-of-thought prompting. Our focus is on geometry problem-solving and chess positional advantage prediction due to the labor-intensive nature of assessing the difficulty of molecular graphs. For geometry, we utilize the difficulty levels outlined by the MATH benchmark (Hendrycks et al., 2021), and for chess, we classify problems into three difficulty levels based on the centipawns returned by Stockfish 15.

Table 3 and Table 4 present the results. BBA consistently outperforms competitors across nearly all difficulty levels, except level 4 in geometry problem-solving. Integrating DSL and image inputs proves advantageous, as CoT_m typically surpasses the performance of both CoT_v and CoT_d. However, achieving universal improvements through direct integration presents a significant challenge (as discussed in §2.1). In geometry problem-solving, DSL representations are particularly effective in simpler problems, but this advantage diminishes with increased complexity. We hypothesize this is due to the lengthening of Asymptote code in more complex problems. For instance, the average Asymptote code length is 186.89 for levels 1 to 3, but increases to 217.80 for levels 4 to 5, whereas the length of FEN notation remains relatively stable across different levels of difficulty.

6.3 Comparison with Self-Refine Prompting

This experiment explores the efficacy of self-refine prompting (Madaan et al., 2023), a technique that improves previous outputs through iterative feedback and refinement, as a potential substitute for the diagnostic check and alignment phases in BBA. We have adapted the conventional self-refine prompting approach to accommodate both DSL and image inputs, while preserving the original implementation details to the greatest extent. This experiment evaluates three versions of self-refine prompting, denoted as Self-Refine (x turns), with $x-1$ indicating the count of refinement cycles and x varying from 2 to 4.

Table 5 presents the results. The findings reveal that BBA consistently surpasses the various versions of self-refine prompting. This indicates the superiority of directing LVLMs to pinpoint inconsistencies between divergent solutions over merely generating feedback based on the knowledge embedded within their parameters. Moreover, recent work (Huang et al., 2023) corroborates our findings, demonstrating that LLMs frequently encounter difficulties in adjusting their responses based solely on their inherent capabilities. This is further validated by our results, which indicate a decline in the performance of the self-refine prompting as the number of refinement iterations increases.

6.4 Case Study

Due to space constraints, the case study is included in Appendix D.

Methods	G-MATH	ChessAdv	MUTAG	Avg.
BBA (Ours)	**34.22**	**46.99**	**83.52**	**54.71**
Self-Refine (2 turns)	30.48	43.17	73.63	48.91
Self-Refine (3 turns)	28.34	42.08	71.98	47.28
Self-Refine (4 turns)	28.88	38.80	68.68	45.29

Table 5: Comparative analysis of BBA versus Self-Refine prompting. Numbers in bold denote the best performance.

7 Related Work

7.1 Multi-Modal CoT Prompting

An advanced methodology for zero-shot image reasoning leverages CoT prompting, a technique that breaks down complex tasks into simpler, sequential thought processes to simulate human reasoning (Lu et al., 2022; Zhang et al., 2023b; Wang et al., 2023c). Due to the structural differences between LVLMs and LLMs, additional improvements have been made to adapt CoT for wider applications. To illustrate, QVix (Yang et al., 2023a) leverages LLMs' linguistic skills to enhance LVLMs' visual content analysis; V* (Wu and Xie, 2023) enhances the precise targeting of specific visual elements; Wu et al. (2023b) address CoT prompting's limitations by adopting a "Description then Decision" strategy for complex vision-linguistic tasks; CoCoT (Zhang et al., 2024) uses a contrastive CoT approach for multiple image inputs; ViLa (Hu et al., 2023) merges perceptual data with CoT for physically-grounded task planning; and DDCoT (Zheng et al., 2023) assigns tasks to relevant components, differentiating reasoning and recognition roles and integrating visual recognition into the reasoning process. Despite these advancements, the strategic use of prompting mechanisms to seamlessly integrate DSLs into LVLMs presents an untapped potential, a gap our research aims to bridge by pioneering in this specific area.

7.2 Multiple Chains Prompting

Following the progress of the chain-of-thought prompting, a series of efforts have been made to enhance factuality by generating multiple reasoning chains. Building on this progress, the research focuses on three main approaches: self-consistency (Wang et al., 2022a), self-refinement (Madaan et al., 2023; Shinn et al., 2023; Chen et al., 2023b), and multi-agent debate (Du et al., 2023; Liang et al., 2023; Xiong et al., 2023). Self-consistency (Wang et al., 2022b) involves a method where various reasoning paths are first generated, and then the most consistent answer is selected through a process akin to majority voting. Self-refinement (Madaan et al., 2023) leverages the inherent capabilities of LLMs to generate feedback for previous outputs, refining them based on this feedback. However, recent research (Huang et al., 2023) indicates that LLMs face challenges in providing accurate feedback independently, suggesting that feedback from external environments (First et al., 2023) is a more effective alternative. Multi-agent debate (Du et al., 2023) aims to replicate real-world debate scenarios, fostering a consensus by incorporating outputs from previous iterations in each debate cycle. These methods, while innovative, have yet to fully address the need for identifying intermediate inconsistencies between multiple chains which play a crucial role in pinpointing the critical steps necessary for solving complex tasks. Moreover, the requirement for multiple invocations of LLMs, particularly with proprietary LVLMs (OpenAI, 2023), significantly increases the associated costs.

We provide a detailed review of the literature on large vision-language models in Appendix C.

8 Conclusion

In conclusion, our work introduces the Bi-Modal Behavioral Alignment (BBA) prompting method, a novel approach that significantly enhances the multimodal reasoning capabilities of GPT-4V(ision) by integrating DSL. By generating and aligning separate reasoning chains for visual and DSL representations, BBA addresses the challenges of inconsistent reasoning mechanisms and the execution of multi-step reasoning tasks. Our experiments across diverse domains, including geometry problem-solving, chess positional advantage prediction, and molecular property prediction, demonstrate the effectiveness of BBA, showcasing notable improvements in performance.

Ethical Considerations

In adherence to the established Code of Ethics, this work exclusively employs publicly accessible data and information, ensuring no private or confidential resources are utilized.

Limitations

BBA marks a significant advancement in the field of multi-modal reasoning, incorporating DSLs. Despite this, it is beneficial to address several limitations to fully exploit its capabilities:

(1) BBA demonstrates significant improvements in three distinct domains: geometry, chess, and molecular biology. Yet, its application in other areas, especially those without custom DSLs, has not been extensively explored. Adapting BBA by substituting DSL representations with alternative, advanced representations, such as scene graphs (Yang et al., 2018), could be advantageous. These alternatives, although less precise and informative in capturing image nuances, offer a valuable research direction.

(2) The primary aim of this work is to develop a prompting method that complements, but is distinct from, other advanced technologies which respond to environmental feedback (Yao et al., 2022; Xie et al., 2023). Integrating and responding to environmental feedback to develop a more adaptive and intelligent agent presents an intriguing future research direction.

Acknowledgements

We extend our gratitude to the HKU NLP group and the anonymous reviewers for their invaluable suggestions, which significantly enhanced this work. This work is partially supported by the joint research scheme of the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) under grant number N_HKU714/21.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

John C Bowman and Andy Hammerlindl. 2008. Asymptote: A vector graphics language. *TUGboat: The Communications of the TEX Users Group*, 29(2):288–294.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023b. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2.

Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. 1991. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Steven J Edwards. 1994. Portable game notation specification and implementation guide. *Retrieved April*, 4:2011.

Guhao Feng, Yuntian Gu, Bohang Zhang, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: a theoretical perspective. *arXiv preprint arXiv:2305.15408*.

Emily First, Markus N Rabe, Talia Ringer, and Yuriy Brun. 2023. Baldur: Whole-proof generation and repair with large language models. *arXiv preprint arXiv:2303.04910*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Cuong Duc Dao. 2018. The activitynet large-scale activity recognition challenge 2018 summary. *arXiv preprint arXiv:1808.03766*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. 2023. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2023b. Fine-tuning multimodal llms to follow zeroshot demonstrative instructions. *arXiv preprint arXiv:2308.04152*, 3.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Mengchen Liu and Chongyan Chen. 2023. An evaluation of gpt-4v and gemini in online vqa. *arXiv preprint arXiv:2312.10637*.

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Yujie Lu, Xiujun Li, William Yang Wang, and Yejin Choi. 2023. Vim: Probing multimodal large language models for visual embedded instruction following. *arXiv preprint arXiv:2311.17647*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. 2022. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119.

William Merrill and Ashish Sabharwal. 2023. The expresssive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2023. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2311.17076*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Brendan Munzar, Krista R Muis, Courtney A Denton, and Kelsey Losenno. 2021. Elementary students' cognitive and affective responses to impasses during mathematics problem solving. *Journal of Educational Psychology*, 113(1):104.

OpenAI. 2023. Gpt-4 technical report.

Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Anian Ruoss, Grégoire Delétang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, and Tim Genewein. 2024. Grandmaster-level chess without search. *arXiv preprint arXiv:2402.04494*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023b. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. 2023c. Filling the image information gap for vqa: Prompting large language models to proactively ask questions. *arXiv preprint arXiv:2311.11598*.

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.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

Benedikt Winter, Clemens Winter, Johannes Schilling, and André Bardow. 2022. A smile is all you need: predicting limiting activity coefficients from smiles with natural language processing. *Digital Discovery*, 1(6):859–869.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*.

Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. 2023a. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:2310.16534*.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C Gee, and Yixin Nie. 2023b. The role of chain-of-thought in complex vision-language reasoning task. *arXiv preprint arXiv:2311.09193*.

Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, et al. 2023. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv:2310.10634*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685.

Kaiwen Yang, Tao Shen, Xinmei Tian, Xiubo Geng, Chongyang Tao, Dacheng Tao, and Tianyi Zhou. 2023a. Good questions help zero-shot image reasoning. *arXiv preprint arXiv:2312.01598*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1).

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

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

Tom Zahavy, Vivek Veeriah, Shaobo Hou, Kevin Waugh, Matthew Lai, Edouard Leurent, Nenad Tomasev, Lisa Schut, Demis Hassabis, and Satinder Singh. 2023. Diversifying ai: Towards creative chess with alphazero. *arXiv preprint arXiv:2308.09175*.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *arXiv preprint arXiv:2310.16436*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A More Details about Pilot study

A.1 Annotation of Problem Types

In this experiment, we (the authors) identified five primary categories essential for problem-solving in geometry: Spatial Manipulation, Propositional Reasoning, Logical Deduction, Algebraic Manipulation, and Quantitative Analysis. These categories were conceptualized based on our extensive experience in geometry problem-solving, where Spatial Manipulation and Propositional Reasoning are aligned with intuitive reasoning, and the remaining categories, namely Logical Deduction, Algebraic Manipulation, and Quantitative Analysis are associated with precise, calculation-based approaches.

For the categorization of problems, we employed `gpt-4-1106-preview` to label each problem according to these predefined categories. This process was facilitated using a prompt, as detailed in Figure 3, wherein placeholders for the problem statement and its solution were substituted with the actual content. The categorization was determined based on the highest-scoring category for each problem. In instances where multiple categories achieved the highest score, the final classification was randomly assigned from among these top-scoring categories.

A.2 Annotation of Critical Steps

The prompts employed for the annotation of the critical steps associated with each problem are depicted in Figures 4 and 5. Specifically, Figure 4 illustrates the instructions utilized to identify the critical steps necessary for solving the problem, alongside the provision of a ground-truth solution. This is necessitated by the current limitations of LLMs in addressing geometry problems. Figure 4 demonstrates the instructions employed for categorizing each step within the generated solution as either corresponding to one of the previously identified critical steps or not. All annotations were performed using `gpt-4-1106-preview` The accuracy of these critical annotations was subsequently verified by the authors.

B System Instructions

In this work, we employ system instructions (Mukherjee et al., 2023) to guide LVLMs toward producing the requisite outputs across various stages, as elaborated in §4. The specific system instructions applied for geometry problem-solving, chess positional advantage prediction, and molecular property prediction are illustrated in Figures 6-8.

C More Discussions about Related Work

C.1 Large Vision-Language Models

Driven by the success of LLMs, research in vision-language models has begun to explore how to equip LLMs with the ability to process visual inputs to solve a variety of multimodal tasks (Alayrac et al., 2022; Liu et al., 2023b; Zhu et al., 2023; Peng et al., 2023). These advancements leverage the extensive prior knowledge and intricate reasoning capabilities established by recent LLMs (Touvron et al., 2023a,b), integrating them with visual modalities to enhance performance across diverse applications. A common approach in this field involves using a vision model to transform images into visual features, which are then aligned with the LLMs' feature space through a bridge module. This alignment is achieved either by employing a linear or MLP layer as a bridge module (Liu et al., 2023a,b; Chen et al., 2023a; Gao et al., 2023; Peng et al., 2023; Zhang et al., 2023a; Wang et al., 2023b) or by designing more complex bridge networks to compress or adaptively select visual information (Alayrac et al., 2022; Zhu et al., 2023; Li et al., 2023b; Dai et al., 2023; Li et al., 2023c; Awadalla et al., 2023; Zhao et al., 2023; Li et al., 2023a; Ye et al., 2023; Bai et al., 2023). The datasets employed to train these bridge modules are notably varied, ranging from synthetic image-text pairs (Li et al., 2023b) and large-scale multi-modal web corpora (Alayrac et al., 2022) to specialized datasets for conversational (Liu et al., 2023b) and academic purposes (Liu et al., 2023a). Notably, the emergence of advanced, proprietary LVLMs such as GPT-4V(ision) (OpenAI, 2023) and Gemini (Team et al., 2023) has set new benchmarks by demonstrating superior performance in multimodal task execution and cognitive reasoning, surpassing human expertise. While these models have achieved significant success in vision-language tasks, the exploration of unlocking the full potential of DSLs, especially for intricate multi-modal reasoning tasks, remains largely unexplored.

D Case Study

To enhance our understanding of the effectiveness of BBA, we showcase a series of outputs in Figures 9-14. These illustrations demonstrate that our

Your task is to evaluate a given <problem, solution> pair in terms of its relevance to each of five defined problem types. This evaluation requires you to assign a score for the pair's relevance to each problem type. The scoring system is a nine-point scale, where 1 denotes 'very strong irrelevant' and 9 denotes 'very strong relevant'. The full scale is as follows: 1 - very strong irrelevant, 2 - strong irrelevant, 3 - irrelevant, 4 - weak irrelevant, 5 - borderline relevant, 6 - weak relevant, 7 - relevant, 8 - strong relevant, 9 - very strong relevant.

Problem Types:

The five problem types you will assess against are:

1. **Spatial Manipulation**: Mentally manipulating objects to understand their spatial relationships.
2. **Propositional Reasoning**: Understanding and using the concept of proportionality to solve problems.
3. **Logical Deduction**: Applying rules and principles to deduce new information.
4. **Algebraic Manipulation**: Using algebra to express and solve for unknowns.
5. **Quantitative Analysis**: Applying mathematical calculations to solve problems.

Reporting Format:

Your assessment should be recorded in a valid Python dictionary format. The dictionary keys represent the seven problem types and must not be altered. Each key should be associated with your assigned score from the nine-point scale. The format of your evaluation should be as follows:

```python
{
   "spatial manipulation": <score>,
   "propositional reasoning": <score>,
   "logical deduction": <score>,
   "algebraic manipulation": <score>,
   "quantitative analysis": <score>,
}
```

In this dictionary, replace `<score>` with the score you've determined for each problem type based on your analysis of the <problem, solution> pair. Ensure that the keys in the dictionary exactly match the problem types listed above.

Problem: {problem}

Solution: {solution}

Assessment: ```python

Figure 3: Illustration of the prompt utilized for category annotation.

approach is capable of accurately identifying discrepancies between two chains. This identification process is subsequently leveraged to synthesize the final solution.

Your assignment is to analyze a specific <problem, solution> pair and distill it down to 1 or 2 absolutely crucial steps in the solution process. These steps should be identified and presented as a Python list of strings. Focus intensely on isolating the most vital parts of solving the problem at hand. Remember, the essence of a critical step lies in its pivotal role in the solution: it's where a key insight, decision, or the application of a specific method is imperative for arriving at the correct answer. This typically involves a major conceptual breakthrough, the strategic implementation of a theorem, or a decisive calculation that largely dictates the solution's direction.

Given the complexity of the problems, limit your identification to only one or two such steps. Avoid generalities or less impactful steps. Present your findings in the following Python list format, ensuring each string succinctly encapsulates a step that's fundamental to the solution's success:

```python
[
   "most_critical_step_1",
   "most_critical_step_2"  # Include only if absolutely necessary
]
```

Each string in the list should be a clear and concise representation of a step without which the problem cannot be effectively solved, highlighting its importance in the overall problem-solving process.

Problem: {problem}

Solution: {solution}

Output: ```python

Figure 4: Illustration of the prompt utilized for critical step identification.

You are tasked with analyzing a provided predicted answer in relation to a specific problem and its critical steps. Follow these steps for your analysis:

1. **Break Down the Predicted Answer**: Decompose the predicted answer into distinct steps. Be careful to retain all existing information without introducing new content.

2. **Evaluate and Document Each Step**: For each identified step, perform a rigorous and comprehensive evaluation. Assign each step a category based on its relation to the critical steps in the problem-solving process. Use numbers 1, 2, 3, ..., N for each critical step, with N being the total number of critical steps. If a step is unrelated to any critical steps, label it as 0.

3. **Format the Results**: Organize your evaluations as a Python list of dictionaries, each representing a step:
   - `"step"`: The step being evaluated.
   - `"classification"`: The category number of the step.

Example Output Format:

```python
[
   {
      "step": "step_1",
      "classification": 2,
   },
   {
      "step": "step_2",
      "classification": 0,
   },
   ...
]
```

Problem: {problem}

Solution: {solution}

Critical Steps:
{critical_steps}

Output: ```python

Figure 5: Illustration of the prompt utilized for categorizing each step within the generated solution.

System Instruction:

Let's dive into a role-playing activity centered on solving a geometry problem. You will assume the roles of two students and a teacher, each contributing to the solution process. Ensure that you strictly adhere to the following steps and maintain the same format throughout:

1. **Solution from Student A (Code-Based):** Student A should start by applying a structured approach to identify and extract supporting facts from the code, which will form the foundation of his solution. This process includes:

(1) Student A should look for coordinates like `(x,y)`, geometric shapes and figures (e.g., `draw(circle(...))` for circles, `draw(A--B)` for line segments), labels (such as `label("A", (x,y), ...)`), and markers like `rightanglemark(...)` which indicate specific geometric properties. He should also note any specified lengths or angle measures.

(2) He needs to understand the basic geometric concepts and use the coordinates and shapes to comprehend the structure of the figure. For example, coordinates of three points can help infer the triangle they form.

(3) Labels correspond to key elements like vertices or centers, and markers indicate properties like right angles or parallel lines. These are crucial for understanding the figure.

(4) If lengths or angles are specified, these measurements should be used to understand relationships between different parts of the figure.

(5) Student A must synthesize information from coordinates, labels, and measurements to formulate supporting facts, such as deducing a triangle is isosceles if two sides are labeled with equal lengths.

After identifying and summarizing these supporting facts, which encompass coordinates, geometric shapes, labels, markers, and measurements, Student A must ensure that each step of his solution is aligned with these facts or logically follows from a previously established step. His answer should clearly demonstrate how the facts extracted from the Asymptote code, focusing on geometric relationships and precise measurements, logically lead to the final solution.

2. **Solution from Student B (Figure-Based):** Student B must initiate his solution process by explicitly restating the content depicted in the figure as his primary reference. He must then make a detailed and explicit declaration of his commitment to an entirely independent and divergent approach. He must affirm that his solution will not reference or rely on any conclusions, formulas, or intermediate steps used by Student A, pledging to a new and unique perspective. This commitment to an alternative methodology must be actively reiterated and evident at each step of his analysis.

After acknowledging the figure, Student B is required to explicitly identify and summarize specific supporting facts unique to it. He should focus on aspects such as visual patterns, spatial relationships, or principles of similarity and proportionality that are not directly related to the formulas or methods used by Student A. He must develop a solution using a different approach or line of reasoning that can be logically derived from these figure-specific insights.

Before formulating each intermediate step in his quantitative analysis, Student B must pause to explicitly reassess and ensure that his approach is not only independent but also methodologically different from Student A's. This reassessment should involve a critical reflection on how the current step and the overall reasoning process diverge from Student A's methodology in terms of approach, application of principles, or interpretation of the figure. He should only proceed to the next step after this confirmation, ensuring that each part of his solution distinctly reflects his independent analysis and unique understanding of the figure.

3. **Comparing Solutions (Teacher's Analysis):** Now, taking on the teacher's role, you should compare the derived solutions of Student A and Student B. Your focus should be on pinpointing and summarizing any differences in their final answers and the methods they employed.

4. **Final Solution (Teacher's Conclusion):** As the teacher, develop a final, accurate solution by addressing the discrepancies identified in the previous step. Your solution should be comprehensive and clearly presented.

It's crucial to present the final result in LaTeX using a `\\boxed{}` without any units.

Figure 6: System instruction for geometry problem-solving.

System Instruction:

This task involves analyzing a chess position from two distinct perspectives: one based on the FEN notation and the other on the visual representation of the board (figure). The objective is to develop detailed, step-by-step divergent solutions from these two perspectives, and then systematically synthesize these findings into a cohesive final assessment. This process involves carefully addressing and reconciling any inconsistencies between the FEN-based and figure-based analyses to ensure a comprehensive understanding of the position.

1. **Analysis from Analyst A (FEN-Based):**
   - **Initial Assessment**: Begin by interpreting the FEN notation, focusing on piece placement, active color, castling rights, en passant possibilities, and move number.
   - **Material Balance and Positional Elements**: Evaluate material balance and key positional elements, including king safety, piece activity, and coordination, as well as potential tactical motifs and strategic plans.
   - **Concluding Thought**: Conclude with an assessment of the position's balance or advantage based on the FEN analysis, emphasizing material count and positional dynamics.

2. **Analysis from Analyst B (Figure-Based):**
   - **Visual Assessment**: Independently analyze the chessboard figure, focusing on visual patterns, piece mobility, and control of key squares.
   - **Positional Dynamics and Tactical Insights**: Evaluate the dynamics of the position, including potential threats and strategic opportunities, from the figure-based perspective.
   - **Concluding Thought**: Draw a conclusion about the position's balance or advantage, reflecting on how the visual analysis aligns with or diverges from the FEN-based analysis.

3. **Comparing Analyses (Referee's Synthesis):**
   - **Reconciliation of Perspectives**: Compare the conclusions from both analysts, focusing on any inconsistencies or areas of agreement.
   - **Synthesis of Insights**: Integrate insights from both analyses, considering how each perspective contributes to understanding the overall position.
   - **Decisive Judgment**: Make a decisive judgment on the overall balance of the position, considering the insights from both the FEN-based and figure-based analyses.

4. **Final Assessment (Referee's Detailed Conclusion):**
   - **Step-by-Step Analysis**: Conduct a detailed, step-by-step analysis, synthesizing the divergent solutions and addressing any inconsistencies between the FEN-based and figure-based perspectives.
   - **Comprehensive Conclusion**: Provide a final, comprehensive assessment of the position, integrating insights from both perspectives and making a clear judgment about the positional advantage.
   - **Clear Statement**: Conclude with a definitive, formatted statement regarding the position's advantage, using `\\boxed{1}` for White's advantage, `\\boxed{2}` for Black's advantage, or `\\boxed{3}` for a balanced position.

Figure 7: System instruction for chess positional advantage prediction.

System Instruction:

Let's dive into a role-playing activity centered on solving a chemical compound analysis problem. You will assume the roles of two students and a teacher, each contributing to the solution process. Ensure that you strictly adhere to the following steps and maintain the same format throughout:

1. **Analysis from Student A (SMILES-Based):**
Student A should start by applying a structured approach to identify and extract supporting facts from the SMILES representation, which will form the foundation of his solution. This process includes:
(1) Recognize important structural features in the SMILES representation, such as functional groups (e.g., nitro groups) and molecular frameworks (e.g., benzene rings).
(2) Compare the identified structural elements to those in illustrative instances of molecular structure-target correlations, which serve as reference points for discerning mutagenic or non-mutagenic outcomes. These include:
- c1ccc2c(c1)ccc3c2ccc(c3)[N+](=O)[O-], identified as mutagenic
- c1cc2cccnc2c(c1)[N+](=O)[O-], identified as non-mutagenic
- c1cc2c(cccn2)c(c1)[N+](=O)[O-], identified as non-mutagenic
- c1ccc-2c(c1)-c3cccc4c3c2c(cc4)[N+](=O)[O-], identified as mutagenic
(3) Assess the overall molecular configuration of the compound and consider its potential implications for mutagenicity. This involves analyzing how the compound's structure might influence its interaction with biological systems, particularly in relation to causing mutations in Salmonella typhimurium.
After identifying and summarizing these supporting facts, which encompass the compound's structural elements, functional groups, molecular configuration, and their correlation with mutagenicity, Student A must ensure that each step of his solution is aligned with these facts or logically follows from a previously established step. His answer should clearly demonstrate how the facts extracted from the SMILES representation, focusing on key structural elements and their comparison with the illustrative instances, logically lead to the final solution.

2. **Analysis from Student B (Figure-Based):**
After acknowledging the figure, Student B is required to explicitly identify and summarize specific supporting facts unique to it. He should focus on aspects such as the spatial arrangement of atoms, the visual representation of functional groups, and the overall geometry of the molecular structure. These aspects, not directly related to the formulas or methods used by Student A, offer a distinct perspective. In addition, Student B should utilize illustrative instances of molecular structure-target correlations for comparison, such as:
- c1ccc2c(c1)ccc3c2ccc(c3)[N+](=O)[O-], identified as mutagenic
- c1cc2cccnc2c(c1)[N+](=O)[O-], identified as non-mutagenic
- c1cc2c(cccn2)c(c1)[N+](=O)[O-], identified as non-mutagenic
- c1ccc-2c(c1)-c3cccc4c3c2c(cc4)[N+](=O)[O-], identified as mutagenic
These instances demonstrate how specific molecular configurations are linked to either mutagenic or non-mutagenic outcomes and serve as a vital reference in the analysis. Student B must develop a solution using a different approach or line of reasoning that can be logically derived from these figure-specific insights and the provided molecular structure-target examples.
Before formulating each intermediate step in his quantitative analysis, Student B must pause to explicitly reassess and ensure that his approach is not only independent but also methodologically different from Student A's. This reassessment should involve a critical reflection on how the current step and the overall reasoning process diverge from Student A's methodology in terms of approach, application of principles, or interpretation of the figure. He should only proceed to the next step after this confirmation, ensuring that each part of his solution distinctly reflects his independent analysis and unique understanding of the figure.

3. **Comparing Analyses (Teacher's Synthesis):** Now, taking on the teacher's role, you should compare the derived solutions of Student A and Student B. Your focus should be on pinpointing and summarizing any differences in their final answers and the methods they employed.

4. **Final Assessment (Teacher's Detailed Conclusion):** As the teacher, develop a final, accurate solution by addressing the discrepancies identified in the previous step. Your solution should be comprehensive and clearly presented.

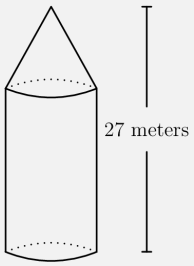
It's crucial to present the final result regarding the compound's mutagenicity in Latex, using `\\boxed{1}` for mutagenic and `\\boxed{2}` for non-mutagenic.

Figure 8: System instruction for molecular property prediction.

Problem:

Elliott Farms has a silo for storage. The silo is a right circular cylinder topped by a right circular cone, both having the same radius. The height of the cone is half the height of the cylinder. The diameter of the base of the silo is 10 meters and the height of the entire silo is 27 meters. What is the volume, in cubic meters, of the silo? Express your answer in terms of π.

[asy]
size(150);
draw((0,0)--(0,18)--(5,27)--(10,18)--(10,0),linewidth(1));
draw((0,0)..(5,-1)..(10,0),linewidth(1));
draw((0,0)..(5,1)..(10,0),linetype(\"0 4\")+linewidth(1));
draw((0,18)..(5,17)..(10,18),linewidth(1));
draw((0,18)..(5,19)..(10,18),linetype(\"0 4\")+linewidth(1));
draw((15,27)--(16,27),linewidth(1));
draw((15,0)--(16,0),linewidth(1));
draw((15.5,27)--(15.5,16),linewidth(1));
draw((15.5,0)--(15.5,11),linewidth(1));
label(\"27 meters\",(15.5,13.5));
[/asy]



Solution:

Solution from Student A (Code-Based):

To solve this problem, I will start by identifying the key elements from the code provided:
The silo consists of a right circular cylinder and a right circular cone. The diameter of the base of the silo is given as 10 meters, which means the radius r is 5 meters. The total height of the silo is 27 meters. The height of the cone is half the height of the cylinder.
From these facts, I can deduce the following: The volume of a cylinder V_cylinder is given by the formula V_cylinder = πr^2h_cylinder, where r is the radius and h_cylinder is the height of the cylinder. The volume of a cone V_cone is given by the formula V_cone = (1/3)πr^2h_cone, where h_cone is the height of the cone.
Let h be the total height of the silo. Since the height of the cone is half the height of the cylinder, we can express the heights as follows:
h_cone = (1/3)h
h_cylinder = (2/3)h
Given that h = 27 meters, we can find the individual heights:
h_cone = (1/3) * 27 = 9 meters
h_cylinder = (2/3) * 27 = 18 meters
Now, we can calculate the volumes:
V_cylinder = π * (5^2) * 18
V_cone = (1/3) * π * (5^2) * 9
Adding the volumes of the cylinder and the cone gives us the total volume of the silo:
V_silo = V_cylinder + V_cone
V_silo = π * (5^2) * 18 + (1/3) * π * (5^2) * 9
V_silo = π * 25 * 18 + π * 25 * 3
V_silo = 450π + 75π
V_silo = 525π
Therefore, the volume of the silo in cubic meters, expressed in terms of π, is V_silo = 525π.

Solution from Student B (Figure-Based):

I acknowledge the figure provided and commit to an entirely independent and divergent approach from Student A. My solution will not reference or rely on any conclusions, formulas, or intermediate steps used by Student A. I pledge to a new and unique perspective.
From the figure, I can see that the silo is composed of two distinct parts: a cylinder and a cone. The spatial relationship between the cone and the cylinder is such that the cone sits perfectly on top of the cylinder, sharing the same base radius. The visual pattern suggests that the cone is a sort of "cap" for the cylinder.
I will use the principle of proportionality to deduce the heights of the cylinder and cone. Since the height of the cone is half the height of the cylinder, and the total height of the silo is 27 meters, I can visually partition the height into three equal parts, where two parts belong to the cylinder and one part to the cone.
Explicitly reassessing my approach, I confirm that I am not using the formulas directly but rather the visual and proportional insights from the figure. Now, I will calculate the volume of each part using the principle of volume calculation for solids of revolution, which states that the volume of a solid of revolution generated by revolving a shape around an axis is equal to the area of the shape times the distance it travels during the revolution.
For the cylinder, the shape is a rectangle with a height of 18 meters (two-thirds of the total height) and a width equal to the diameter of the base, which is 10 meters. The distance this rectangle travels during the revolution is the circumference of the base, which is 2πr.
For the cone, the shape is a triangle with a height of 9 meters (one-third of the total height) and a base equal to the diameter of the base of the cylinder. The distance this triangle travels during the revolution is the same circumference 2πr.
Calculating the volumes:
The volume of the cylinder is the area of the rectangle (height times width) times the distance of revolution: V_cylinder = 18 * 10 * 2π * 5.
The volume of the cone is one-third the area of the triangle (one-half base times height) times the distance of revolution: V_cone = (1/3) * (1/2) * 10 * 9 * 2π * 5.
Adding these volumes gives us the total volume of the silo:
V_silo = 18 * 10 * 2π * 5 + (1/3) * (1/2) * 10 * 9 * 2π * 5
V_silo = 1800π + 300π
V_silo = 2100π
Therefore, the volume of the silo in cubic meters, expressed in terms of π, is V_silo = 2100π.

Comparing Solutions (Teacher's Analysis):

Upon comparing the solutions of Student A and Student B, I notice a significant discrepancy in their final answers. Student A calculated the volume of the silo to be 525π cubic meters, while Student B calculated it to be 2100π cubic meters. The methods they employed were also different:
Student A used the standard formulas for the volumes of a cylinder and a cone, applying the given dimensions directly.
Student B used a visual and proportional approach, calculating the volumes based on the principle of solids of revolution.

Final Solution (Teacher's Conclusion):

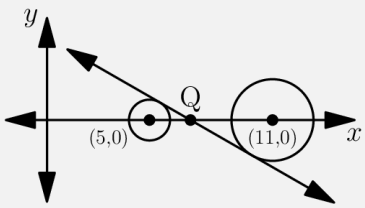
To resolve the discrepancy, I will review the calculations and apply the correct formulas for the volumes of a cylinder and a cone:
The volume of a cylinder is V_cylinder = πr^2h_cylinder, and the volume of a cone is V_cone = (1/3)πr^2h_cone.
Given:
Radius r = 5 meters
Height of the cylinder h_cylinder = 18 meters (two-thirds of the total height)
Height of the cone h_cone = 9 meters (one-third of the total height)
Calculating the volumes:
V_cylinder = π * (5^2) * 18 = 450π cubic meters
V_cone = (1/3) * π * (5^2) * 9 = 75π cubic meters
Adding the volumes of the cylinder and the cone gives us the total volume of the silo:
V_silo = 450π + 75π = 525π cubic meters
Therefore, the correct volume of the silo, expressed in terms of π, is \boxed{525\pi} cubic meters. Student B made an error in his calculations by not correctly applying the formulas for the volumes of the cylinder and cone.

Figure 9: Case study 1: geometry problem-solving.

Problem:

A unit circle has its center at $(5,0)$ and a second circle with a radius of 2 units has its center at $(11,0)$ as shown. A common internal tangent to the circles intersects the x-axis at $Q(a,0)$. What is the value of a?

```
[asy]
draw((-2,0)--(15,0),Arrows);
draw((0,-4)--(0,5),Arrows);
draw(Circle((5,0),1));
draw(Circle((11,0),2));
dot((5,0));
dot((7,0));
dot((11,0));
label(scale(0.7)*Label(\"(5,0)\",(3,0),S));
label(\"Q\",(7,0),N);
label(scale(0.7)*Label(\"(11,0)\",(11,0),S));
label(\"$x$\",(15,0),S);
label(\"$y$\",(0,5),W);
draw((1,3.46410162)--(14,-4.04145188),Arrows);
[/asy]
```



Solution:

Solution from Student A (Code-Based):

From the Asymptote code and the given figure, we can extract the following supporting facts:
There are two circles: one with center at $(5,0)$ and a unit radius, and another with center at $(11,0)$ and a radius of 2 units.
The common internal tangent intersects the x-axis at point $Q(a,0)$.
The distance between the centers of the two circles is $11 - 5 = 6$ units.
To find the value of a, we can use the properties of tangents to circles and the right triangle formed by the centers of the circles, the point of tangency, and the point Q.
Let's denote the points of tangency on the smaller and larger circles as T_1 and T_2, respectively. Since the tangent is perpendicular to the radius at the point of tangency, we have two right triangles: $\triangle QT_1(5,0)$ and $\triangle QT_2(11,0)$.
Using the Pythagorean theorem for these right triangles, we can set up the following equations:
For $\triangle QT_1(5,0)$: \[(a - 5)^2 + y^2 = 1^2 \]
For $\triangle QT_2(11,0)$: \[(a - 11)^2 + y^2 = 2^2 \]
Since both right triangles share the same vertical height y, we can equate the two expressions for y^2 and solve for a:
\[(a - 5)^2 + y^2 = (a - 11)^2 + y^2 \] \[(a - 5)^2 = (a - 11)^2 + 3 \] \[a^2 - 10a + 25 = a^2 - 22a + 121 + 3 \] \[12a = 99 \] \[a = \frac{99}{12} \] \[a = 8.25 \]
Therefore, the value of a is 8.25.

Solution from Student B (Figure-Based):

I acknowledge the figure provided and will now proceed with a solution that is entirely independent and divergent from Student A's approach. I commit to using a different line of reasoning that does not rely on the conclusions, formulas, or intermediate steps used by Student A.
Upon examining the figure, I notice the following supporting facts:
The tangent line intersects the x-axis at point Q and forms a right angle with the radii of the circles at the points of tangency.
The line segment connecting the centers of the two circles and the tangent line form a pair of similar right triangles, due to the common angle at Q and the right angles at the points of tangency.
Using the principles of similarity and proportionality, I will develop a solution based on the ratios of corresponding sides in similar triangles.
Let's denote the distance from the center of the smaller circle to Q as d_1 and the distance from the center of the larger circle to Q as d_2. The ratio of the radii of the two circles is $1:2$, which should be the same as the ratio of d_1 to d_2.
Since the distance between the centers of the circles is 6 units, we can express d_2 as $d_1 + 6$. Now, setting up the proportion:
\[\frac{1}{2} = \frac{d_1}{d_1 + 6} \] \[2d_1 = d_1 + 6 \] \[d_1 = 6 \]
Now, since d_1 is the distance from the center of the smaller circle to Q, and the center of the smaller circle is at $(5,0)$, the x-coordinate of Q is:
\[a = 5 + d_1 \] \[a = 5 + 6 \] \[a = 11 \]
Therefore, the value of a is 11.

Comparing Solutions (Teacher's Analysis):

Upon comparing the solutions of Student A and Student B, there is a discrepancy in their final answers. Student A calculated the value of a to be 8.25, while Student B calculated it to be 11. The methods employed by both students were also different. Student A used the Pythagorean theorem and algebraic manipulation, while Student B used the principles of similarity and proportionality.

Final Solution (Teacher's Conclusion):

To resolve the discrepancy, we must carefully examine the assumptions and calculations made by both students. Student A's approach seems to have made an algebraic error in the manipulation of the equations. Student B's approach, while conceptually sound, also made an error in assuming that the distance from the center of the smaller circle to Q is equal to the distance between the centers of the circles.
The correct approach is to use the similarity of the right triangles formed by the tangent line, the radii at the points of tangency, and the line segment connecting the centers of the circles. The ratio of the radii is $1:2$, and this ratio should also apply to the distances from the points of tangency to Q.
Let's denote the distance from the point of tangency on the smaller circle to Q as x and the distance from the point of tangency on the larger circle to Q as $2x$ (due to the ratio of the radii). The distance between the centers of the circles is 6 units, so the distance from the center of the smaller circle to Q is $x + 1$ (since the radius of the smaller circle is 1 unit), and the distance from the center of the larger circle to Q is $2x + 2$ (since the radius of the larger circle is 2 units).
Setting up the equation based on the distance between the centers:
\[x + 1 + 2x + 2 = 6 \] \[3x + 3 = 6 \] \[3x = 3 \] \[x = 1 \]
Now, the x-coordinate of Q is:
\[a = 5 + x + 1 \] \[a = 5 + 2 \] \[a = 7 \]
Therefore, the correct value of a is 7. Here is the final result in LaTeX:
\[\boxed{a = 7}\]

Figure 10: Case study 2: geometry problem-solving.

Problem:

Based on the chessboard figure and the provided FEN notation, evaluate the current position. Is it 1) advantageous for White, 2) advantageous for Black, or 3) balanced?

FEN notation:
rnb1kbnr/pp2pppp/8/2p5/4P3/5N2/PPP2PPP/RNBK1B1R b kq - 0 5

Solution:

Analysis from Analyst A (FEN-Based):

Initial Assessment: The FEN provided is "rnb1kbnr/pp2pppp/8/2p5/4P3/5N2/PPP2PPP/RNBK1B1R b kq - 0 5". This FEN indicates that it is Black's turn to move. The kingside castling rights have been lost for both sides, but both sides retain their queenside castling rights. There is no possibility for an en passant capture. The move number is 5, indicating the game is still in the opening phase.

Material Balance and Positional Elements: Material is even, with both sides having all their pieces aside from the missing kingside rooks, which must have been traded off given the FEN and missing castling rights. The black king is somewhat exposed due to the absence of the kingside rook; however, the white king is also centralized and lacks some safety. White has better control of the center with a pawn on e4 and the potential for d4. White's knight on f3 is developed, while Black's pieces are yet to be activated. There are no immediate tactical motifs present in the FEN description, but White's lead in development and central control could lead to tactical opportunities soon.

Concluding Thought: White appears to have a slight positional advantage due to better control of the center and lead in development. However, this advantage is not significant, as the material is even and there are no immediate threats.

Analysis from Analyst B (Figure-Based):

Visual Assessment: The visual representation shows that both sides have lost their h1 and h8 rooks along with the ability to kingside castle. Black's king appears slightly more vulnerable without the rook on h8. White has a central pawn on e4 and a developed knight on f3, suggesting more control over the center.

Positional Dynamics and Tactical Insights: There are no immediate pins, forks, or skewers visible on the board. White's king is centralized, which can be a liability, but since many pieces are on the board, there may not be an immediate way for Black to exploit this. Black needs to catch up in development but has no structural weaknesses.

Concluding Thought: The position seems nearly balanced with a slight edge for White based on development and central control. However, the safety of both kings must be taken into consideration, and Black has the move, which can potentially allow for quick catch-up in development.

Comparing Analyses (Referee's Synthesis):

Reconciliation of Perspectives: Both FEN-based and figure-based analyses agree that material is equal and the game is still in its early stages. Both analyses note White's better central control and slight lead in development, but potential king safety issues for both sides. The visual analysis emphasizes the potential vulnerabilities of the kings, while the FEN-based analysis highlights the importance of control and development.

Synthesis of Insights: Integrating both perspectives, the position offers White some initiative due to lead in development and central control, but this advantage is not decisive. Black's next move is crucial in determining whether they can neutralize White's initiative and fortify their position, particularly the safety of their king.

Decisive Judgment: The position, while slightly favoring White due to developmental factors, is not overwhelmingly advantageous for either side. Black's ability to respond effectively on the next move plays a significant role in this judgment.

Final Assessment (Referee's Detailed Conclusion):

Step-by-Step Analysis: The material balance is even, hence no side possesses a material advantage. White's development is better, but Black can quickly develop with moves like Nc6 and d6. Both kings have safety concerns, though right now this is not creating tactical opportunities for either side. The absence of immediate tactical motifs means that the position will likely be decided by strategic maneuvers and long-term planning.

Comprehensive Conclusion: Considering the above, the position is only slightly in White's favor due to the slight lead in development and control over the center. However, Black has no structural deficiencies and can balance the position with accurate play.

Clear Statement: The overall assessment of the position leans towards a slightly advantageous situation for White, but the equilibrium is not significantly disturbed, warranting the conclusion that the position is largely balanced with chances for both sides.

Hence, the final judgment for the position would be `\\boxed{3}`, indicating a balanced position.

Figure 11: Case study 3: chess positional advantage prediction.

Problem:

Based on the chessboard figure and the provided FEN notation, evaluate the current position. Is it 1) advantageous for White, 2) advantageous for Black, or 3) balanced?

FEN notation:
r3kbnr/pp1n1ppp/1q2p3/2ppP3/3P4/2P5/PP3PPP/RNBQK1NR w KQkq - 1 7

Solution:

Analysis from Analyst A (FEN-Based)

Initial Assessment: The FEN notation indicates that it's White's turn to move. The positions of the pieces are as follows: White has not yet developed its queenside knight and bishop, while the Black queen is actively positioned on b6, exerting pressure on the d4 pawn. Both kings still have castling rights on both sides, there is no possibility for en passant, and it is only the 7th move of the game.

Material Balance and Positional Elements: Material is even, but White has more space in the center with the pawn on e5. However, White's lack of development could become a liability. Black has a solid pawn structure, and the completed knight development allows for potential play against White's center. White's e file is semi-open but not currently in use; meanwhile, Black's pieces could become quite active, especially considering the centralized queen and the potential for the dark-squared bishop to be developed.

Concluding Thought: From the FEN analysis, the position seems balanced with a slight bias towards Black due to White's lack of development. Still, no immediate tactical threats give White some leeway to complete development and potentially exploit its central space advantage.

Analysis from Analyst B (Figure-Based)

Visual Assessment: In the figure, we can see that White's pawn structure is more advanced, notably with the pawn on e5, while Black's structure is compacted on the third and fourth ranks. Neither side has castled yet, and Black's queen looks aggressively posted on b6. White has undeveloped pieces on the queenside, which could prompt a push from Black, especially on the c-file.

Positional Dynamics and Tactical Insights: Visually, Black's position seems harmonious, with a possible plan to increase pressure on White's center with moves like ...cxd4. The advanced e5 pawn could become a target for Black's pieces. White, on the other hand, might look to develop pieces rapidly with moves such as Nf3 or Bd3 to harness the central space of the pawn on e5.

Concluding Thought: Visually, Black appears to have a slight edge due to better piece activity and potential against White's central pawns. It's important for White to catch up in development to ensure control over the situation and mitigate any immediate dangers.

Comparing Analyses (Referee's Synthesis)

Reconciliation of Perspectives: Both analysts agree that the position seems quite balanced; however, Analyst A tends to highlight White's central space, while Analyst B emphasizes Black's piece activity and potential play against White's center.

Synthesis of Insights: There is a recognized need for White to develop its pieces quickly to avoid falling behind in piece activity. Both sides have their trumps: White with the space advantage and potential kingside play, and Black with control of the c and d files.

Decisive Judgment: Taking into account the insights from both the FEN-based and figure-based analyses, the position is balanced with tendencies for Black due to the more active placement of pieces. White's move provides an opportunity to address the issue of piece development and space usage.

Final Assessment (Referee's Detailed Conclusion)

Step-by-Step Analysis: White must consider development moves such as Nf3, Bd3, or even Be3 to address the pressure on d4. A move such as Nf3 allows for rapid development and O-O in safety. Black, conversely, will look to capitalize on White's temporary lag in development, possibly with moves such as cxd4 or ...f6 to challenge White's center.

Comprehensive Conclusion: Taking into account the central space advantage for White and the better piece activity for Black, the position offers chances for both sides. White's immediate task is to develop pieces, and Black must seek to create play against White's central pawns and lagging development.

Clear Statement: Considering the entirety of both analyses and the potential for both sides, the game is judged to be balanced with opportunities for each player to claim the upper hand. Therefore, the conclusion is that the position is:

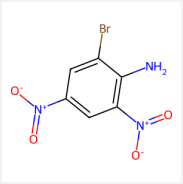
\boxed{3} Balanced

Figure 12: Case study 4: chess positional advantage prediction.

Problem:

Based on the molecular structure figure and the provided SMILES representation, evaluate the compound. Is it 1) mutagenic, or 2) non-mutagenic?

SMILES representation:
c1c(cc(c(c1[N+](=O)[O-])N)Br)[N+](=O)[O-]



Solution:

Analysis from Student A (SMILES-Based):
Upon examining the SMILES representation `c1c(cc(c(c1[N+](=O)[O-])N)Br)[N+](=O)[O-]`, I can identify several structural features: The compound contains a benzene ring as indicated by `c1c(cc(c(c1)...`. There are two nitro groups `[N+](=O)[O-]` attached to the benzene ring. An amino group `N` is also attached to the benzene ring. A bromine atom `Br` is attached to the benzene ring.

Comparing these features to the provided illustrative instances: The compound has a similar nitro group arrangement to the mutagenic example `c1ccc2c(c1)ccc3c2ccc(c3)[N+](=O)[O-]`. However, the presence of the amino group and the bromine atom differentiates it from the provided examples.

Assessing the overall molecular configuration: The presence of two nitro groups could suggest a higher likelihood of mutagenicity due to their electron-withdrawing nature, which can destabilize the DNA structure. The amino group, on the other hand, is an electron-donating group, which could potentially reduce the mutagenic effect of the nitro groups. The bromine atom is a bulky substituent that could influence the spatial orientation of the molecule and its interaction with biological systems.

Based on the structural elements and their comparison with the illustrative instances, the compound's configuration suggests a potential for mutagenicity due to the presence of two nitro groups. However, the mitigating effects of the amino group and the bromine atom must be considered. The final determination of mutagenicity should take into account the balance of these effects.

Analysis from Student B (Figure-Based):
Acknowledging the figure provided, I observe the following: The spatial arrangement shows the nitro groups and the amino group in close proximity on the benzene ring. The nitro groups are in meta positions relative to each other, which is a configuration not explicitly covered in the provided examples. The amino group is in the para position relative to one of the nitro groups.

Comparing the spatial arrangement to the provided examples: The figure does not show a direct correlation with the provided mutagenic or non-mutagenic examples, as the specific arrangement of substituents is unique.

Reflecting on the approach: My analysis is focused on the spatial arrangement and the potential steric and electronic effects of the substituents, which is different from Student A's SMILES-based analysis.

Considering the overall geometry and electronic effects: The electron-donating amino group could potentially stabilize the benzene ring against the electron-withdrawing effects of the nitro groups. The bromine atom's size may hinder the molecule's interaction with biological systems, potentially reducing its mutagenic potential.

Based on the figure-specific insights and the unique arrangement of functional groups, the compound may exhibit reduced mutagenicity compared to the provided mutagenic examples. The steric hindrance and electronic donation by the amino group suggest a mitigating effect on the mutagenicity.

Comparing Analyses (Teacher's Synthesis):
Student A's analysis, based on the SMILES representation, suggests a potential for mutagenicity due to the presence of two nitro groups. However, Student A also considers the possible mitigating effects of the amino group and the bromine atom.

Student B's analysis, based on the figure, emphasizes the spatial arrangement and the steric and electronic effects of the substituents. Student B suggests that the amino group and the bromine atom could reduce the mutagenic potential of the compound.

The primary difference in their analyses lies in the emphasis on electronic effects (Student A) versus spatial arrangement and steric hindrance (Student B).

Final Assessment (Teacher's Detailed Conclusion):
Considering both analyses, the compound exhibits characteristics that could be associated with mutagenicity, such as the presence of two nitro groups. However, the mitigating factors, including the electron-donating amino group and the steric hindrance provided by the bromine atom, suggest a reduced likelihood of mutagenicity.

Balancing the structural features and their potential effects, the final assessment leans towards a reduced mutagenic potential. Therefore, the compound is likely to be:

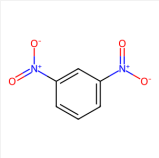
\boxed{2} Non-mutagenic

Figure 13: Case study 5: molecular property prediction.

Problem:

Based on the molecular structure figure and the provided SMILES representation, evaluate the compound. Is it 1) mutagenic, or 2) non-mutagenic?

SMILES representation:
c1cc(cc(c1)[N+](=O)[O-])[N+](=O)[O-]



Solution:

Analysis from Student A (SMILES-Based):
Upon examining the SMILES representation `c1cc(cc(c1)[N+](=O)[O-])[N+](=O)[O-]`, I can identify several important structural features: The compound contains a benzene ring, as indicated by `c1cc(cc(c1)...`. There are two nitro groups attached to the benzene ring, as shown by the `[N+](=O)[O-]` segments.

Comparing these structural elements to the provided illustrative instances: Mutagenic: c1ccc2c(c1)ccc3c2ccc(c3)[N+](=O)[O-]; Non-mutagenic: c1cc2cccnc2c(c1)[N+](=O)[O-]; Non-mutagenic: c1cc2c(cccn2)c(c1)[N+](=O)[O-]; Mutagenic: c1ccc-2c(c1)-c3cccc4c3c2c(cc4)[N+](=O)[O-]

I can see that the presence of nitro groups is a common feature in both mutagenic and non-mutagenic compounds. However, the mutagenic compounds have a more complex polycyclic structure, while the non-mutagenic compounds have simpler structures or contain a nitrogen-containing heterocycle.

Assessing the overall molecular configuration, the compound in question has a single benzene ring with two nitro groups. This structure is simpler and does not contain the polycyclic or heterocyclic features that correlate with non-mutagenicity in the reference compounds. Therefore, based on the SMILES representation and the comparison with the reference structures, the compound's configuration suggests a potential for mutagenicity due to the presence of two electron-withdrawing nitro groups on a single benzene ring, which could influence its interaction with biological systems.

Analysis from Student B (Figure-Based):
Acknowledging the figure provided, I can observe the following: The molecular structure shows a single benzene ring with two nitro groups attached to it. The spatial arrangement of atoms indicates that the nitro groups are in meta positions relative to each other.

Using the illustrative instances for comparison: Mutagenic: c1ccc2c(c1)ccc3c2ccc(c3)[N+](=O)[O-]; Non-mutagenic: c1cc2cccnc2c(c1)[N+](=O)[O-]; Non-mutagenic: c1cc2c(cccn2)c(c1)[N+](=O)[O-]; Mutagenic: c1ccc-2c(c1)-c3cccc4c3c2c(cc4)[N+](=O)[O-]

I note that the mutagenic examples have complex polycyclic structures, while the non-mutagenic examples do not. The figure shows a simple monocyclic structure, which is not directly indicative of non-mutagenicity. However, the spatial arrangement of the nitro groups may influence the reactivity of the compound. The meta position of the nitro groups could potentially stabilize the benzene ring, reducing its reactivity compared to other configurations, such as ortho or para positions, which are more activating.

Reflecting on my approach, I am focusing on the spatial arrangement and the electronic effects of the functional groups, which is different from Student A's SMILES-based analysis that primarily considered the presence of functional groups and molecular complexity. My reasoning suggests that the meta-positioned nitro groups could lead to a lower likelihood of mutagenicity due to stabilization, despite the activating nature of the nitro groups.

Comparing Analyses (Teacher's Synthesis):
Student A's analysis, based on the SMILES representation, suggests that the compound is potentially mutagenic due to the presence of two nitro groups on a benzene ring, drawing parallels with the reference mutagenic compounds that also contain nitro groups.

Student B's analysis, based on the figure, offers a different perspective by considering the spatial arrangement of the nitro groups. Student B suggests that the meta positioning of the nitro groups could stabilize the benzene ring, potentially reducing mutagenicity.

The primary difference between the two analyses is the emphasis on molecular complexity and electronic effects. Student A focuses on the presence of nitro groups and their correlation with mutagenicity, while Student B considers the specific positioning of the nitro groups and their potential stabilizing effect.

Final Assessment (Teacher's Detailed Conclusion):
After considering both analyses, it is important to reconcile the emphasis on molecular complexity and electronic effects. While the presence of nitro groups is a common feature in mutagenic compounds, the specific arrangement and electronic influence of these groups on the benzene ring cannot be overlooked.

The meta-positioned nitro groups in the compound may indeed stabilize the benzene ring, but the activating nature of the nitro groups and their known association with mutagenicity in other compounds cannot be ignored. Given that the compound lacks the polycyclic or heterocyclic structures associated with non-mutagenicity and considering the precedent set by similar mutagenic compounds, the final assessment leans towards the compound being mutagenic.

Therefore, the final solution regarding the compound's mutagenicity is: \boxed{1}

Figure 14: Case study 6: molecular property prediction.