CoTZero: Annotation-Free Human-Like Vision Reasoning via Hierarchical Synthetic CoT

Chengyi Du
University of Electronic Science and Technology of China
Shanghai Artificial Intelligence Laboratory
duchengyi1224@gmail.com

Yazhe Niu *
Shanghai Artificial Intelligence Laboratory
The Chinese University of Hong Kong MMLab
niuyazhe314@outlook.com

Dazhong Shen
The College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
shendazhong@nuaa.edu.cn

Luxin Xu
University of Electronic Science and Technology of China
xul022332@gmail.com

Abstract

Recent advances in vision–language models (VLMs) have markedly improved image–text alignment, yet they still fall short of human-like visual reasoning. A key limitation is that many VLMs rely on surface correlations rather than building logically coherent structured representations, which often leads to missed higher-level semantic structure and non-causal relational understanding, hindering compositional and verifiable reasoning. To address these limitations by introducing human models into the reasoning process, we propose CoTZero, an annotation-free paradigm with two components: **(i) a dual-stage data synthesis approach and (ii) a cognition-aligned training method**. In the first component we draw inspiration from neurocognitive accounts of *compositional productivity* and *global-to-local analysis*. In the bottom-up stage, CoTZero extracts atomic visual primitives and incrementally composes them into diverse, structured question–reasoning forms. In the top-down stage, it enforces hierarchical reasoning by using coarse global structure to guide the interpretation of local details and causal relations. In the cognition-aligned training component, built on the synthesized CoT data, we introduce **Cognitively Coherent Verifiable Rewards** (CCVR) in Reinforcement Fine-Tuning (RFT) to further strengthen VLMs' hierarchical reasoning and generalization, providing stepwise feedback on reasoning coherence and factual correctness. Experiments show that CoTZero achieves an F1 score of 83.33% on our multi-level semantic inconsistency benchmark with lexical-perturbation negatives, across both in-domain and out-of-domain settings. Ablations confirm that each component contributes to more interpretable and human-aligned visual reasoning.

1 Introduction

Vision-Language Models (VLMs) Liu et al. (2023); OpenAI (2023); Team (2025); Zhu et al. (2025); Bai et al. (2025) have achieved notable progress in tasks such as image captioning Dong et al. (2024); Wang et al. (2024a); Lu et al. (2025), visual question answering(VQA) Park & Kim (2023); Antol et al. (2015); Mañas et al. (2024), and text-to-image Narasimhaswamy et al. (2024); Li et al. (2024b) generation. Yet they still struggle with the fundamental challenge of visual reasoning-particularly when faced with complex scenarios demanding hierarchical analysis Huang et al. (2025). While these models excel at surface-level image-text associations, their reasoning capabilities remain far

*corresponding author

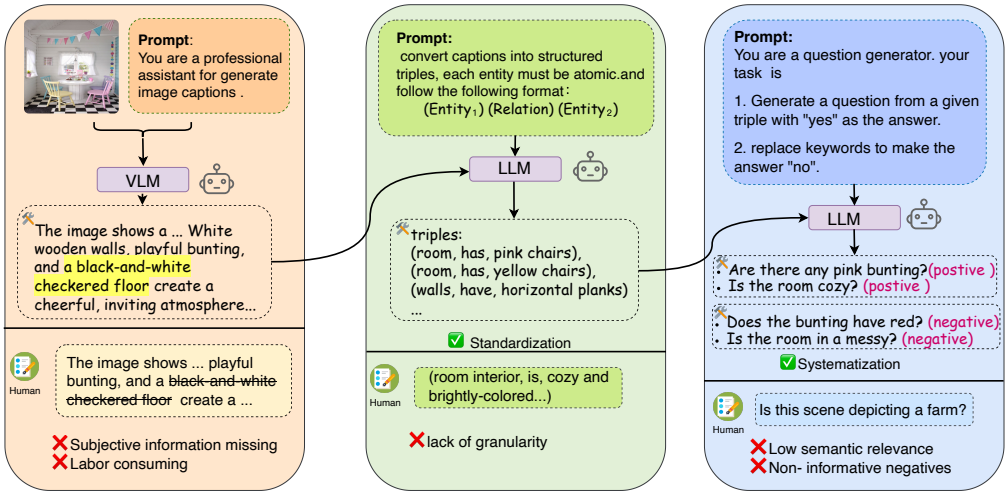

Figure 1: Our annotation-free data generation pipeline. Starting from external image inputs, a VLM produces rich captions, which are structured into (entity, relation, entity) triples by an LLM. These triples are then transformed into yes/no QA pairs through controlled prompting. This pipeline enables scalable, fine-grained, and semantically consistent supervision without human annotation.

behind human-like performance. This persistent gap arises because current VLMs rely on statistical pattern recognition rather than the construction of human-like, causally structured models of the world. To bridge the gap between these human traits and AI, some works attempt to teach models through synthetic data. However, as illustrated in Figure 1, current visual reasoning data generation frequently involves subjective manual annotation or yields unstructured, low-granularity knowledge. These methods are not only labor-intensive but also fail to capture the deep causal structure required for human-like reasoning. Other efforts focus on direct model generation; yet, prevalent Chain-of-Thought (CoT) methods for visual tasks primarily generate linear or unstructured thought sequences. As illustrated in Figure 2, this lack of hierarchy causes reasoning to suffer from hallucination, redundancy, and information missing. Without a structured mechanism to mirror the Bayesian brain's drive to minimize surprise (prediction error) through stagewise verification, these "unstructured" thoughts remain unreliable and lack the interpretability necessary for complex visual scenarios.

Human intelligence is distinguished by two profound cognitive capabilities. First, In human cognition, productivity is at the core of compositionality Lake et al. (2016); Hockett (1960), enabling the mind to construct an infinite number of thoughts and diverse reasoning paths from a finite set of primitives. This flexibility enables humans to navigate multiple itinerant "thought trajectories" that, while varied, are all anchored by causal coherence to reach the same correct conclusion. Second, human perception follows a global-to-local trajectory Oliva et al. (2006); Navon (1977), where an initial analysis of global layouts guides the interpretation of fine-grained local details. Motivated by these two cognitive principles, we introduce CoTZero, an annotation-free paradigm that consists of a dual-stage data synthesis approach and a cognition-aligned training method. Drawing on the "analysis-by-synthesis" paradigm Bever & Poeppel (2010); Lake et al. (2016); Mermelstein & Eyden (1964), our data approach represents visual scenes by modeling the causal process that generated question-reasoning pairs. Specifically, in the bottom-up stage, CoTZero operationalizes compositional productivity Lake et al. (2016); Hockett (1960); Marc et al. (2002) — the human capacity to construct infinite representations from a finite set of primitives. By extracting atomic visual elements and structured triples, our data approach incrementally composes these units into diverse question forms, enabling the emergence of potentially unbounded reasoning traces. In the top-down stage, CoTZero adheres to the cognitive principle of global-to-local analysis Oliva et al. (2006); Navon (1977), a trajectory supported by the hierarchical architecture of human vision. Neuroscientific studies Felleman & Van Essen (1991); Bull & Zhang (2021) reveal that the human visual cortex processes information through successive stages, progressing from low-level feature detection in the primary visual cortex (V1) to high-level integration and category recognition in the inferotemporal cortex (IT), –namely "seeing the forest before the trees" Navon (1977). By operationalizing this hierarchi-

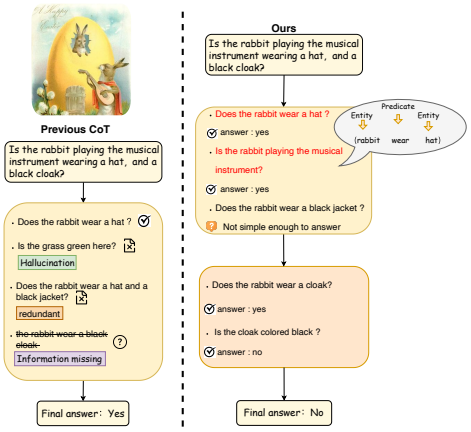

Figure 2: Comparison between our CoT process and previous CoT process.

In contrast, we design a rule-based reward function that jointly considers edit distance and semantic similarity between the model-generated reasoning chain and the reference CoT, thereby capturing both semantic coherence and hierarchical alignment. This reward design provides a learning signal that closely mirrors the way humans incrementally construct and validate their understanding of complex visual scenes. We detail this process in Section 3.3.The model first undergoes a cold-start phase on a public dataset to instill a cognitive blueprint.

cal logic, CoTZero systematically decomposes complex queries into nested, verifiable subproblems, ensuring that the global structure of a scene provides the necessary causal constraints to guide the interpretation of atomic components. Integrating this dual-stage data synthesis approach, CoTZero ensures global coherence and precise local understanding, thereby overcoming the unstructured and pattern-driven processing limitations of current VLMs.

Building on the synthesized hierarchical CoT trajectories produced by our dual-stage approach, we propose a cognition-aligned training method that enhances reinforcement fine-tuning with **Cognitively Coherent Verifiable Rewards** (CCVR). Unlike conventional approaches that evaluate only final outputs like using IoU scores in object grounding Du & Jin (2025), accuracy in visual question answering Mañas et al. (2024), or success rates in navigation tasks Zhao et al. (2023), recent work has begun to adopt process-supervised reward models (PRMs) that provide intermediate-step supervision for logical reasoning Shao et al. (2024), these PRMs typically rely on LLM-based judges or learned reward models for scoring

It is then trained with Supervised Fine-Tuning (SFT) on the dataset generated by our dual-stage approach. Subsequently, the model is reinforcement fine-tuned using the Group Relative Policy Optimization (GRPO) algorithm Shao et al. (2024), with Cognitively Coherent Verifiable Rewards (CCVR) as the reward mechanism. This CCVR-guided GRPO stage improves accuracy by 15.90% over the baseline. We hope these insights and advancements unlock deeper compositional reasoning in VLMs, bridging the gap between powerful data-driven training and hierarchical human-like reasoning.

2 Related Work

2.1 Advancing VLM Problem-Solving Performance

In the evolution of VLMs, Chain-of-Thought (CoT) technology is regarded as the core driving force for achieving the transition from foundational ”visual perception” to high-order ”cognitive reasoning” Zhou et al. (2025). While early research primarily relied on prompt engineering Zhang et al. (2024); Wei et al. (2023) to stimulate reasoning potential, the current research focus has shifted toward enhancing problem-solving abilities via training. Specifically, training methodologies have advanced from initial imitation learning, such as the text-visual dependency constraints introduced by Multimodal-CoT Zhang et al. (2024), to phased curriculum learning exemplified by the decomposition-alignment-integration stages of LLaVA-CoT Xu et al. (2025b) and LlamaV-o1 Thawakar et al. (2025). Recent efforts have further adopted preference learning algorithms like GRPO Shao et al. (2024) and DPO Rafailov et al. (2024) to optimize logical consistency and align reasoning paths with visual facts. Furthermore, reasoning mechanisms are transitioning from traditional linear greedy search toward non-linear structures based on Tree of Thoughts (ToT) Yao et al.

(2023) and Monte Carlo Tree Search (MCTS), which allow models to perform backtracking and multi-path exploration. The latest progress emphasizes a dynamic reasoning loop ("Think with Image") Su et al. (2025); Zhou et al. (2025), where the model continuously re-examines visual evidence according to evolving inferential needs through endogenous attention refocusing Gao et al. (2025); Yang et al. (2025); Qi et al. (2025) or exogenous tool calls Gupta & Kembhavi (2022). thereby effectively alleviating hallucinations in long-chain reasoning.

2.2 Reinforcement Learning and Reward Mechanism Design for Reasoning in VLMs

While pre-trained models perform well on many tasks, they often fall short in complex reasoning and human-aligned generation; thus, post-training has progressed from supervised fine-tuning (SFT) to preference-based Ouyang et al. (2022); Xu et al. (2025a), where Reinforcement Learning from Human Feedback (RLHF) optimizes policies via a learned reward model (e.g., PPO Schulman et al. (2017)), and DPO Rafailov et al. (2024) provides a simpler alternative by directly learning from pairwise preferences without explicit reward modeling. Central to this evolution is the design of the reward model, which has shifted from providing sparse, end-to-end feedback to delivering detailed, process-oriented supervision. Early alignment efforts relied on Outcome Reward Models (ORM) Luong et al. (2024); Kazemnejad et al. (2025), where feedback is concentrated solely on the final solution; however, this approach faces a severe credit assignment problem Xu et al. (2025a), as the model struggles to identify which specific intermediate steps contributed to the final result. To overcome this bottleneck, the research frontier has moved toward Process Reward Models (PRM) Hwang et al. (2024); Wang et al. (2024b), which provide dense, step-wise rewards that encourage models to master human-like reasoning trajectories through trial-and-error.

3 Method

3.1 Overview

3.2 dual-stage data synthesis approach

As shown in Figure 1, dual-stage data synthesis approach requires only image inputs, thereby eliminating the need for additional human annotations. This automated process efficiently produces structured question-reasoning data tailored for enhancing the vision reasoning capabilities of models. We jointly leverage visual and language models to generate structured chain-of-thought (CoT) training data. This helps the model better capture fine-grained reasoning steps that are often overlooked in long-form multimodal inputs. This enhances the model's capability in tasks such as visual question answering (VQA) and multimodal reasoning.

At the initial stage of our pipeline, a Vision-Language Model (VLM) is employed to extract semantically rich captions from raw visual inputs. These captions are then processed by a LLM to generate triples, where each triple is structured as

$$(\mathcal{E}_i, \mathcal{R}_{\text{attr}}/\mathcal{R}_{\text{verb}}/\mathcal{R}_{\text{locate}}/\mathcal{R}_{\text{exist}}, \ldots, \mathcal{E}_j) \tag{1}$$

Each triple consists of two entities $\mathcal{E}_{i,j}$ and a relation $\mathcal{R}$ that links them, representing the key attributes and relationships within the visual input. Our triples typically contain the simplest relationships possible, maintaining a structured format. This simplicity ensures that each triple remains easily interpretable and directly relevant to downstream tasks. By avoiding overly complex or nested relationships, our approach facilitates straightforward reasoning and efficient generation process. After obtaining triples, we first generate the simplest atomic questions based on individual triples, capturing basic attributes or object relations. These atomic questions serve as the foundational building blocks. To construct semantically relevant negative samples, we further generate lexically altered negatives by replacing key lexical elements—typically the entity or attribute that determines the answer—with alternatives that render the question incompatible with the image content. This process ensures that the negative questions remain grammatically correct and contextually plausible, yet contradict the visual semantics, thereby enhancing the model's ability to perform fine-grained visual reasoning.See Appendix A for details.

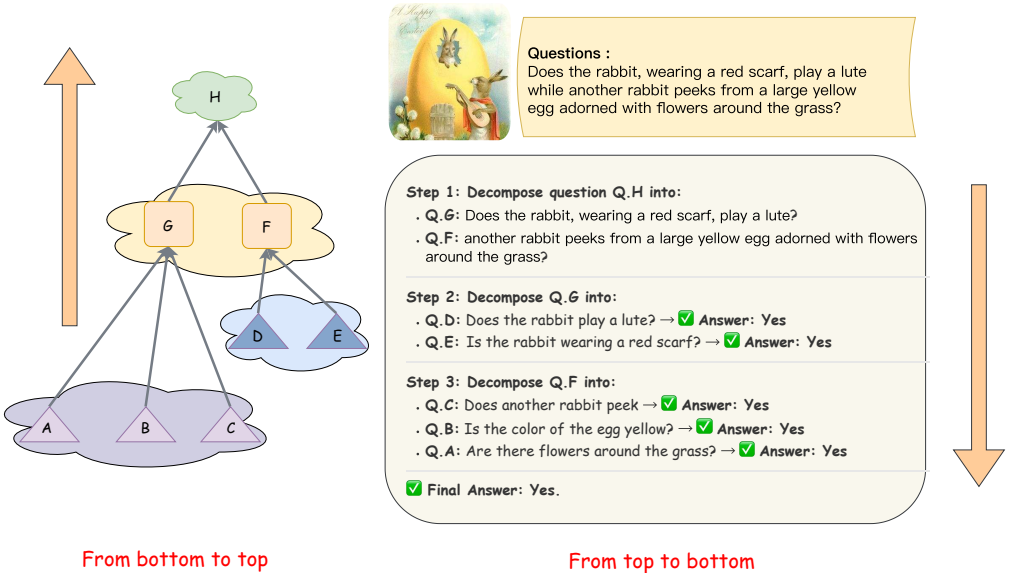

Figure 3: Atomic-level questions derived from image captions are merged bottom-up into higher-level reasoning chains based on semantic similarity, forming a question hierarchy. In turn, complex questions are decomposed top-down to generate training data with multi-granularity supervision.

Bottom-Up Merging to Build the Question Tree. We start by generating atomic questions from basic attributes and relationships extracted from data. These atomic questions serve as the fundamental units. Using a similarity-based merging strategy, we iteratively group similar questions to form intermediate-level questions. This merging process continues until we obtain complex, long-text questions representing high-level reasoning tasks.The merging criterion is determined by the embedding similarity between questions, calculated using pre-trained language models. Specifically, we compute the cosine similarity between sentence embeddings using the following formula:

$$\text{cosine}(s_i, s_j) = \frac{s_i \cdot s_j}{\|s_i\|\|s_j\| + \epsilon} \tag{2}$$

Here, s_i and s_j represent the embedding vectors of two questions, and the small constant $\epsilon = e^{-9}$ is added to prevent division by zero. To ensure diversity, the number of questions merged at each step is chosen randomly from a predefined range. This progressive merging forms a hierarchical question tree, where each level encapsulates a different granularity of information.

Top-Down Decomposition to Generate Training Data. Once the question tree is constructed, we utilize it in a reverse manner to generate data, as illustrated in Figure 3. Starting from the complex long-text questions at the top, we systematically decompose them into intermediate and atomic questions. This top-down decomposition mirrors the reasoning process where complex queries are broken down into simpler subproblems. By representing the reasoning path in a hierarchical structure, the model learns to navigate from abstract, high-level concepts to concrete, fundamental details, thereby enhancing its capacity for comprehensive question understanding.

By leveraging this dual approach — constructing questions in a bottom-up manner while generating training data in a top-down fashion — we ensure that the model is exposed to both complex reasoning and fundamental atomic units. This combination fosters robust question-reasoning capabilities, allowing the model to better handle multi-step reasoning and long-text analysis.

3.3 A COGNITION-ALIGNED TRAINING METHOD

While preliminary SFT provides basic reasoning, it often fails on complex queries by relying on superficial patterns. We thus adopt RFT with CCVR, leveraging our hierarchical CoT trajectories to ensure structured and logically consistent reasoning.

3.3.1 GRPO

We utilize the Group Relative Policy Optimization (GRPO) Shao et al. (2024) method to fine-tune large language models (LLMs) for enhanced reasoning performance in complex tasks. GRPO is designed to optimize model behavior by leveraging comparisons among multiple responses generated from the same prompt, rather than solely evaluating individual outputs.

The GRPO training begins with the generation of multiple response sequences for each input prompt. These responses are then evaluated using a reward function, which assigns scores based on the quality and relevance of each response. To capture relative performance, the advantage of each response is computed by normalizing the reward values within the group, following the formula:

$$\hat{A}^{i,t} = \frac{r_i - \mu_r}{\sigma_r} \tag{3}$$

where r_i denotes the reward of the i-th response, while μ_r and σ_r represent the mean and standard deviation of the rewards across the group, respectively. This normalized advantage helps to highlight how each response compares to others in the same batch. To prevent the model from deviating excessively from the reference policy, the KL divergence between the current policy and the reference policy is estimated using the following formula:

$$D_{\mathrm{KL}}[\pi_\theta \| \pi_{\text{ref}}] = \frac{\pi_{\text{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} - \log \frac{\pi_{\text{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} - 1, \tag{4}$$

The overall loss function in GRPO combines the advantage function and the KL divergence penalty as follows:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{(q,\{o_i\})} \left[\sum_{i=1}^{G} \sum_{t=1}^{|o_i|} \left(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\text{ref}}(o_{i,t} \mid q, o_{i,<t})} \hat{A}_{i,t} - \beta D_{\mathrm{KL}}[\pi_\theta \| \pi_{\text{ref}}] \right) \right] \tag{5}$$

Here, β acts as a hyperparameter controlling the penalty for divergence, ensuring that the updated policy remains close to the reference policy. This approach facilitates more stable optimization, allowing the model to adapt to complex reasoning tasks without sacrificing consistency or performance. By integrating the relative comparison of multiple outputs and maintaining a balanced update through KL regularization, GRPO enhances the robustness and efficacy of LLM fine-tuning.

3.3.2 COGNITIVELY COHERENT VERIFIABLE REWARDS (CCVR)

Building on the GRPO framework, we introduce Cognitively Coherent Verifiable Rewards (CCVR), designed to evaluate both the final output and intermediate reasoning steps. The complete optimization procedure is summarized in Algorithm 1. Unlike GRPO, which typically focuses on optimizing generative models through reinforcement learning by maximizing a reward function that evaluates the quality of final responses. Our approach incorporates structured reasoning evaluation directly into the reward function. Our rule-based reward function decomposes the entire reasoning process into structured steps, using predefined rules to capture both the accuracy of intermediate reasoning and the logical consistency of the final output. To comprehensively assess response quality, the model integrates three core components: Format Reward, Answer Reward, and Process Reward. These components address structural correctness, output accuracy, and reasoning coherence, respectively, ensuring a balanced evaluation that prioritizes both accurate answers and consistent reasoning. The overall reward r is calculated as a combination of the three components:

$$r = \lambda_{\text{format}} \cdot r_{\text{format}} + \lambda_{\text{format}} \cdot r_{\text{answer}} + \lambda_{\text{format}} \cdot r_{\text{process}} \tag{6}$$

Here, r_{format}, r_{answer}, and r_{process} represent the format, answer, and process rewards, respectively. The λ_{format}, λ_{format}, and λ_{format} control the relative importance of each component. This combined reward structure ensures that the model not only generates accurate final answers but also adheres to structured reasoning and formatting.

Format and Answer Rewards. The Format Reward evaluates structural correctness by verifying the presence and sequence of <think> and <answer> tags, ensuring model interpretability. The Answer Reward ensures alignment with ground-truth labels by rewarding explicit, conclusive answers. To mitigate reward hacking, it prioritizes single, unambiguous conclusions over contradictory outputs, better aligning the model with human-like decision-making

Process reward. The process reward is a critical component that evaluates the coherence and logical consistency of the reasoning steps. In structured reasoning tasks, maintaining a consistent logical pathway from input to final output is essential. This reward component quantifies the alignment between the generated reasoning process and a reference reasoning chain. It is specifically designed to capture both semantic similarity and logical sequence coherence, which are crucial for accurately modeling multi-step reasoning tasks. The process reward combines two sub-scores with linear weighting: the Semantic Score and the Edit Distance Score:

$$r_{\text{process}} = \lambda \cdot S_{\text{semantic}} + (1 - \lambda) \cdot S_{\text{edit}} \tag{7}$$

Here, λ controls the balance between the semantic similarity and the edit distance components. The semantic score quantifies the similarity between the generated reasoning steps and the reference reasoning steps. To calculate this score, we first encode the sentences from both the generated and reference reasoning using an embedding model, which converts each sentence into a representation. We then compute the similarity between each pair of sentences using cosine similarity. The semantic score is calculated as the proportion of sentence pairs whose similarity exceeds a threshold δ:

$$S_{\text{semantic}} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}\left(\text{Similarity}(x_i, y_j) \geq \delta\right)}{\max(m, n)} \tag{8}$$

In this formula, m and n denote the number of sentences in the generated reasoning and reference reasoning, respectively, and $\mathbb{1}$ is an indicator function that returns 1 if the similarity exceeds δ, and 0 otherwise. This score captures the extent to which the generated reasoning steps are semantically aligned with the expected logical flow, thereby ensuring that both content accuracy and reasoning coherence are assessed. To compute the soft edit distance between a generated reasoning chain $G = \{g_1, \ldots, g_m\}$and a reference chain $T = \{t_1, \ldots, t_n\}$, we define a dynamic programming matrix $D \in \mathbb{R}^{(m+1)\times(n+1)}$, where $D_{i,j}$ represents the minimum number of operations required to align the first i sentences of G with the first j sentences of T.We initialize the borders as:

$$D_{0,j} = j,\ \forall j \in [0, n], \qquad D_{i,0} = i,\ \forall i \in [0, m]. \tag{9}$$

The recurrence is defined for $i > 0$ and $j > 0$ (i.e., excluding the boundary cases).

$$D_{i,j} = \begin{cases} D_{i-1,j-1}, & \text{if } \text{sim}(g_i, t_j) \geq \theta, \\ \min \begin{cases} D_{i-1,j} + 1 \\ D_{i,j-1} + 1 \\ D_{i-1,j-1} + 1 \end{cases}, & \text{otherwise} \end{cases} \qquad S_{\text{edit}} = 1 - \frac{D_{m,n}}{n}. \tag{10}$$

Here, $\text{sim}(g_i, t_j) \in [0, 1]$ denotes the cosine similarity between sentence embeddings, and θ is the similarity threshold for considering two sentences semantically equivalent. $D_{m,n}$ is the total soft edit distance, and n is the number of reference sentences used for normalization.

Algorithm 1: Reinforcement Fine-Tuning with Cognitively Coherent Verifiable Rewards

Input: Initial policy model π_θ (from a pretrained VLM);
Dataset $\mathcal{D} = \{(x_i, y_i, r_i^*)\}$, where x_i: prompt, y_i: ground-truth answer, r_i^*: reference reasoning chain
Output: Updated policy model π_θ
for *each training step* **do**
  Sample (x, y, r^*) from dataset $\mathcal{D}$;
  Generate K trajectories $\{\tau_k\}_{k=1}^K$ using $\pi_\theta(\tau \mid x)$;
  for *each trajectory* τ_k **do**
    Compute format score s_{format} based on presence and order of <think> and <answer>;
    Compute answer score s_{answer} by comparing answer in <answer> section with y;
    Extract reasoning steps from <think> section and compute process score:;
    - Semantic similarity s_{sem} via Sentence-BERT;
    - Edit similarity s_{edit} via normalized edit distance;
    - $s_{\text{proc}} \leftarrow \lambda s_{\text{sem}} + (1-\lambda) s_{\text{edit}}$;
    Compute final reward:;
    $s_k \leftarrow \lambda_{\text{format}} s_{\text{format}} + \lambda_{\text{answer}} s_{\text{answer}} + \lambda_{\text{proc}} s_{\text{proc}}$;
  Compute baseline reward: $\bar{s} \leftarrow \frac{1}{K} \sum_{k=1}^K s_k$;
  Compute policy gradient:;
  $\nabla_\theta \mathcal{L} \leftarrow \frac{1}{K} \sum_{k=1}^K (s_k - \bar{s}) \nabla_\theta \log \pi_\theta(\tau_k \mid x)$;
  Update π_θ using gradient descent;
return π_θ

4 Experiment

4.1 Experiment Settings

Dataset and Metrics. We first constructed a training set of 1,000 image-question pairs, which was subsequently expanded to 4,000 samples to achieve consistent performance improvements. For evaluation, we curated a multi-source benchmark consisting of 100 high-quality samples uniformly selected from five datasets. This test set is structured with multiple difficulty levels and fine-grained, lexically altered negative samples to assess the model's reasoning robustness under challenging conditions. Performance is measured using Accuracy (ACC) and F1 Score across total, in-domain, and out-of-domain settings. Comprehensive details on the composition of the data set, source distributions and construction protocols are provided in Appendix B

4.2 Main Results

Table 1 compares different training strategies on our annotation-free, hierarchically structured CoTZero dataset (see Sec. D for the experimental setup and evaluation protocol). Conventional fine-tuning alone yields only moderate improvements and remains insufficient for robust and interpretable reasoning. In contrast, leveraging our training method with CCVR gains in overall performance and produces consistent improvements across total, in-domain, and out-of-domain evaluations, indicating better generalization and more reliable reasoning behavior.

Compared to the baseline, our full method improves total F1 by nearly 10%, with over 20% improvement in in-domain accuracy. Out-of-domain F1 also rises substantially, demonstrating enhanced generalization. These results underscore the effectiveness of our synthetic reasoning data in teaching structured, multi-step thought processes. Moreover, CCVR further reinforce the model's ability to produce verifiable and logically coherent reasoning chains. Overall, our framework outperforms all other configurations, validating the synergy between structured data generation and reasoning-aware optimization. Additional qualitative results are provided in Appendix C.

Table 1: Performance comparison under different training strategies.

Method	**Total**		**In-domain**		**Out-of-domain**	
	ACC	F1	ACC	F1	ACC	F1
Qwen-VL 2.5 (3B)	65.31	71.98	56.41	66.67	67.52	73.30
+ SFT (LoRA)	72.04	75.47	62.16	66.67	74.50	77.65
+ Cold Start	66.67	73.04	51.35	59.09	70.47	76.34
+ CCVR	77.42	80.19	81.08	82.05	76.51	79.77
+ Cold Start + SFT + CCVR (*Best*)	**81.11**	**81.95**	76.68	75.68	81.21	83.33

Table 2: Impact of Different Negative Sample Strategies for CoT Data Generation.

Strategy Setting	**ACC (%)**	**F1 (%)**
Baseline (no Negative Samples)	65.30	71.98
Cross-image negatives	51.61	67.68
Lexically altered negatives	**65.59**	**74.80**

4.3 Ablation Results

4.3.1 Ablation Study of the Synthesis Method

To understand the contribution of our proposed training data components, we conduct a comprehensive ablation study on negative sample construction and atomic QA composition strategies, as summarized in Table 2 and Table 3.

Table 2 investigates the impact of different negative sample strategies. Without any negative samples, the model exhibits limited baseline performance. Introducing cross-image negatives where questions from one image are used as negatives for another significantly degrades the outcome. This suggests that such negatives lack semantic relevance and may introduce misleading noise. In contrast, lexically altered questions as negatives better preserve the semantic structure while introducing controlled perturbations, enabling the model to effectively learn meaningful reasoning boundaries.

Table 3 focuses on atomic QA composition strategies under the lexically altered negative setting. Without composition, the model lacks structural guidance. Fixed template combinations offer a moderate improvement but suffer from deterministic patterns and limited diversity. In contrast, our full method *dual-stage data synthesis approach* achieves the best performance (F1: 84.09%), highlighting the benefit of structurally diverse and multi-granular supervision. The hierarchical nature of this composition strategy encourages reasoning that better align with human cognition. These findings validate that both semantically meaningful negative samples and structurally diverse reasoning paths are critical to developing interpretable and robust visual reasoning systems.

4.3.2 Ablation Study of Reward Components

Results in table 4 demonstrate that removing any individual CCVR component—format, answer, semantic, or edit distance—leads to a consistent performance drop across all metrics and domains. The full CCVR model consistently outperforms the baseline across total, in-domain, and out-of-domain splits. Removing any single reward term leads to a performance drop, indicating that the reward components are complementary. Overall, the ablation supports that CCVR's stepwise reward is important for learning well-structured multi-step reasoning and maintaining robustness.

Table 3: Effect of Atomic QA Composition Strategies

Atomic QA Composition Strategy	**ACC (%)**	**F1 (%)**
Baseline (no Composition)	65.31	74.98
Fixed template combination	77.42	79.21
Dual-stage data synthesis approach	**81.21**	**84.09**

Table 4: Ablation study of cognitively coherent verifiable reward (CCVR). Removing any individual reward component leads to a consistent performance drop across all domains.

Method	Total		In-domain		Out-of-domain	
	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)
Baseline	65.31	71.98	56.41	66.67	67.52	73.30
CCVR (Ours)	**77.96**	**79.60**	**72.97**	**75.00**	**79.19**	**80.75**
No format score	75.81	76.92	83.78	83.33	73.83	75.47
No answer score	64.52	69.44	59.46	66.67	65.77	70.18
No semantic score	72.58	76.92	72.97	77.27	72.48	76.84
No edit distance score	76.34	79.63	72.97	77.27	77.18	80.23

5 Conclusion and Discussion

Inspired by findings in human cognitive science, we introduce CoTZero, an annotation-free framework that hierarchically parses visual scenes and improves vision–language reasoning via dual-stage data synthesis and cognition-aligned training. CoTZero autonomously generates structured CoT data without human labels and achieves strong performance in both in-domain and out-of-domain settings. Future directions include: (1) exploring more efficient RL optimization algorithms under Cognitively Coherent Verifiable Rewards (CCVR); (2) scaling up synthetic data generation and RL training to further improve VLM reasoning robustness; and (3) extending the hierarchical reasoning paradigm to broader multimodal understanding and reasoning. These advancements could unlock deeper compositional reasoning in VLMs, bridging the gap between pattern recognition and genuine scene comprehension.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. Technical report, Alibaba Group, 2025. URL https://arxiv.org/abs/2511.21631.

Thomas Bever and David Poeppel. Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics*, 4:174–200, 09 2010. doi: 10.5964/bioling.8783.

David R. Bull and Fan Zhang. Chapter 2 - the human visual system. In David R. Bull and Fan Zhang (eds.), *Intelligent Image and Video Compression (Second Edition)*, pp. 17–58. Academic Press, Oxford, second edition edition, 2021. ISBN 978-0-12-820353-8. doi: https://doi.org/10.1016/B978-0-12-820353-8.00011-6. URL https://www.sciencedirect.com/science/article/pii/B9780128203538000116.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption, 2024. URL https://arxiv.org/abs/2405.19092.

Chengyi Du and Keyan Jin. Multi-object grounding via hierarchical contrastive siamese transformers. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2025.

Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991. doi: 10.1093/cercor/1.1.1-a.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought, 2025. URL https://arxiv.org/abs/2411.19488.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training, 2022. URL https://arxiv.org/abs/2211.11559.

Charles F. Hockett. The origin of speech. *Scientific American*, 203(3):88–96, 1960. doi: 10.1038/scientificamerican0960-88. URL https://doi.org/10.1038/scientificamerican0960-88.

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding, 2025. URL https://arxiv.org/abs/2502.11492.

Hyeonbin Hwang, Doyoung Kim, Seungone Kim, Seonghyeon Ye, and Minjoon Seo. Self-explore: Enhancing mathematical reasoning in language models with fine-grained rewards, 2024. URL https://arxiv.org/abs/2404.10346.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms, 2025. URL https://arxiv.org/abs/2410.01679.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people, 2016. URL https://arxiv.org/abs/1604.00289.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. Vlrewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*, 2024a.

Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans, 2024b. URL https://arxiv.org/abs/2404.01294.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning, 2025. URL https://arxiv.org/abs/2412.08614.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning, 2024. URL https://arxiv.org/abs/2401.08967.

Hauser Marc, D. Chomsky, Noam Fitch, and W. Tecumseh. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(22):1569–1579, 2002.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models, 2024. URL https://arxiv.org/abs/2310.02567.

Paul Mermelstein and Murray Eyden. A system for automatic recognition of handwritten words. In *Proceedings of the October 27-29, 1964, Fall Joint Computer Conference, Part I*, AFIPS '64 (Fall, part I), pp. 333–342, New York, NY, USA, 1964. Association for Computing Machinery. ISBN 9781450378895. doi: 10.1145/1464052.1464081. URL https://doi.org/10.1145/1464052.1464081.

Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2468–2479. IEEE, June 2024. doi: 10.1109/cvpr52733.2024.00239. URL http://dx.doi.org/10.1109/CVPR52733.2024.00239.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(77)90012-3. URL https://www.sciencedirect.com/science/article/pii/0010028577900123.

Aude Oliva, Antonio Torralba, and Philippe G. Schyns. Hybrid images. *ACM Trans. Graph.*, 25(3): 527–532, July 2006. ISSN 0730-0301. doi: 10.1145/1141911.1141919. URL https://doi.org/10.1145/1141911.1141919.

Open Diffusion AI. Laion-2b-en aesthetic square cleaned. Hugging Face Dataset, 2025. URL https://huggingface.co/datasets/opendiffusionai/laion2b-en-aesthetic-square-cleaned. Subset of LAION-2B-EN-Aesthetic-Square with watermark & duplicate removal (200 K images).

OpenAI. GPT-4V(ision) System Card. Technical report, OpenAI, September 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf. System card / technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Sang-Min Park and Young-Gab Kim. Visual language integration: A survey and open challenges. *Comput. Sci. Rev.*, 48(C), May 2023. ISSN 1574-0137. doi: 10.1016/j.cosrev.2023.100548. URL https://doi.org/10.1016/j.cosrev.2023.100548.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogcom: A visual language model with chain-of-manipulations reasoning, 2025. URL https://arxiv.org/abs/2402.04236.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Christoph Schuhmann and Peter Bevan. 220k-gpt4vision-captions-from-lvis. https://huggingface.co/datasets/laion/220k-GPT4Vision-captions-from-LIVIS, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers, 2025. URL https://arxiv.org/abs/2506.23918.

LLaVA Team. Llava: Large language and vision assistant, 2025. https://llava-vl.github.io/.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Llamav-o1: Rethinking step-by-step visual reasoning in llms, 2025. URL https://arxiv.org/abs/2501.06186.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models, 2024a. URL https://arxiv.org/abs/2407.00634.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2024b. URL https://arxiv.org/abs/2312.08935.

Xiangyu Wang. LAION-Art. Hugging Face Dataset, 2023. URL https://huggingface.co/datasets/fantasyfish/laion-art. 20.9k image-text pairs (aesthetic scores) in Parquet format.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. Towards large reasoning models: A survey of reinforced reasoning with large language models, 2025a. URL https://arxiv.org/abs/2501.09686.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2025b. URL https://arxiv.org/abs/2411.10440.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in mllm reasoning, 2025. URL https://arxiv.org/abs/2507.03019.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024. URL https://arxiv.org/abs/2302.00923.

Chongyang Zhao, Yuankai Qi, and Qi Wu. Mind the gap: Improving success rate of vision-and-language navigation by revisiting oracle success routes, 2023. URL https://arxiv.org/abs/2308.03244.

Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, Fei Luo, Xiaohua Chen, Xiaoshuai Hao, Hehan Li, Andi Zhang, Wenxuan Wang, Kaiyan Zhang, Guoli Jia, Lingling Li, Zhiwu Lu, Yang Lu, and Yike Guo. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models, 2025. URL https://arxiv.org/abs/2509.25373.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. Technical report, Shanghai AILab, 2025. URL https://arxiv.org/abs/2504.10479.

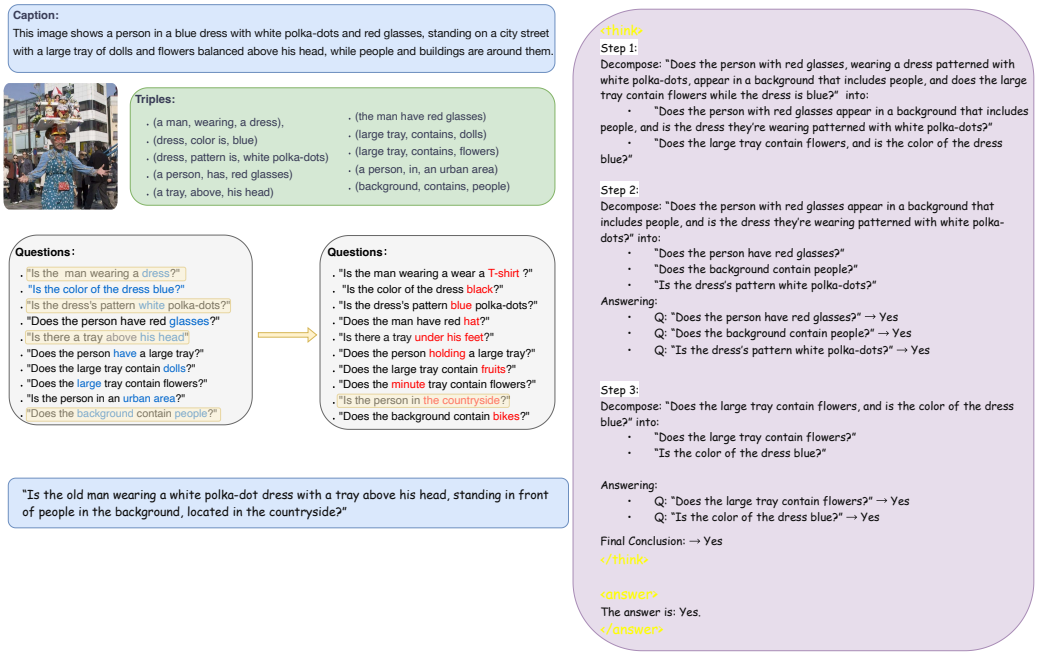

Figure 4: Data generation details.

A Implementation Details of Data Synthesis

As illustrated in Fig. 4, we construct our training dataset through a multi-stage pipeline: **Caption Generation.** We use the *Doubao-1.5-Vision-Pro* model to generate concise and factual image descriptions. The system is prompted to use fewer than 150 English words, describe only visible content in the present tense, avoid subjective speculation or technical jargon, and begin with phrases like "The image" or "This is".

Triple Extraction. We employ *DeepSeek-V3* to extract semantic triples from the generated captions. Each triple is represented as $(\text{Entity}_1, \text{Relation}, \text{Entity}_2)$ and categorized into one of eight relation types: action verbs, state verbs, possession verbs, spatial/location verbs, causality/effect verbs, temporal verbs, quantitative verbs, and perception verbs.

Atomic Question Generation and Negative Sampling. For each triple, we generate an atomic yes/no question. To construct hard negative samples, we replace key tokens in the question such that the answer flips from "yes" to "no". Each modified question is validated to ensure grammaticality, structural consistency, and successful answer reversal.

Compound Question Construction. Multiple atomic questions are merged into fluent compound questions using *DeepSeek-V3*. A natural language prompt is applied to generate semantically coherent outputs without overusing simple conjunctions such as "and".

B Dataset Details

Dataset.We first constructed a training set of 1,000 image-question pairs by sampling a subset of images from FantasyFish, and then expanded it to 4,000 samples by adding images from additional source. These images were annotated using our proposed data generation pipeline, which produces both semantically correct and fine-grained negative questions by modifying key terms based on visual content. For evaluation, we curated a 100-image test set by selecting 20 high-quality samples from each of the following datasets: FantasyFish Wang (2023), OpenDiffusion Open Diffusion AI (2025), Night2Day Isola et al. (2017) (night subset), MMInstruction Li et al. (2024a), and LVIS

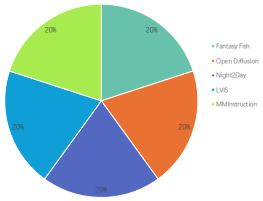

Figure 5: The composition of our test set.

Schuhmann & Bevan (2023). For each image, we generated negative samples by replacing key terms in the original questions, ensuring that the modified questions remained fluent but semantically incorrect. The test set was structured into multiple difficulty levels, with higher levels involving a greater number of positive distractors. This design allows us to assess the model's ability to identify subtle semantic inconsistencies under increasingly challenging conditions across both in-domain and out-of-domain samples.

B.1 TRAINING DATASET

Our training dataset initially consists of 1,000 samples generated using our proposed Annotation-free Hierarchical CoT Generation pipeline, leveraging images from the Fantasy Fish dataset. We then expand the dataset to 4,000 samples, observing consistent improvements in the model's performance across various aspects.

B.2 TEST DATASET

Figure 5 illustrates the composition of our evaluation set, which is uniformly sampled from five sources: FantasyFish Wang (2023), OpenDiffusion Open Diffusion AI (2025), Night2Day (night subset) Isola et al. (2017), MMInstruction Li et al. (2024a), and LVIS Schuhmann & Bevan (2023), each contributing 20% of the total. This balanced mixture ensures diversity in object categories, lighting conditions, and instruction-following scenarios, supporting a comprehensive evaluation of our model's reasoning and perception capabilities.

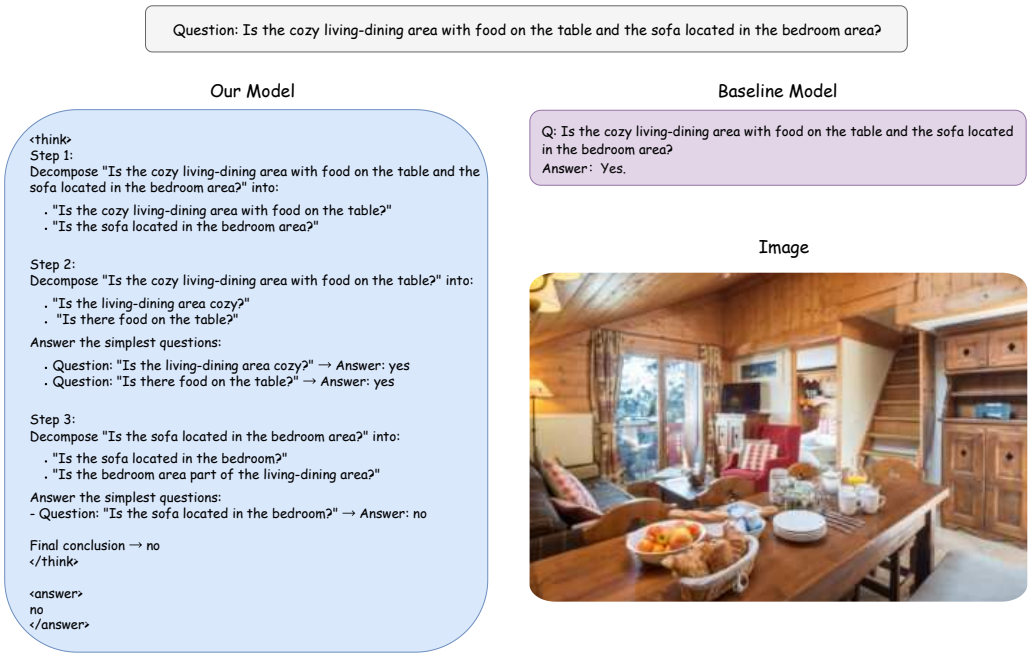

Figure 6: Comparison showing that our model performs step-by-step reasoning to verify details, while the baseline provides a direct answer without decomposition.

C Additional Quantitative Results

To better understand the differences between our approach and the baseline, we present the comparison in Figure 6. The baseline model directly outputs a single answer without reasoning, which can lead to errors when the question requires multi-step verification across different visual elements. In contrast, our model decomposes the complex question into sub-questions, systematically verifies each aspect, and then integrates the results to reach a final answer. This structured reasoning enables our model to handle questions involving multiple objects and spatial relationships more reliably, reducing hallucinations and improving consistency in multi-step visual question answering.

D Experiment Configuration

Experiment configuration. We first fine-tuned the Qwen2.5-VL-3B model using full-parameter tuning or LoRA with a rank of 64. The training is conducted for 3 epochs with a global batch size of 64 on 8 A800 GPUs, using gradient accumulation to simulate large-batch training. The learning rate of SFT is set to 2e-5, with a weight decay of 0.1 and cosine learning rate scheduling. A warmup ratio of 0.03 was applied. Besides supervised fine-tuning, we further optimize the model via reinforcement learning with the GRPO algorithm. The learning rate was set to 1e-6 with a weight decay of 1e-2. An initial KL coefficient of 0.001 was used for regularization. Reinforcement learning was performed for 3 episodes with a batch size of 64. The model is optimized using DeepSpeed ZeRO3.