STELAR-VISION: Self-Topology-Aware Efficient Learning for Aligned Reasoning in Vision

Chen Li, Han Zhang, Zhantao Yang, Fangyi Chen, Zihan Wang, Anudeepsekhar Bolimera, Marios Savvides

Carnegie Mellon University
{chenli4, hanz3, zhantaoy, fangyic, zihanw4, abolimer, marioss}@andrew.cmu.edu

Abstract

Vision-language models (VLMs) have made significant strides in reasoning, yet they often struggle with complex multimodal tasks and tend to generate overly verbose outputs. A key limitation is their reliance on chain-of-thought (CoT) reasoning, despite many tasks benefiting from alternative topologies like trees or graphs. To address this, we introduce STELAR-Vision, a training framework for topology-aware reasoning. At its core is TopoAug, a synthetic data pipeline that enriches training with diverse topological structures. Using supervised fine-tuning and reinforcement learning, we post-train Qwen2VL models with both accuracy and efficiency in mind. Additionally, we propose Frugal Learning, which reduces output length with minimal accuracy loss. On MATH-V and VLM_S2H, STELAR-Vision improves accuracy by 9.7% over its base model and surpasses the larger Qwen2VL-72B-Instruct by 7.3%. On five out-of-distribution benchmarks, it outperforms Phi-4-Multimodal-Instruct by up to 28.4% and LLaMA-3.2-11B-Vision-Instruct by up to 13.2%, demonstrating strong generalization. Compared to Chain-Only training, our approach achieves 4.3% higher overall accuracy on in-distribution datasets and consistently outperforms across all OOD benchmarks.

Datasets — https://huggingface.co/collections/Stellar-Neuron/stelar-vision-collection

1 Introduction

Recent advances in large language models (LLMs) have significantly improved reasoning capabilities, with models like GPT-o3 achieving strong performance on complex mathematical and scientific tasks. This progress has extended into the multimodal domain through vision-language models (VLMs) such as GPT-4o (Hurst et al. 2024), GPT-4o-mini (OpenAI 2024), and Qwen2.5-VL (Bai et al. 2025). Despite the recent advances, there is still room of improvement in open-sourced VLMs when tackling complex vision-based reasoning tasks (e.g., math and science questions), and the path to enhance their abilities under an affordable training budget remains under-explored.

To address this, we begin by analyzing VLMs' reasoning behaviors and find that the popular models, both open-source and closed-source, tend to default to the chain-of-thought (CoT) (Wei et al. 2023) generation. This behavior reflects the prevailing trend in their training data, which is overwhelmingly dominated by CoT-style reasoning samples. However, our empirical analysis reveals that *different questions benefit from different reasoning topologies, such as Chain, Tree, or Graph structures* (Figure 1). The benefits of diverse reasoning topologies have yet been well studied or effectively incorporated into existing training pipelines. Moreover, even state-of-the-art reasoning VLMs tend to generate unnecessarily verbose responses, i.e., "overthinking", which increases the computational cost and makes real-time applications less viable. We find that there is a correlation between the topological reasoning structures and the output sequence length, thus providing an insight to the overthinking problem created by the CoT reasoning.

We introduce Self-Topology-Aware-Efficient-Learning for Aligned Reasoning in Vision, **STELAR-Vision**, a training framework for topology-aware vision-language reasoning. The central to this framework is **TopoAug**, a synthetic data generation pipeline that produces augmented reasoning topologies with diverse structures, including $\{Chain, Tree, Graph\}$. Specifically, we generate a set of question-answer responses, where each question is repeatedly answered by different topological reasoning structures. For each question, we assign a preferred topology through an automated annotation process. The generated responses and their labels are used to post-train VLMs via supervised fine-tuning (SFT) and Reinforcement Learning (RL) (Meng, Xia, and Chen 2024). Regarding the learning process, we observe that RL further amplifies the performance gains introduced by topological augmentation. Moreover, we propose to further increase efficiency with Frugal Learning by promoting shorter responses. We show that only by leveraging the benefits of augmented reasoning topologies, can the model generate concise responses and maintain high accuracy with minimal performance drops, while the counterpart model trained only on CoT-style data fails to gain efficiency without incurring a greater loss in accuracy.

The augmented topologies expand the exploration space, allowing RL to discover higher-quality optima and leading to superior performance improvement. In contrast, the model trained with CoT-style data only sees diminishing returns. We also test our models on five out-of-distribution

Figure 1: **Limitations of the Popular Chain-of-Thought Reasoning Structures.** The widely adopted Chain-of-Thought (CoT) reasoning paradigm (in green) often results in unnecessarily verbose reasoning processes, as demonstrated in the first example. Under CoT reasoning, the model redundantly counts each cube, whereas with $Graph$ topology (in blue), it quickly identifies the key point of the question. In the bottom-row example, CoT reasoning begins with a detailed examination of each subplot but ultimately arrives at an incorrect answer. In contrast, $Tree$ topology (in red) initiates reasoning with a high-level overview before delving into specific features. In both scenarios, CoT-style reasoning proves suboptimal.

datasets, and it consistently outperform its base model, while consuming fewer tokens. which suggests strong generalization of our proposed STELAR-Vision.

Our contributions are summarized as follows:

- We propose STELAR-Vision, a training framework explicitly trained for topology-aware reasoning. It leverages diverse reasoning topologies such as chains, trees, and graphs, aligns reasoning paths with question characteristics, and enables adaptive and efficient multimodal inference.
- We introduce TopoAug, a data generation pipeline that automatically produces diverse topological reasoning and annotates optimal structures per question. We also integrate Frugal Learning into the learning framework, achieving reductions in output length with minimal accuracy tradeoff.
- By conducting experiments with post-training supervision and reinforcement learning, STELAR-Vision improves accuracy by 9.7% over its base model and its larger variant Qwen2VL-72B-Instruct by 7.3%. On the out-of-distribution dataset, it surpasses The Frugal Learning variant reduces output length by 18.1% while maintaining comparable accuracy.

2 Related Work

2.1 Topological Reasoning in Language and Vision Models

Chain-of-Thought (CoT) prompting (Wei et al. 2023) is a widely used reasoning strategy in LLMs and VLMs, guiding models to generate step-by-step solutions. However, its

linear structure may not suit all tasks. To address this, Tree-of-Thought (ToT) (Yao et al. 2023) enables branching exploration, while Graph-of-Thought (GoT) (Besta et al. 2024) supports iterative and global reasoning. Both improve performance on complex tasks like TSP, algorithmic problem-solving, and multi-stage decision-making.

These methods, however, often rely on rule-based topology generations through prompt engineering or sampling, and are limited to language-only settings. In contrast, our framework automatically generates diverse topological structures and trains a VLM to adaptively select the optimal one per instance during decoding, enabling more flexible and generalizable reasoning.

2.2 Reinforcement Learning for LLM and VLM Reasoning

Reinforcement learning (RL) is a key technique for aligning LLMs and VLMs with desired behaviors in reasoning and preference modeling. Approaches like RLHF (Stiennon et al. 2022; Ouyang et al. 2022) and Constitutional AI (Bai et al. 2022) enable models to acquire complex reasoning and ethical behaviors via preference optimization.

Reward-based methods such as PPO (Schulman et al. 2017), RPO (Yin et al. 2024), and GRPO (Shao et al. 2024) rely on explicit rewards, while reward-free approaches like DPO (Rafailov et al. 2024), SimPO (Meng, Xia, and Chen 2024), and ORPO (Hong, Lee, and Thorne 2024) achieve comparable results without reward modeling. These methods are widely applied to mathematical reasoning, long-horizon tasks, and instruction tuning.

In VLMs, RL has been used to enhance structured reasoning and safety-critical applications. VLM-RL (Huang et al. 2024) improves decision-making in autonomous driving, MedVLM-R1 (Pan et al. 2025) ensures safety in medical imaging, and RLVR (Chen et al. 2025) boosts OOD generalization in tasks like visual counting and open-ended QA.

Building on these insights, we show that combining topology-aware data generation with RL (e.g., SimPO) improves both accuracy and efficiency. Topological diversity expands the exploration space, increasing the likelihood of discovering stronger reasoning strategies during RL.

2.3 Curriculum Learning and Structured Reasoning

Curriculum learning (Bengio et al. 2009) trains models on progressively harder tasks, aiding structured reasoning. Recent works extend this idea: Xi et al. (Xi et al. 2024) use reverse curricula for LLMs, Zhao et al. (Zhao et al. 2024) propose Auto-CEI for ability-aligned sampling, and Ma et al. (Ma, Jiang, and Huang 2025) design logic-guided curricula for in-context learning. Fine-tune-CoT (Ho, Schmid, and Yun 2023) learns CoT from larger reasoning models.

In VLMs, LlamaV o1 (Thawakar et al. 2025) schedules visual examples for step-wise reasoning, and VLM-R1 (Shen et al. 2025) integrates curriculum-aware GRPO for stability and interpretability. However, most curricula are static and predefined.

In contrast, our approach is an end-to-end learning via a dataset from TopoAug, where synthetic reasoning paths evolve with model performance. Coupled with RL and topological diversity, our training framework promotes structured reasoning in a more scalable and generalizable way.

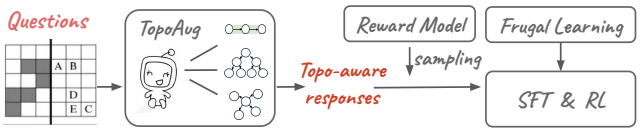

Figure 2: An overview of the STELAR-Vision framework

3 Method

In this section, we first construct topology-aware responses on two mathematical datasets. We then investigate the relationship between the topological reasoning and response accuracy. Finally, we present the topology-aware training framework shown in Figure 2.

3.1 Constructing Topology-Aware Responses

Data We use two math datasets: MATH-V (Wang et al. 2024a) (3,040 visual problems) and VLM_S2H (Park et al. 2025) (7,000 logic puzzles), each pairing images with questions, giving diverse samples requiring high-level reasoning

TopoAug We generate **Topology-Aware** responses using topologies $T = \{Chain, Tree, Graph\}$, prompted via Qwen2-VL-7B-Instruct (Wang et al. 2024b) and GPT-4o-Mini (OpenAI 2024a) with extensive degrees of freedom in maximum depth, number of children, and number of neighbors. Please see the supplementary for detailed prompts.

Topology and Outcome Labels Each response r has an **Outcome Label** $\mathcal{H}_r \in \{0, 1\}$, and is assigned label 1 if correct and 0 otherwise. Each question-topology pair gets a **Topology Label** $\mathcal{F}_{q,t} = \frac{N_{\text{correct}}(q,t)}{N_{\text{total}}(q,t)}$ based on accuracy, where $N_{\text{correct}}(q, t)$ is the number of correct responses using topology t for question q, and $N_{\text{total}}(q, t)$ is the total number of responses generated using t.

Problem Difficulty Segmentation Problems are labeled Easy (>85th percentile), Hard (<15th), or Medium based on topology score distributions.

3.2 Analysis: Topological Reasoning Structures

We first show that different questions are better solved using distinct reasoning topologies. We evaluate the two aforementioned vision-language models (VLMs). We exclude the newer QWEN2.5-VL(Qwen et al. 2025) series from our analysis due to its relatively unstable performance in generating diverse reasoning topologies.

We first prompt the models to generate responses using their default reasoning behavior. Next, we explicitly instruct them to reason using three distinct topologies: *Chain*, *Tree*, and *Graph*. We compute a topology-wise *Win Rate* to assess the performance of each topological reasoning structure, which is defined below. We then conduct a subject category-wise study on win rate.

Win Rate: We measure across the entire dataset and calculate the percentage of occurrence where a topology t is the best performing reasoning structure among the three topology types in Equation 1:

$$\text{Win Rate}(t) = \frac{\sum_{q \in Q} \mathbb{1}_t(\arg\max_{t' \in T} \mathcal{F}_{q,t'})}{N_Q} \quad (1)$$

where N_Q is the total number of questions, and $\mathbb{1}$ is the indicator function. For each question, the topology with the highest Topology Label $\mathcal{F}_{q,t}$ wins. We first analyze which topological reasoning structure works the best for the questions. The Win Rate statistics is presented in Table 1.

	Chain	Tree	Graph
Win Rate	49%	28%	23%

Table 1: Comparison of Win Rates across reasoning topologies.

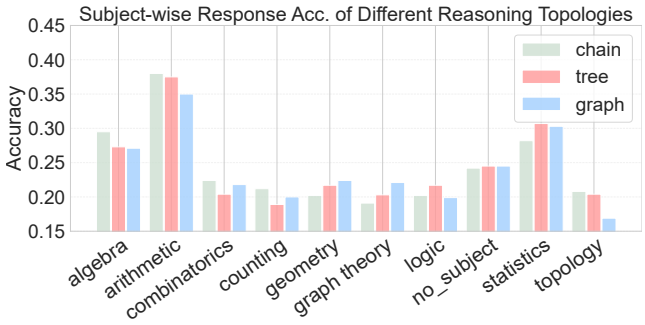

Figure 3: **Comparison of topology accuracy across subjects:** Accuracy of $Chain$, $Tree$, and $Graph$ reasoning topological structures per subject of MATH-V dataset. $Chain$ remains the best overall reasoning structure, while $Tree$, and $Graph$ perform better in at reasoning subjects such as "graph theory" or "statistics".

Our analysis in Table 1 reveals that $Chain$ structure reasoning remains the best overall reasoning pattern in both VLMs, regardless of model size. However, $Tree$ and $Graph$ collectively account for more than half of the winning responses. We conduct a detailed investigation of each topology's performance across subject domains by following the 10 subject categories in the MATH-V dataset and measuring the accuracy of each reasoning topology per subject, as illustrated in Figure 3. And thus we arrive at a conclusion that different problems benefit from different topologies, and identifying the optimal ones yields accuracy gains. This aligns with theoretical expectations—while straightforward tasks are adequately addressed by CoT-style reasoning, more complex or structurally intricate problems require richer topological representations to capture their underlying relationships effectively.

Topology-Wise Generation Length We further investigate the reasoning generation token length distribution within the TopoAug dataset. As illustrated in Figure 4, while most generations exhibit an average length of approximately 550 tokens, the $Chain$ topology produces the longest generations with a right-skewed distribution that favors extended reasoning processes. In contrast, both $Tree$ and $Graph$ topologies demonstrate similar token length distributions with overall a lower length, indicating comparable reasoning verbosity across these structured approaches.

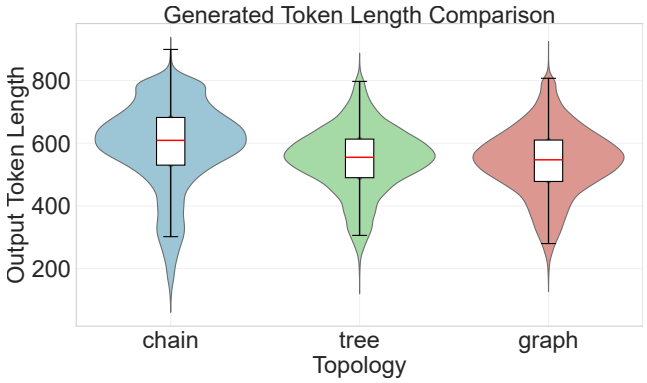

Figure 4: Distribution of generated reasoning token length of $Chain$, $Tree$, and $Graph$ topological structures in TopoAug Dataset. The box within each violin plot represents the median, and 25% and 75% percentile thresholds.

3.3 STELAR-Vision Post-Training

Our finding propels us to make two reasonable assumptions:

- **Assumption 1:** A model trained with topologically diverse reasoning structures—without increasing data volume—can achieve higher reasoning accuracy by learning to adaptively identify the best reasoning structure for each problem at test time.
- **Assumption 2:** Building up Assumption 1, a learning mechanism can be designed to encourage concise yet accurate outputs, enhancing inference efficiency with minimal performance loss.

We device a Post-training framework that consists of two phases: Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL).

Phases 1: Supervised Fine-tuning We collect the data generated via TopoAug and combine it with three additional datasets, OKVQA(Marino et al. 2019), A-OKVQA(Schwenk et al. 2022), and LLaVA150k-Instruct (Liu et al. 2023). We include them as general VQA tasks to preserve the model's basic VQA capabilities. Note that the additional 3 VQA datasets are used without any topological augmentation to avoid altering their original structure and maintain generalization. We perform SFT on TopoAug-generated data mixed with general VQA datasets, using a three-step filtering process: (1) balanced sampling from Easy, Medium, and Hard problems, (2) keeping only responses with positive outcome labels, and (3) rejection sampling with a 7B Outcome Reward Model (ORM) (Ouyang

et al. 2022), trained with both topology and outcome labels to select higher-quality samples with greater win potential.

The fine-tuning uses LoRA (Hu et al. 2021) with next-token prediction (NTP), minimizing the loss

$$\mathcal{L}_{\text{NTP}} = -\sum_{t=1}^{T} \log P_\theta(y_t \mid y_{<t}, x) \tag{2}$$

where x is the input, $y = (y_1, \ldots, y_T)$ the target, and P_θ the model's predicted token probabilities.

Phases 2: Reinforcement Learning We follow prior work (Chu et al. 2025; Zhai et al. 2024) by initializing reinforcement learning (RL) from an SFT checkpoint, which improves both in-distribution accuracy and OOD generalization. We adopt SimPO (Meng, Xia, and Chen 2024) for its simplicity and alignment with test-time behavior. Its objective encourages preferred responses over less-preferred ones, and is defined as

$$\mathcal{L}_{\text{SimPO}}(\pi_\theta) = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\frac{\beta}{|y_w|}\log\pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|}\log\pi_\theta(y_l \mid x) - \gamma\right)\right] \tag{3}$$

where x is the input, y_w and y_l are the preferred and less-preferred responses, π_θ is the policy, σ is the sigmoid, and β, γ are temperature and margin parameters.

We compare two RL setups: one trained on TopoAug-based preference pairs and one on Chain-only data of equal size. Correct responses are treated as the preferred responses. To prevent data leakage, we remove topology prompts during training, so the model must infer optimal structures at test time.

Frugal Learning While recent work has explored efficient reasoning in LLMs (Arora and Zanette 2025; Aggarwal and Welleck 2025), these approaches rely on additional reward models, lack cost-controllability, and do not focus on vision-based reasoning. To improve test-time reasoning efficiency, we propose *Frugal Learning*, which trains a compact variant called STELAR-V-Short. We propose and compare two training strategies:

Variant 1: STELAR-Vision-Short†. We first introduce a filter that identifies "short and correct" responses as preferred targets, where the outcome Label $\mathcal{H}_r = 1$ and the generated token length falls below a 25% percentile threshold. The model undergoes training with both supervised fine-tuning (SFT) and reinforcement learning (RL). During RL training, we designate "short and correct" responses as winners and incorrect responses as losers in the preference optimization process. We also train a model with Chain-Only data with the same process and we denote it as **Chain-Only-Short†**.

Variant 2: STELAR-Vision-Short‡. A natural alternative approach to encourage shorter and more efficient responses involves explicitly penalizing lengthy outputs. Building upon STELAR-Vision-Short†, we investigate an alternative configuration, STELAR-Vision-Short‡, which maintains the preference for "short and correct" responses while treating both "incorrect responses" and "correct yet lengthy" responses as losers during RL preference training. This design encourages the model to avoid excessive token generation while maintaining accuracy in reaching the correct final answer.

4 Experiments

4.1 Experimental Setup

Datasets We use In-Distribution (ID) datasets from the Method Section 3, splitting them into train/test set. We include 85K training samples from MATH-V and 160K samples from VLM_S2H. OKVQA and A-OKVQA provide 18k and 20K VQA pairs for training, with A-OKVQA adding rationales to training. We also use 17k instruction-following examples from LLaVA-150k for multimodal training. performed on the test sets of MATH-V and VLM_S2H.

For Out-of-Distribution (OOD) evaluation, we use five recent benchmarks with three math oriented datasets, Geometry3K (Lu et al. 2021), We-Math (Qiao et al. 2024) and PolyMath (Gupta et al. 2024), a STEM dataset SciBench (Wang et al. 2024c) and a generic logic reasoning dataset LogicVista (Xiao et al. 2024). In Table 3, we show the evaluation datasets as well as their respective sizes. Collectively, we gather 2,425 evaluation samples.

Models Our base model is Qwen2VL-7B-Instruct (Wang et al. 2024b), chosen for its stable generation of all three topologies in 74% of cases. We exclude the newer Qwen2.5VL-7B-Instruct (Team 2024) for its instability in producing $Tree$ and $Graph$ structure reasoning. We also compare our models with additional open-source and proprietary models.

Evaluation Metrics We report **Accuracy** as the main metric: Accuracy $= \frac{\sum_{q\in Q}\text{Answer}(q)=\text{GT}(q)}{N_Q}$, where GT stands for the ground-truth answer. To assess efficiency under Frugal Learning, we also consider **Generated Token Length** as a metric for reasoning efficiency.

All experiments are conducted on eight NVIDIA A100/H100 GPUs with 80GB of memory each. Supervised fine-tuning (SFT) on a 7B model with 50K–60K samples takes approximately 5–7 hours, while reinforcement learning (RL) on the same scale requires 8–10 hours, depending on system variability.

4.2 Overall Evaluation Results

We use STELAR-Vision to denote models trained with both phases of post-training, the -SFT suffix indicates models trained with supervised fine-tuning only, and -RL-ONLY indicates reinforcement learning directly from the base model without SFT. Results are averaged across 3 random seeds.

Table 2 shows results on the in-distribution datasets MATH-V and VLM_S2H as well as OOD datasets. STELAR-Vision achieves the highest in-distribution accuracy, significantly outperforming its base model Qwen2VL-7B-Instruct by **9.7%**, and surpassing larger model LLaMA-3.2-11B by **10.0%** and Qwen2VL-72B-Instruct by **7.3%**.

Model	In-Distribution Accuracy (%)			Out-of-Distribution Accuracy (%)				
	VLM_S2H	MATH-V	Overall	Geometry3K	We-Math	PolyMath	SciBench	LogicVista
GPT-4o (OpenAI 2024b)	32.0	28.0	30.7	57.0	66.4	25.0	31.1	34.6
LLaVA-v1.6-Mistral-7B (Liu et al. 2024)	26.0	8.0	18.0	20.6	26.0	9.2	3.4	18.5
Llama-3.2-11B-Vision-Instruct (Meta 2024)	22.0	10.0	18.0	35.0	37.8	22.2	10.7	24.8
MiniCPMv2.6-8B (Yao et al. 2024)	1.5	13.0	18.7	45.0	50.2	14.4	8.5	20.7
Phi-4-multimodal-5.6B-instruct (Abouelenin et al. 2025)	23.0	11.0	22.0	8.4	35.8	10.2	10.2	6.7
InternVL3-9B (Zhu et al. 2025)	25.0	21.0	27.3	41.2	51.4	21.6	20.3	32.6
Qwen2VL-72B-Instruct (Yang et al. 2024)	21.0	20.0	20.7	50.2	60.6	13.0	25.4	28.8
Qwen2VL-7B-Instruct (Yang et al. 2024)	21.0	13.0	18.3	35.2	46.6	16.0	10.7	17.0
Chain-Only	25.0	21.0	23.7	31.4	42.2	17.2	10.7	25.4
STELAR-Vision-SFT	28.0	**24.0**	26.7	**44.4**	47.4	24.8	9.0	**33.3**
STELAR-Vision-RL-ONLY	24.0	23.0	23.7	32.8	39.0	**26.0**	**17.5**	23.9
STELAR-Vision	**31.0**	22.0	**28.0**	36.8	**51.0**	23.8	12.4	29.0

Table 2: Quantitative Evaluation. STELAR-Vision achieves strong gains across both in-distribution and out-of-distribution reasoning benchmarks. On ID datasets, it outperforms its base model Qwen2VL-7B-Instruct by **9.7%**, and even surpasses the larger Qwen2VL-72B-Instruct by **7.3%**. On OOD benchmarks, it exceeds Phi-4-multimodal-instruct by up to **36%** and LLaMA-3.2-11B-Vision-Instruct by up to **13.2%**. Compared to Chain-Only training, STELAR-Vision achieves up to **13%** higher accuracy, highlighting the power of topological augmentation.

Dataset	Subject	Question Type	Sample Size
VLM_S2H	Math	multiple-choice	200
MATH-V	Math	free-form, multiple-choice	100
Geometry3K	Math	multiple-choice	500
We-Math	Math	multiple-choice	500
PolyMath	Math	multiple-choice	500
SciBench	STEM	free-form	177
LogicVista	Generic	multiple-choice	448
Total			2425

Table 3: Summary of Evaluation Datasets. VLM_S2H and MATH-V are in-distribution datasets, while others are out-of-distribution

Results on out-of-distribution (OOD) benchmarks highlight the strong generalization of STELAR-Vision. It outperforms its base model Qwen2VL-7B-Instruct on all five OOD datasets and trails GPT-4o by only 1.2% on PolyMath. Compared to recently popular model Phi-4-Multimodal-instruct, STELAR-Vision achieves significantly higher accuracy across all OOD tasks, including a **+28.4%** gain on the spatial reasoning benchmark Geometry3K. It also surpasses other strong open-source models such as LLaVA-v1.6, LLaMA-3.2-11B-Vision, MiniCPMv2.6-8B, and InternVL3-9B on most benchmarks, demonstrating the impact of our TopoAug-enhanced training. STELAR-Vision also outperforms baseline on LogicVista, which includes non-math tasks like 26.4% diagram and 18% spatial.

Interestingly, the `-SFT` and `-RL-ONLY` variants sometimes outperform the full model. We attribute this to: (1) `-SFT` potentially overfitting due to memorization (Chu et al. 2025), and (2) `-RL-ONLY` occasionally lacking alignment. Topological augmentation appears to mitigate both limitations by improving generalization and alignment.

Finally, we find that Chain-Only models, despite moderate in-distribution gains, fall short on OOD tasks, further validating the necessity of topology-aware training. See Section 4.3 for details.

4.3 Ablation Studies

We perform ablation studies comparing our models to counterparts trained solely on chain-based reasoning data. As shown in Table 4, **STELAR-Vision** consistently outperforms Chain-Only variants on in-distribution (ID) datasets, improving accuracy from 23.7% to 28% (+**4.3%**). Table 2 further confirms this trend across benchmarks. On out-of-distribution (OOD) datasets, **STELAR-Vision** achieves up to **8.8%** higher accuracy than Chain-Only models.

These findings collectively demonstrate that **STELAR-Vision** benefits not only from data distillation or pattern memorization but also from a genuine ability to adaptively select optimal reasoning topologies based on the task.

4.4 Efficiency Gains from Frugal Learning

We assess the impact of Frugal Learning on both in-distribution (ID) and out-of-distribution (OOD) performance, as shown in Table 5. The difference between STELAR-Vision-Short† and STELAR-Vision-Short‡ is detailed in Section 3.3.

As shown in Table 5, STELAR-Vision-Short† reduces token length by 101 (ID) and 24.5 (OOD), while still outperforming Qwen2VL-7B-Instruct by 2.5%. However, it shows a 2.9% accuracy drop compared to full STELAR-Vision. In contrast, STELAR-Vision-Short‡ yields inconsistent length reduction and greater performance loss, likely due to conflicting optimization signals that penalize correct outputs.

Chain-Only-Short† also shortens outputs but suffers from even larger accuracy drop. Notably, Chain-Only models fine-tuned with RL often generate overly verbose responses, even with Frugal Learning—highlighting the advantage of TopoAug in promoting concise and effective reasoning.

Model	VLM_S2H	MATH-V	Overall
Qwen2VL-7B-Instruct	21.0	13.0	18.3
Chain-Only-SFT	18.5	19.0	18.7
Chain-Only-RL-ONLY	23.5	15.0	20.7
Chain-Only	25.0	21.0	23.7
STELAR-Vision-SFT	28.0	**24.0**	26.7
STELAR-Vision-RL-ONLY	24.0	23.0	23.7
STELAR-Vision	**31.0**	22.0	**28.0**

Table 4: Impact of TopoAug Dataset and Training Methods. We present an ablation study on the in-distribution VLM_S2H and Math-V datasets to compare the performance of our models against counterparts trained exclusively on chain-based reasoning data across all training methods. STELAR-Vision consistently outperforms all Chain-Only variants across all ID datasets-specifically it improves the highest variant Chain-Only from 25% to 31% by 6%, and boosts overall accuracy by 4.3%, highlighting the effectiveness of topological augmentation.

Model	Accuracy (%)	Gen. Token Length	
		ID	OOD
Qwen2VL-7B-Instruct	26.2	613.5	543.3
Chain-Only	28.7	878.4	742.6
Chain-Only-Short†	23.9	843.1	713.0
STELAR-Vision-SFT	26.7	604.4	483.3
STELAR-Vision	**31.6**	556.7	523.4
STELAR-Vision-Short†	28.7	**455.7**	**498.6**
STELAR-Vision-Short‡	21.9	538.9	555.9

Table 5: Comparison of accuracy and generated token length across models: STELAR-Vision improves performance while using fewer generation tokens. Frugal learning further improves generation efficiency.

Model	Dataset	Tree	Graph	Chain
w/o STELAR-Vision	Overall	-	-	100.00
w/ STELAR-Vision	ID	14.3	9.7	76.0
	We-Math	63.0	7.4	29.6
	Geometry3K	96.4	3.0	0.6
	LogicVista	22.7	15.6	61.7
	PolyMATH	54.0	14.8	31.2
	SciBench	54.2	23.2	22.6

Table 6: Impact of Training on Test-time Topology Selection. Percentage of reasoning topologies autonomously selected by each model on our evaluation datasets, without explicit prompting. ID denotes the in-distribution test split.

4.5 Why Our Method Works?

Our method achieves substantial improvements on both in-distribution and out-of-distribution benchmarks. We hypothesize two key reasons for this success. First, for models that already possess the ability to generate diverse reasoning topologies but lack the capability to select the optimal topology for a given problem (e.g., Qwen2VL-7B-Instruct), our approach enables the model to adaptively choose the most effective structure. Second, for models that might not inherently support diverse topologies, our framework could instill this ability through guided supervision and reinforcement learning, making our methodology broadly applicable. However, due to computational constraints, we will leave systematic study to isolate this effect for future work.

Empirical results support these hypotheses: STELAR-Vision consistently outperforms the Chain-only baseline across both in-distribution and out-of-distribution tasks, as shown in Table 2. Moreover, Table 6 reveals a notable post-training increase in the generation of tree and graph structures, which correlates with accuracy gains—demonstrating the utility of topological augmentation.

On out-of-distribution benchmarks, we observe that datasets such as LogicVista favor simpler, more generic reasoning and hence exhibit a higher frequency of chain-based reasoning. In contrast, datasets requiring more complex reasoning, such as Geometry3K and SciBench, show an increased prevalence of tree and graph topologies. This distribution indicates that the model does not merely memorize patterns but has genuinely learned to select the most suitable topology based on the nature of the problem.

5 Conclusion and Discussion

In this work, we propose STELAR-Vision, a training framework that enables VLM's topology-aware reasoning ability via generated responses. STELAR-Vision enhances vision-language reasoning by leveraging diverse topological structures, achieving a 9.7% accuracy improvement over its base model and outperforming its larger variant Qwen2VL-72B-Instruct by 7.3%. The Frugal Learning reduces output length by 18.1% while maintaining comparable accuracy, surpassing Chain-Only baselines in both efficiency and task effectiveness. STELAR-Vision demonstrates strong generalization capabilities across OOD tasks, as evidenced by five diverse datasets spanning multiple domains. Compared to Chain-Only training, our method achieves 4.3% higher overall accuracy on in-distribution datasets and delivers consistent performance improvements across all OOD evaluations. These results highlight the benefits of topological augmentation and the model's ability to adapt to a variety of reasoning challenges through dynamic reasoning structures.

Despite these promising results, STELAR-Vision relies on predefined topology types, and the dynamic relationship between problem structure and optimal reasoning topology remains underexplored. Future work will focus on enabling scalable, end-to-end topology induction and extending applicability to more advanced multimodal reasoning tasks.

References

Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Aggarwal, P.; and Welleck, S. 2025. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. arXiv:2503.04697.

Arora, D.; and Zanette, A. 2025. Training Language Models to Reason Efficiently. arXiv:2502.04463.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48. ACM.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; and Hoefler, T. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16): 17682–17690.

Chen, L.; Li, L.; Zhao, H.; Song, Y.; Vinci; Kong, L.; Liu, Q.; and Chang, B. 2025. RLVR in Vision Language Models: Findings, Questions and Directions. *Notion Post*.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. arXiv:2501.17161.

Gupta, H.; Verma, S.; Anantheswaran, U.; Scaria, K.; Parmar, M.; Mishra, S.; and Baral, C. 2024. Polymath: A Challenging Multi-modal Mathematical Reasoning Benchmark. *arXiv preprint arXiv:2410.14702*.

Ho, N.; Schmid, L.; and Yun, S.-Y. 2023. Large Language Models Are Reasoning Teachers. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14852–14882. Toronto, Canada: Association for Computational Linguistics.

Hong, J.; Lee, N.; and Thorne, J. 2024. ORPO: Monolithic Preference Optimization without Reference Model. arXiv:2403.07691.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Huang, Z.; Sheng, Z.; Qu, Y.; You, J.; and Chen, S. 2024. VLM-RL: A Unified Vision Language Models and Reinforcement Learning Framework for Safe Autonomous Driving. arXiv:2412.15544.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Lu, P.; Gong, R.; Jiang, S.; Qiu, L.; Huang, S.; Liang, X.; and Zhu, S. 2021. Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning. *CoRR*, abs/2105.04165.

Ma, X.; Jiang, W.; and Huang, H. 2025. Problem-Solving Logic Guided Curriculum In-Context Learning for LLMs Complex Reasoning. arXiv:2502.15401.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. arXiv:1906.00067.

Meng, Y.; Xia, M.; and Chen, D. 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward. arXiv:2405.14734.

Meta. 2024. Llama-3.2-11B-Vision-Instruct.

OpenAI. 2024a. GPT-4o mini: Advancing Cost-Efficient Intelligence. Accessed: March 8, 2025.

OpenAI. 2024b. GPT-4o Technical Report. Accessed: Mar. 8, 2025.

OpenAI. 2024. OpenAI API: o4-mini Model Documentation. https://platform.openai.com/docs/models/o4-mini. Accessed May 2024.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning. arXiv:2502.19634.

Park, S.; Panigrahi, A.; Cheng, Y.; Yu, D.; Goyal, A.; and Arora, S. 2025. Generalizing from SIMPLE to HARD Visual Reasoning: Can We Mitigate Modality Imbalance in VLMs? *arXiv preprint arXiv:2501.02669*.

Qiao, R.; Tan, Q.; Dong, G.; Wu, M.; Sun, C.; Song, X.; Gongque, Z.; Lei, S.; Wei, Z.; Zhang, M.; Qiao, R.; Zhang, Y.; Zong, X.; Xu, Y.; Diao, M.; Bao, Z.; Li, C.; and Zhang, H. 2024. We-Math: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning? *CoRR*, abs/2407.01284.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. arXiv:2206.01718.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Shen, H.; Zhang, Z.; Zhang, Q.; Xu, R.; and Zhao, T. 2025. VLM-R1: A stable and generalizable R1-style Large Vision-Language Model. https://github.com/om-ai-lab/VLM-R1. Accessed: 2025-02-15.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. 2022. Learning to summarize from human feedback. arXiv:2009.01325.

Team, Q. 2024. Qwen2.5-VL-7B-Instruct. Accessed: Mar. 8, 2025.

Thawakar, O.; Dissanayake, D.; More, K.; Thawkar, R.; Heakl, A.; Ahsan, N.; Li, Y.; Zumri, M.; Lahoud, J.; Anwer, R. M.; Cholakkal, H.; Laptev, I.; Shah, M.; Khan, F. S.; and Khan, S. 2025. LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs. arXiv:2501.06186.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Zhan, M.; and Li, H. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2024c. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Proceedings of the Forty-First International Conference on Machine Learning*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Xi, Z.; Chen, W.; Hong, B.; Jin, S.; Zheng, R.; He, W.; Ding, Y.; Liu, S.; Guo, X.; Wang, J.; Guo, H.; Shen, W.; Fan, X.; Zhou, Y.; Dou, S.; Wang, X.; Zhang, X.; Sun, P.; Gui, T.; Zhang, Q.; and Huang, X. 2024. Training Large Language Models for Reasoning through Reverse Curriculum Reinforcement Learning. arXiv:2402.05808.

Xiao, Y.; Sun, E.; Liu, T.; and Wang, W. 2024. LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts. arXiv:2407.04973.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

Yin, Y.; Wang, Z.; Gu, Y.; Huang, H.; Chen, W.; and Zhou, M. 2024. Relative Preference Optimization: Enhancing LLM Alignment through Contrasting Responses across Identical and Diverse Prompts. arXiv:2402.10958.

Zhai, Y.; Bai, H.; Lin, Z.; Pan, J.; Tong, S.; Zhou, Y.; Suhr, A.; Xie, S.; LeCun, Y.; Ma, Y.; and Levine, S. 2024. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. arXiv:2405.10292.

Zhao, Z.; Dong, H.; Saha, A.; Xiong, C.; and Sahoo, D. 2024. Automatic Curriculum Expert Iteration for Reliable LLM Reasoning. arXiv:2410.07627.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; Gao, Z.; Cui, E.; Wang, X.; Cao, Y.; Liu, Y.; Wei, X.; Zhang, H.; Wang, H.; Xu, W.; Li, H.; Wang, J.; Deng, N.; Li, S.; He, Y.; Jiang, T.; Luo, J.; Wang, Y.; He, C.; Shi, B.; Zhang, X.; Shao, W.; He, J.; Xiong, Y.; Qu, W.; Sun, P.; Jiao, P.; Lv, H.; Wu, L.; Zhang, K.; Deng, H.; Ge, J.; Chen, K.; Wang, L.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.