Vision–language models (VLMs) have achieved impressive results across multimodal tasks, yet they continue to struggle with visual reasoning. Studies reveal frequent failures in both perception and reasoning, even on relatively simple tasks [2, 6, 9, 10]. As illustrated in Figure 1, VLMs may propose rules that violate the task constraints: in this case, the rule "contains candle or candles" is incorrectly satisfied by a negative support image. Such errors highlight the gap between pattern recognition and systematic reasoning in VLMs.

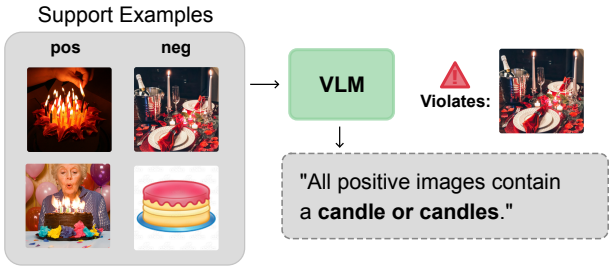

Figure 1: **VLMs cannot perform inductive logic learning faithfully**, failing to capture the logical structure of *"cake **and** candles on cake"*.

Recent work attempts to address this through test-time scaling, where models "think" longer via extended chain-of-thought generation [5]. While effective in some cases, this approach is computationally expensive and prone to contradictions or repetitive loops [4, 7]. An alternative lies in neuro-symbolic AI, which integrates neural perception with symbolic reasoning [3, 8]. Program synthesis [1], for example, is able to induce symbolic rules that are interpretable and logically consistent. However, for visual reasoning, such methods usually depend on domain-specific detectors, limiting their generality [8].

We therefore propose **combining VLMs with program synthesis** to address these limitations. Rather than embedding reasoning inside the VLM, we use it as part of a domain specific language, producing structured visual descriptions that can be combined with symbolic functions. Our approach, Vision Language Programs (VLP), operates directly on images, enforces logical consistency with task constraints, and remains fully interpretable.

Table 1: **VLP vs. direct prompting.** Accuracy (%). First column shows averages. VLP is model-agnostic and boosts performance across backbones, with largest gains on logic-heavy benchmarks.

Model	Avg.	Bongard-OW	Bongard-HOI	COCOLogic	CLEVR-Hans3
InternVL3-8B	60.4	56.3 ±2.3	60.8 ±1.7	66.3 ±1.0	58.3 ±7.6
w/ VLP	**72.8** (+12.4)	63.7 ±2.4 (+7.4)	68.9 ±1.3 (+8.1)	72.5 ±5.1 (+6.2)	86.1 ±6.1 (+27.8)
InternVL3-14B	61.7	63.3 ±1.9	67.1 ±1.4	64.2 ±2.1	52.2 ±6.4
w/ VLP	**69.2** (+7.5)	61.5 ±1.3 (-1.8)	70.6 ±1.7 (+3.5)	71.2 ±1.0 (+7.0)	73.3 ±5.4 (+21.1)
Kimi-VL-A3B	55.4	58.1 ±0.6	57.5 ±2.1	66.7 ±2.6	39.4 ±6.9
w/ VLP	**62.5** (+7.1)	57.7 ±1.9 (-0.4)	57.1 ±2.3 (-0.4)	59.7 ±2.5 (-7.0)	75.6 ±2.1 (+36.2)
Qwen2.5-VL-7B	60.6	61.8 ±1.0	67.6 ±0.4	61.7 ±3.3	51.1 ±2.1
w/ VLP	**67.7** (+7.1)	59.6 ±1.9 (-2.2)	67.2 ±0.6 (-0.4)	67.9 ±0.6 (+6.2)	76.1 ±1.6 (+25.0)

Experiments on synthetic and real-world data show that even small VLMs, when used in this programmatic setting, outperform direct prompting, particularly on tasks requiring complex logical rules (cf. Table 1). The result of our method is a VLP that encodes a visual rule, combining VLM-based and symbolic functions. E.g., for the example in Figure 1 we obtain:

```
(and (exists_object (get_objects IMG) cake)(exists_object (get_actions IMG) candles_on_cake))
```

In conclusion, our hybrid approach leverages VLM priors while enabling systematic reasoning, a crucial step toward models that can both obtain high performance and provide transparent explanations to humans.

References

[1] Sumit Gulwani, Oleksandr Polozov, Rishabh Singh, et al. Program synthesis. *Foundations and Trends® in Programming Languages*, 4(1-2):1–119, 2017.

[2] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv Preprint:2505.23678*, 2025.

[3] Hikaru Shindo, Viktor Pfanschilling, Devendra Singh Dhami, and Kristian Kersting. Learning differentiable logic programs for abstract visual reasoning. *Mach. Learn.*, 2024.

[4] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

[5] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[6] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

[7] Guojun Wu. It's not that simple. an analysis of simple test-time scaling. *arXiv preprint arXiv:2507.14419*, 2025.

[8] Antonia Wüst, Wolfgang Stammer, Quentin Delfosse, Devendra Singh Dhami, and Kristian Kersting. Pix2code: Learning to compose neural visual concepts as programs. In *The 40th Conference on Uncertainty in Artificial Intelligence*, .

[9] Antonia Wüst, Tim Tobiasch, Lukas Helff, Inga Ibs, Wolfgang Stammer, Devendra Singh Dhami, Constantin A Rothkopf, and Kristian Kersting. Bongard in wonderland: Visual puzzles that still make ai go mad? In *Forty-second International Conference on Machine Learning*, .

[10] Yizhe Zhang, He Bai, Ruixiang Zhang, Jiatao Gu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. How far are we from intelligent visual deductive reasoning? *ICLR 2024 Workshop: How Far Are We From AGI*, 2024.