

Beyond Pass@1: Self-Play with Variational Problem Synthesis Sustains RLVR (SvS)

ICLR 2026



Arxiv: <https://arxiv.org/abs/2508.14029>

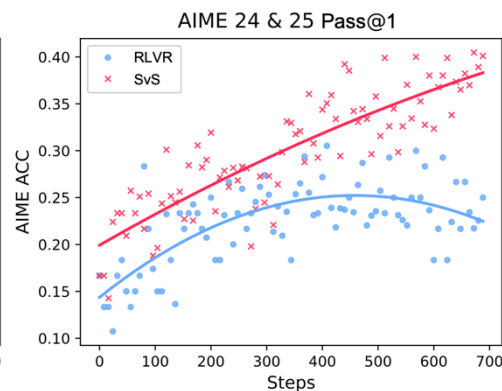
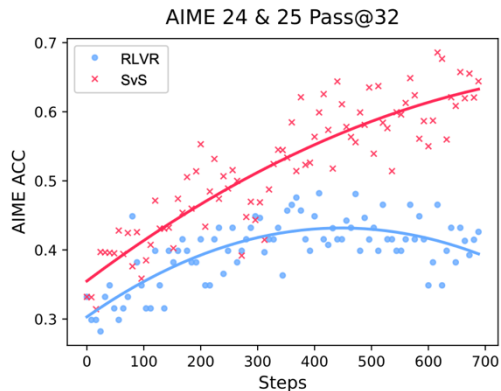


Github: <https://github.com/MasterVito/SvS>

Motivation for SvS

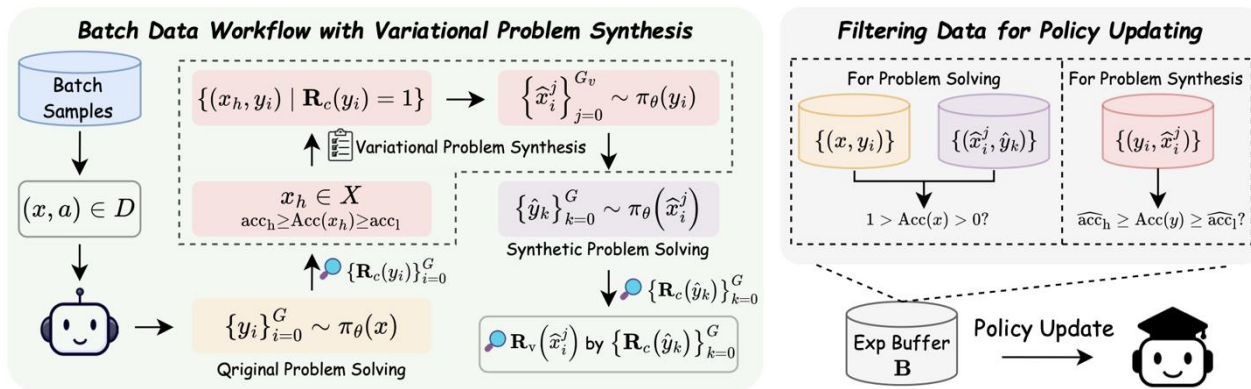
- Standard RLVR fails to improve **Pass@k** due to limited exploration, often caused by **entropy collapse**.
- Preserving **data diversity** is the most straightforward solution.
- SvS: Leveraging **online self-play** data augmentation to enhance model generalizability in RLVR.

Qwen2.5-32B-IT on DAPO-Math-17k:



Overview

- Self-play with Variational problem Synthesis (SvS) strategy for RLVR.
- Variational problems are synthesized from correct solutions to challenging tasks.
- The variational problems are constrained to share the same reference answers as the original ones.
- Each step include: (1) Original Problem Solving, (2) Variational Problem Synthesis (3) Synthetic Problem Solving.



Example of SvS Augmentation

- The challenging problems are identified with **group accuracy** between acc_l (0.125) and acc_h (0.5).
- We filter verified **correct solutions** to generate multiple (n=8) synthetic problems.
- The policy generates a group of solutions to each synthetic problems.

Original Challenging Problem:

Let Γ be the set of functions $f_n : \mathbb{C} \rightarrow \mathbb{C}$ defined by $f_n(z) = \frac{\pi^2}{24}z^2 - \frac{1}{n^2}z + \frac{589}{20} \cdot \frac{1}{n^3}$ for $n \in \mathbb{N}$. Let Λ be the set of all roots with positive imaginary part of the elements of Γ . Find the sum of all elements of Λ , rounding the real and imaginary parts to the nearest integer.

GT Answer: $2 + 22i$



A Correct Solution:

We start by investigating $f_n(z) = \frac{\pi^2}{24} \cdot z^2 - \frac{1}{n^2}z + \frac{589}{20} \times \frac{1}{n^3}$.

This is quadratic polynomial: $f_n(z) = a_n z^2 + b_n z + c_n$.

where $a_n = \frac{\pi^2}{24}$, $b_n = -\frac{1}{n^2}$, $c_n = \frac{589}{20} \cdot \frac{1}{n^3}$.

...

Imaginary part of $S \approx 22.14$, rounds to 22.

Hence its final form, encapsulated required format, is $2 + 22i$.

Original Problem Solving

Synthetic Variational Problems

Consider the polynomial functions defined by: $f_n(z) = \frac{\pi^2}{24} \cdot z^2 - \frac{1}{n^2}z + \frac{589}{20} \times \frac{1}{n^3}$, for each $n \in \mathbb{N}$. For each function, evaluate the root that has a positive imaginary part. Round the real and imaginary parts of each of these roots to the nearest integer first, and then compute the sum of all these rounded roots.

Acc: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0] GT Answer: $1 + 16i$ | $R_V = 0.0$ **X** (Failed synthesis)

Let us examine the function defined by $f_n(z) = \frac{\pi^2}{24} \cdot z^2 - \frac{1}{n^2}z + \frac{589}{20} \cdot \frac{1}{n^3}$ for $n \in \mathbb{N}$. Evaluate all roots of $f_n(z) = 0$. Determine the sum of all those roots whose imaginary parts are positive, expressed in the set Λ , and sum these over all n in the natural numbers. Express your final answer with both real and imaginary parts rounded to the nearest integer. Present the answer in the appropriate rounded form.

Acc: [0.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 1.0] GT Answer: $2 + 22i$ | $R_V = 1.0$ **✓**

...

Let's consider the family of functions $f_n(z) = \frac{\pi^2}{24} \cdot z^2 - \frac{1}{n^2}z + \frac{589}{20} \times \frac{1}{n^3}$, where $n \in \mathbb{N}$. We are tasked with finding the sum of all roots with positive imaginary parts that belong to the set Λ , as $f_n(z) = 0$ for all natural n . Express your answer after rounding the real and imaginary parts to the nearest integers, presenting the sum in a suitable form.

The sum of a series of complex numbers can be found by summing the real and imaginary parts separately.

Acc: [1.0, 1.0, 1.0, 1.0, 0.0, 1.0, 1.0, 1.0] GT Answer: $2 + 22i$ | $R_V = 0.0$ **X** (With hints, oversimple)

Variational Problem Synthesis & Synthetic Problem Solving

indicates policy accuracy on the synthetic problems, verified using the original problem's reference answer.

Reward Shaping for Problem Synthesis

- Correctness: The synthetic problems should **preserve the original reference answer**.
 - In our implementation, we guide the policy to solve the synthetic problems and verify whether it produces the original reference answer as a proxy. **At least one** such solution is required to validate each synthetic problem.
- Helpfulness: The synthetic problems should **remain challenging** for the policy at the current iteration.
- Final reward design for problem synthesis:

$$\mathbf{R}_v(\hat{x}_i^j) = \mathbb{I} \left(\hat{acc}_l \leq \text{Acc}(\hat{x}_i^j, a) \leq \hat{acc}_h \right)$$

1/8, maintain correctness

5/8, maintain helpfulness

Experimental Setup

- Training Models: Qwen2.5-Instruct (3B, 32B), LLaMA-3.1-Instruct-8B.
- Datasets:
 - For all models: MATH-12k.
 - For the 32B models, we additionally run the experiments on DAPO-Math-17k.
- RL Algorithm: **GRPO + Dynamic Sampling + Clip-Higher + Token-level Reward.**
- Benchmarks: GSM8k, MATH-500, Minerva-Math, Olympiad-Bench, Gaokao23, AMC23, AIME24 & 25, Beyond-AIME, Math24o, OlymMath.

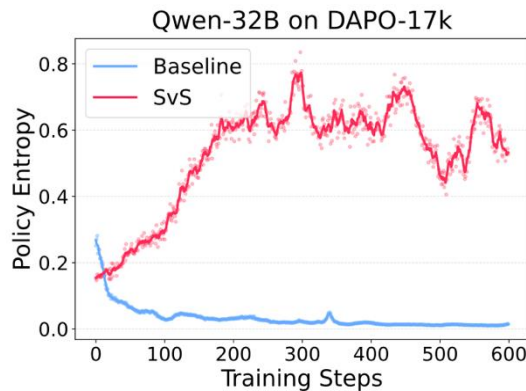
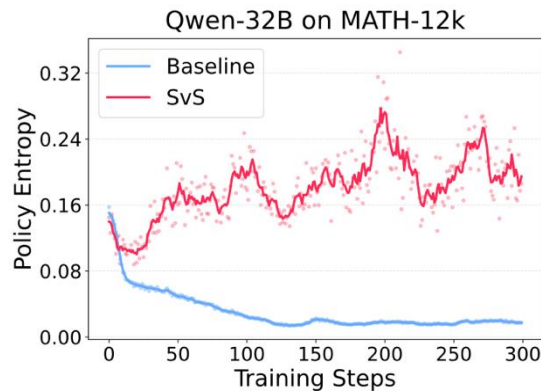
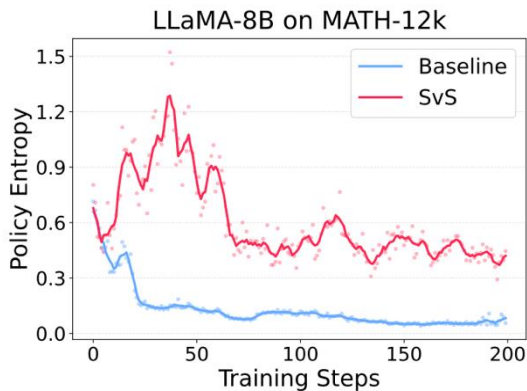
Main Results

- The SvS training demonstrate **much better performance** in comparison with standard RLVR training.
 - Pass@32: AIME 24 +18.3%; AIME 25 + 22.3%.
- Performance drop on OlymMath arises from answer formats that are o.o.d to the training data.

Model	Pass@1							Pass@32						
	AIME24	AIME25	BAIME	Math24o	OlymE	OlymH	Avg.	AIME24	AIME25	BAIME	Math24o	OlymE	OlymH	Avg.
<i>Open-Source Models</i>														
Qwen2.5-32B	4.3	1.2	2.4	8.0	3.7	1.6	3.5	38.9	15.6	18.7	34.0	24.6	15.2	24.5
Qwen2.5-32B-IT	10.0	13.0	7.4	26.0	8.6	2.0	11.2	40.2	34.6	24.0	67.8	35.2	9.5	35.2
SimpleRL-32B	22.1	13.9	8.3	25.5	9.4	3.7	13.8	62.0	38.5	27.4	69.9	42.5	19.4	43.3
ORZ-32B	24.2	26.3	10.9	16.1	12.2	1.1	15.1	55.7	47.0	29.4	58.0	45.9	12.3	41.4
<i>MATH-12k</i>														
→ RLVR	22.2	15.8	11.5	34.5	11.7	4.1	16.6	47.4	36.4	29.2	66.0	36.2	16.4	38.6
→ SvS	30.3	21.7	13.8	42.7	20.1	3.3	22.0	63.6	55.1	41.5	79.2	63.6	24.8	54.6
Δ	+8.1	+5.9	+2.3	+8.2	+8.4	-0.8	+5.4	+16.2	+18.7	+12.3	+13.2	+27.4	+8.4	+16.0
<i>DAPO-17k</i>														
→ RLVR	28.8	30.0	14.0	39.6	17.9	4.8	22.5	52.5	42.4	35.9	71.2	47.1	18.3	44.6
→ SvS	39.3	40.5	19.2	44.1	21.8	2.7	27.9	70.8	65.2	45.9	76.5	43.4	16.7	53.1
Δ	+10.5	+10.5	+5.2	+4.5	+3.9	-2.1	+5.4	+18.3	+22.8	+10.0	+5.3	-3.7	-1.6	+8.5

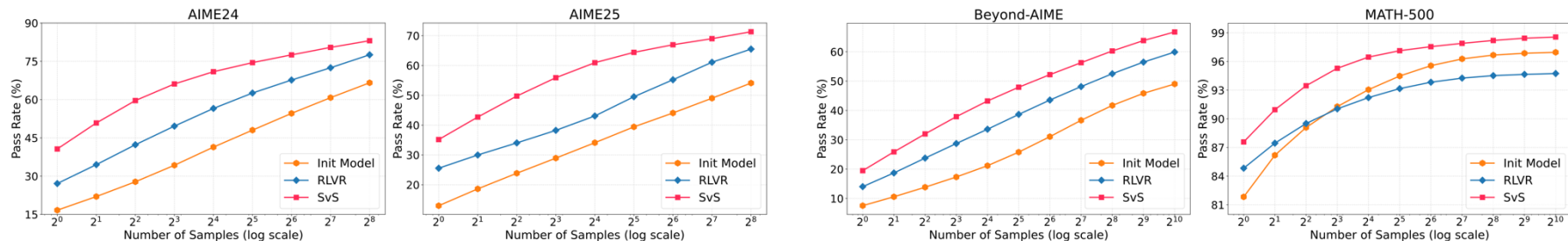
SvS Stably Maintains Policy Entropy in Training

- Training entropy and generation diversity are preserved because the **policy is forced to solve new problems at each RL step.**



SvS Pushes the Reasoning Boundary of the Policy

- SvS training augmentation consistently yields substantially stronger Pass@k performance than standard RLVR and the initial model, even as k increases to 1,024.
- On MATH-500, RLVR fails to outperform the initial policy at Pass@1024, **whereas SvS achieves so.**



SvS generalizes effectively to code generation

- We run Qwen2.5-7B-Instruct on code generation tasks **without tuning any hyperparameters**.
- Correct code programs are used to synthesize the corresponding coding task descriptions.
- Even show more promising results than on the mathematical reasoning tasks.

