SPECS: Decoupling Multimodal Learning via Self-distilled Preference-based Cold Start

Chen Kun

CASIA

September 28, 2025

Slides powered by Claude Code

Outline

The Challenge: Multimodal RL with Cold Start

Current Paradigm:

- MLLM-r1 approaches apply RL to vision language models
- Standard cold start uses SFT on reasoning data
- Joint learning of reasoning + format + content

Problems:

× SFT-based cold start causes overfitting

× Poor out-of-distribution generalization

× Affects downstream RL performance

Generalization Factor (GF) Metric

Problem: How to quantify model generalization under different cold start methods?

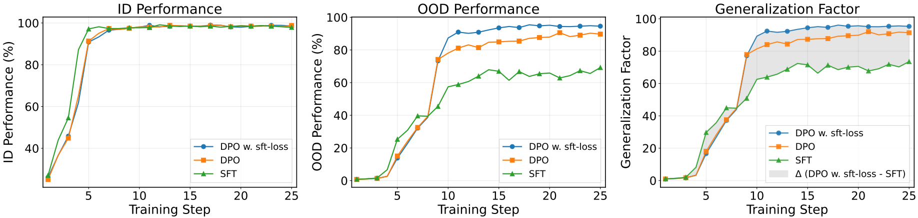
Solution: Generalization Factor (GF)

$$\Gamma_\tau(n) = \frac{(1+\beta^2)\, G^{\tau}_{\mathrm{OOD}}(n)\, G^{\tau}_{\mathrm{ID}}(n)}{\beta^2\, G^{\tau}_{\mathrm{ID}}(n) + G^{\tau}_{\mathrm{OOD}}(n)} \tag{1}$$

Where:

- $G^{\tau}_{\mathrm{ID}}(n)$ = In-distribution performance gain
- $G^{\tau}_{\mathrm{OOD}}(n)$ = Out-of-distribution performance gain
- $\beta = 2$ (emphasizes OOD performance)
- $\tau \in \{pref, sup\}$ (preference vs supervised data)

Empirical Findings: DPO vs SFT



Key Observations:

- SFT: Fast ID convergence, poor OOD performance
- DPO: Slower ID start, better OOD generalization
- DPO+SFT: Best overall generalization (highest GF)

Key Insight: Decoupling Learning Objectives

Our Hypothesis

Cold start should focus on shallow, surface-form learning (format, structure, style) rather than deep reasoning content to avoid premature convergence.

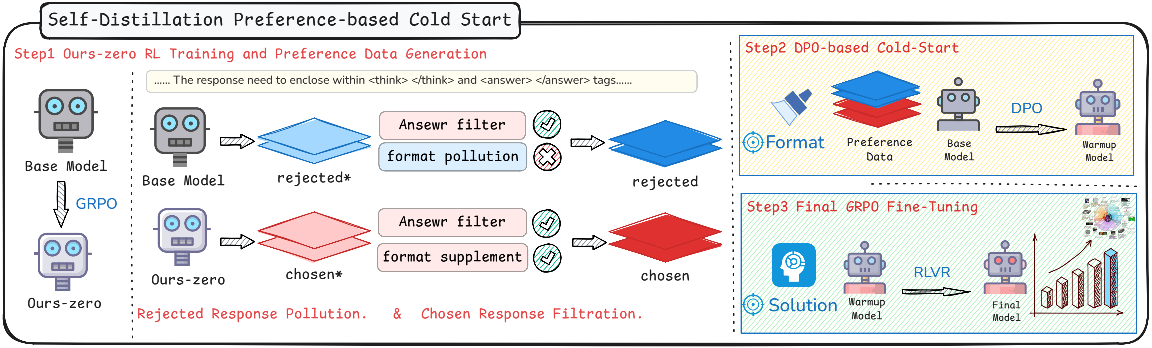
Traditional Approach:

- Joint learning of all objectives
- SFT on reasoning data
- High in-distribution performance
- Poor generalization

SPECS Approach:

- Decoupled learning stages
- Preference-based cold start
- Better generalization
- Superior RL initialization

SPECS: Three-Stage Framework



Three Stages:

1. **Self-Distillation:** Generate preference data focusing on format
2. **DPO Pre-Alignment:** Train on preference data for cold start
3. **Final GRPO:** RL fine-tuning with enhanced initialization

Stage 1: Self-Distillation for Preference Data

Objective: Create preference pairs that differ in format, not correctness

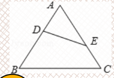
Process:

1. Create exploratory model $\pi_{\text{GRPO-zero}}$ via initial RL
2. Generate responses from both $\pi_{\text{GRPO-zero}}$ and π_{base}
3. **Chosen (y^+):** Correct reasoning + correct format
4. **Rejected (y^-):** Correct answer + corrupted format

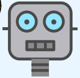
Format Corruptions:

- Remove all tags (<think>, <answer>)
- Remove specific tag pairs
- Replace tags with plain text

Stage 1: Data Example

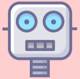


As shown in the figure, it is known that ∠AED = ∠B, DE = 6; AB = 10, AE = 5, the length of BC is ?



Base Model

Since ∠AED = ∠B and ∠A = ∠A, we have △ADE ∼ △ACB. So, $\frac{AE}{AB} = \frac{DE}{BC}$. Given AB = 10, AE = 5, and DE = 6, we can calculate $\frac{5}{10} = \frac{6}{BC}$. Therefore, $BC = \frac{12}{1} . Thus, the answer is 12.



Ours-zero

<think>Given that ∠AED = ∠B and ∠DAE = ∠CAB, we can conclude that △ADE ∼ △ABC (AA similarity criterion). Therefore, the ratios of corresponding sides are equal, which gives us:

\[\frac{DE}{BC} = \frac{AE}{AB} \implies \frac{6}{BC} = \frac{5}{10} \implies \frac{6}{BC} = \frac{1}{2} \implies BC = 12 \]

Thus, the length of BC is 12.</think>
<answer> The answer is $\boxed{12}$</answer>

Key Feature: Both responses have correct answers, but differ in format compliance

Stage 2: DPO-based Pre-Alignment

Objective: Pre-align model on format preferences before RL

Hybrid Loss Function:

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{DPO}} + \lambda \mathcal{L}_{\text{SFT}} \tag{2}$$

Where:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{3}$$

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}\left[\log \pi_\theta(y_w|x)\right] \tag{4}$$

Result: Model learns format compliance without content overfitting

Stage 3: Final GRPO Fine-tuning

Objective: RL training with format-aware initialization

Composite Reward:

$$R_{\text{total}}(o, q) = R_{\text{format}}(o) + R_{\text{acc}}(o, q) \tag{5}$$

- $R_{\text{format}}(o) = 0.5$ for correct format
- $R_{\text{acc}}(o, q) = 1.0$ for correct answer, 0 otherwise

Advantage: Model starts with format mastery, focuses RL on reasoning quality

Experimental Setup

Base Model: Qwen2.5-VL-7B

Training Data:

- Stage 1 & 3: Orsta47K + virl39K datasets
- Stage 2: 9K self-distilled preference pairs

Evaluation Benchmarks:

- MEGA-Bench (general multimodal reasoning)
- MMMU (multimodal understanding)
- MathVista, MATH-Vision, MathVerse (mathematical reasoning)

Baselines: General VLMs and reasoning-specialized models

Main Results: MEGA-Bench Performance

Model	**Science**	**Metrics**	**Planning**	**Perception**	**Coding**	**Core**
QwenVL-2.5-7B	36.75	41.64	16.32	41.24	28.93	35.07
InternVL3-8B	35.21	<u>49.60</u>	17.10	42.76	<u>34.05</u>	36.02
MM-Eureka-7B	<u>37.25</u>	46.39	16.64	39.71	28.75	35.96
VL-Rethinker-7B	36.82	46.90	17.83	42.02	29.87	37.25
Orsta-7B	36.91	41.66	17.83	<u>43.84</u>	32.82	<u>38.31</u>
Ours-7B	**38.73**	**51.87**	<u>18.76</u>	**44.58**	**34.14**	**39.17**
Δ (Improvement)	+2.0	+4.0	+4.6	+3.3	+5.2	**+4.1**

Key Results: Consistent improvements across all categories, **4.1% gain** on MEGA-Bench Core

Mathematical Reasoning Results

Model	MMMU	MathVision	MathVista	MathVerse	Overall
QwenVL-2.5-7B	54.2	25.40	63.70	38.20	45.38
R1-Onevision	49.67	**29.90**	64.1	40.0	45.92
MM-Eureka-7B	55.55	26.90	73.00	47.58	50.76
VL-Rethinker-7B	56.7	29.70	73.60	**48.98**	52.25
Ours-7B	**56.78**	29.50	**75.90**	48.73	**52.73**
Δ (Improvement)	+2.5	+4.1	**+12.2**	+10.5	+7.3

Standout Result: 12.2% improvement on MathVista benchmark

Ablation Studies

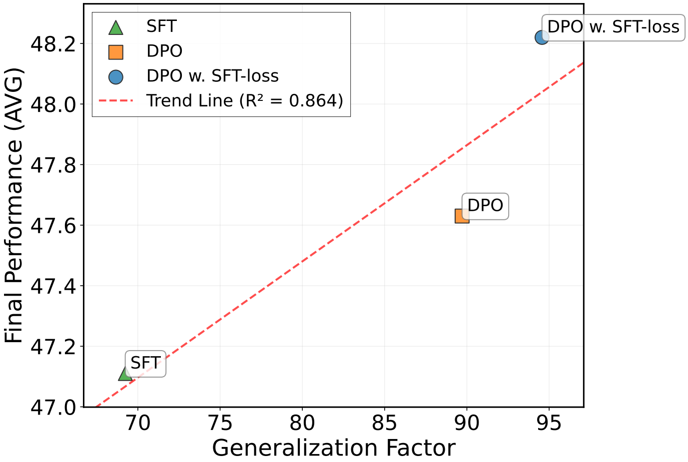
Approach	MEGA	MMMU	MathVista	MathVision	AVG
Qwen32B Distillation	29.87	56.67	71.50	28.03	46.43
Qwen72B Distillation	37.30	**58.56**	73.30	28.91	48.98
Self Distillation	**39.17**	56.78	**75.90**	**29.50**	**50.02**
Coupled Data	38.76	55.44	73.10	28.65	48.68
Decoupled Data	**39.17**	**56.78**	**75.90**	**29.50**	**50.02**

Key Findings:

Self-distillation outperforms external teacher models

Decoupled data strategy superior to coupled approach

GF Correlation with Final Performance



Key Insight: Strong correlation between Generalization Factor during cold start and final RL performance validates our approach

Training Efficiency & Stability

Observations:

- DPO cold start provides higher initial performance
- Faster convergence and higher final ceiling
- More stable training (smoother loss curves)

Hybrid Loss Analysis: Effect of λ Parameter

Combined Loss Function: $\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{DPO}} + \lambda \mathcal{L}_{\text{SFT}}$

Key Observations:

- $\lambda = 0$: Largest margin between chosen/rejected, but both rewards decrease
- $\lambda = 0.5, 1.0$: Ensures chosen rewards increase during training
- Optimal balance maintains preference signal while preventing distribution drift

Exploration Capability Analysis

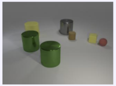
Rollout Branching Factor (RBF): Measures generation diversity during RL

Model	Sample Size 120	Sample Size 240	Sample Size 480
Base Model	1.7686	1.9166	1.9206
SFT Cold Start	1.8086	1.8961	1.9336
SPECS	**1.8216**	**1.9268**	**1.9596**

Key Finding: SPECS maintains higher exploration diversity, enabling better solution discovery during RL

Pass@K Performance: SPECS shows superior performance across Pass@8 and Pass@32 on MMMU, indicating greater initial capabilities for alignment tasks

QA Case Analysis: Case 001



Question: Hint: Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end. Question: Subtract all yellow matte blocks. Subtract all tiny brown cylinders. How many objects are left?

Ground Truth: 5

Qwen2.5-VL-7B response: 6

Ours-7B response

<think>To solve the problem, first identify all objects in the image:
1. Yellow matte cylinder
2. Green metallic cylinder (left)
3. Green metallic cylinder (front)
4. Silver metallic cylinder
5. Tiny brown cylinder
6. Yellow matte block
7. Red matte sphere

Next, follow the instructions:
1. Subtract all yellow matte blocks: Removing the yellow matte block leaves 6 objects.
2. Subtract all tiny brown cylinders: Removing the tiny brown cylinder leaves 5 objects.

Thus, the remaining objects are 5.
</think>
<answer>The answer is $\boxed{5}$</answer>

Qwen2.5-VL-7B response

To solve this, we first identify all objects in the image: there are 8 objects in total. The yellow matte block is one object, and the tiny brown cylinder is another. Subtracting these two objects leaves us with $8 - 1 - 1 = 6$ objects.
So, the number of objects left is 6.

Key Contributions

1. **SPECS Framework:** Novel three-stage cold start strategy that decouples format learning from reasoning
2. **Generalization Factor:** New metric to quantify model generalization under different training methods
3. **Decoupled Learning:** Demonstrates importance of separating shallow vs. deep learning objectives
4. **Strong Results:** Consistent improvements across benchmarks (4.1% MEGA-Bench, 12.2% MathVista)

Why SPECS Works

The Decoupling Principle

By separating format learning (cold start) from reasoning learning (RL), we:

- Avoid premature convergence on in-distribution solutions
- Improve exploration during RL phase
- Achieve better generalization and higher performance ceiling

Practical Benefits:

More stable RL training

Reduced ”stuckness” on format issues

Better out-of-distribution performance

Limitations & Future Work

Current Limitations:

- Experiments focused on multimodal domain
- Limited diversity in OOD benchmarks
- Three-stage pipeline adds complexity

Future Directions:

- Extend to text-only reasoning tasks
- Explore other preference data generation methods
- Investigate optimal balance between stages
- Apply to other model architectures

Questions?

Thank you for your attention!

SPECS: Decoupling Multimodal Learning via Self-distilled Preference-based Cold Start

Questions & Discussion