

Appendix

A Limitations and Discussions

In this work, we studied how to effectively cultivate slow-thinking via Reinforcement Learning. We proposed Selective Sample Replay to mitigate the vanishing advantages in GRPO, and employed forced rethinking to foster deliberate reasoning. While we achieve state-of-the-art results on math-related benchmarks, our models still lag behind human expert performance on more general multimodal benchmarks like EMMA and MEGA-Bench. This reveals that our model is still limited in high-quality training queries. While we show that a direct RL approach without costly distillation can outperform existing RL-based VLMs that involve costly distillations, it remains an open question in what conditions SFT can indeed help the subsequent RL phase for VLMs.

B Training and Implementations

B.1 Training Dataset

Our initial seed query set was constructed by aggregating publicly available multimodal datasets [Yang et al., 2025, Meng et al., 2025, Kembhavi et al., 2016, Saikh et al., 2022, Du et al., 2025] with novel queries gathered from the web. This aggregated dataset exhibits a broad topical diversity, as visually represented in Fig. 8. Given our reliance on rule-based reward mechanisms for subsequent Reinforcement Learning (RL) training, a crucial first step involved filtering the seed queries. We retained only those queries with reference answers that were programmatically verifiable by our defined rules. From this verifiable subset, an augmented query set was systematically generated through the rephrasing of questions and permutation of multi-choice options. This augmentation strategy was designed to facilitate knowledge re-occurrence and reinforce learning across variations of the same core information. This rigorous data preparation pipeline culminated in a final training set comprising 38,870 queries.

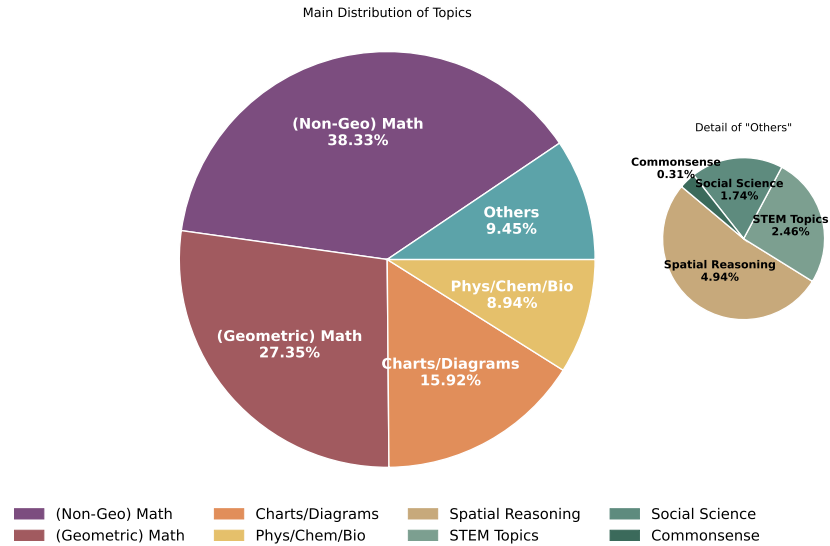


Figure 8: Our training data contains a diverse collection of topics, including eight major categories.

Utilizing this comprehensive query set, we proceeded to train models at different scales. To ensure efficient training and leverage each model’s inherent strengths, we selected subsets of queries tailored to their initial capabilities. Specifically, for each model scale, we curated a training subset consisting of queries where the initial checkpoint of that model demonstrated a non-zero PassRate@8. This selection criterion ensured that the models were trained on queries falling within their potential

Algorithm 1 Selective Sample Replay (SSR)

```
1: Input: Buffer  $\mathcal{B}_{\text{replay}}$ , raw training batch  $\mathcal{D}_{\text{raw}} = \{(x_i, y_i, \hat{A}_i)\}$ , intensity  $\alpha \geq 0$ .
2: Output: Training batch  $\mathcal{D}_{\text{train}}$ , updated buffer  $\mathcal{B}_{\text{replay}}$ 
3: Let  $N_{\text{batch}} = |\mathcal{D}_{\text{raw}}|$ 
4: Initialize list for effective current samples  $\mathcal{D}_{\text{effective}} \leftarrow \emptyset$ 
5: for each sample  $(x_i, y_i, \hat{A}_i)$  in  $\mathcal{D}_{\text{raw}}$  do
6:   Add  $(x_i, y_i, \hat{A}_i)$  to  $\mathcal{D}_{\text{effective}}$  when  $|\hat{A}_i| > 0$ 
7: end for
8: Update buffer:  $\mathcal{B}_{\text{replay}} \leftarrow \mathcal{B}_{\text{replay}} \cup \mathcal{D}_{\text{effective}}$ 
9: Let  $n_{\text{effective}} = |\mathcal{D}_{\text{effective}}|$ 
10: Calculate number of samples needed from buffer:  $n_{\text{from\_buffer}} = \max(0, N_{\text{batch}} - n_{\text{effective}})$ 
11: Initialize list for samples from buffer  $\mathcal{D}_{\text{from\_buffer}} \leftarrow \emptyset$ 
12: if  $n_{\text{from\_buffer}} > 0$  then
13:   Calculate sampling probabilities  $P(\text{select } j)$  for all  $j \in \mathcal{B}_{\text{replay}}$  according to Eq. 1
14:   Form  $\mathcal{D}_{\text{from\_buffer}}$  by drawing  $n_{\text{from\_buffer}}$  samples from  $\mathcal{B}_{\text{replay}}$ 
15: end if
16:  $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{effective}} \cup \mathcal{D}_{\text{from\_buffer}}$ 
```

500 competence range, allowing the RL process to refine and enhance existing, albeit nascent, abilities
501 rather than attempting to instill knowledge from scratch.

502 B.2 Algorithms

503 We provide a diagram for Selective Sample Replay in Alg. 1

504 B.3 Implementations and Training Details

505 Our VL-Rethinker-72B was trained using OpenRLHF for a maximum of 3 epochs on 8 sets of
506 $8 \times \text{A800}(80\text{G})$ for approximately 60 hours. The final checkpoint was selected based on the mean
507 reward achieved on a held-out validation set. We employed a near on-policy RL paradigm, where
508 the behavior policy was synchronized with the improvement policy after every 1024 queries, which
509 we define as an episode. The replay buffer for SSR persisted for the duration of each episode before
510 being cleared. For each query, we sampled 8 responses. The training batch size was set to 512
511 query-response pairs. We accept at most two correct rethinking trajectories for each query. We set the
512 priority hyperparameter in SSR to $\alpha = 1.0$ in the experiments. We released code, models and our
513 high-quality 39K dataset to support further research.

514 B.4 Prompts Used for Training

Default Instruction Prompt

```
{question}
Please reason step by step, and put your final answer within \boxed{}
```

515
516 During the first stage RL training with SSR, we use the default instruction prompt as above.

Rethinking Instruction Prompt

```
{question}
Guidelines:

Please think step by step, and **regularly perform self-questioning, self-
verification, self-correction to check your ongoing reasoning**, using
connectives such as "Wait a moment", "Wait, does it seem right?", etc. Remember
to put your final answer within \boxed{}
```

517

During the Forced Rethinking training stage, we use the above prompt to encourage self-reflection, and use three types of rethinking textual triggers.

Rethinking Triggers

```
self_questioning = "\n\nWait, does it seem right?"
self_correction = "\n\nWait, there might be a mistake"
self_verification = "\n\nWait, let's double check"
```

C Additional Experiments and Analysis

C.1 Experiments

We conducted an ablation on the data compositions. Our training queries are comprised of three major genres: math-related vision-language queries, science-related queries and text-only ones. We conducted ablation studies on these components. As shown in Table. 3, removing text-only queries does not cause significant differences. As we further remove queries from the broader scientific domains, we observe a more pronounced drop in performance. This significant reduction underscores the importance of scientific data in improving the model’s general reasoning ability.

Model	RL-Algo	Data	MathVision	MathVista	MathVerse	MMMU-Pro	EMMA
VL-Rethinker-7B	SSR	16K	32.3	74.9	54.2	41.7	29.7
w/o ‘Forced-Rethinking’	SSR	16K	29.8	72.4	53.2	40.9	29.5
- no SSR	Filter	16K	28.5	72.0	50.0	40.0	26.9
- no SSR& Filter	GRPO	16K	26.0	70.9	51.4	38.8	26.2
- no Text	SSR	13K	29.1	73.5	53.5	41.1	28.7
- no Science&Text	SSR	11K	28.0	71.6	50.3	39.7	28.0

Table 3: Ablation Results to show the impact of SSR and Data Mix.

We performed a sensitivity analysis to the prioritization intensity hyperparameter in SSR. We plot the final averaged eval reward of models trained with different hyperparameters. As shown in Fig. 9, the performance does not vary significantly regarding different hyperparameters, and the best value is around 1.0. We observe that models trained with different hyperparameters will always reach similar peaks sooner or later.

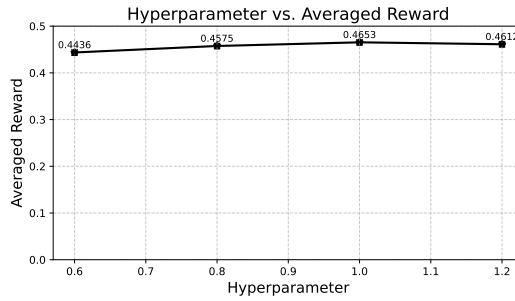


Figure 9: Sensitivity Analysis of the Hyperparameter in SSR

C.2 Cases

Below we show a few cases where VL-Rethinker shows the ability of rethinking. This ability can help achieve correct answers more often than not, but sometimes it still fails with challenging questions.

Listing 1: successful case1

538 Q: Hint: Please answer the question requiring an integer answer and
539 ↪ provide the final value, e.g., 1, 2, 3.
540 What is the highest amount this class measures?
541 GT: $\boxed{600}$
542 A: To determine the highest amount this beaker measures, let's analyze
543 ↪ the image step by step:
544
545 1. The beaker is labeled " 600ml " at the bottom, which indicates its
546 ↪ total capacity.
547
548 2. The markings on the side of the beaker show volume levels: 100, 200,
549 ↪ 300, and 400 ml.
550
551 3. The highest marking visible on the beaker is 400 ml.
552
553 Wait a moment, does the highest marking represent the total capacity?
554 ↪ No, it doesn't. The beaker is labeled " 600ml " at the bottom,
555 ↪ which means it can hold up to 600 ml of liquid.
556
557 Wait, does it seem right? Yes, it does. The beaker is designed to hold
558 ↪ 600 ml, even though the highest marking is 400 ml. The remaining
559 ↪ capacity above 400 ml is not marked but is still part of the
560 ↪ beaker's total capacity.
561
562 Therefore, the highest amount this beaker measures is 600 ml.
563 $\boxed{600}$
564
565

Listing 2: successful case2

566 Q: In the triangle with hypotenuse length 20, find
567 ↪ y. \nChoices:\nA:4\nB:6\nC:8\nD:10
568 GT: \boxed{D}
569 A: To find the length of y in the given right triangle with a
570 ↪ hypotenuse of 20 and an angle of 60° , we can use
571 ↪ trigonometric relationships.
572 1. **Identify the trigonometric relationship:**
573 - In a right triangle, the sine of an angle is the ratio of the length
574 ↪ of the opposite side to the hypotenuse.
575 - Here, $\sin(60^\circ) = \frac{y}{20}$.
576
577 2. **Recall the value of $\sin(60^\circ)$:**
578 - $\sin(60^\circ) = \frac{\sqrt{3}}{2}$.
579
580 3. **Set up the equation:**
581 $\sin(60^\circ) = \frac{y}{20} \implies \frac{\sqrt{3}}{2} = \frac{y}{20}$
582 ↪ $y = 20 \cdot \frac{\sqrt{3}}{2} = 10\sqrt{3}$
583
584 **Wait a moment**, let's double check the problem statement and the
585 ↪ choices. **The problem asks for y and the choices are 4,**
586 ↪ **6, 8, 10. It seems there might be a mistake in the problem setup**
587 ↪ **or the choices provided. However, based on the trigonometric**
588 ↪ **calculation, the correct value of y is $10\sqrt{3}$**
589 ↪ **which is approximately 17.32, not matching any of the given**
590 ↪ **choices.**
591
592 Given the choices, the closest integer value to $10\sqrt{3}$ is
593 ↪ 10, but it is not exactly 10.
594
595 **Wait, does it seem right?** The problem might have a different context
596 ↪ or the choices might be incorrect.
597
598 Since the problem asks for the closest value and 10 is the closest to
599 ↪ $10\sqrt{3}$, the answer is: \boxed{D}
600
601

Listing 3: failure case

```

603
604
605 Q: In the figure there are nine regions inside the circles. The numbers
606   ↪ 1 to 9 should be written in the regions so that the sum of the
607   ↪ numbers in each circle is exactly 11. Which number has to go in
608   ↪ the region with the question mark?<image1>
609
610 GT: \boxed{6}
611 A: To solve this problem, we need to place the numbers 1 to 9 in the
612   ↪ nine regions such that the sum of the numbers in each circle is
613   ↪ exactly 11. Let's denote the regions as follows:
614
615 - \(\ A \) is the region in the top left circle.
616
617 - \(\ B \) is the region in the top middle circle.
618
619 - \(\ C \) is the region in the top right circle.
620
621 - \(\ D \) is the region in the bottom left circle.
622
623 - \(\ E \) is the region in the bottom middle circle.
624
625 - \(\ F \) is the region in the bottom right circle.
626
627 - \(\ G \) is the region in the middle left circle.
628
629 - \(\ H \) is the region in the middle right circle.
630
631 - \(\ I \) is the region in the middle circle (the one with the
632   ↪ question mark).
633
634
635 The circles and their regions can be described as follows:
636 - The top left circle contains \(\ A, G, D \).
637 - The top middle circle contains \(\ G, B, H \).
638 - The top right circle contains \(\ H, C, F \).
639 - The bottom left circle contains \(\ D, E, I \).
640 - The bottom right circle contains \(\ I, F, E \).
641
642 We need to satisfy the following equations:
643 1. \(\ A + G + D = 11 \)
644 2. \(\ G + B + H = 11 \)
645 3. \(\ H + C + F = 11 \)
646 4. \(\ D + E + I = 11 \)
647 5. \(\ I + F + E = 11 \)
648
649 Let's start by adding all five equations:
650 \(\ (A + G + D) + (G + B + H) + (H + C + F) + (D + E + I) + (I + F + E) \)
651   ↪ = 55 \)
652 This simplifies to:
653 \(\ A + B + C + 2D + 2E + 2F + 2G + 2H + 2I = 55 \)
654 Since \(\ A + B + C + D + E + F + G + H + I = 45 \) (the sum of the
655   ↪ numbers 1 to 9), we can substitute \(\ 45 + D + E + F + G + H + I \)
656   ↪ = 55 \), which simplifies to:
657 \(\ D + E + F + G + H + I = 10 \)
658
659 Now, let's try to find a solution by trial and error, keeping in mind
660   ↪ that each circle must sum to 11.
661
662 Assume \(\ I = 1 \):

```

663 Then $(D + E + 1 = 11)$ and $(1 + F + E = 11)$, so $(D + E =$
 664 $\rightarrow 10)$ and $(F + E = 10)$. This means $(D = F)$, which is
 665 \rightarrow not possible since all numbers must be distinct.
 666
 667 Assume $(I = 2)$:
 668 Then $(D + E + 2 = 11)$ and $(2 + F + E = 11)$, so $(D + E = 9$
 669 $\rightarrow)$ and $(F + E = 9)$. This means $(D = F)$, which is not
 670 \rightarrow possible.
 671
 672 Assume $(I = 3)$:
 673
 674 Then $(D + E + 3 = 11)$ and $(3 + F + E = 11)$, so $(D + E = 8$
 675 $\rightarrow)$ and $(F + E = 8)$. This means $(D = F)$, which is not
 676 \rightarrow possible.
 677
 678 Assume $(I = 4)$:
 679 Then $(D + E + 4 = 11)$ and $(4 + F + E = 11)$, so $(D + E = 7$
 680 $\rightarrow)$ and $(F + E = 7)$. This means $(D = F)$, which is not
 681 \rightarrow possible.
 682
 683 Assume $(I = 5)$:
 684 Then $(D + E + 5 = 11)$ and $(5 + F + E = 11)$, so $(D + E = 6$
 685 $\rightarrow)$ and $(F + E = 6)$. This means $(D = F)$, which is not
 686 \rightarrow possible.
 687
 688 Assume $(I = 6)$:
 689 Then $(D + E + 6 = 11)$ and $(6 + F + E = 11)$, so $(D + E = 5$
 690 $\rightarrow)$ and $(F + E = 5)$. This means $(D = F)$, which is not
 691 \rightarrow possible.
 692
 693 Assume $(I = 7)$:
 694 Then $(D + E + 7 = 11)$ and $(7 + F + E = 11)$, so $(D + E = 4$
 695 $\rightarrow)$ and $(F + E = 4)$. This means $(D = F)$, which is not
 696 \rightarrow possible.
 697
 698 Assume $(I = 8)$:
 699 Then $(D + E + 8 = 11)$ and $(8 + F + E = 11)$, so $(D + E = 3$
 700 $\rightarrow)$ and $(F + E = 3)$. This means $(D = F)$, which is not
 701 \rightarrow possible.
 702
 703 Assume $(I = 9)$:
 704 Then $(D + E + 9 = 11)$ and $(9 + F + E = 11)$, so $(D + E = 2$
 705 $\rightarrow)$ and $(F + E = 2)$. This means $(D = F)$, which is not
 706 \rightarrow possible.
 707
 708 After trying all possibilities, we find that the only number that works
 709 \rightarrow is $(I = 1)$.
 710
 711 ****Wait, does it seem right? Let's verify:****
 712 If $(I = 1)$, then $(D + E = 10)$ and $(F + E = 10)$. Let's
 713 \rightarrow try $(D = 6)$ and $(E = 4)$, and $(F = 6)$ and $(E =$
 714 $\rightarrow 4)$. This is not possible. Let's try $(D = 5)$ and $(E = 5$
 715 $\rightarrow)$, and $(F = 5)$ and $(E = 5)$. This is not possible.
 716 \rightarrow Let's try $(D = 3)$ and $(E = 7)$, and $(F = 3)$ and
 717 $\rightarrow (E = 7)$. This is not possible. Let's try $(D = 2)$ and
 718 $\rightarrow (E = 8)$, and $(F = 2)$ and $(E = 8)$. This is not
 719 \rightarrow possible. Let's try $(D = 1)$ and $(E = 9)$, and $(F = 1$
 720 $\rightarrow)$ and $(E = 9)$. This is not possible.
 721
 722 After all these trials, we find that the only number that works is $(I$
 723 $\rightarrow = 1)$.
 724
 725 Therefore, the number that has to go in the region with the question
 726 \rightarrow mark is $(\boxed{1})$.
 727