DreamPRM-1.5: Unlocking the Potential of Each Instance for Multimodal Process Reward Model Training

Qi Cao **Pengtao Xie**
University of California, San Diego
{q9cao,p1xie}@ucsd.edu

Project Page: https://github.com/coder-qicao/DreamPRM-1.5

Abstract

Training multimodal process reward models (PRMs) is challenged by distribution shifts and noisy data. We introduce DreamPRM-1.5, an instance-reweighted framework that adaptively adjusts the importance of each training example via bi-level optimization. We design two complementary strategies: Instance Table, effective for smaller datasets, and Instance Net, scalable to larger ones. Integrated into test-time scaling, DreamPRM-1.5 achieves 84.6 accuracy on the MMMU benchmark, surpassing GPT-5.

1 Introduction

Recent advances in reasoning [20] have substantially boosted the performance of large language models (LLMs)[1, 4, 23, 17], with Process Reward Models (PRMs)[10, 8] enabling step-level supervision and more reliable selection of reasoning trajectories. Extending PRMs to multimodal LLMs (MLLMs)[27, 9] is thus a natural progression. However, multimodal inputs couple high-dimensional visual features with discrete language tokens, enlarging the input space and intensifying *distribution shifts*[21]. At the same time, multimodal reasoning data face severe *quality imbalance*[28, 13], where noisy or trivial samples dilute the benefits of effective training. Consequently, directly applying text-only PRM methods [26, 14] yields limited gains due to poor generalization [5].

To address this problem, DreamPRM [2] introduced a domain-reweighted framework, where Process Reward Models were fine-tuned with dataset-level weights to emphasize high-quality domains while suppressing noisy ones. At the meta-level, these weights were updated through validation-driven aggregation losses [19, 6, 22, 7, 11], enabling more robust and generalizable multimodal reasoning.

Building on this foundation, we propose DreamPRM-1.5, which extends the idea of reweighting from the domain level to the individual instance level. Instead of treating each dataset uniformly, DreamPRM-1.5 assigns adaptive weights to every training example, thereby amplifying the impact of informative samples while down-weighting noisy or trivial ones. This finer-grained reweighting significantly enhances the effectiveness of PRM training, unlocking the full potential of each data instance and significantly improve performance.

Our contributions are summarized as follows:

- We propose DreamPRM-1.5, an *instance-reweighted* multimodal process reward model training framework that dynamically adjusts the weight of each individual data example. To realize instance-level reweighting, we further design two complementary training paradigms: Instance Table, which maintains more activated parameters during training and proves effective for smaller datasets; and Instance Net, which employs a lightweight parameterization with stronger generalization ability, making it more suitable for large-scale training sets.

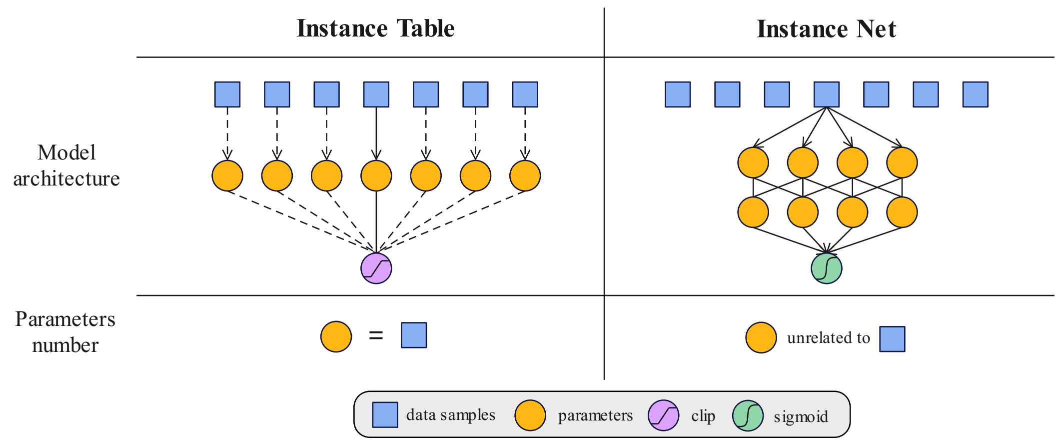

Figure 1: Comparison of two model designs for instance reweighting in DreamPRM-1.5. Instance Table assigns an explicit learnable weight to each training sample, offering strong per-instance flexibility but scaling with dataset size. Instance Net parameterizes instance weights via a lightweight MLP appended to the PRM, maintaining a fixed parameter size independent of dataset scale and providing better generalization.

- By integrating DreamPRM-1.5 into test-time scaling, we achieve a new state-of-the-art accuracy of 84.6 on the validation set of MMMU benchmark, further advancing the performance frontier of the strongest existing model, GPT-5-mini.

2 The Proposed Instance-reweighting Method

Training multimodal PRMs is difficult due to (1) data quality imbalance and (2) mismatch between training and inference. We propose DreamPRM-1.5, which learns instance weights via a bi-level framework adapted from DreamPRM [2] . The lower level updates PRM parameters with instance-reweighted training, while the upper level optimizes instance weights on a meta dataset.

2.1 Bi-level Optimization for Instance-reweighhting

Lower-level optimization. Given a PRM $\mathcal{V}$ parameterized by ϕ and instance weights parameterized by α, the training loss on training sample x is

$$\mathcal{L}_{tr}(\phi, \alpha, x) = \alpha \sum_{i=1}^{n} \mathcal{L}_{CrossEntropy}(\mathcal{V}_\phi(x, \hat{y}_i), p_i), \tag{1}$$

where p_i is the step-wise supervision for prefix $\hat{y}_i$. The overall objective across N instances is a weighted sum, yielding $\phi^*(\alpha) = \arg\min_\phi \mathcal{L}_{tr}(\phi, \alpha, x)$, where $x \in \mathcal{D}_{tr}$ is from training set $\mathcal{D}_{tr}$. Only ϕ is optimized here, with α fixed.

Upper-level optimization. We then optimize α on meta learning dataset $\mathcal{D}_{meta}$ with a meta loss that mimics PRM inference. For each generated solution $\hat{y}$, we compute an aggregated score $\mathcal{A}(\mathcal{V}_{\phi^*(\alpha)}(x, \hat{y}))$ and compare it with ground truth $r(\hat{y}, y) \in \{0, 1\}$:

$$\mathcal{L}_{meta}(\mathcal{D}_{meta}, \phi^*(\alpha)) = \sum_{(x,y)\in\mathcal{D}_{meta}} \mathcal{L}_{MSE}(\sigma(\mathcal{A}(\mathcal{V}_{\phi^*(\alpha)}(x, \hat{y}))), r(\hat{y}, y)). \tag{2}$$

This gradient-based update refines instance weights α, enabling DreamPRM-1.5 to adaptively emphasize needed data examples.

2.2 Model Design for Instance-reweighting

Instance table. A straightforward way to assign weights at the instance level is to maintain a lookup table, where each training sample x is associated with a learnable weight α_x. In this formulation, the number of parameters is equal to the number of training samples. The key advantage of this approach is its ability to fully exploit the potential of each individual example, often yielding strong results even on relatively small datasets (see experiments). To prevent extreme values, we apply a clipping function that constrains all weights within a fixed range; any value outside this interval is automatically projected to its boundary.

Instance net. An alternative strategy is to parameterize the instance weight via a lightweight network. Specifically, we append a small MLP after the final layer of the PRM, which dynamically predicts a weight for each input based on its representation. Unlike Instance Table, this approach has a fixed number of parameters regardless of the dataset size (typically far fewer than the number of samples), making it both scalable and generalizable. To ensure stability, we apply a sigmoid activation at the MLP output, keeping the predicted weights within a normalized range.

3 Experiments

In this section, we describe the implementation details of DreamPRM-1.5 and present the main experimental results.

3.1 Implementation Details

Model. We adopt InternVL3-1B [30] as the base model for training the PRM. This state-of-the-art, small-scale multimodal model is pretrained on general vision–language understanding tasks, and we fine-tune it to obtain the final checkpoint. For inference, we use GPT-5-mini [16] as the underlying MLLM. GPT-5-mini is a lightweight variant of the state-of-the-art reasoning model GPT-5, offering a favorable balance between cost efficiency and competitive performance.

Generative reward model. We employ a generative reward model to assign scores to individual reasoning steps. Specifically, we adapt the system prompt from VisualPRM [24] (See Appendix A), which instructs the model to output either + or – for each step in the response. The score is then computed as the softmax probability of the + token. A higher probability indicates greater model confidence in the correctness of the step, and thus corresponds to a higher reward score.

Training and meta datasets. For training, we construct two datasets. Specifically, we sample 12k examples from VisualPRM-400k [24] to train the Instance Table variant of instance reweighting, and 100k examples from the same source to train the Instance Net variant. We also conduct a rule-based check to ensure there is no overlap between training set and test set.

For the meta set, we adopt MMMU-Pro [29] (standard 4-option split), while excluding its validation split to avoid overlap. We further use GPT-5-mini to generate four candidate responses for each question, forming the meta-evaluation set used for weight updates. There are about 1.2k data points in meta set.

To maintain balance between positive and negative supervision, we filter both the training and meta datasets to ensure an approximately equal number of positive and negative samples.

Cold-start initialization. Prior to bi-level optimization, we perform a cold-start fine-tuning stage. Specifically, we sample 20k examples from VisualPRM and conduct one epoch of supervised fine-tuning (SFT). The resulting checkpoint is then used as the initialization for bi-level optimization. This step ensures that the base model learns to follow the system prompt and reliably generate + and – tokens, which are essential for subsequent optimization.

Multi-turn fine-tuning. We cast process supervision as a multi-turn dialogue task to better exploit the generative capabilities of MLLMs. Given a multimodal input question x, the first turn includes the question and its initial reasoning step $\hat{y}_1$, while each subsequent turn introduces the next step in the reasoning trajectory. This formulation allows the model to incrementally process and evaluate reasoning steps in a conversational manner.

Aggregation function loss. Following DreamPRM [2], we adopt an aggregation loss for the meta-learning of the generative reward model. Specifically, we apply a mean aggregation function to

Table 1: MMMU accuracy (%) on leading models and our DreamPRM-1.5 variants. Gray numbers indicate absolute gains over the base GPT-5-mini w/ thinking (80.0).

Category	**Model / Method**	**Accuracy**
Leaderboard (external, top-performing models)		
	GPT-5 w/ thinking [16]	84.2
	Gemini 2.5 Pro Deep-Think [18]	84.0
	o3 [15]	82.9
Test-time Scaling (built on GPT-5-mini w/ thinking)		
	Base: GPT-5-mini w/ thinking	80.0
	VanillaPRM — No Selection	79.1 (-0.9)
	Self-consistency [25]	81.4 (+1.4)
	VisualPRM [24]	80.5 (+0.5)
	DreamPRM-1.5 — Instance Table	**84.6 (+4.6)**
	DreamPRM-1.5 — Instance Net	83.6 (+3.6)

average the step-level scores, and optimize the model using the mean squared error (MSE) between the aggregated score and the ground-truth binary label. To encourage the model to generate both + and -, we compute the score from the logit of + when the ground-truth label is positive, and from the logit of - when the ground-truth label is negative.

3.2 Main Results

Benchmark evaluation. We evaluate the performance of our methods on MMMU validation set [29]. MMMU is a recently introduced benchmark designed to evaluate multimodal models on large-scale, multi-disciplinary tasks that require college-level subject knowledge and deliberate reasoning. It contains carefully curated multimodal questions sourced from college exams, quizzes, and textbooks, spanning six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. In total, the benchmark covers 30 subjects across 183 subfields, with questions paired with 30 diverse image types, including charts, diagrams, maps, tables, music scores, and chemical structures.

Table 1 summarizes the results on the MMMU benchmark. We first report the leading proprietary models on the leaderboard, including GPT-5 with thinking, Gemini 2.5 Pro Deep-Think, and o3. Using GPT-5-mini with thinking as our base model (80.0 accuracy), DreamPRM-1.5 significantly improves performance through instance-level reweighting. Both Instance Table and Instance Net variants achieve substantial gains of +4.6 and +3.6, respectively, demonstrating the effectiveness of our approach.

Baseline comparasion. We compare DreamPRM-1.5 against several representative baselines built on GPT-5-mini with thinking. The No Selection baseline uses the same subset of data as DreamPRM-1.5 but without bi-level optimization, thus directly reflecting the importance of instance-reweighting. VisualPRM [24] trains a PRM on the same data domains but with the full 400k dataset, which is substantially larger than ours, providing a strong data-scale baseline. Self-consistency [25] is the classical test-time scaling method, widely regarded as a robust baseline for reasoning tasks. As shown in Table 1, all these baselines underperform compared to DreamPRM-1.5. This confirms that data quality imbalance indeed harms PRM training, and further highlights the effectiveness of our instance-reweighted approach, which consistently achieves superior accuracy even with fewer training examples.

4 Conclusion

In this paper, we presented DreamPRM-1.5, an instance-reweighted multimodal PRM framework that extends domain-level reweighting in DreamPRM to the level of individual training examples. Through bi-level optimization, DreamPRM-1.5 dynamically learns instance weights that emphasize informative samples while down-weighting noisy or trivial ones. We further explored two complementary implementations—Instance Table and Instance Net—that trade off per-sample expressiveness and scalability. Extensive experiments on the MMMU benchmark demonstrated that DreamPRM-1.5 significantly improves GPT-5-mini, achieving state-of-the-art performance. Importantly, the PRM itself is trained solely on VisualPRM-400k, with the meta set only used to guide instance reweighting. These findings highlight the effectiveness of fine-grained instance reweighting and open new directions for robust multimodal reasoning.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Qi Cao, Ruiyi Wang, Ruiyi Zhang, Sai Ashish Somayajula, and Pengtao Xie. Dreamprm: Domain-reweighted process reward model for multimodal reasoning, 2025.

[3] Sang Keun Choe, Willie Neiswanger, Pengtao Xie, and Eric Xing. Betty: An automatic differentiation library for multilevel optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[5] Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. Progressive multimodal reasoning via active retrieval, 2024.

[6] Simin Fan, Matteo Pagliardini, and Martin Jaggi. Doge: Domain reweighting with generalization estimation, 2024.

[7] Chelsea Finn, P. Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

[8] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making large language models better reasoners with step-aware verifier, 2023.

[9] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025.

[10] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[13] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

[14] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision, 2024.

[15] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024.

[16] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David

Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[17] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[18] Alex Reid et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens, 2024.

[19] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting, 2019.

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

[21] Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, and Weimin Zhang. How to bridge the gap between modalities: Survey on multimodal large language model, 2025.

[22] Daouda Sow, Herbert Woisetschläger, Saikiran Bulusu, Shiqiang Wang, Hans-Arno Jacobsen, and Yingbin Liang. Dynamic loss-based sample reweighting for improved large language model pretraining, 2025.

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[24] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. Visualprm: An effective process reward model for multimodal reasoning, 2025.

[25] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[26] Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision, 2024.

[27] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey, 2023.

[28] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024.

[29] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024.

[30] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

A System Prompt for Generative Reward Model.

The Generative Reward Model leverages a system prompt adapted from VisualPRM [24].

```
You are an advanced AI assistant, designed to serve as a process
supervision model.  In this task, I will provide a problem statement
followed by the first step of the solution process.  For each subsequent
turn, I will give you a new step in the solution.  Your role is to assess
whether the solution process is correct up to the current step.- In
the **first round**, I will input the problem and the first step of the
solution process.- In **each subsequent round**, I will provide the next
step in the solution.  For each step, you should:- Respond with **+** if
you believe the solution process is correct up to this step.- Respond
with **-** if you detect any issues or errors in the process up to this
step.  Please note:- Only respond with **+** or **-**.  Do not provide
any additional explanations, comments, or justifications.  Your task is
to verify the accuracy and correctness of each step in the given solution
process.
```

B Hyperparameter Settings.

For the lower-level optimization, we perform one inner gradient step per outer update (*unroll steps* = 1), using the AdamW optimizer [12] with a learning rate of 5×10^{-5} and weight decay of 10^{-2}.

For the upper-level optimization, we also adopt AdamW. In the Instance Table setting, we use a learning rate of 5×10^{-3} with weight decay 10^{-3}; in the Instance Net setting, we use a learning rate of 5×10^{-4} with weight decay 10^{-3}, and set the hidden dimension of the network to 10.

Both levels employ a cosine learning rate schedule with linear warm-up, where the warm-up phase corresponds to 5% of the total training steps. Overall, DreamPRM-1.5 is fine-tuned for 100,000 iterations. The framework is implemented using Betty [3], and full training requires approximately 72 hours on a single NVIDIA A100 GPU.