

TINYEYE: SHARPENING VISUAL REASONING OF TINY MODELS WITH OFFLINE POLICY OPTIMIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Multimodal reasoning with small vision–language models (VLMs) is increasingly important in real-world applications, yet their limited capacity makes optimization and alignment especially challenging. In this paper, we propose a holistic framework for offline policy optimization to sharpen the visual reasoning capabilities of small models. At its foundation is TinyEye-Data, a large-scale corpus of two million reasoning trajectories distilled from state-of-the-art VLMs across 68 verifiable tasks, which provides diverse and reliable binary supervision entirely in the offline setting. We instantiate the framework through a four-stage pipeline: (1) native-resolution warm-up for robust vision–language alignment, (2) instruction tuning on TinyEye-Data to establish a broad reasoning foundation, (3) annealed rejection sampling to mine hard cases and refine supervision, and (4) Discriminative Direct Preference Optimization (DDPO), a new margin-based objective that formulates policy learning as reward classification and resolves the likelihood displacement issues of DPO. Stages (3) and (4) together form the core of verifiable offline reinforcement learning, where rejection sampling refines signals and DDPO optimizes the policy against them. The resulting model, TinyEye-2B, achieves state-of-the-art results across diverse reasoning benchmarks, reaching 50.3% on MMMU, 55.2% on MathVerse, and 63.9% on HallBench, outperforming other models of comparable scale by significant margins.

1 INTRODUCTION

Post-training has become a cornerstone in the development of visual–language models (VLMs), enabling them to follow instructions and perform long chain-of-thought reasoning. Supervised fine-tuning (SFT) with curated instruction following datasets is among the earliest and most widely adopted approaches (Liu et al., 2023; Li et al., 2024a; Chen et al., 2023; Bai et al., 2023; 2025; Li et al., 2025b). More recently, the community has witnessed a paradigm shift beyond SFT toward methods that directly incorporate reward or preference signals. Direct Preference Optimization (Rafailov et al., 2023; Zhang et al., 2024b) (DPO) aligns models by contrasting preferred and non-preferred responses, while reinforcement learning (RL) methods such as PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), and GSPO (Zheng et al., 2025) explicitly optimize model policies against reward feedback obtained from real-time sampled rollouts. Together, these three families of techniques define the core toolkit for post-training VLMs.

Despite these successes, such approaches also face inherent limitations, especially under constraints on model parameter size. Supervised fine-tuning (SFT) relies solely on positive examples and ignores negative evidence, which can lead to overfitting and a collapse in output diversity (Li et al., 2024b), particularly when the model’s capacity is limited (Kaplan et al., 2020; Hoffmann et al., 2022). Direct Preference Optimization (DPO) avoids explicit reward modeling but suffers from likelihood displacement (Cho et al., 2025; Yang et al., 2025b): the additive margin in its loss may suppress the probability of negative responses but fails to sufficiently reinforce positive ones (Cho et al., 2025), sometimes even penalizing desirable behaviors (Pal et al., 2024a). Reinforcement learning (RL) methods, in contrast, provide stronger optimization signals but come with high computational costs, poor sample efficiency due to limited rollout reuse, and training instability that makes them difficult to tune. When reinforcement signals are high-variance and unstable, capacity-limited LLMs are especially prone to entropy collapse (Cui et al., 2025; Le et al., 2025), often resulting in low-diversity, repetitive outputs or even endless tandem loops during on-policy training. Recent

studies (Shumailov et al., 2024; Wenger, 2024) also indicate that if synthetic or self-generated data are reused recursively without careful filtering, models tend to first forget rare, long-tail knowledge and eventually distort even common pretrained knowledge—a process known as model collapse, which degrades both the realism and diversity of the model’s capabilities.

Multimodal reasoning with small models has many appealing real world applications requiring edge deployment. However, the above drawbacks are particularly pronounced in such settings, where naive reinforcement learning often exploits superficial shortcuts (Sarch et al., 2025; Xia et al., 2025) rather than learning grounded reasoning, leading to poor generalization. DPO is also highly sensitive to data quality (Xu et al., 2024): low-margin or noisy preference pairs can weaken parameter updates, and additive margin inside its loss objective may further suppress the probability of positive responses (Pal et al., 2024a), resulting in degraded and unstable alignment. For this reason, filtering such pairs is strongly advised (Morimura et al., 2024). Altogether, these observations underscore the urgent need for an alignment method that is discriminative, stable under limited capacity, and efficient without relying on costly on-policy rollouts.

We propose a unified offline policy optimization framework that elicits strong multimodal reasoning capabilities in compact visual–language models. The framework contains three integral components: large-scale offline distillation from diverse reasoning trajectories, offline rejection sampling to mine and refine supervision signals, and a final stage of offline policy optimization through Discriminative Direct Preference Optimization (DDPO). Central to this design is TinyEye-Data, a verifiable dataset that enables broad coverage for distillation and provides the foundation for learning entirely in the offline setting. Together, these components form a coherent pipeline: distillation supplies rich and diverse reasoning signals, rejection sampling sharpens them into high-quality positive–negative supervision, and DDPO optimizes over the curated pairs with a margin-based discriminative objective. This coupling ensures that each stage reinforces the next, yielding stable and scalable alignment without relying on unstable on-policy rollouts.

We instantiate the framework through a four-stage pipeline as shown in Figure 1. In Stage-1, a native-resolution warm-up adapts a SigLIP2 (Tschannen et al., 2025) backbone into a ViT variant capable of flexible variable-resolution processing and robust vision–language alignment. In Stage-2, we conduct large-scale instruction tuning over TinyEye-Data, incorporating both curated multimodal corpora and trajectories distilled from strong teacher models such as MiMo-VL (Yue et al., 2025) and Key-VL (Team et al., 2025c), thereby establishing a broad reasoning foundation in an offline manner. In Stage-3, we apply annealed rejection sampling to mine hard cases, filter degenerate outputs, and amplify supervision density through an additional round of tuning. Finally, in Stage-4, we introduce DDPO, a discriminative objective that enforces a margin-based update rule. In the binary reward setting common to multimodal reasoning, DDPO guarantees that gradient descent always increases the probability of preferred responses while suppressing negatives, thus resolving the likelihood displacement issue of DPO (Razin et al., 2025) and providing stable optimization.

In conclusion, our work makes three main contributions:

- We propose a holistic framework for **offline policy optimization** that integrates large-scale distillation, rejection sampling, and discriminative policy optimization, enabling stable reasoning alignment for small multimodal models.
- We construct **TinyEye-Data**, a large-scale verifiable corpus of two million reasoning trajectories distilled from diverse teacher models across 68 tasks.
- We introduce **DDPO**, a discriminative policy optimization objective that provably guarantees monotonic reinforcement of positives while suppressing negatives, directly resolving the likelihood displacement (Cho et al., 2025; Yang et al., 2025b) issue of DPO.
- We demonstrate strong empirical performance: our model, **TinyEye-2B**, achieves state-of-the-art results on multimodal and math reasoning benchmarks, surpassing strong open-weight baselines such as InternVL3-2B (Zhu et al., 2025) and Ovis2-2B (Lu et al., 2024).

2 METHOD

In this section, we introduce the proposed framework. We first introduce TinyEye-Data, a curated dataset that emphasizes verifiable supervision and multi-teacher distillation to provide both relia-

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

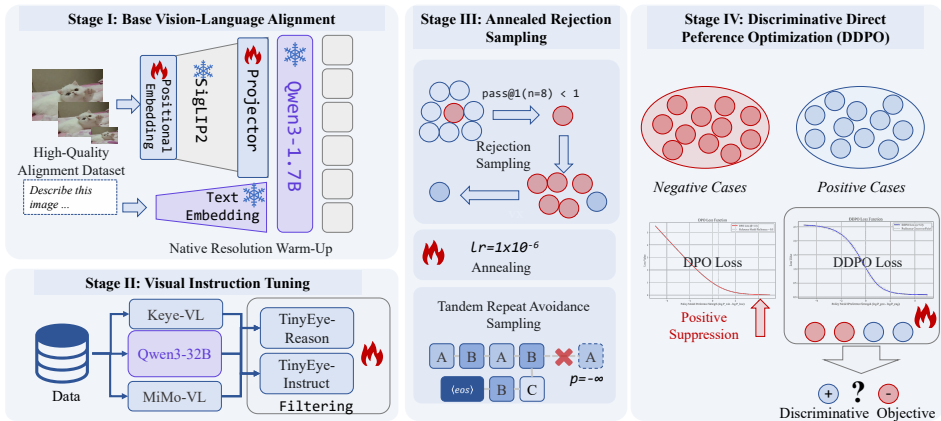


Figure 1: An overview of the proposed offline policy optimization pipeline.

bility and broad coverage across multimodal reasoning tasks. Building on this dataset, TinyEye is trained through the instantiated four-stage pipeline as illustrated in Figure 1.

2.1 TINYEYE-DATA

Effective offline policy optimization requires diverse and verifiable reasoning data: diversity ensures that compact models encounter a broad spectrum of tasks, while verifiability guarantees stable signals for subsequent rejection sampling and discriminative optimization. To this end, we construct TinyEye-Data, a heterogeneous pool spanning text, vision, mathematics, STEM, puzzles, and documents, designed to expose small models to both core skills and long-tail reasoning challenges.

We aim to enable robust visual reasoning under limited model capacities. Our core strategy is to construct compact and verifiable supervision through a base–tail paradigm: a solid base of essential skills complemented by a highly diverse tail of long-tail tasks. This design follows Li et al. (2025b) on the principle of diversity first, then quality, where we deliberately maximize coverage of reasoning styles and domains before applying strict distillation and cleansing. By combining multi-teacher offline distillation with curated verification signals, TinyEye-Data equips small models with reliable supervision while mitigating noise and collapse risks at the source.

Data construction. Instead of indiscriminately aggregating large-scale instructions, we curate a dataset pool called TinyEye-Data comprising 179 datasets across 14 categories. Collection proceeds along two complementary lines: TinyEye-Instruct and TinyEye-Reason, both following the base–tail philosophy.

TinyEye-Instruct provides the foundation for multimodal instruction tuning. Its base consolidates key skills through math and geometry (17 datasets), chart and table reasoning (11), and scientific diagrams (11). The tail then pushes diversity by incorporating long-tail tasks: 27 perception VQA datasets, 11 OCR and document-understanding datasets, and 4 captioning corpora, along with 5 text-only QA datasets to stabilize instruction following. In total, TinyEye-Instruct integrates 86 datasets spanning mathematics, structured data, diagrams, perception, documents, captioning, and text. Quality is safeguarded with filtering strategies such as majority-vote distillation on AOPS-Instruct (Mahdavi et al., 2025). A detailed breakdown is shown in Table 6.

TinyEye-Reason extends the same diversity-first design with explicit emphasis on reasoning trajectories. Its base consists of canonical math corpora (AOPS-Instruct, MATH, AIME, etc.), while the tail greatly broadens coverage: 10 game and puzzle datasets, 11 STEM domains (chemistry, medicine, finance, philosophy, etc.), and 38 general VQA datasets grounding reasoning in visual understanding. Multimodal reasoning is further strengthened by 17 math-related datasets, 7 puzzle-style datasets, and 5 STEM-oriented corpora requiring joint symbolic–visual reasoning. Some datasets by themselves contain highly diverse sub-domains. For instance, StackExchange-longtail (Lambert et al., 2023) contains 2 million longtail examples covering 171 topics. In total, TinyEye-Reason covers 93 datasets across seven categories, ranging from Olympiad-level proofs to pediatric medicine and counterfactual history. The composition is summarized in Table 7.

Some datasets naturally serve both instruction tuning and reasoning distillation. Across the pool, supervision signals fall into two types: verifiable targets (equations, multiple-choice) strictly validated via exact matching, and natural-language targets judged by VLM-as-a-Judge to retain expressive flexibility. This layered design ensures both verifiability and diversity, enabling compact models to internalize stable reasoning patterns without sacrificing breadth of coverage.

Multi-teacher offline distillation. Building on this supervised dataset, we further distill high-quality reasoning trajectories. Relying on a single teacher model could often produce noisy trajectories that “accidentally” or “luckily retry” lead to the correct answer. Accordingly, we assign teachers by task type: MiMo-VL-7B-RL for math-related visual questions, Keye-VL-8B-Preview for other VQA tasks, and Qwen3-32B for text-only prompts. For each question, we generate 8 independent trajectories with temperature sampling—the 8 trajectories allow us to compute pass@8 as a statistical measure of dataset difficulty, which can further guide the subsequent training stages.

Reward Labeling. To guarantee the quality of distilled trajectories, we adopt task-specific reward surrogates. For equations and multiple-choice problems, we apply `math-verify` Kydlíček to perform symbolic and numerical equivalence checks, providing a near noise-free supervision signal. For natural-language responses, we use a lightweight VLM-as-a-Judge based on Qwen2.5-VL-3B-Instruct (Wang et al., 2024b; Team, 2025), which can jointly interpret image and text contexts to assess semantic consistency. This prevents cases where an answer is factually correct but phrased inconsistently, and it reduces the risk of introducing image-text misalignment hallucinations during subsequent training. In this way, we enforce strict correctness on verifiable tasks while ensuring semantic robustness on open-ended ones, thereby improving the reliability and controllability of the distilled corpus.

Cleaning and aggregation. To block degenerate supervision signal, we apply two key post-distillation steps. First, we remove long, low-variation tandem repeats using a linear-time detector (Algorithm 2), preventing the model from imitating collapse-like patterns present in teacher outputs. Second, among the remaining candidates we perform shortest-chain aggregation, keeping for each prompt only the shortest valid chain of thought. This strategy instills an explicit preference for efficient reasoning without sacrificing correctness, and attenuates exhaustive or enumerative traces.

After the above construction and filtering, the resulting corpus is overall more compact, contains substantially fewer repetitive patterns, and preserves higher sample-level diversity, making it better aligned with the limited learning capacity of small models.

2.2 VISION-LANGUAGE ALIGNMENT AND INSTRUCTION TUNING

Stage-1: Vision-Language Alignment. To build a base model for our task, we consider a native resolution VLM where we adopt *Native Resolution Warm-Up* to adapt the vision encoder to varied input sizes. The proposed process performs vision-language alignment with a frozen language backbone. In particular, we use Qwen3-1.7B (Yang et al., 2025a) as the language backbone and the SigLIP2-So400M (Tschannen et al., 2025) vision encoder, which has a modified 14×14 patch size (originally 16×16) and a fixed input resolution of 448×448 . To enable the native resolution capabilities, we represent each input image as a patch sequence $V \in \mathbb{R}^{N \times d}$. The absolute positional embeddings P_{abs} are interpolated to the target grid size (H, W) , yielding $\tilde{P} \in \mathbb{R}^{N \times d}$. Following Keye-VL (Team et al., 2025c), we apply a learnable two-dimensional rotary positional embedding (RoPE) layer with parameters ϕ :

$$V' = \text{RoPE}_{\phi}(V + \tilde{P}). \quad (1)$$

Unlike previous approaches that continue pretraining the vision tower independently with large-scale contrastive learning, we warm up the ViT to handle native-resolution inputs while simultaneously performing vision-language alignment. During this stage, both the vision encoder and the LLM remain frozen, while only the projector module and the learnable positional encoding layers are trained. We use ALLaVA (Chen et al., 2024a) and LLaVA-ReCap-CC3M (Li et al., 2024a) datasets, and optimize with a standard cross-entropy loss for captioning and question answering.

Stage-2: Instruction Tuning. Building upon the aligned base model, we perform large-scale multimodal instruction tuning to equip the model with robust task-following and reasoning capabilities. Training is conducted on the curated TinyEye-Instruct and TinyEye-Reason datasets, which combine broad task coverage with compact and verifiable supervision across image captioning, visual question answering, mathematical reasoning, and general knowledge tasks.

We optimize the model with the standard autoregressive cross-entropy loss over the instruction–response pairs. Given a sequence of target tokens $y = \{y_1, \dots, y_T\}$ conditioned on multimodal input image tokens I and text tokens x , the objective is:

$$\mathcal{L}_{\text{CE}} = - \sum_{t=1}^T \log p_{\theta}(y_t | y_{<t}, I, x), \quad (2)$$

where p_{θ} denotes the model’s conditional probability distribution. This stage transitions the model from merely aligned representations to effective instruction following, thereby establishing a stable foundation for subsequent hard-case mining and policy optimization.

2.3 VERIFIABLE OFFLINE REINFORCEMENT LEARNING

Stage-3: Annealing Rejection Sampling Fine-Tuning. In Stage 3, we shift the training focus from saturated prompts to more informative borderline or hard cases. Concretely, we estimate the teacher’s success rate for each prompt as $\text{pass}@1_{(n=8)}(x)$. Prompts with $\text{pass}@1_{(n=8)}(x) = 1$ are considered saturated and removed, while the remaining subset $\mathcal{H} = \{x : \text{pass}@1_{(n=8)}(x) < 1\}$ defines the hard set for annealed resampling. On \mathcal{H} , we sample $K = 4$ candidates per prompt, thereby concentrating supervision on outputs that lie closer to the error boundary and thus carry richer alignment signal.

When generating such candidates, small-capacity models are prone to collapse into low-entropy tandem repetitions (e.g., endlessly concatenating short fragments). To mitigate this, we introduce Tandem Repeat Avoidance Sampling (TRAS). During decoding, given the partial sequence $y_{1:t}$, a linear-time detector $D(\cdot)$ identifies newly formed concatenated repeats. Once detected, the offending tokens are added to the banned set B_{t+1} , which is passed to a `LogitsProcessor` interface of Transformers library to mask degenerate continuations. To maintain efficiency, we store recent repeat patterns in an LRU buffer and only perform prefix matching against this buffer, dynamically refreshing recency. In this way, TRAS explicitly disrupts nascent repetition loops while preserving normal decoding. The pseudo-code of TRAS can be found in Algorithm 3.

After candidate generation, we score outputs using the same reward annotation method as in Stage 2 and apply the principle of Rejection Sampling Fine-Tuning (RFT). Specifically, only high-quality “positive” samples, as judged by the reward annotation method, are retained for supervised fine-tuning, while low-scoring “negative” samples are discarded. The underlying rationale is that by repeatedly generate-filter-finetune, the policy distribution of model is gradually shifted from the policy of base model toward a reward-aligned distribution. Compared with on-policy reinforcement learning methods that penalize negative samples, RFT better accommodates the fragility of small models, offering greater stability, simpler optimization, and reduced risks from directly using noisy negative feedback.

During fine-tuning, we further anneal the learning rate and train solely on positives from \mathcal{H} , ensuring stable convergence on genuinely challenging cases. By discarding saturated prompts and curating preference pairs nearer to the decision boundary, this stage delivers a tempered supervision signal: it suppresses degenerate repetitions, promotes exploration without collapse, and steers the student model more reliably toward alignment objectives. The retained positive–negative pairs also serve as valuable inputs for the discriminative optimization in Stage 4.

Stage-4: Discriminative Direct Preference Optimization. Since the reward signal in our task is inherently binary, which means each response to a given prompt is labeled either correct or incorrect, we aim to tailor the contrastive learning objective of DPO such that optimization directly focuses on the binary classification of reward labels. To this end, we reformulate the original DPO’s preference modeling problem into an explicit binary discrimination task: the model must reliably classify responses into correct or incorrect categories, making the objective more consistent with the supervision structure of binary rewards in visual reasoning tasks. Motivated by this, we propose *Discriminative Direct Preference Optimization (DDPO)*, which reinterprets the relative likelihood differences in DPO as binary classification probabilities, and optimizes an interval-based objective designed for this two-class reward setting.

We first define the binary classification probability for a single response as

$$r(x, y) = \sigma(\Delta(x, y)), \quad \Delta(x, y) = \ell_{\theta}(y|x) - \ell_{\text{ref}}(y|x), \quad (3)$$

where $\ell_\theta(y|x)$ is the length-normalized log-likelihood under the current policy and $\ell_{\text{ref}}(y|x)$ is the log-likelihood under a reference model. Through the sigmoid mapping, $\Delta(x, y)$ is interpreted as the probability that the response is correct.

For a preference pair (y^+, y^-) , the margin signal is given by the probability difference

$$d = \sigma(\Delta(x, y^+)) - \sigma(\Delta(x, y^-)). \tag{4}$$

A natural way to optimize this margin is with a hinge loss

$$L_{\text{hinge}}(\theta) = \max(0, 1 - d). \tag{5}$$

To enable smooth gradient-based optimization, we adopt the log-sigmoid as a differentiable surrogate, yielding the Discriminative Direct Preference Optimization (DDPO) objective

$$L_{\text{DDPO}}(\theta) = -\log \sigma(\alpha d), \tag{6}$$

where $\alpha > 0$ controls the sharpness of the margin. This formulation establishes a binary discriminative training objective, the policy is thus trained to decide, for each prompt, which response should be assigned the positive reward label and which the negative one. By introducing the sigmoid smoothing, DDPO avoids the likelihood displacement present in standard DPO, while providing sign-preserving gradient signals for both positives and negatives. From a perspective of sequence-level policy optimization, DDPO objective can view as a offline surrogate of GSPO method which concentrate on optimizing margin of advantages¹. In practice, this stage consolidates the supervision: the model not only imitates high-quality responses but also learns to reliably separate them from poor ones, achieving more robust alignment under binary reward settings.

3 EXPERIMENTS

3.1 EXPERIMENT SETTINGS

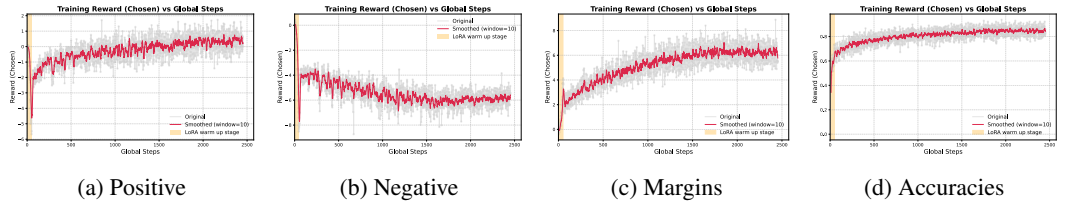


Figure 2: Train rewards v.s. global steps. A 10-step moving average is used to visualize. Yellow areas are warm-up stages of random initialized parameters of LoRA layers.

Training Details. For hyperparameters of training, please refer to Appendix C. Particularly, as illustrated in Figure 2, the reward dynamics during LoRA-based DDPO training exhibit a clear separation between positive and negative samples. After the warm-up of random initialized parameters of LoRA layers, the reward for positive examples (Figure 2a) steadily increases and stabilizes over time, indicating that the policy consistently strengthens its preference for high-quality responses. In contrast, the reward for negative examples (Figure 2b) progressively decreases, reflecting the model’s ability to suppress low-quality responses. Consequently, the reward margin between positives and negatives (Figure 2c) grows throughout training, demonstrating that DDPO not only imitates preferred behaviors but also enlarges their separation from undesirable ones. This trend is further confirmed by the accuracy curve (Figure 2d), which rises toward saturation, showing that the policy becomes increasingly reliable in distinguishing good from poor responses.

Benchmarks. We evaluate our 2B model across three main tasks to show its multiple capability: multimodal reasoning, textual reasoning and general multimodal question answering. For multimodal reasoning, we choose MMMU (val) (Yue et al., 2023), MathVision (Wang et al., 2024a),

¹For proof, please refer to Appendix B.

Models	MMM(U(Val)	MathVision	MathVista	MathVerse	LogicVista	WeMath
TinyEye-2B-Stage4 (Ours)	<u>50.3</u>	32.6	<u>67.9</u>	55.2	41.7	52.4
Ovis2-2B (Lu et al., 2024)	45.6	17.7	64.1	29.4	34.7	9.9
InternVL3-2B (Zhu et al., 2025)	48.6	21.7	57.0	25.3	36.9	22.4
Qwen2.5-VL-3B (Bai et al., 2025)	51.2	<u>21.9</u>	61.2	31.2	<u>40.3</u>	<u>22.9</u>
Ovis2-4B (Lu et al., 2024)	49.0	21.5	69.6	<u>38.5</u>	35.3	16.9
GPT-5 (OpenAI, 2025)	81.8	72.0	81.9	81.2	70.0	71.1
Gemini-2.5-Pro (DeepMind, 2025)	74.7	69.1	80.9	76.9	73.8	78.0
GLM-4.5V-108B (Team et al., 2025a)	75.4	65.6	84.6	72.1	62.4	68.8
Qwen2.5-VL-72B (Bai et al., 2025)	68.2	39.3	74.2	47.3	55.7	49.1
InternVL3-78B (Zhu et al., 2025)	72.2	43.1	79.0	51.0	55.9	46.1

Table 1: Main results on multimodal reasoning datasets. We choose the vision-only part of MathVerse. Some results are from InternVL-series (Zhu et al., 2025) technical report and OpenCompass (Contributors, 2023) platform. We **highlight** the best performance and underline the second best of smaller VLMs in this table.

Models	AIME24	AIME25	MATH500	GPQA-D	MMLU-Pro
TinyEye-2B-Stage4 (Ours)	33.3	36.7	88.0	43.9	58.5
GLM-4.1V-9B (Team et al., 2025a)	36.2	32.0	81.8	50.3	57.1
Kimi-VL-2506-16B (Team et al., 2025d)	54.0	39.1	91.8	42.3	68.5
GLM-4.5V-106B (Team et al., 2025a)	72.8	80.1	94.2	56.6	69.7
Qwen2.5-VL-72B (Bai et al., 2025)	15.0	13.3	82.8	49.0	71.2
GPT-5 (OpenAI, 2025)	90.0	94.6	97.8	85.7	85.6
Gemini-2.5-Pro (DeepMind, 2025)	92.0	88.0	95.2	86.4	84.1
Grok-4 (xAI, 2025)	-	90.0	96.2	88.1	85.3
Claude-Sonnet-4 (Anthropic, 2025)	-	76.3	93.8	74.5	83.8

Table 2: Main results on textual math reasoning tasks. We apply some VLMs evaluated on these benchmarks, and use the score from their technical report.

MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024a), LogicVista (Xiao et al., 2024) and WeMath (Qiao et al., 2024). For textual reasoning, we select AIME24&AIME25 (math-ai, 2024; OpenCompass, 2025), MATH500 (Hendrycks et al., 2021), GPQA-Diamond (Rein et al., 2023) and MMLU-pro (Wang et al., 2024c). For general multimodal question answering tasks, we select MMBench (MMB) (Liu et al., 2024b), MMStar (Chen et al., 2024b), MMVet (Yu et al., 2023), AI2D (Kembhavi et al., 2016), HallBench (Guan et al., 2024) and BLINK (Fu et al., 2024).

Baselines. Several VLMs of the similar scales (2B-4B) are selected to compare with our model. Ovis2-2B, 4B (Lu et al., 2024), InternVL3-2B (Zhu et al., 2025) and Qwen2.5-VL-3B (Bai et al., 2025) are chosen. We also report the scores of some advanced VLMs like GPT-series (OpenAI, 2025; Hurst et al., 2024) and Gemini-series (DeepMind, 2025). Details are in experiment tables.

3.2 EVALUATION RESULTS

Results on multimodal reasoning benchmarks. Despite having only 2B parameters, our model, TinyEye-2B-Stage4 achieves strong performance across a wide range of multimodal reasoning benchmarks. It nearly outperforms all existing models of similar or smaller scale and even rivaling much larger ones. As shown in Table 1, TinyEye-2B-Stage4 achieves 55.2 on MathVerse, significantly surpassing other similar-scale models. It is also the only lightweight model to exceed 30 on MathVision (32.6) and sets a new state-of-the-art (SOTA) among small-scale VLMs on WeMath with a score of 52.4, outperforming models with up to $40\times$ more parameters, such as InternVL3-78B and GLM-4.5V-108B. These results demonstrate that TinyEye-2B-Stage4 not only closes the performance gap with large-scale VLMs but also offers a highly efficient and scalable solution for multimodal reasoning tasks under resource constraints.

Results on textual reasoning benchmarks. Evaluation results are displayed in Table 2. Here we report VLMs with 9B parameters or more with official results. Our model achieves strong performance across several textual reasoning benchmarks. It reaches 88.0 on MATH500, outperforming larger VLMs like GLM-4.1V-9B (81.8) and Qwen2.5-VL-72B (82.8). On AIME25, it surpasses GLM-4.1V-9B by +4.7 points. It also performs competitively on GPQA-D (43.9) and MMLU-Pro

Models	MMBench	MMStar	MMVet	AI2D	HallBench	BLINK
TinyEye-2B-Stage4 (Ours)	81.5	61.3	73.0	79.7	63.9	44.3
Ovis2-2B Lu et al. (2024)	78.6	56.7	58.3	<u>82.7</u>	50.2	47.9
InternVL3-2B Zhu et al. (2025)	80.8	60.7	62.2	78.7	42.5	<u>50.3</u>
Qwen2.5-VL-3B Bai et al. (2025)	79.1	55.9	63.2	81.6	46.3	47.6
Ovis2-4B Lu et al. (2024)	83.6	61.9	<u>65.5</u>	85.7	<u>53.8</u>	53.0
GPT4o Hurst et al. (2024)	82.1	64.7	69.1	84.6	55.0	68.0
Claude3.5 Sonnet Anthropic (2024)	83.4	65.1	70.1	81.2	55.2	-
Gemini-2-flash DeepMind (2024)	83.0	69.4	-	-	-	-
GLM-4.5V-108B Team et al. (2025a)	88.2	75.3	75.2	93.7	64.5	65.3
Qwen2.5-VL-72B Bai et al. (2025)	88.3	70.8	76.2	88.4	55.2	64.4
InternVL3-78B Zhu et al. (2025)	88.8	72.5	81.3	96.0	59.1	66.3

Table 3: Main results on multimodal general datasets. Some results are from InternVL-series ([Zhu et al., 2025](#)) technical report and OpenCompass ([Contributors, 2023](#)) platform. We **highlight** the best performance and underline the second best of smaller VLMs in this table.

Method	MathVista	MathVision	MATH500	Hallbench
TinyEye-2B-Stage2	63.8	28.7	86.7	35.9
TinyEye-2B-Stage3	66.0	30.4	87.2	48.7
TinyEye-2B-Stage4 (DDPO)	67.9	32.6	88.0	63.9
TinyEye-2B-Stage3 + DPO	68.0	31.0	87.4	55.8

Table 4: Performance of TinyEye-2B at different training stages and with policy optimization methods (DPO/DDPO) on four benchmarks.

(58.5), outperforming several larger models. These results depict that our lightweight model can match or even exceed the performance of much larger models on textual reasoning tasks.

Results on general multimodal benchmarks. On general multimodal benchmarks, TinyEye-2B-Stage4 achieves performance close to or better than models twice its size. For example, it outperforms Ovis2-2B and InternVL3-2B across almost all benchmarks. It also outperforms Ovis2-4B on HallBench, and illustrates close results on MMBench and MMStar. These results reveal that our 2B model can deliver competitive general understanding and reasoning capabilities, comparable to 4B-level VLMs.

Case Study. For examples of case study, please refer to Appendix F.

3.3 ABLATION STUDY

Ablation on training stages. We conduct an ablation study across different training stages. As shown in Table 4, model performance improves progressively at each stage. For example, on HallBench, accuracy rises from 35.9% after instruction fine-tuning (Stage 2) to 48.7% with hard-case rejection sampling (Stage 3). Similar improvements are observed on other benchmarks, such as MathVista (63.8% \rightarrow 66.0%) and MathVision (28.7% \rightarrow 30.4%). These results indicate that carefully designed training stages can steadily enhance the model’s reasoning ability across diverse mathematical benchmarks.

Ablation on policy optimization. We further analyze the effect of preference-based policy optimization. As shown in Table 4, Incorporating DPO in default setting of `trl` library after Stage 3 improves MathVista from 66.0% to 68.0% and HallBench from 48.7% to 55.8%. More notably, applying DDPO at Stage 4 leads to consistent and significant improvements, e.g., HallBench accuracy surges to 63.9% (a 28-point gain over Stage 2), and the model achieves the best performance on MathVision (32.6%) and MATH500 (88.0%). These results highlight that DDPO is particularly effective for binary classification tasks like HallBench, while also contributing to robust gains across a variety of mathematical datasets.

Comparison on thinking and non-thinking modes To assess the impact of enabling thinking mode, we perform an ablation study by comparing model performance with and without it. Specifically, we append “no think” to the prompt to disable this mode. As shown in Table 5, the absence of thinking leads to clear degradation on all benchmarks: MathVista drops by 12.2 points, MathVi-

Method	MathVista	MathVision	MATH500
TinyEye-2B-Stage4 (thinking)	67.9	32.6	88.0
TinyEye-2B-Stage4 (no thinking)	55.7	19.7	77.4

Table 5: Think v.s. no think mode.

sion by 12.9, and MATH500 by 10.6. These consistent declines indicate that structured reasoning plays a crucial role in guiding VLMs toward correct solutions, and that thinking mode significantly enhances the model’s ability to tackle reasoning-intensive tasks.

4 RELATED WORKS

Variants of DPO. DPO aligns models by maximizing the log-likelihood margin between preferred and non-preferred samples, but raises issues of reference dependence, likelihood displacement, margin design, gradient stability, and noise robustness. Variants address these in different ways: SimPO (Meng et al., 2024) weakens reliance on reference models with a simplified objective and fixed margin; Alpha-DPO (Wu et al., 2025) introduces instance-adaptive margins to balance alignment and diversity; DPO-Positive (Pal et al., 2024b) enforces “only increase” updates for preferred samples; Refined-DPO (Gallego, 2024) improve noise tolerance via importance weighting; ORPO (Hong et al., 2024) uses win-loss ratios for single-stage alignment.

Offline policy optimizations in VLMs Training. In vision and multimodal settings, V-DPO (Xie et al., 2024) curbs hallucinations through image-text consistency, while MIA-DPO (Liu et al., 2024c) leverages multi-image inputs to mitigate data scarcity. Methods such as VideoDPO (Liu et al., 2024a), VistaDPO (Huang et al., 2025), and TEMPO (Li et al., 2025a) incorporate temporal and semantic constraints for video generation and understanding. VLM applications include LLaVA-Reasoner-DPO (Zhang et al., 2024c) for reasoning, CLIP-DPO (Ouali et al., 2024) for hallucination reduction. At scale, InternVL (Chen et al., 2024c) and Keye-VL (Team et al., 2025e) integrate mixed preference optimization with large multimodal preference datasets, while MiniCPM-V (Yao et al., 2024) applies DPO to learn from AI feedback and improve factual consistency in visual understanding.

Thinking VLMs. With the development of VLMs, prior works develop thinking and reasoning abilities of them for solving more complex tasks. Recent advances in multimodal reasoning have shifted from static instruction-tuning toward reasoning-centric training and adaptive inference. MiMo-VL introduces a four-stage pretraining pipeline with mixed on-policy reinforcement learning to enhance chain-of-thought (CoT) capabilities (Yue et al., 2025). For video reasoning, Keye-VL employs a five-mode cold-start mixture to regulate when and how models should think, while Keye-VL 1.5 integrates Slow-Fast encoding and long-context tuning for temporal understanding (Team et al., 2025c;b). GLM-4.1V-Thinking applies reinforcement learning with curriculum sampling (RLCS) to a 9B VLM (Team et al., 2025a). Meanwhile, GPT-5 adopts a unified routed system, where a lightweight model defers to a deeper “thinking” pathway when complexity demands it (OpenAI, 2025).

5 CONCLUSION

In this work, we propose a unified offline policy optimization framework that equips compact visual-language models with strong multimodal reasoning capabilities. At its core is TinyEye-Data, a large-scale verifiable corpus distilled from diverse teacher models, which supports a four-stage training pipeline combining native-resolution warm-up, large-scale distillation, annealed rejection sampling, and Discriminative Direct Preference Optimization (DDPO). DDPO resolves the likelihood displacement issue of DPO by provably reinforcing positives while suppressing negatives, ensuring stable offline optimization. Our 2B-parameter model, TinyEye-2B, achieves state-of-the-art performance on multimodal and math reasoning benchmarks, surpassing strong open-weight baselines such as InternVL3-2B and Ovis2-2B.

6 ETHICS STATEMENT

This work does not involve any human subjects or sensitive personal data. The usage of all datasets strictly complies with their respective licenses.

Our methods are intended solely for academic and scientific purposes. We do not foresee direct harmful applications, but acknowledge that misuse could occur if applied without proper safeguards. We encourage responsible use of the research outcomes, with attention to fairness, transparency, and legal compliance.

7 REPRODUCIBILITY STATEMENT

We have taken several measures to ensure the reproducibility of our work. All details of the proposed model, preprocessing steps of datasets and algorithms with full hyperparameter settings and training procedures provided are described in the main text. Proofs of theoretical results are given in Appendix B.

REFERENCES

- Anthropic. Claude 3.5 sonnet model card addendum, 2024. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf. 8
- Anthropic. Claude-sonnet-4: Hybrid reasoning for advanced coding and general intelligence. <https://www.anthropic.com/claude/sonnet>, 2025. 7
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv:2308.12966*, 2023. 1
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv:2502.13923*, 2025. 1, 7, 8
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. ALLaVA: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv:2402.11684*, 2024a. 4
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024b. 7
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv:2312.14238*, 2023. 1
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024c. URL <https://arxiv.org/abs/2312.14238>. 9
- Jay Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. Rethinking dpo: The role of rejected responses in preference misalignment. *arXiv preprint arXiv:2506.12725*, 2025. 1, 2
- OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023. 7, 8
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025. 1
- Google DeepMind. Gemini 2.0 flash thinking, 2024. URL <https://deepmind.google/technologies/gemini/flash-thinking/>. 8

- 540 Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long
541 context, and next generation agentic capabilities. *arXiv:2507.10088*, 2025. 7
- 542
- 543 Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A
544 Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but
545 not perceive. In *ECCV*, 2024. 7
- 546 Víctor Gallego. Refined direct preference optimization with synthetic data for behavioral alignment
547 of llms, 2024. URL <https://arxiv.org/abs/2402.08005>. 9
- 548
- 549 Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang
550 Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for
551 entangled language hallucination and visual illusion in large vision-language models. In *CVPR*,
552 2024. 7
- 553 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xi-
554 aodong Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math
555 dataset. *arXiv:2103.03874*, 2021. 7
- 556 Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza
557 Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Train-
558 ing compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 1
- 559
- 560 Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without
561 reference model, 2024. URL <https://arxiv.org/abs/2403.07691>. 9
- 562 Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang,
563 and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for
564 large video models, 2025. URL <https://arxiv.org/abs/2504.13122>. 9
- 565
- 566 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
567 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv:2410.21276*,
568 2024. 7, 8
- 569 Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child,
570 Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language
571 models. *arXiv preprint arXiv:2001.08361*, 2020. 1
- 572
- 573 Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali
574 Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 7
- 575 Hynek Kydlíček. Math-Verify: Math Verification Library. URL [https://github.com/
576 huggingface/math-verify](https://github.com/huggingface/math-verify). 4
- 577
- 578 Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. Huggingface h4
579 stack exchange preference dataset, 2023. URL [https://huggingface.co/datasets/
580 HuggingFaceH4/stack-exchange-preferences](https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences). 3
- 581
- 582 Hung Le, Dai Do, Dung Nguyen, and Svetha Venkatesh. Reasoning under 1 billion: Memory-
583 augmented reinforcement learning for large language models, 2025. URL [https://arxiv.
584 org/abs/2504.02273](https://arxiv.org/abs/2504.02273). 1
- 585
- 586 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li,
587 Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv:2408.03326*,
588 2024a. 1, 4
- 589
- 590 Shicheng Li, Lei Li, Kun Ouyang, Shuhuai Ren, Yuanxin Liu, Yuanxing Zhang, Fuzheng Zhang,
591 Lingpeng Kong, Qi Liu, and Xu Sun. Temple:temporal preference learning of video llms via
592 difficulty scheduling and pre-sft alignment, 2025a. URL [https://arxiv.org/abs/2503.
593 16929](https://arxiv.org/abs/2503.16929). 9
- 594
- 595 Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang,
596 Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from
597 scratch for frontier vision-language models. *arXiv:2501.14818*, 2025b. 1, 3

- 594 Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu
595 Sun. Preserving diversity in supervised fine-tuning of large language models. *arXiv preprint*
596 *arXiv:2408.16673*, 2024b. 1
- 597
598 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*,
599 2023. 1
- 600
601 Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen.
602 Videodpo: Omni-preference alignment for video diffusion generation, 2024a. URL <https://arxiv.org/abs/2412.14167>. 9
603
- 604
605 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan,
606 Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around
607 player? In *ECCV*, 2024b. 7
- 608
609 Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuan-
610 jun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference opti-
611 mization for large vision-language models, 2024c. URL <https://arxiv.org/abs/2410.17637>. 9
- 612
613 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-
614 Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of
615 foundation models in visual contexts. In *ICLR*, 2023. 7
- 616
617 Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis:
618 Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024.
619 2, 7, 8
- 620
621 Sadegh Mahdavi, Muchen Li, Kaiwen Liu, Christos Thrampoulidis, Leonid Sigal, and Renjie
622 Liao. AoPS dataset: Leveraging online olympiad-level math problems for LLMs training and
623 contamination-resistant evaluation, 2025. URL <https://openreview.net/forum?id=Bgz3okeZ7H>. 3
- 624
625 math-ai. Aime 2024 dataset, 2024. URL <https://huggingface.co/datasets/math-ai/aime24>. American Invitational Mathematics Examination 2024 problems (I & II).
626 7
- 627
628 Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a
629 reference-free reward, 2024. URL <https://arxiv.org/abs/2405.14734>. 9
- 630
631 Tetsuro Morimura, Mitsuki Sakamoto, Yuu Jinnai, Kenshi Abe, and Kaito Ariu. Filtered di-
632 rect preference optimization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.),
633 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*,
634 pp. 22729–22770, Miami, Florida, USA, November 2024. Association for Computational Lin-
635 guistics. doi: 10.18653/v1/2024.emnlp-main.1266. URL <https://aclanthology.org/2024.emnlp-main.1266/>. 2
- 636
637 OpenAI. Introducing gpt-5. OpenAI Blog, August 2025. URL <https://openai.com/index/introducing-gpt-5/>. 7, 9
638
- 639
640 OpenCompass. Aime 2025 dataset, 2025. URL <https://huggingface.co/datasets/opencompass/AIME2025>. American Invitational Mathematics Examination 2025 problems
641 (I & II). 7
- 642
643 Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-
644 language models as a source of preference for fixing hallucinations in llms, 2024. URL
645 <https://arxiv.org/abs/2408.10433>. 9
- 646
647 Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White.
Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint*
arXiv:2402.13228, 2024a. 1, 2

- 648 Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White.
649 Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024b. URL <https://arxiv.org/abs/2402.13228>. 9
650
651
- 652 Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma
653 Gongque, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong,
654 Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-Math: Does your large
655 multimodal model achieve human-like mathematical reasoning? *arXiv:2407.01284*, 2024. 7
- 656 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
657 Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*,
658 2023. 1
- 659 Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Un-
660 intentional unalignment: Likelihood displacement in direct preference optimization, 2025. URL
661 <https://arxiv.org/abs/2410.08847>. 2
662
- 663 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien
664 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a
665 benchmark. *arXiv:2311.12022*, 2023. 7
- 666 Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar,
667 and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning, 2025. URL
668 <https://arxiv.org/abs/2505.23678>. 2
669
- 670 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
671 optimization algorithms. *arXiv:1707.06347*, 2017. 1
672
- 673 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
674 Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathemati-
675 cal reasoning in open language models. *arXiv:2402.03300*, 2024. 1
- 676 Ilya Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal.
677 Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759,
678 2024. 2
- 679 GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang,
680 Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.1V-Thinking: Towards versatile multi-
681 modal reasoning with scalable reinforcement learning. *arXiv:2507.01006*, 2025a. 7, 8, 9
682
- 683 Keye Team, Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song,
684 Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai Keye-VL 1.5 technical report.
685 *arXiv:2509.01563*, 2025b. 9
- 686 Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao,
687 Chuan Yi, Da Li, Dunju Zang, et al. Kwai Keye-VL technical report. *arXiv:2507.01949*, 2025c.
688 2, 4, 9
689
- 690 Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen,
691 Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint*
692 *arXiv:2504.07491*, 2025d. 7
- 693 Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling
694 Rao, Chuan Yi, Da Li, Dunju Zang, Fan Yang, Guorui Zhou, Hao Peng, Haojie Ding, Jiaming
695 Huang, Jiangxia Cao, Jiankang Chen, Jingyun Hua, Jin Ouyang, Kaibing Chen, Kaiyu Jiang,
696 Kaiyu Tang, Kun Gai, Shengnan Zhang, Siyang Mao, Sui Huang, Tianke Zhang, Tingting Gao,
697 Wei Chen, Wei Yuan, Xiangyu Wu, Xiao Hu, Xingyu Lu, Yang Zhou, Yi-Fan Zhang, Yiping Yang,
698 Yulong Chen, Zhenhua Wu, Zhenyu Li, Zhixin Ling, Ziming Li, Dehua Ma, Di Xu, Haixuan
699 Gao, Hang Li, Jiawei Guo, Jing Wang, Lejian Ren, Muhao Wei, Qianqian Wang, Qigen Hu,
700 Shiyao Wang, Tao Yu, Xinchun Luo, Yan Li, Yiming Liang, Yuhang Hu, Zeyi Lu, Zhuoran Yang,
701 and Zixing Zhang. Kwai keye-vl technical report, 2025e. URL <https://arxiv.org/abs/2507.01949>. 9

- 702 Qwen Team. Qwen2.5-vl, January 2025. URL [https://qwenlm.github.io/blog/](https://qwenlm.github.io/blog/qwen2.5-vl/)
703 [qwen2.5-vl/](https://qwenlm.github.io/blog/qwen2.5-vl/). 4
704
- 705 Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdul-
706 mohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2:
707 Multilingual vision-language encoders with improved semantic understanding, localization, and
708 dense features. *arXiv:2502.14786*, 2025. 2, 4
- 709 Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and
710 Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In
711 *NeurIPS*, 2024a. 6
712
- 713 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu,
714 Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng
715 Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s
716 perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b. 4
- 717 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming
718 Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max W.F. Ku, Kai Wang, Alex Zhuang,
719 Rongqi ”Richard” Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging
720 multi-task language understanding benchmark. *arXiv:2406.01574*, 2024c. 7
721
- 722 Emily Wenger. AI produces gibberish when trained on too much AI-generated data. *Nature*, 631
723 (8022):742–743, 2024. 2
- 724 Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and
725 Xiangnan He. Alphadpo: Adaptive reward margin for direct preference optimization, 2025. URL
726 <https://arxiv.org/abs/2410.10148>. 9
727
- 728 xAI. Grok 4: The world’s most advanced ai model with native tool use. [https://x.ai/news/](https://x.ai/news/grok-4)
729 [grok-4](https://x.ai/news/grok-4), 2025. 7
- 730 Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating short-
731 cuts in visual reasoning with reinforcement learning, 2025. URL [https://arxiv.org/](https://arxiv.org/abs/2505.14677)
732 [abs/2505.14677](https://arxiv.org/abs/2505.14677). 2
733
- 734 Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal llm logical reasoning
735 benchmark in visual contexts. *arXiv:2407.04973*, 2024. 7
736
- 737 Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large
738 vision language models via vision-guided direct preference optimization, 2024. URL <https://arxiv.org/abs/2411.02712>. 9
739
- 740 Shusheng Xu, Wei Fu, Jiakuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu,
741 and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint*
742 *arXiv:2404.10719*, 2024. 2
- 743 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
744 Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv:2505.09388*, 2025a. 4
745
- 746 Xiliang Yang, Feng Jiang, Qianen Zhang, Lei Zhao, and Xiao Li. Dpo-shift: Shifting the distribution
747 of direct preference optimization. *arXiv preprint arXiv:2502.07599*, 2025b. 1, 2
748
- 749 Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li,
750 Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding
751 Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong
752 Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL [https://arxiv.org/](https://arxiv.org/abs/2408.01800)
753 [abs/2408.01800](https://arxiv.org/abs/2408.01800). 9
- 754 Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang,
755 and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities.
arXiv:2308.02490, 2023. 7

756 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,
757 Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun,
758 Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and
759 Wenhua Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning
760 benchmark for expert agi. In *CVPR*, 2023. 6

761 Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian
762 Li, Liang Zhao, Lei Li, Hao Tian, et al. MiMo-VL technical report. *arXiv:2506.03569*, 2025. 2,
763 9

764 Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou,
765 Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal llm
766 truly see the diagrams in visual math problems? In *ECCV*, 2024a. 7

767 Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chun-
768 yuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large
769 multimodal models from language model reward. *arXiv:2404.01258*, 2024b. 1

770 Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang,
771 Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning.
772 *arXiv preprint arXiv:2410.16198*, 2024c. 9

773 Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang,
774 Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv:2507.18071*,
775 2025. 1

776 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen
777 Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes
778 for open-source multimodal models. *arXiv:2504.10479*, 2025. 2, 7, 8

779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

A THE USE OF LARGE LANGUAGE MODELS (LLMs)

In this work, LLMs were employed in four main ways: (i) to aid and polish the writing for clarity and style; (ii) to assist in retrieval and discovery of related work; (iii) to support research ideation by suggesting possible directions and problem formulations; and (iv) to provide coding assistance, including code generation, debugging, and optimization suggestions.

All outputs were critically reviewed, verified, and refined by the authors. The core scientific ideas, methodology, and contributions remain entirely the authors' own.

B PROOF OF DISCRIMINATIVE DIRECT PREFERENCE OPTIMIZATION

Given a LLM policy π_θ to be optimized, there exists prompt x and a generated candidate sequence y , we denote the length-normalized log-probability as

$$\ell_\theta(y|x) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_\theta(y_t | x, y_{<t}),$$

for a sampled group $\{y_i\}_{i=1}^G$, each sequence receives a reward $r_i = r(x, y_i)$, which is then normalized within the group as

$$\hat{A}_i = \frac{r_i - \bar{r}}{s}, \quad \bar{r} = \frac{1}{G} \sum_{j=1}^G r_j, \quad s = \sqrt{\frac{1}{G} \sum_{j=1}^G (r_j - \bar{r})^2}.$$

The GSPO objective is given by

$$J_{\text{GSPO}}(\theta) = \mathbb{E}_{x, \{y_i\} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G \min \left(s_i(\theta) \hat{A}_i, \text{clip}(s_i(\theta), 1 - \varepsilon, 1 + \varepsilon) \hat{A}_i \right) \right],$$

where $s_i(\theta) = \exp(\ell_\theta(y_i|x) - \ell_{\theta_{\text{old}}}(y_i|x))$ denotes the sequence-level importance ratio.

Specifically, we consider a case of group size is restricted to two and reward values are binary, i.e., $G = 2$, and $r_i \in \{0, 1\}$, containing a positive instance y^+ and a negative instance y^- with corresponding rewards r^+ and r^- , the normalized advantage difference simplifies to

$$\hat{A}^+ - \hat{A}^- = \frac{r^+ - r^-}{s}, \quad s = \frac{|r^+ - r^-|}{2},$$

which further reduces to

$$\hat{A}^+ - \hat{A}^- = 2 \frac{r^+ - r^-}{|r^+ - r^-|} = 2 \text{sign}(r^+ - r^-).$$

Thus, in our binary case, the advantage difference degenerates to the **sign of the reward difference**.

Since the sign function is non-differentiable, we introduce a smooth approximation:

$$\text{sign}(z) \approx 2\sigma(\alpha z) - 1, \quad \sigma(z) = \frac{1}{1 + e^{-z}}, \quad \alpha > 0.$$

Consequently,

$$\hat{A}^+ - \hat{A}^- \approx 4\sigma(\alpha(r^+ - r^-)) - 2.$$

Up to scaling constants, which do not affect the optimum in expectation, we obtain the core approximation:

$$\hat{A}^+ - \hat{A}^- \propto \sigma(\alpha(r^+ - r^-)).$$

To connect rewards with the policy, we adopt the concept of *language models are reward models* from DPO algorithm. To reflect the binary nature of the reward, we use a sigmoid function to map the log-ratio between the policy and the reference model into the binary classification domain:

$$r(x, y) = \sigma(\Delta(x, y)), \quad \Delta(x, y) = \ell_\theta(y|x) - \ell_{\text{ref}}(y|x).$$

Here, the reward is interpreted as the probability of correctness, modeled as a sigmoid transformation of the log-ratio between the policy and a reference model. The margin between positive and negative sequences is then

$$d = r^+ - r^- = \sigma(\Delta^+) - \sigma(\Delta^-).$$

Then, we can reformulated the GSPO objective as

$$J_{\text{GSPO}}(\theta) = \mathbb{E}[(\hat{A}^+ - \hat{A}^-) \cdot f(\theta)].$$

In our reformulation, the GSPO objective naturally connects to the margin-based classification paradigm underlying support vector machines (SVMs). Specifically, the difference term $\hat{A}^+ - \hat{A}^-$ functions as a margin signal that determines the correct ordering between y^+ and y^- . In the classical SVM framework, such a margin is enforced through the hinge loss,

$$\ell_{\text{hinge}}(y^+, y^-) = \max(0, 1 - (\hat{A}^+ - \hat{A}^-)),$$

which penalizes violations of the desired ranking $y^+ \succ y^-$. However, the hinge loss is non-smooth and thus less amenable to gradient-based optimization in large-scale neural models. To address this, we adopt a smooth surrogate analogous to replacing the hinge loss with a softplus/logistic formulation, yielding

$$L_{\text{DDPO}}(\theta) = -\log \sigma(\alpha(\sigma(\Delta^+) - \sigma(\Delta^-))).$$

This objective defines **Discriminative Direct Preference Optimization (DDPO)**.

Let’s see the case of Direct Preference Optimization (DPO), the objective is defined as

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(\beta(\Delta^+ - \Delta^-))],$$

where $\Delta^+ = \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}$ and $\Delta^- = \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$. The gradients with respect to the transformed logits are

$$\frac{\partial L}{\partial \Delta^+} = -\beta(1 - \sigma(\beta(\Delta^+ - \Delta^-))), \quad \frac{\partial L}{\partial \Delta^-} = \beta(1 - \sigma(\beta(\Delta^+ - \Delta^-))).$$

While the expressions exhibit a superficially consistent sign pattern—negative for the preferred response and positive for the dispreferred response—the directional signal is only preserved relative to the difference $(\Delta^+ - \Delta^-)$. Specifically, the objective is invariant under additive shifts, i.e., for any constant c , replacing $\Delta^+ \mapsto \Delta^+ + c$ and $\Delta^- \mapsto \Delta^- + c$ leaves both the loss and its gradients unchanged. This additive shift invariance implies that the optimization dynamics depend solely on relative differences and are insensitive to the absolute scale of the logits.

Such invariance introduces non-identifiability in the learned implicit reward: the model cannot distinguish between reward functions that differ by a constant offset. In practice, this can lead to gradient degeneracy, particularly in binary reward settings. For example, simultaneously increasing or decreasing both Δ^+ and Δ^- has no effect on the loss, producing vanishing or ambiguous gradient signals and potentially stalling optimization. Thus, although the partial derivatives of DPO appear to have fixed signs, they do not exhibit true sign-preserving behavior in an absolute sense, but only in a relative one.

By contrast, Discriminative Direct Preference Optimization (DDPO) introduces a nonlinear transformation prior to differencing, yielding the objective

$$L_{\text{DDPO}}(\theta) = -\log \sigma(\alpha(\sigma(\Delta^+) - \sigma(\Delta^-))),$$

with gradients

$$\frac{\partial L}{\partial \Delta^+} = -\alpha \sigma'(\Delta^+) \sigma(-\alpha d), \quad \frac{\partial L}{\partial \Delta^-} = \alpha \sigma'(\Delta^-) \sigma(-\alpha d),$$

918 where $d = \sigma(\Delta^+) - \sigma(\Delta^-)$. Since $\sigma'(\cdot) > 0$ and $\sigma(-\alpha d) > 0$, the gradient signs are
919 strictly preserved— Δ^+ is always pushed upward, and Δ^- always downward—independent of ad-
920 ditive shifts. This design eliminates the degeneracy observed in DPO and guarantees stable, non-
921 degenerate updates even under binary reward signals.

922 923 C TRAINING DETAILS 924

925 All experiments were conducted on 4 machines, each equipped with 8 A100-80GB GPUs, for a total
926 of 32 GPUs. The training framework leveraged **Accelerate** with **DeepSpeed ZeRO-2**, employing
927 **bf16 mixed precision**, **FlashAttention-2**, and distributed data parallelism. Unless otherwise spec-
928 ified, the maximum input length was set to 8K tokens, and **neat packing** was enabled to improve
929 batch efficiency. The optimizer used was **AdamW**, with a **cosine decay** learning rate schedule, a
930 warmup ratio of 0.1, and gradient clipping set to 1.0. The per-GPU batch size was 1, with gra-
931 dient accumulation over 4 steps, resulting in an effective global batch size of 128. The specific
932 configurations for each training stage are as follows:

- 933 • **Stage 1 (Vision–Language Alignment):** The language model and vision encoder were frozen,
934 while only the projection layer and positional encodings were trained for 1 epoch. The maxi-
935 mum sequence length was 8K, and the learning rate was set to $5e-5$.
- 936 • **Stage 2 (Instruction Fine-tuning):** All parameters were unfrozen and trained for 1 epoch with
937 a learning rate of $5e-5$. Gradient checkpointing was enabled.
- 938 • **Stage 3 (Annealed Rejection Sampling Fine-tuning):** Training was conducted on the hard-
939 case subset \mathcal{H} , with the sequence length extended to 16K. The learning rate was reduced to $5e-6$
940 for annealing, while other configurations remained consistent with the previous stages.
- 941 • **Stage 4 (DDPO):** LoRA-based parameter-efficient fine-tuning was applied to the projection
942 layer and feed-forward layers. α of DDPO set to 1.0. The frozen base model was used as the
943 reference policy. Training lasted 1 epoch with a learning rate of $5e-5$, a maximum sequence
944 length of 8K, and the use of memory-optimized gradient checkpointing from Unsloth.
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

Category	Datasets
Math & Geometry QA	CLEVR-Math, GEOS, GeoQA, Geometry3K, IconQA, MapQA, Super-CLEVR, TabMWP, UniGeo, Geo170k, geo3k, GeomVerse, MM-MathInstruct, MMathCoT, multimath-300k, geoqa-plus, gqa
Chart & Table QA	ChartQA, Chart2Text, DVQA, HiTab, LRV-Chart, LRV-Normal-filtered, TabMWP, ROBUT-WTQ, ROBUT-WikiSQL, ROBUT-SQA, FigureQA
Diagram & Science QA	AI2D, Diagram-Image-to-Text, ScienceQA, Infographic-VQA, Infographic, InterGPS, Raven, MMSci, Gaokao-Datasets, K12-VQA, MM-Chem-Exam
General & Perception VQA	A-OKVQA, Visual7W, VizWiz, ST-VQA, TallyQA, VSR, Screen2Words, VisText, VisualMRC, VQA-RAD, SROIE, PMC-VQA, Infographic-VQA-LLaVA, HateMemes, WebSight, GUI-Grounding, ShareGPT4o, ShareGPT4V-COCO, ShareGPT4V-Knowledge, ShareGPT4V-LLaVA, ShareGPT4V-SAM, LLaVA-Wild-12K, LLaVA-Wild-39K, LLaVAR-GPT4-20K, VisionFlan-filtered, Mulberry, VisualWebInstruct
Document & OCR	Chrome-Writing, ORAND-Car-A, UReader-CAP, UReader-IE, Docmatix-Multi-20k-en, TextOCR, TextCaps, IAM, IIT5K, Rendered-Text, HME100K
Image Caption for Pretrain	Allava-Instruct-VFlan, Allava-Instruct-Laion, Allava-Instruct-Chinese, LLaVA-ReCap-CC3M
Pure-text QA	WildChat, MegaScience, Hermes3, ToolAce, AOPS-Instruct-filtered

Table 6: Composition of the **TinyEye-Instruct** dataset pool. We organize datasets into categories including math & geometry QA, chart & table QA, diagram & science QA, General & Perception VQA, Document & OCR, Image Caption for Pretrain, and Pure-text QA.

Category	Datasets
Pure-text Math	AOPS-Instruct-filtered, OpenThoughts, Calculus, MATH, AIME1983-2023
Pure-text Game & Puzzle	24Points, Chinese-chess, Chinese-go, Japanese-shogi, Chinese-riddle, English-riddle, corporate-governance-reasoning, RelatLogic-Reasoning, Reasoning-gym, Stocks
Pure-text STEM	Chemistry-molecule-MCQ, Chemistry-reaction-MCQ, StackExchange-longtail, MegaScience, WildSci, Philosophy-Ethics-Morality, quantitative-finance-reasoning, Pediatric-medical-reasoning, counterfactual-history-reasoning, academic-chain, OlympiadExtended
General VQA	chartqa, CLEVR, coco, docvqa, ocr-vqa, pisc, sam, share-textvqa, sqa, textvqa, vg, web-celebrity, web-landmark, wikiart, A-OKVQA, multihiertt, DVQA, infovqa, IRV-chart, plotQA, PMC-VQA, TQA, VizWiz, VQA-AS, VQA-RAD, VQA2.0, chart2Text, CharQA, RoBUT-SQA, sharegpt4o, sharegpt4v, visual7w, ArxivQA, EXAMS-V, GVLQA, OpenSpaces, Spacellava, VisualWebInstruct, FigureQA
Multi-modal Math	CLEVR-Math, GEOS, GeoQA, Geometry3K, IconQA, MapQA, Super-CLEVR, TabMWP, UniGeo, Geo170k, geo3k, GeomVerse, MM-MathInstruct, MMathCoT, multimath-300k, geoqa-plus, gqa
Multi-modal Game & Puzzle	VisualPuzzles, seekworld, PuzzleVQA, AlgoPuzzleVQA, GameQA, VisualSphinx, Zebra-CoT
Multi-modal STEM	scienceqa, ai2d, MMK12, MM-Chem-Exam, K12-VQA

Table 7: Composition of the **TinyEye-Reason** dataset pool. We organize datasets into categories including Pure-text Math, Pure-text STEM, Pure-text STEM, General VQA, Multi-modal Math, Multi-modal Game & Puzzle, and Multi-modal STEM.

D DATASET DETAILS

Here we list the details of datasets we use in Table 6 and Table 7.

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

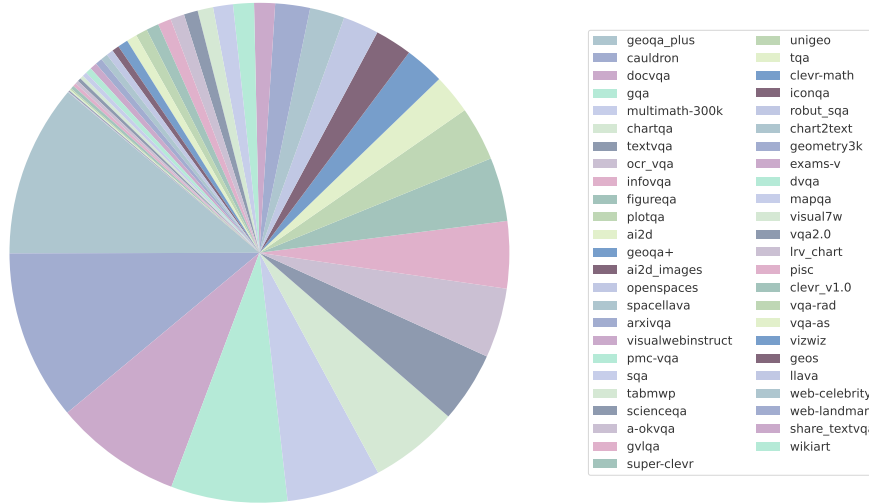


Figure 3: The distribution of the category of General VQA datasets inside **TinyEye-Reason**

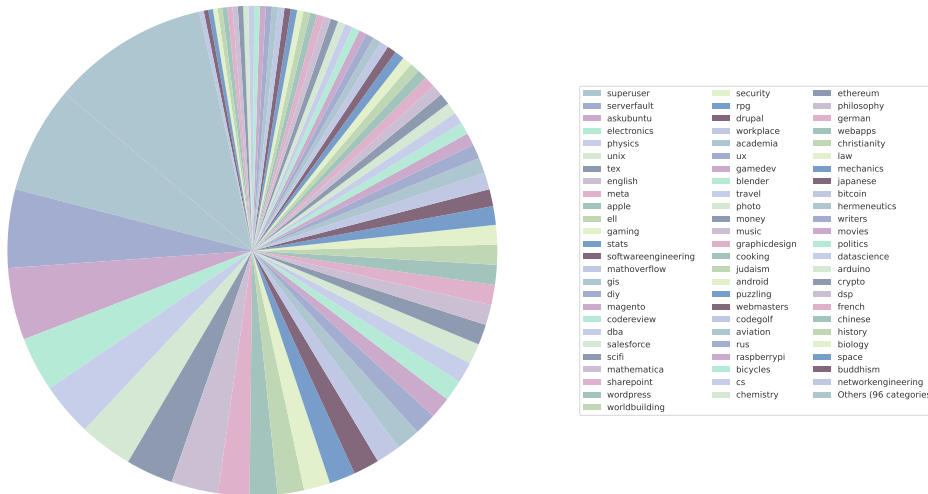


Figure 4: The distribution of the topics of longtail domain QA from StackExchange-longtail dataset inside **TinyEye-Reason**

1080 E PSEUDO-CODES

1081

1082

1083 **Algorithm 1:** ISTANDEMREPEATATK(S, k, n, m)

1084

1084 **Input:** string S of length L , step $k \in \mathbb{N}$, repeats $n \geq 2$, probes m

1085

1085 **Output:** True/False

1086

1086 **if** $k < 1$ **or** $n < 2$ **or** $m < 1$ **or** $m > k - 1$ **or** $k > \lfloor L/n \rfloor$ **then return False;**

1087

1087 $J \leftarrow \text{SAMPLEWITHOUTREPLACEMENT}(\{1, 2, \dots, k - 1\}, m);$

1088

1088 **foreach** $j \in J$ **do**

1089

1089 $v \leftarrow S[L - j]$ **for** $i \leftarrow 1$ **to** $n - 1$ **do**

1090

1090 \quad **if** $S[L - j - k \cdot i] \neq v$ **then**

1091

1091 $\quad \quad$ **return False**

1092

1092 **return True**

1093

1094

1094 **Algorithm 2:** ISTANDEMREPEAT(S, k_{\min}, n, m)

1095

1095 **Input:** string S of length L , min step k_{\min} , repeats $n \geq 2$, probes m

1096

1096 **Output:** True/False

1097

1098

1098 $K_{\max} \leftarrow \left\lfloor \frac{L}{n} \right\rfloor;$

1099

1099 **for** $k \leftarrow k_{\min}$ **to** $K_{\max} - 1$ **do**

1100

1100 \quad **if** ISTANDEMREPEATATK(S, k, n, m) **then**

1101

1101 $\quad \quad$ **return True**

1102

1102 **return False**

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

```

1134 Algorithm 3: Tandem Repeat Avoidance Sampling with LRU Mechanism
1135
1136 Input: Current token sequence  $seq$ , logits  $scores$ 
1137 Output: Updated logits  $scores$ 
1138  $p \leftarrow \text{DETECTTANDEMREPEATTAIL}(seq, k_{\min}, k_{\max}, n)$ ;
1139 if  $p \neq \emptyset$  then
1140   if  $p \notin \text{LRU buffer}$  then
1141     insert  $p$  into LRU buffer (as most recent);
1142     if  $\text{size}(\text{LRU buffer}) > \text{max capacity}$  then
1143       | evict least-recently-used pattern from LRU buffer;
1144   else
1145     | move  $p$  to most recent position in LRU buffer;
1146    $B \leftarrow \emptyset$ ; // tokens to ban
1147   foreach pattern  $p$  in LRU buffer do
1148      $m \leftarrow \text{LONGESTPREFIXMATCHLEN}(p, seq)$ ;
1149     if  $m > 0$  then
1150       | move  $p$  to most recent position in LRU buffer; // refresh recency
1151     if  $0 < m < |p|$  then
1152       |  $B \leftarrow B \cup \{p[m]\}$ ; // ban next token of prefix
1153     if  $m = |p|$  then
1154       |  $B \leftarrow B \cup \{p[0]\}$ ; // ban restart head
1155   if  $B \neq \emptyset$  then
1156     if top- $k$  constraint enabled then
1157       |  $B \leftarrow B \cap \text{TOPK}(scores)$ 
1158     if  $B \neq \emptyset$  then
1159       if hard ban then
1160         |  $scores[B] \leftarrow -\infty$ 
1161       else
1162         |  $scores[B] \leftarrow scores[B] - \lambda$ 
1163     return  $scores$ ;

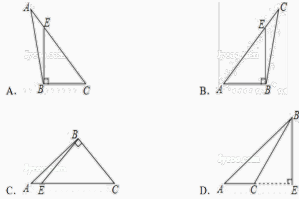
```

1167 F CASE STUDY

1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

9. 下列四个图形中, 线段 BE 是 $\triangle ABC$ 的高的是()



Solve this problem.

To determine which figure shows that line segment BE is the height of $\triangle ABC$, we need to refer to the **definition of the height of a triangle**: the height is the perpendicular segment drawn from a vertex to the opposite side (or its extension), with the foot of the perpendicular lying on the opposite side.

Analysis of each option:

- Option A: In $\triangle ABC$, point E lies on AC , but BE is not perpendicular to AC (there is no right angle mark indicating perpendicularity), so BE is not the height.
- Option B: $\angle ABC = 90^\circ$ (right angle), indicating $AB \perp BC$, but BE is not necessarily perpendicular to AC , and the position of E does not satisfy the condition of "perpendicular to the opposite side," so BE is not the height.
- Option C: There is no indication that BE is perpendicular to AC (no right angle mark), so BE is not the height.
- Option D: Point E lies on AC , and $BE \perp AC$ (indicated by the right angle mark at E), meaning BE is the perpendicular segment from vertex B to side AC , satisfying the definition of the height of a triangle.

D

<think>

The current task for the user is to determine which figure shows that line segment BE is the height of triangle ABC . First, recall the definition of a height in a triangle: the height is the perpendicular segment drawn from a vertex to the opposite side (or its extension), with the foot of the perpendicular being on the opposite side. Now, let's examine each option:

Option A: In triangle ABC , point E is on AC , and BE is not perpendicular to AC , so BE is not the height.

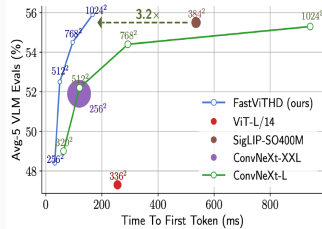
Option B: In triangle ABC , point B is a right angle, meaning $\angle ABC=90^\circ$, so $AB \perp BC$. However, BE is not necessarily perpendicular to AC , and the position of E does not indicate that BE is perpendicular to AC , so it is not the height.

Option C: In triangle ABC , point E is on AC , but there is no indication that BE is perpendicular to AC , so it is not the height.

Option D: In triangle ABC , point E is on AC , and BE is perpendicular to AC (as indicated by the right angle mark at E), so BE is the perpendicular segment from vertex B to side AC , making BE the height of triangle ABC .

Therefore, the correct option is D .

</think>



What does '3.2x' mean in the chart?

To determine the meaning of "3.2X" in the figure, it is necessary to analyze the annotations in the chart:

- Locate the annotation "3.2X": In the figure, "3.2X" is marked next to a data point, indicating that the value corresponding to this point is related to another point.
 - Observe the relationship between data points: The green line (ConvNext-L) and the blue line (FastViT-HD) have corresponding data points. By comparing the values of these two lines, it can be found that the value of the green line's data point is approximately 3.2 times the value of the blue line's data point (for example, if the blue line corresponds to a value of 256, the green line corresponds to a value of 768, and $768 \div 256 \approx 3$), which may be an approximation, but the annotation "3.2X" clearly indicates that the value of this point is 3.2 times the value of another point).
- Therefore, "3.2X" in the figure indicates that the value corresponding to this data point is 3.2 times the value of another data point.

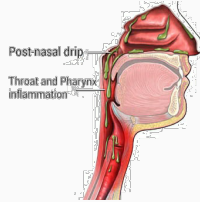
The value corresponding to this data point is 3.2 times the value of another data point

<think>

The user now needs to understand the meaning of "3.2X" in the chart. First, look at the annotation in the chart: "3.2X" is marked next to a point, indicating that the value corresponding to this point is 3.2 times the value of another point. For example, looking at the green line (ConvNext-L), the point corresponding to "3.2X" should be compared with the point on the blue line (FastViT-HD). By observing the values, the point on the blue line is approximately 256, and the point on the green line is approximately 768, so $768/256 = 3$, which might be an approximation, but the annotation "3.2X" indicates that the value of this point is 3.2 times the value of another point. Therefore, "3.2X" means that the value corresponding to this point is 3.2 times the value of another point.

</think>

1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

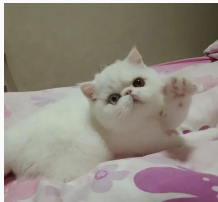


Which hospital department should I go to for an appointment?

To determine which hospital department to visit for an appointment, we analyze the information in the image. The image highlights “Throat and Pharynx inflammation,” which refers to conditions affecting the throat and pharynx. In hospitals, the **ENT (Ear, Nose, and Throat)** department specializes in diagnosing and treating conditions related to these areas.

<think>
The user now needs to determine which department to visit for an appointment. First, look at the image, which shows "Throat and Pharynx inflammation," indicating that the issue involves the throat and pharynx. Therefore, it is necessary to find a department related to the throat and pharynx. Typically, in hospitals, ENT (Ear, Nose, and Throat) departments handle such issues. Thus, the corresponding department should be the ENT department.
</think>

ENT (Ear, Nose, and Throat) department



Describe this image.

The image depicts a white cat with large, round eyes and fluffy fur, lying comfortably on a bed. The cat's body is oriented diagonally across the frame, with its head raised slightly as if looking directly at the camera. Its front paws are extended outward, giving a sense of curiosity or playfulness. The cat rests on a blanket with a pink floral pattern, which contrasts subtly against the neutral-toned wall in the background. The overall atmosphere is cozy and intimate, capturing a moment of rest in a domestic setting.

<think>
Okay, let me start by breaking down the question. The user wants a detailed description of the image based on the provided image. The key points mentioned are the white cat with large eyes and fluffy fur, lying on a pink-patterned blanket. It's in a cozy indoor setting. First, I need to focus on the main subject: the cat. The color is white, which is a primary color, so that's straightforward. The mention of "large eyes" and "round pupils" suggests it's looking directly at something or someone, maybe the camera. The fluffy fur adds texture and detail, making the cat look soft. The position on the blanket indicates comfort, perhaps indicating a relaxed environment. The background is simple, neutral tones, which helps the cat stand out. The pink pattern on the blanket adds a touch of color contrast, making the scene more visually appealing. The overall impression is one of warmth and coziness, typical of a pet in a home setting. I should make sure not to include any assumptions beyond what's given. The image doesn't mention any other animals or objects, so I can stick to the cat and the blanket. No need for extra details like the time of day unless specified, but the setting is indoors and neutral, so daylight isn't an issue here. The main elements are the cat, its features, the blanket, and the background. I need to present these in a clear, logical order, starting with the subject (the cat), then its physical characteristics, followed by the setting (blanket and background). Make sure each part flows naturally into the next without being redundant. Avoid any sensory language unless necessary, but since the question is about describing, including colors and textures would be helpful. Double-check that all key points from the original image are covered: white fur, large eyes, fluffy, lying on pink-patterned blanket, indoor, neutral background. Yes, that's all there. Alright, time to put it all together in a coherent description.
</think>

1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

Direct materials	\$15,000
Direct labor	25,000
Factory depreciation expense	45,000
Factory utilities expense	2,000
Payroll staff's salary	15,000

Extract texts from this image.

<think>

Okay, let's see. The user wants me to extract the text from the provided image. The image shows a list of costs related to manufacturing. The items are: Direct materials at \$15,000, Direct labor \$25,000, Factory depreciation expense \$45,000, Factory utilities expense \$2,000, and Payroll staff's salary \$15,000. I need to make sure I copy each line exactly as presented. Let me check each entry again to avoid any mistakes. Direct materials is 15k, then direct labor 25k, followed by factory depreciation 45k, factory utilities 2k, and payroll staff's salary 15k. No commas or extra spaces mentioned in the original text, so just copy them verbatim. Alright, that should be it.

</think>

The extracted text from the image is as follows:

Direct materials - \$15,000
Direct labor - \$25,000
Factory depreciation expense - \$45,000
Factory utilities expense - \$2,000
Payroll staff's salary - \$15,000