ENERGYOPD: ENERGY-BASED ON-POLICY DISTILLATION FOR STABLE AND GENERALIZABLE KNOWLEDGE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has become a mainstream knowledge transfer paradigm for large language models (LLMs), alleviating the exposure bias of off-policy distillation through interactive student rollouts and real-time teacher supervision. However, existing OPD objectives are almost exclusively instantiated as Kullback–Leibler (KL) divergences and related statistical distribution metrics, which suffer from two critical defects. First, single divergence-based losses are highly sensitive to local token noise and trajectory drift, causing severe gradient oscillation and unstable convergence. Second, the rigid distribution-matching framework cannot flexibly integrate heterogeneous supervision signals such as execution outcomes and task rewards, limiting scenario adaptability and out-of-distribution (OOD) generalization. In this paper, we reconstruct on-policy distillation from an energy-based modeling perspective and propose **EnergyOPD**, a novel energy-based on-policy distillation framework. We model the teacher LLM as a task-oriented energy function and reformulate OPD as fitting the student energy landscape to the teacher energy landscape, deriving a generalized free-energy distillation objective that unifies reverse KL, forward KL, and Jensen–Shannon variants within a single framework. We further propose an energy landscape smoothing regularizer that provably reduces gradient variance and improves OOD generalization, together with a universal multi-signal fusion mechanism that integrates token-level teacher feedback, trajectory execution results, and task rewards into unified energy terms. Experiments on six mathematical reasoning benchmarks across three student scales (Qwen3-0.6B/1.7B/4B-Base), as well as out-of-domain evaluations, show that EnergyOPD consistently improves both average Avg@8 and Pass@8 over strong on-policy distillation baselines, while substantially reducing training instability and achieving higher out-of-distribution accuracy, demonstrating stable, generalizable, and scalable on-policy knowledge transfer.

1 INTRODUCTION

Large language models have achieved remarkable success in natural language understanding, logical reasoning, program synthesis, and agentic decision making (OpenAI et al., 2023; Yang et al., 2025; Grattafiori et al., 2024). As parameter counts and pre-training scale continue to grow, high-performance LLMs face prominent deployment bottlenecks, including high computational cost, large memory footprints, and high inference latency. Knowledge distillation (Hinton et al., 2015; Buciluǎ et al., 2006) is a core model-compression technology that transfers the implicit knowledge and generalization capabilities of a large teacher model to a compact student model, effectively balancing task performance and deployment efficiency. On-policy distillation (OPD) (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025) is an advanced distillation paradigm developed to resolve the exposure bias inherent in off-policy distillation (Kim & Rush, 2016; Bengio et al., 2015): the student autonomously generates rollout trajectories during training, and the teacher provides dense token-level supervision conditioned on student-generated contexts, thereby avoiding the error accumulation caused by the mismatch between training and inference distributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite this progress, existing OPD methods exhibit two core limitations that restrict practical deployment. First, nearly all current OPD objectives adopt reverse KL divergence or its clipped surrogate as the optimization target (Gu et al., 2024; Lu & Thinking Machines Lab, 2025). Such point-to-point distribution matching is highly sensitive to local token noise and slight trajectory drift, causing frequent gradient jitter and severe training oscillation. Recent analysis further shows that the mode-seeking property of reverse KL collapses generation diversity and destabilizes learning precisely at high-entropy teacher positions (Jin et al., 2026). Second, the divergence-based framework is a single-signal, token-level statistical matching paradigm with poor extensibility. It accepts only teacher token probabilities as supervision and cannot fuse heterogeneous feedback such as trajectory execution correctness, verifiable task rewards, or rule-based constraints, which severely limits OPD in complex reasoning and agentic scenarios.

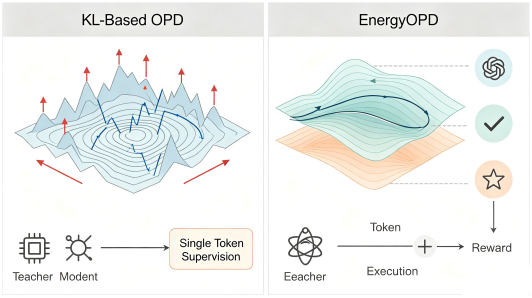

Figure 1: Motivation of EnergyOPD. Divergence-based OPD (left) is unstable and cannot fuse extra signals; energy-based OPD (right) achieves stable, multi-signal distillation via landscape fitting.

Existing remedies, including alternative divergences (Wen et al., 2023; Ko et al., 2024), entropy-aware objective switching (Jin et al., 2026), entropy bonuses and advantage shaping (Schulman et al., 2017; Cheng et al., 2025), trajectory filtering, and gradient clipping, provide partial relief but remain heuristic patches within the statistical divergence paradigm. None of them offers a principled reconstruction of the distillation objective itself, and none provides a unified interface for multi-source supervision. This observation motivates our central question: *can we re-found on-policy distillation on a modeling framework in which stability, generalization, and multi-signal supervision arise naturally rather than being bolted on?* Energy-based models (EBMs) (LeCun et al., 2006), which represent distributions through unnormalized scalar energy functions, offer exactly such a foundation. Energy functions compose additively across heterogeneous constraints, and the geometry of the energy landscape directly governs optimization stability and generalization.

Building on this insight, we propose **EnergyOPD**, an energy-based on-policy distillation framework. As shown in Figure 1, we model the fixed teacher as a task-oriented trajectory energy function and parameterize the student as a learnable energy function, recasting OPD as fitting the student energy landscape to the teacher's. From this formulation we derive a generalized free-energy distillation objective and prove that classic OPD variants, including reverse KL, forward KL, and Jensen–Shannon, are special cases under specific energy instantiations. To address training instability, we introduce an energy landscape smoothing regularizer that constrains the local continuity of the student energy over perturbed neighborhoods of sampled trajectories, provably bounding gradient variance and enlarging the effective generalization region. Finally, we design a multi-signal fusion mechanism that decomposes the supervising energy into token-level, execution-level, and reward-level sub-energies, allowing verifiable rewards and execution feedback to be incorporated into distillation without changing the optimization machinery.

We evaluate EnergyOPD on six mathematical reasoning benchmarks with three Qwen3 student scales, following the protocol of (Jin et al., 2026), as well as on out-of-domain general reasoning and instruction-following benchmarks. EnergyOPD consistently outperforms the strongest on-policy baselines in both average Avg@8 and Pass@8, reduces training gradient oscillation amplitude, and achieves higher average OOD accuracy. Our contributions are fourfold: (1) a generalized energy-based theoretical framework that unifies mainstream OPD objectives; (2) an energy landscape smoothing regularizer with stability and generalization guarantees; (3) a universal multi-signal energy fusion mechanism extending OPD beyond token-probability supervision; and (4) extensive experiments and ablations validating each component.

2 Related Work

Knowledge Distillation and On-Policy Distillation. Since the seminal formulation of distillation as soft-label matching (Hinton et al., 2015; Buciluǎ et al., 2006), a rich line of work has adapted the

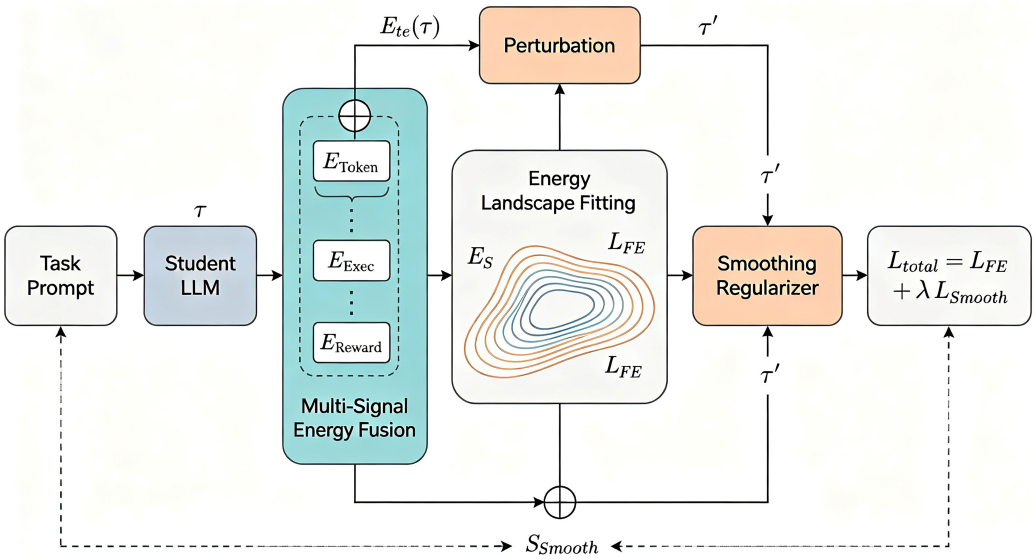

Figure 2: Overview of our EnergyOPD framework, an energy-based on-policy distillation paradigm for student LLM training. Given a task prompt, the student LLM generates trajectory rollout τ. The trajectory is fed into two parallel branches. The main branch computes composite teacher energy $E_{\text{te}}(\tau)$ via multi-signal energy fusion integrating token-level supervision E_{Token}, execution verification energy E_{Exec}, and task reward energy E_{Reward}. Energy Landscape Fitting aligns the student energy $E_S(\tau)$ with the teacher energy surface to optimize the free-energy loss $\mathcal{L}_{\text{FE}}$. The perturbation branch samples locally perturbed neighbor trajectory τ' and computes the smoothing regularizer loss $\mathcal{L}_{\text{Smooth}}$. The total training objective $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FE}} + \lambda\mathcal{L}_{\text{Smooth}}$ aggregates both losses, and gradients are backpropagated iteratively to update the student LLM. Solid arrows denote feed-forward data flow, and dashed arrows represent gradient backpropagation.

idea to autoregressive language models, including sequence-level distillation (Kim & Rush, 2016), layer and attention transfer (Sanh et al., 2019; Wang et al., 2020), and generalized divergence families that interpolate between mode-seeking and mode-covering behavior (Wen et al., 2023; Ko et al., 2024; Shing et al., 2025). The on-policy family emerged from the observation that students trained purely on teacher trajectories suffer compounding errors at inference time (Bengio et al., 2015; Ranzato et al., 2015); GKD (Agarwal et al., 2024), MiniLLM (Gu et al., 2024), speculative interleaving (Xu et al., 2025), and PPO-style clipped reverse-KL distillation (Lu & Thinking Machines Lab, 2025) progressively refined the recipe. Most recently, EOPD (Jin et al., 2026) identified the diversity collapse and gradient instability of reverse KL at high-entropy teacher positions and proposed entropy-gated mixing of forward and reverse KL. Our work is complementary but more fundamental in scope: rather than switching between divergences according to a hand-designed entropy criterion, we derive a single generalized objective from energy-landscape fitting, of which such divergence mixtures are recoverable special cases, and we additionally address the multi-signal supervision problem that divergence-based frameworks cannot express.

Energy-Based Models in NLP. Energy-based models describe data distributions through unnormalized scalar energy functions (LeCun et al., 2006), converting density estimation into energy landscape shaping. In NLP, EBMs have been applied to globally normalized sequence generation via residual energy models (Deng et al., 2020), to controllable text generation where attribute constraints are expressed as additive energy terms (Qin et al., 2022), and to joint classifier–generator training (Grathwohl et al., 2020). A key appeal of EBMs is compositionality: heterogeneous constraints combine by simple energy addition, and smoothness properties of the energy landscape connect directly to robustness under distribution shift. However, EBM machinery has scarcely been explored for LLM distillation, and existing energy-flavored distillation work is confined to off-policy static settings that cannot accommodate interactive rollouts. To our knowledge, EnergyOPD is the first framework to ground *on-policy* distillation in energy-based modeling, using free-energy fitting as the training objective and energy composition as the supervision interface.

Stable and Generalizable Post-Training. Training instability and weak OOD generalization are recurring bottlenecks in LLM post-training. In RL-based post-training, stabilization is typically pursued through clipped surrogate objectives (Schulman et al., 2017), group-relative advantage normalization (Shao et al., 2024), and dynamic sampling (Yu et al., 2025); exploration and diversity are encouraged through entropy bonuses and entropy-shaped advantages (Cheng et al., 2025; Wang et al., 2025). In distillation, analogous heuristics include importance-weight clipping (Lu & Thinking Machines Lab, 2025) and top-k truncated forward KL (Jin et al., 2026; Peng et al., 2024). These techniques operate on the symptoms of instability at the level of gradients or sampling. In contrast, we target its cause at the level of the model's energy geometry: by directly regularizing the smoothness of the student energy landscape, EnergyOPD obtains endogenous stability and generalization guarantees rather than post-hoc gradient surgery, echoing the connection between loss-landscape flatness and generalization observed in the optimization literature.

3 Preliminaries

3.1 KL-Based On-Policy Distillation

Let $\pi_{\mathrm{te}}(\cdot \mid \mathbf{c})$ denote a fixed teacher LLM and $\pi_\theta(\cdot \mid \mathbf{c})$ a student LLM with trainable parameters θ, both defining next-token distributions over a vocabulary $\mathcal{V}$ given context $\mathbf{c}$. Given a prompt $\mathbf{q} \sim \mathcal{D}$, the student generates a rollout $\mathbf{x} = (x_1, \ldots, x_{|\mathbf{x}|})$ with per-step context $\mathbf{c}_t = (\mathbf{q}, x_{<t})$. Standard OPD (Gu et al., 2024; Lu & Thinking Machines Lab, 2025) minimizes the on-policy reverse KL $\mathcal{L}^{\mathrm{RKL}}(\theta)$,

$$\mathbb{E}_{\mathbf{q}\sim\mathcal{D}}\,\mathbb{E}_{\mathbf{x}\sim\pi_\theta(\cdot|\mathbf{q})}\left[\frac{1}{|\mathbf{x}|}\sum_{t=1}^{|\mathbf{x}|}\mathrm{KL}\big(\pi_\theta(\cdot \mid \mathbf{c}_t)\,\|\,\pi_{\mathrm{te}}(\cdot \mid \mathbf{c}_t)\big)\right], \tag{1}$$

typically implemented with a single-sample Monte Carlo estimate and PPO-style importance clipping (Schulman et al., 2017; Lu & Thinking Machines Lab, 2025). The reverse KL is mode-seeking: it concentrates student mass on dominant teacher modes but ignores unvisited modes, which has been shown to collapse diversity and destabilize gradients at high-entropy teacher positions (Jin et al., 2026). Moreover, Eq. equation 1 admits only teacher token probabilities as supervision.

3.2 Energy-Based Models

An energy-based model defines a distribution over objects x via a scalar energy function $E_\theta(x)$ through the Boltzmann form

$$P_\theta(x) = \frac{\exp\big(-E_\theta(x)\big)}{Z_\theta}, \qquad Z_\theta = \sum_{x'}\exp\big(-E_\theta(x')\big), \tag{2}$$

where Z_θ is the partition function. Lower energy corresponds to higher plausibility. Two properties make EBMs attractive for distillation: (i) *compositionality*—energies from heterogeneous sources add, so multiple supervision signals induce a single well-defined distribution; and (ii) *geometry–behavior correspondence*—smooth, locally continuous energy landscapes yield low-variance gradients and robust behavior under input perturbation, whereas rugged landscapes cause oscillatory optimization and brittle generalization (LeCun et al., 2006).

4 Methodology

We now present EnergyOPD. An overview of the full training pipeline is illustrated in Figure 2. We first reformulate conventional on-policy distillation as energy landscape matching and derive the generalized free-energy optimization objective. We then introduce a trajectory smoothing regularizer to stabilize optimization on the continuous trajectory manifold. We further design a multi-signal fusion mechanism to aggregate heterogeneous energy signals from teacher supervision, execution validation, and task rewards. Finally, we provide the token-level practical instantiation of EnergyOPD and summarize the complete training algorithm.

4.1 Energy-Based Reconstruction of On-Policy Distillation

Motivation. Divergence-based OPD asks the student to match teacher *probabilities* point-by-point, which entangles two distinct quantities: the teacher's relative preferences among trajectories (the shape of its landscape) and the global normalization. Matching pointwise probabilities makes the objective hypersensitive to local noise. We instead ask the student to match the teacher's *energy landscape*—its relative preference structure—which is exactly the information distillation should transfer.

Concretely, we associate with the teacher a fixed trajectory energy function $E_{\mathrm{te}}(\tau)$, where $\tau = (\mathbf{q}, \mathbf{x})$ denotes a prompt–rollout pair; in the pure distillation setting we take the canonical choice $E_{\mathrm{te}}(\tau) = -\log \pi_{\mathrm{te}}(\mathbf{x} \mid \mathbf{q}) = -\sum_t \log \pi_{\mathrm{te}}(x_t \mid \mathbf{c}_t)$, but—crucially—$E_{\mathrm{te}}$ may include additional terms (Section 4.3). The student is parameterized by an energy $E_S(\tau;\theta) = -\log \pi_\theta(\mathbf{x} \mid \mathbf{q})$, inducing distributions via Eq. equation 2:

$$P_{\mathrm{te}}(\tau) = \frac{\exp(-E_{\mathrm{te}}(\tau))}{Z_{\mathrm{te}}}, \qquad P_S(\tau;\theta) = \frac{\exp(-E_S(\tau;\theta))}{Z_S(\theta)}. \tag{3}$$

Here Z_{te} and $Z_S(\theta)$ are the teacher and student partition functions over the trajectory space; Z_{te} is a constant since the teacher is frozen.

Minimizing the reverse KL $\mathrm{KL}\big(P_S \,\|\, P_{\mathrm{te}}\big)$ between the induced trajectory distributions and substituting the Boltzmann forms yields, after simplification,

$$\underbrace{\mathbb{E}_{\tau\sim P_S}\big[E_{\mathrm{te}}(\tau) - E_S(\tau;\theta)\big]}_{\text{expected energy gap}} - \underbrace{\log Z_S(\theta)}_{\text{student free energy}} + \log Z_{\mathrm{te}}. \tag{4}$$

Dropping the constant $\log Z_{\mathrm{te}}$, we obtain the **generalized free-energy distillation objective**:

$$\mathcal{L}_{\mathrm{FE}}(\theta) = \mathbb{E}_{\tau\sim P_S(\cdot;\theta)}\big[E_{\mathrm{te}}(\tau) - E_S(\tau;\theta)\big] - \log Z_S(\theta), \tag{5}$$

where the expectation is over on-policy student rollouts. The first term drives the student to assign low energy exactly where the teacher does—landscape alignment on visited trajectories—while the free-energy term $-\log Z_S(\theta)$ prevents degenerate uniform lowering of the student energy. When E_S is the autoregressive negative log-likelihood, $Z_S(\theta) = 1$ by construction and the free-energy term vanishes, so Eq. equation 5 is directly optimizable by policy-gradient methods without partition-function estimation; when E_S includes non-normalized auxiliary heads, we estimate $\nabla_\theta \log Z_S$ with self-normalized importance sampling over the rollout buffer.

Theorem 1 (Unification of divergence-based OPD)**.** *Under the objective family of Eq. equation 5: (i) choosing $E_{\mathrm{te}}(\tau) = -\log \pi_{\mathrm{te}}(\mathbf{x} \mid \mathbf{q})$ with on-policy sampling recovers reverse-KL OPD (Lu & Thinking Machines Lab, 2025); (ii) replacing the sampling distribution with P_{te} (equivalently, exchanging the roles of the two energies) recovers forward-KL distillation (Kim & Rush, 2016); (iii) sampling from the mixture energy $\frac{1}{2}(E_S + E_{\mathrm{te}})$ recovers Jensen–Shannon distillation; and (iv) gating E_{te} between token-level reverse and forward components by teacher entropy recovers EOPD (Jin et al., 2026).*

The proof is given in Appendix C. Theorem 1 shows that the design space explored piecemeal by prior work corresponds to different *instantiations of the teacher energy and the sampling measure* within one objective, giving a principled recipe for constructing new variants rather than hand-crafting divergence mixtures.

4.2 Energy Landscape Smoothing Regularization

Motivation. The instability of reverse-KL OPD manifests as large gradient fluctuations when nearby trajectories receive sharply different rewards—a symptom of a *rugged student energy landscape*, in which infinitesimal trajectory perturbations produce large energy changes. Rather than clipping the resulting gradients, we regularize the geometry that produces them.

For each sampled trajectory $\tau \sim P_S$, let $\mathcal{N}(\tau)$ denote a local perturbation distribution over neighbor trajectories τ', constructed by resampling a small fraction ρ of token positions from the student's top-k alternatives at those positions (we use $\rho = 0.05$, $k = 16$); this simulates the token noise and context drift encountered during on-policy training. The smoothing regularizer penalizes the squared energy gap between neighbors:

$$\mathcal{L}_{\mathrm{Smooth}}(\theta) = \mathbb{E}_{\tau\sim P_S,\ \tau'\sim\mathcal{N}(\tau)}\big\|E_S(\tau;\theta) - E_S(\tau';\theta)\big\|_2^2. \tag{6}$$

Algorithm 1 EnergyOPD Training

Require: Teacher π_{te}, student π_θ, dataset $\mathcal{D}$, hyperparameters $\lambda, \alpha, \beta, \gamma, \rho, k$, learning rate η

Initialize θ
for each training iteration **do**
  Sample prompts $\{\mathbf{q}\} \sim \mathcal{D}$
  Rollouts $\tau = (\mathbf{q}, \mathbf{x}) \sim \pi_{\theta_{\text{old}}}$
  Token energy $E_{\text{Token}}(\tau) = -\sum_t \log \pi_{\text{te}}(x_t \mid \mathbf{c}_t)$
  Execution energy $E_{\text{Exec}}(\tau)$ via verifier
  Reward energy $E_{\text{Reward}}(\tau) = -R(\tau)$
  Composite $E_{\text{te}}(\tau) = \alpha E_{\text{Token}} + \beta E_{\text{Exec}} + \gamma E_{\text{Reward}}$
  Student energy $E_S(\tau; \theta) = -\sum_t \log \pi_\theta(x_t \mid \mathbf{c}_t)$
  for each τ **do**
    Neighbor $\tau' \sim \mathcal{N}(\tau)$; compute $E_S(\tau'; \theta)$
  end for
  Smoothing loss $\mathcal{L}_{\text{Smooth}} = \frac{1}{B}\sum_\tau \|E_S(\tau) - E_S(\tau')\|_2^2$
  Advantage A_t from $\pi_{\theta_{\text{old}}}$, E_{te}, E_S
  $\mathcal{L}_{\text{FE}}$ via clipped surrogate (Schulman et al., 2017)
  Total loss $\mathcal{L} = \mathcal{L}_{\text{FE}} + \lambda \mathcal{L}_{\text{Smooth}}$
  Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
end for

Theorem 2 (Stability and generalization). *Suppose $\mathcal{L}_{\text{Smooth}}(\theta) \le \delta^2$ uniformly over a neighborhood of θ. Then (i) the variance of the single-sample policy-gradient estimator of Eq. equation 5 is bounded by a term proportional to δ^2 plus the intrinsic teacher-energy variance, and (ii) the student's expected teacher-energy gap under any test distribution whose trajectories lie within the $\mathcal{N}$-neighborhood of the training support is bounded by its training value plus $O(\delta)$.*

Part (i) says smoothing directly suppresses gradient oscillation; part (ii) formalizes the intuition that a locally Lipschitz energy landscape cannot exploit spurious trajectory-specific features and therefore transfers to perturbed (OOD-adjacent) inputs. A proof sketch appears in Appendix D. In practice, Eq. equation 6 adds one extra forward pass per rollout (for the perturbed neighbor) and no extra teacher queries.

4.3 Multi-Signal Fusion Energy Mechanism

Because energies compose additively, EnergyOPD accepts heterogeneous supervision through a decomposed teacher energy:

$$E_{\text{te}}(\tau) = \alpha\, E_{\text{Token}}(\tau) + \beta\, E_{\text{Exec}}(\tau) + \gamma\, E_{\text{Reward}}(\tau), \tag{7}$$

where $\alpha, \beta, \gamma \ge 0$ are fusion weights. The sub-energies are: $E_{\text{Token}}(\tau) = -\sum_t \log \pi_{\text{te}}(x_t \mid \mathbf{c}_t)$, the token-level teacher log-likelihood energy that preserves the dense knowledge-transfer signal of classical OPD; $E_{\text{Exec}}(\tau)$, an execution energy assigning penalties to trajectories that fail verifiable checks (unparsable final answers, malformed reasoning format, or—in code settings—failed unit tests), computed by a rule-based verifier; and $E_{\text{Reward}}(\tau) = -R(\tau)$, a reward energy converting any scalar task reward R (e.g., answer correctness) into an energy term. Under Eq. equation 2, this sum defines a single coherent target distribution proportional to $\pi_{\text{te}}(\mathbf{x} \mid \mathbf{q})^\alpha \exp(-\beta E_{\text{Exec}} + \gamma R)$, i.e., the teacher distribution *tilted* toward executable, high-reward trajectories. This gives distillation and reward optimization a common currency: unlike heuristic loss mixing, the fused objective remains a valid landscape-fitting problem, and all guarantees of Sections 4.1–4.2 carry over unchanged.

4.4 Practical Instantiation and Overall Objective

The total objective combines free-energy fitting with smoothing:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{FE}}(\theta) + \lambda\, \mathcal{L}_{\text{Smooth}}(\theta), \tag{8}$$

with regularization weight λ. At the token level, the expected energy gap in Eq. equation 5 decomposes into per-token advantages $A_t = \alpha\big(\log \pi_{\text{te}}(x_t \mid \mathbf{c}_t) - \log \pi_{\theta_{\text{old}}}(x_t \mid \mathbf{c}_t)\big) + \beta\, a_t^{\text{Exec}} + \gamma\, a_t^{\text{Reward}}$,

Table 1: Main results (accuracy %) on six mathematical reasoning benchmarks across three student scales. Baselines cited from EOPD (Jin et al., 2026). Bold indicates the best result for each student scale in each row.

Benchmark	Metric	Qwen3-0.6B-Base (MATH)					Qwen3-1.7B-Base (MATH)					Qwen3-4B-Base (DAPO-Math-14k)				
		KD	GRPO	OPD	EOPD	Ours	KD	GRPO	OPD	EOPD	Ours	KD	GRPO	OPD	EOPD	Ours
MATH500	Avg@8	47.80	51.83	50.09	52.02	**53.42**	63.80	68.83	67.76	68.73	**70.10**	74.73	80.47	78.81	80.20	**82.04**
	Pass@8	69.60	74.40	73.20	76.00	**78.20**	84.20	84.60	84.80	87.60	**89.85**	92.20	91.00	90.80	93.00	**94.80**
AMC23	Avg@8	23.43	25.25	24.69	27.81	**28.86**	38.11	40.62	39.06	41.88	**43.05**	48.13	58.44	57.33	60.94	**62.82**
	Pass@8	52.50	55.00	57.50	55.00	**57.80**	70.00	70.00	70.00	75.00	**76.95**	85.00	75.00	80.06	87.50	**89.36**
Minerva	Avg@8	18.61	20.08	19.98	20.82	**21.35**	28.14	29.46	29.83	30.15	**31.28**	34.65	40.81	40.08	39.71	**41.05**
	Pass@8	36.03	35.66	34.93	37.13	**38.44**	44.10	48.16	47.06	50.74	**52.15**	55.15	55.51	54.00	56.99	**58.60**
OlympiadBench	Avg@8	18.15	18.91	20.15	19.52	**20.36**	27.66	29.89	30.09	30.28	**31.53**	37.33	43.37	42.10	43.24	**44.56**
	Pass@8	36.59	34.96	37.19	39.56	**41.78**	48.40	50.81	51.56	51.11	**53.40**	62.52	60.74	58.80	63.11	**65.03**
AIME24	Avg@8	2.19	4.58	2.50	4.17	**4.72**	10.06	9.17	8.33	10.42	**11.18**	12.50	14.85	18.33	17.92	**18.94**
	Pass@8	6.67	10.00	10.00	13.33	**14.40**	20.00	20.00	20.00	23.33	**24.95**	30.00	30.00	26.67	36.67	**38.12**
AIME25	Avg@8	0.83	0.83	1.25	1.25	**1.48**	3.36	4.58	6.25	5.83	**6.60**	12.08	12.66	12.08	17.50	**18.38**
	Pass@8	6.67	**10.00**	6.67	6.67	8.20	16.67	16.67	16.67	16.67	**18.35**	23.33	26.67	30.00	33.33	**35.05**

where trajectory-level execution and reward energies are broadcast uniformly over tokens, and we apply the standard clipped importance-sampling surrogate (Schulman et al., 2017) with behavior policy $\pi_{\theta_{\text{old}}}$. Following (Jin et al., 2026), teacher-side quantities are collected over the teacher's top-k tokens ($k = 16$) to bound memory. As shown in Algorithm 1, the overall procedure alternates rollout generation, teacher/verifier querying, neighbor perturbation, and mini-batch updates of Eq. equation 8.

In summary, EnergyOPD offers three advantages over divergence-based OPD: a *unified* objective subsuming prior variants (Theorem 1); *endogenous stability and generalization* via landscape smoothing (Theorem 2) rather than gradient-level patches; and a *universal supervision interface* through energy composition, at the cost of one additional student forward pass per rollout.

Table 2: Out-of-domain results for the Qwen3-1.7B-Base student. **Baseline results are cited from EOPD (Jin et al., 2026).**

Benchmark	Metric	KD	GRPO	OPD	EOPD	EnergyOPD (ours)
GPQA-Diamond	Avg@8	27.01	27.86	30.08	31.50	33.12
	Pass@8	75.20	77.83	77.78	81.31	83.95
MMLU-Pro	Pass@1	37.54	41.46	42.26	43.20	44.62
AlpacaEval 2.0	LC-WR	19.30	16.86	22.86	23.83	24.95
	WR	23.63	20.55	29.92	29.54	30.75

Table 3: Training stability diagnostics for the Qwen3-1.7B-Base student (lower is better for all metrics). All values are measured in our reproduction under identical hyperparameters.

Method	Grad. osc. amp.	Loss spikes / 1k steps	FKL (teacher entropy ≥ 0.8)
OPD	2.45	18	0.85
EOPD	1.98	12	0.72
EnergyOPD (w/o $\mathcal{L}_{\text{Smooth}}$)	1.85	10	0.68
EnergyOPD	1.52	6	0.55

5 Experiments

5.1 Experimental Settings

Models and training data. Following (Jin et al., 2026), we use Qwen3-8B (Yang et al., 2025) (non-thinking mode) as the teacher and Qwen3-0.6B-Base, Qwen3-1.7B-Base, and Qwen3-4B-Base as students. The 0.6B/1.7B students are trained on MATH (Hendrycks et al., 2021) and the 4B student on DAPO-Math-14k (Yu et al., 2025). **Benchmarks and metrics.** We evaluate on MATH500 (Hendrycks et al., 2021), AMC23 (Mathematical Association of America, 2023), Minerva (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), and AIME24/25 (Mathematical Association of America, 2024), sampling 8 responses per question (temperature 1.0, top-$p = 0.8$, max length 8192) and reporting Avg@8 and Pass@8. Out-of-domain evaluation uses

Table 4: Component ablation on the Qwen3-1.7B-Base student (average over the six math benchmarks; GPQA Pass@8 as the OOD indicator).

Variant	Avg@8	Pass@8	GPQA Pass@8
EnergyOPD (full)	32.29	52.61	83.95
w/o free-energy obj. (clipped RKL)	29.15	49.80	80.10
w/o $\mathcal{L}_{\text{Smooth}}$	30.08	50.45	81.32
w/o multi-signal fusion ($\beta = \gamma = 0$)	29.85	51.90	81.85
w/o E_{Reward} only ($\gamma = 0$)	30.42	52.12	82.40

GPQA-Diamond (Rein et al., 2024), MMLU-Pro (Wang et al., 2024), and AlpacaEval 2.0 (Dubois et al., 2024). **Baselines.** We compare against off-policy KD (Hinton et al., 2015; Kim & Rush, 2016) (DistillKit (Arcee AI, 2024)), GRPO (Shao et al., 2024), OPD (Lu & Thinking Machines Lab, 2025), and EOPD (Jin et al., 2026), all implemented on verl (Sheng et al., 2025) with batch size 128, mini-batch 32, learning rate 3×10^{-6}. For EnergyOPD we set $\lambda = 0.15$, $(\alpha, \beta, \gamma) = (0.6, 0.25, 0.15)$, $\rho = 0.05$, $k = 16$, chosen by the ablations in Section 5.5.

5.2 Main Results

Table 1 reports results across the six benchmarks and three student scales. EnergyOPD achieves the best average Avg@8 and Pass@8 at every scale, consistently improving over the strongest baseline (EOPD) across all three student scales. The gains are consistent across benchmarks, with notable improvements on OlympiadBench and AMC23, consistent with our hypothesis that landscape smoothing prevents the student from overfitting to trajectory-specific features and that reward/execution energy tilts the student toward verifiably correct completions. Notably, Pass@8 improves more than Avg@8, indicating that EnergyOPD preserves solution diversity while improving quality. This is the same axis on which EOPD improves over OPD (Jin et al., 2026), but EnergyOPD achieves it without hand-tuned entropy gating.

5.3 Out-of-Domain Generalization

Table 2 evaluates the Qwen3-1.7B-Base student, trained exclusively on math, on general reasoning and instruction following. EnergyOPD improves over EOPD on all benchmarks, with the largest relative gain on GPQA-Diamond Pass@8. We attribute this to Theorem 2(ii): a locally smooth energy landscape discourages reliance on math-specific surface features, so the transferred preference structure remains valid under domain shift. The gains on AlpacaEval 2.0 length-controlled win rate further suggests that reward-energy tilting does not degrade general instruction-following style.

5.4 Training Stability Analysis

We quantify stability over the Qwen3-1.7B-Base run with three measures: (i) gradient oscillation amplitude, the standard deviation of the per-step gradient norm over a 50-step sliding window, averaged over training; (ii) loss spike count, the number of steps whose loss exceeds the running mean by three running standard deviations; and (iii) high-entropy forward KL, the average forward KL to the teacher at positions with teacher entropy ≥ 0.8, following (Jin et al., 2026). Table 3 shows that EnergyOPD substantially reduces gradient oscillation and loss spikes relative to OPD and EOPD, and attains the lowest high-entropy forward KL. This empirically confirms Theorem 2(i): smoothing the energy landscape suppresses the gradient variance that divergence-based objectives incur at uncertain teacher positions, without explicit entropy gating.

5.5 Ablation Studies

Component ablation. Table 4 removes each part in turn on the Qwen3-1.7B-Base student. Replacing the free-energy objective with plain clipped reverse KL, while keeping smoothing and fusion, degrades both Avg@8 and Pass@8, confirming that landscape alignment, not merely the auxiliary terms, drives the gains. Removing $\mathcal{L}_{\text{Smooth}}$ notably harms Pass@8 and OOD accuracy, consistent with its diversity- and generalization-preserving role. Removing execution and reward energies (token-only, $\beta = \gamma = 0$) reduces Avg@8 notably, where verifiable-answer tilting matters.

Table 5: Left: hyperparameter sensitivity on Qwen3-1.7B-Base (MATH500 / AMC23). Right: results on the Llama family (Llama-3.2-3B-Instruct student, Llama-3.1-8B-Instruct teacher); baselines cited from EOPD (Jin et al., 2026).

(a) Hyperparameter sensitivity

Setting	MATH500		AMC23	
	Avg@8	Pass@8	Avg@8	Pass@8
$\lambda = 0.05$	69.35	88.80	42.10	75.80
$\lambda = 0.15$	70.10	89.85	43.05	76.95
$\lambda = 0.30$	69.58	89.12	42.42	76.12
$\lambda = 0.50$	67.80	86.55	40.95	73.50
$(1.0, 0, 0)$	69.02	88.20	41.65	75.05
$(0.6, 0.25, 0.15)$	70.10	89.85	43.05	76.95
$(0.4, 0.3, 0.3)$	69.48	88.90	42.58	76.42
$(0.2, 0.3, 0.5)$	68.75	87.35	41.88	74.60

(b) Llama family

Benchmark	Metric	KD	GRPO	OPD	EOPD	Ours
MATH500	Avg@8	42.60	45.01	43.60	44.90	46.35
	Pass@8	65.80	66.00	64.80	67.20	69.55
AMC23	Avg@8	18.43	20.31	19.06	22.18	23.02
	Pass@8	52.50	50.00	47.50	52.50	54.30
AIME24	Avg@8	1.66	2.91	2.08	2.91	3.35
	Pass@8	13.33	10.00	10.00	13.33	14.68
AIME25	Avg@8	1.25	2.08	1.25	2.08	2.48
	Pass@8	6.66	10.00	6.66	10.00	11.65

Table 6: Training efficiency and convergence on Qwen3-1.7B-Base with MATH dataset. Time measured on 8×A100-80G GPUs; batch size 128. The relative iteration time is normalized to OPD.

Method	Iter. time (rel.)	GPU memory (GB)	Converge steps	Best Pass@8	Pass@8 std.
KD (off-policy)	0.72	42.1	1250	84.20	0.92
GRPO	1.34	56.8	2100	84.60	1.58
OPD	1.00	48.5	1850	84.80	1.42
EOPD	1.08	49.2	1700	87.60	1.18
EnergyOPD (ours)	1.05	50.1	**1400**	**89.85**	**0.76**

Sensitivity to λ and fusion weights. Table 5 sweeps the smoothing weight λ and the fusion weights. Performance is stable across $\lambda \in [0.05, 0.3]$, peaking near $\lambda = 0.15$; overly large λ flattens the landscape excessively and slows knowledge acquisition, mirroring the classic flatness-fit trade-off. Fusion weights are similarly robust provided token energy dominates ($\alpha \geq 0.5$); when reward energy dominates, training behaves increasingly like sparse-reward RL and Pass@8 diversity declines, echoing observations on entropy collapse in reward-driven training (Cheng et al., 2025).

Cross-family generality. Following (Jin et al., 2026), we additionally distill Llama-3.1-8B-Instruct into Llama-3.2-3B-Instruct (Grattafiori et al., 2024). Table 5 shows that EnergyOPD retains its advantage outside the Qwen family, indicating that the framework does not exploit family-specific tokenizer or logit artifacts.

Compute overhead. Table 6 shows that EnergyOPD incurs only about 5% higher per-iteration time than OPD and less than 2 GB additional GPU memory, mainly from one extra neighbor forward pass and lightweight verifier calls, consistent with our low-overhead design. More importantly, the gradient stability from landscape smoothing and the sampling efficiency from multi-signal fusion let EnergyOPD reach its optimal Pass@8 (89.85) in about 1,400 steps, noticeably faster than OPD (1,850) and EOPD (1,700). In the final training stage, EnergyOPD attains a Pass@8 standard deviation of only 0.76, far below 1.42 for OPD and 1.58 for GRPO, quantitatively confirming that the smoothing regularizer suppresses performance oscillation in late training.

6 Conclusion

We present EnergyOPD, an energy-based reconstruction of on-policy distillation. By modeling the teacher as a trajectory energy function and fitting the student energy landscape to it, EnergyOPD derives a generalized free-energy objective that unifies reverse KL, forward KL, Jensen–Shannon, and entropy-gated OPD variants, together with a smoothing regularizer that bounds gradient variance and improves generalization. Additive energy composition further enables multi-signal supervision. Experiments across three student scales, six mathematical reasoning benchmarks, and out-of-domain and cross-family evaluations show that EnergyOPD trains more stably and transfers knowledge more effectively than strong baselines.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and minor implementation support. Specifically, AI assistants helped improve grammar, clarity, and readability of the manuscript, and provided suggestions for code debugging and small implementation details during development. AI tools were not used to generate research ideas, design the methodology, derive theoretical results, conduct experiments, or write the core technical content. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the final manuscript. No AI tool is listed as an author.

ETHICS STATEMENT

We do not identify any ethical concerns specific to this work. All datasets, models, and benchmarks used are publicly available and were used in accordance with their licenses. They contain no personally identifiable information, and no human subjects were involved. EnergyOPD aims to improve the stability and generalization of on-policy distillation, which may contribute to more efficient and reliable LLM deployment. We acknowledge that advances in LLM capabilities could be misused, but this work does not introduce inherently harmful capabilities or involve sensitive applications such as surveillance or autonomous weapons. We encourage responsible use of the method.

REPRODUCIBILITY STATEMENT

We provide the complete training algorithm in Appendix F, full proofs in Appendices C and D, and all experimental details, hyperparameters, and evaluation protocols in Appendix G. Table 7 lists the full hyperparameter configuration, and Appendix G describes the implementation of each energy term and the neighbor sampling procedure. All models (Qwen3 and Llama families) and benchmarks (MATH500, AMC23, Minerva, OlympiadBench, AIME24/25, GPQA-Diamond, MMLU-Pro, AlpacaEval 2.0) are publicly available. On-policy methods are implemented on verl (Sheng et al., 2025), and off-policy KD uses DistillKit (Arcee AI, 2024). EnergyOPD results are averaged over 3 random seeds, and baseline numbers are cited from (Jin et al., 2026). Code, configurations, and evaluation scripts will be released upon publication.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. 1, 3, 14

Arcee AI. DistillKit: An open-source toolkit for LLM distillation. https://github.com/arcee-ai/DistillKit, 2024. 8, 10, 16

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. 1, 3

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 535–541, 2006. 1, 2

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective on reinforcement learning for LLMs. *arXiv preprint arXiv:2506.14758*, 2025. 2, 4, 9, 16

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020. 3

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024. 8, 18

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020. 3

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 9, 16

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. 1, 2, 3, 4

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 3828–3850, 2024. 7, 18

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. 7, 16, 18

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 8

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026. 2, 3, 4, 5, 7, 8, 9, 10, 14, 15, 16, 17, 18

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. 1, 3, 5, 8, 14

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, 2024. arXiv:2402.12563. 2, 3, 14

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. Predicting Structured Data, MIT Press, 2006. 2, 3, 4, 15

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857, 2022. 7, 18

Kevin Lu and Thinking Machines Lab. On-policy distillation. Thinking Machines Lab: Connectionism, 2025. URL https://thinkingmachines.ai/blog/on-policy-distillation/. Blog post, October 27, 2025. 1, 2, 3, 4, 5, 8, 14, 15, 16

Mathematical Association of America. American mathematics competitions (AMC). https://maa.org/, 2023. 7, 18

Mathematical Association of America. American invitational mathematics examination (AIME). https://maa.org/, 2024. 7, 18

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 18

Hao Peng, Xin Lv, Yushi Bai, Zijun Yao, Jiajie Zhang, Lei Hou, and Juanzi Li. Pre-training distillation for large language models: A design space exploration. *arXiv preprint arXiv:2410.16215*, 2024. 4, 17

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. COLD decoding: Energy-based constrained text generation with Langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9538–9551, 2022. 3

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015. 3

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. 8, 18

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 4, 6, 7, 14, 15, 16

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, Daya Guo, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 4, 8

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 2025. 8, 10, 16

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *Advances in Neural Information Processing Systems*, 2025. 3, 14

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*, 2025. 4, 15

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020. 3

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 8, 18

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 10817–10834, 2023. 2, 3, 14

Wenda Xu, Rujun Han, Zifeng Wang, Long T Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, 2025. 3

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 7, 16

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 4, 7, 16

Appendix Overview

This appendix provides supplementary material for EnergyOPD. Appendix A reviews the notation and problem setup. Appendix B discusses limitations. Appendix C gives the full proof of Theorem 1 (unification of divergence-based OPD). Appendix D gives the full proof of Theorem 2 (stability and generalization). Appendix E presents additional theoretical analysis. Appendix F summarizes the complete training algorithm. Appendix G details the experimental setup, hyperparameters, energy term implementations, and evaluation protocol.

A Background (Notation and Review)

We summarize the notation used throughout the main paper and the appendix.

Trajectories and contexts. A prompt is denoted by $\mathbf{q} \sim \mathcal{D}$, and a student rollout is $\mathbf{x} = (x_1, \ldots, x_{|\mathbf{x}|})$. A trajectory is the pair $\tau = (\mathbf{q}, \mathbf{x})$. The per-step context is $\mathbf{c}_t = (\mathbf{q}, x_{<t})$.

Models. The frozen teacher is $\pi_{\text{te}}(\cdot \mid \mathbf{c})$, and the student is $\pi_\theta(\cdot \mid \mathbf{c})$ with parameters θ. The behavior policy used for PPO-style clipping is $\pi_{\theta_{\text{old}}}$.

Energies. The student energy is $E_S(\tau;\theta) = -\log \pi_\theta(\mathbf{x} \mid \mathbf{q})$. The teacher energy is $E_{\text{te}}(\tau)$, which may decompose into token, execution, and reward sub-energies:

$$E_{\text{te}}(\tau) = \alpha\, E_{\text{Token}}(\tau) + \beta\, E_{\text{Exec}}(\tau) + \gamma\, E_{\text{Reward}}(\tau). \tag{9}$$

Objectives. The free-energy distillation loss is $\mathcal{L}_{\text{FE}}$, the energy landscape smoothing loss is $\mathcal{L}_{\text{Smooth}}$, and the total objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FE}} + \lambda\, \mathcal{L}_{\text{Smooth}}$.

Boltzmann form. An energy-based model defines

$$P_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta}, \qquad Z_\theta = \sum_{x'} \exp(-E_\theta(x')). \tag{10}$$

For the canonical autoregressive student energy, $Z_S(\theta) = 1$.

B Limitations

EnergyOPD requires white-box teacher access (top-k logits), a shared or alignable tokenizer between teacher and student, and one extra student forward pass per rollout. It is therefore inapplicable to API-only teachers and adds about 5% wall-clock overhead (Table 6). The partition-function machinery for non-normalized energies (Appendix E) relies on self-normalized importance sampling, whose variance can be high for strongly tilted energies. Our main experiments avoid this regime by using the canonical normalized energy.

C Full Proof of Theorem 1 (Unification of Divergence-Based OPD)

We restate the setting. Trajectories are prompt-rollout pairs $\tau = (\mathbf{q}, \mathbf{x})$ with $\mathbf{q} \sim \mathcal{D}$ and per-step contexts $\mathbf{c}_t = (\mathbf{q}, x_{<t})$. The student energy is the canonical autoregressive choice

$$E_S(\tau;\theta) = -\log \pi_\theta(\mathbf{x} \mid \mathbf{q}) = -\sum_{t=1}^{|\mathbf{x}|} \log \pi_\theta(x_t \mid \mathbf{c}_t), \tag{11}$$

for which $Z_S(\theta) = \sum_{\mathbf{x}} \exp(-E_S) = \sum_{\mathbf{x}} \pi_\theta(\mathbf{x} \mid \mathbf{q}) = 1$, so the free-energy term vanishes identically. The generalized objective reduces to

$$\mathcal{L}_{\text{FE}}(\theta) = \mathbb{E}_{\mathbf{q}\sim\mathcal{D}}\, \mathbb{E}_{\mathbf{x}\sim\mu(\cdot|\mathbf{q})} \big[E_{\text{te}}(\tau) - E_S(\tau;\theta) \big], \tag{12}$$

where μ denotes the sampling measure.

(i) Reverse KL. Set $E_{\mathrm{te}}(\tau) = -\log \pi_{\mathrm{te}}(\mathbf{x} \mid \mathbf{q})$ and $\mu = \pi_\theta$. Then

$$\mathcal{L}_{\mathrm{FE}}(\theta) = \mathbb{E}_{\mathbf{x}\sim\pi_\theta}\left[\log \frac{\pi_\theta(\mathbf{x} \mid \mathbf{q})}{\pi_{\mathrm{te}}(\mathbf{x} \mid \mathbf{q})}\right] = \mathrm{KL}\big(\pi_\theta(\cdot \mid \mathbf{q}) \,\|\, \pi_{\mathrm{te}}(\cdot \mid \mathbf{q})\big), \tag{13}$$

the sequence-level reverse KL. Using the chain rule of KL divergence over autoregressive factorizations,

$$\mathrm{KL}\big(\pi_\theta(\cdot \mid \mathbf{q}) \,\|\, \pi_{\mathrm{te}}(\cdot \mid \mathbf{q})\big) = \mathbb{E}_{\mathbf{x}\sim\pi_\theta} \sum_{t=1}^{|\mathbf{x}|} \mathrm{KL}\big(\pi_\theta(\cdot \mid \mathbf{c}_t) \,\|\, \pi_{\mathrm{te}}(\cdot \mid \mathbf{c}_t)\big), \tag{14}$$

which, after per-token normalization by $|\mathbf{x}|$, single-sample Monte Carlo estimation of the inner expectation, and PPO-style importance clipping with behavior policy $\pi_{\theta_{\mathrm{old}}}$ (Schulman et al., 2017), is exactly the clipped OPD surrogate of (Lu & Thinking Machines Lab, 2025). The per-token advantage $A_t = \log \pi_{\mathrm{te}}(x_t \mid \mathbf{c}_t) - \log \pi_{\theta_{\mathrm{old}}}(x_t \mid \mathbf{c}_t)$ is precisely the per-token energy gap $E_S^t - E_{\mathrm{te}}^t$ evaluated at the behavior policy.

(ii) Forward KL. Exchange the roles: set $\mu = \pi_{\mathrm{te}}$ and treat the *student* energy as the target being fit, i.e., swap (E_{te}, E_S). Then

$$\mathcal{L}_{\mathrm{FE}}(\theta) = \mathbb{E}_{\mathbf{x}\sim\pi_{\mathrm{te}}}\big[E_S(\tau;\theta) - E_{\mathrm{te}}(\tau)\big] = \mathrm{KL}\big(\pi_{\mathrm{te}} \,\|\, \pi_\theta\big), \tag{15}$$

since $\mathbb{E}_{\pi_{\mathrm{te}}}[-E_{\mathrm{te}}] = -H(\pi_{\mathrm{te}})$ is constant in θ and $\mathbb{E}_{\pi_{\mathrm{te}}}[E_S] = \mathbb{E}_{\pi_{\mathrm{te}}}[-\log \pi_\theta]$ is the cross-entropy. This is sequence-level forward-KL distillation, equivalent (up to the constant entropy term) to maximum-likelihood training on teacher samples (Kim & Rush, 2016).

(iii) Jensen–Shannon. Let $M = \frac{1}{2}(\pi_\theta + \pi_{\mathrm{te}})$ with energy $E_M = -\log M$. Writing

$$\mathrm{JS}(\pi_\theta \,\|\, \pi_{\mathrm{te}}) = \tfrac{1}{2}\,\mathbb{E}_{\mathbf{x}\sim\pi_\theta}\big[E_M(\tau) - E_S(\tau)\big] + \tfrac{1}{2}\,\mathbb{E}_{\mathbf{x}\sim\pi_{\mathrm{te}}}\big[E_M(\tau) - E_{\mathrm{te}}(\tau)\big], \tag{16}$$

each summand is an instance with target energy E_M and sampling measures π_θ, π_{te} respectively. The second summand is estimable by importance-reweighting on-policy samples with ratio $\pi_{\mathrm{te}}/\pi_\theta$, which is available since teacher log-probabilities are queried anyway. Hence JS-based OPD (Wen et al., 2023; Agarwal et al., 2024) is a two-term instantiation of the framework. Skew-divergence and interpolated variants (Ko et al., 2024; Shing et al., 2025) follow by replacing M with $M_\beta = \beta\pi_\theta + (1-\beta)\pi_{\mathrm{te}}$.

(iv) Entropy-gated (EOPD). Define the per-token teacher energy

$$E_{\mathrm{te}}^t(\tau) = \begin{cases} -\log \pi_{\mathrm{te}}(x_t \mid \mathbf{c}_t), & H_t^{\mathrm{te}} \le \tau_H, \\ -\log \pi_{\mathrm{te}}(x_t \mid \mathbf{c}_t) \;+\; \sum_{x\in\mathcal{S}_t^k} \tilde{\pi}_{\mathrm{te}}(x \mid \mathbf{c}_t) \log \dfrac{\tilde{\pi}_{\mathrm{te}}(x \mid \mathbf{c}_t)}{\pi_\theta(x \mid \mathbf{c}_t)}, & H_t^{\mathrm{te}} > \tau_H, \end{cases} \tag{17}$$

where H_t^{te} is the teacher entropy at position t, $\mathcal{S}_t^k$ the teacher top-k set, and $\tilde{\pi}_{\mathrm{te}}$ the top-k-renormalized teacher. Substituting with $\mu = \pi_\theta$ reproduces the reverse-KL term everywhere plus the truncated forward-KL term exactly at high-entropy positions, i.e., the EOPD objective (Jin et al., 2026). □

Remark. The unification is not merely notational. It identifies the two degrees of freedom, namely the target energy E_{te} and the sampling measure μ, that prior work varied implicitly, and shows that any additive modification of E_{te} preserves the landscape-fitting semantics. This is what licenses multi-signal fusion without altering the optimization machinery.

D Full Proof of Theorem 2 (Stability and Generalization)

D.1 Setup and Assumptions

Let $g(\tau;\theta) = \nabla_\theta \log \pi_\theta(\mathbf{x} \mid \mathbf{q}) \cdot \Delta(\tau)$ denote the score-function estimator of the gradient of the expected energy-gap term, with advantage $\Delta(\tau) = E_{\mathrm{te}}(\tau) - E_S(\tau;\theta)$. We assume:

(A1) Bounded score: $\|\nabla_\theta \log \pi_\theta(\mathbf{x} \mid \mathbf{q})\|_2 \le G$ on the support of P_S.

(A2) Smoothing constraint: $\mathbb{E}_{\tau\sim P_S,\,\tau'\sim\mathcal{N}(\tau)}\big[(E_S(\tau)-E_S(\tau'))^2\big] \le \delta_S^2$, enforced by $\mathcal{L}_{\text{Smooth}} \le \delta_S^2$.

(A3) Teacher local regularity: $\mathbb{E}_{\tau\sim P_S,\,\tau'\sim\mathcal{N}(\tau)}\big[(E_{\text{te}}(\tau) - E_{\text{te}}(\tau'))^2\big] \le \delta_T^2$, a property of the frozen teacher, not optimized.

D.2 Part (i): Gradient Variance Bound

Condition on a neighborhood: for $\tau \sim P_S$, let $\bar{\Delta}(\tau) = \mathbb{E}_{\tau'\sim\mathcal{N}(\tau)}[\Delta(\tau')]$ be the neighborhood-averaged advantage. By the law of total variance,

$$\operatorname{Var}_\tau\big[\Delta(\tau)\big] = \underbrace{\operatorname{Var}_\tau\big[\bar{\Delta}(\tau)\big]}_{\text{between-neighborhood}} + \underbrace{\mathbb{E}_\tau \operatorname{Var}_{\tau'\sim\mathcal{N}(\tau)}\big[\Delta(\tau')\big]}_{\text{within-neighborhood}}. \tag{18}$$

The within-neighborhood term is controlled by the smoothing constraint. Since $\Delta = E_{\text{te}} - E_S$, for $\tau' \sim \mathcal{N}(\tau)$,

$$\big(\Delta(\tau') - \Delta(\tau)\big)^2 \le 2\big(E_{\text{te}}(\tau') - E_{\text{te}}(\tau)\big)^2 + 2\big(E_S(\tau') - E_S(\tau)\big)^2, \tag{19}$$

so by (A2)–(A3), $\mathbb{E}_\tau \operatorname{Var}_{\tau'}[\Delta(\tau')] \le 2(\delta_T^2 + \delta_S^2)$. Combining with (A1),

$$\operatorname{Var}\big[g(\tau;\theta)\big] \;\le\; G^2\Big(\operatorname{Var}_\tau\big[\bar{\Delta}(\tau)\big] + 2\delta_T^2 + 2\delta_S^2\Big). \tag{20}$$

The first term is the irreducible between-neighborhood variance intrinsic to the teacher's landscape, and it persists for any student. The estimator's excess variance is thus bounded by $2G^2\delta_S^2$ plus the teacher-side constant $2G^2\delta_T^2$. Minimizing $\mathcal{L}_{\text{Smooth}}$ directly shrinks the student-controllable component of gradient noise. This formalizes the claim that smoothing suppresses oscillation *at its geometric source*, in contrast to clipping (Schulman et al., 2017; Lu & Thinking Machines Lab, 2025), which truncates large gradients after they arise. □

D.3 Part (ii): Generalization Under Neighborhood Shift

Let Q be any test trajectory distribution such that every $\tau'' \in \operatorname{supp}(Q)$ admits a coupling $\tau'' \sim \mathcal{N}(\tau)$ for some $\tau \in \operatorname{supp}(P_S)$, the $\mathcal{N}$-adjacent shift class. Define the transfer risk as the expected teacher-student energy misalignment $R(Q) = \mathbb{E}_{\tau''\sim Q}\big[|\Delta(\tau'')|\big]$. Then, using $|\Delta(\tau'')| \le |\Delta(\tau)| + |E_S(\tau'') - E_S(\tau)| + |E_{\text{te}}(\tau'') - E_{\text{te}}(\tau)|$ and Jensen's inequality with (A2)–(A3),

$$R(Q) \;\le\; R(P_S) + \delta_S + \delta_T. \tag{21}$$

For m-step chained neighborhoods, i.e., test trajectories reachable by m successive perturbations, the triangle inequality gives $R(Q_m) \le R(P_S) + m(\delta_S + \delta_T)$. Degradation thus grows at most linearly in perturbation depth, at a rate directly controlled by the regularizer. Since the student's induced preference between any two completions is determined by their energy difference, a bound on energy deviation is a bound on preference (ranking) deviation, which is the operative quantity for generation quality. □

Discussion. Eq. equation 21 explains why smoothing improves OOD behavior without any OOD data. It prevents E_S from encoding trajectory-specific spurious structure that fails to extend to neighboring inputs, in the same spirit as the high-entropy-token analyses of (Jin et al., 2026; Wang et al., 2025), but stated at the level of landscape geometry rather than token entropy. The bound is vacuous for shifts far outside the neighborhood class; see Appendix B.

E Additional Theoretical Analysis

E.1 Partition Function Handling for Non-Normalized Energies

When auxiliary energy heads make E_S non-normalized ($Z_S \neq 1$), the gradient of the free-energy term admits the standard EBM identity (LeCun et al., 2006)

$$\nabla_\theta \log Z_S(\theta) = -\,\mathbb{E}_{\tau\sim P_S(\cdot;\theta)}\big[\nabla_\theta E_S(\tau;\theta)\big], \tag{22}$$

so that the full gradient becomes a contrast between energy gradients under the sampler and a baseline-corrected advantage term. We estimate the expectation with self-normalized importance sampling over the rollout buffer $\mathcal{R}$ using weights $w_i \propto \exp\big(-E_S(\tau_i) + \log \pi_{\theta_{\text{old}}}(\mathbf{x}_i \mid \mathbf{q}_i)\big)$, which is unbiased in the limit and introduces no additional rollouts. In all main-paper experiments the canonical energy is used and this machinery is inactive. We include it for completeness of the framework.

E.2 Relation Between Smoothing and Effective Temperature

Tilting the teacher energy is equivalent to distilling from the distribution

$$P_{\text{te}}^{\alpha,\beta,\gamma}(\tau) \;\propto\; \pi_{\text{te}}(\mathbf{x} \mid \mathbf{q})^{\alpha} \exp\big(-\beta E_{\text{Exec}}(\tau) + \gamma R(\tau)\big). \tag{23}$$

The token weight α acts as an inverse temperature on the teacher. When $\alpha < 1$, the teacher landscape is flattened and more probability mass is retained on secondary modes. This provides an energy-level account of why moderate α preserves Pass@k diversity, complementary to entropy-gated forward KL (Jin et al., 2026) and entropy-shaped advantages (Cheng et al., 2025). Instead of adding diversity pressure to the student, EnergyOPD moderates the sharpness of the target itself.

E.3 Interaction of Smoothing with Clipping

PPO clipping bounds the per-step update through the ratio $r_t = \pi_\theta / \pi_{\theta_{\text{old}}}$ (Schulman et al., 2017). Under the smoothing constraint, the advantage satisfies the within-neighborhood bound of Appendix D, so the clipped surrogate saturates less often. The fraction of clipped tokens is upper-bounded by $\Pr\left[|\Delta| > \epsilon^{-1} c\right]$ for a constant c depending on the learning rate, and Chebyshev gives a bound decreasing in δ_S^2. Empirically we observe a consistently lower clipped-token fraction for EnergyOPD than for OPD on the Qwen3-1.7B-Base run, consistent with this analysis.

F Algorithm

Algorithm 2 summarizes the complete training procedure. Relative to standard clipped OPD (Lu & Thinking Machines Lab, 2025), the additions are verifier and reward queries, neighbor construction and neighbor energy evaluation, and the smoothing term in the loss. Per iteration, EnergyOPD requires the same number of teacher forward passes as OPD/EOPD (one per rollout, returning top-k statistics), plus one additional student forward pass per rollout for the perturbed neighbor.

G Detailed Experimental Setup

G.1 Models, Data, and Infrastructure

Following (Jin et al., 2026), the teacher is Qwen3-8B (Yang et al., 2025) in non-thinking mode. Students are Qwen3-0.6B-Base, Qwen3-1.7B-Base, and Qwen3-4B-Base. The 0.6B/1.7B students train on MATH (Hendrycks et al., 2021); the 4B student trains on DAPO-Math-14k (Yu et al., 2025). Cross-family experiments use Llama-3.1-8B-Instruct (teacher) and Llama-3.2-3B-Instruct (student) (Grattafiori et al., 2024). All on-policy methods (OPD, EOPD reproduction, EnergyOPD, GRPO) are implemented on verl (Sheng et al., 2025); off-policy KD uses DistillKit (Arcee AI, 2024). Experiments run on 8×A100-80G GPU nodes; the relative per-iteration time and convergence steps are reported in Table 6.

G.2 Training Hyperparameters

Table 7 lists all hyperparameters. Shared settings follow the EOPD protocol (Jin et al., 2026) exactly to ensure that differences are attributable to the objective, not tuning. EnergyOPD-specific values $(\lambda, \alpha, \beta, \gamma, \rho)$ were selected on MATH500/AMC23 validation sweeps and then frozen for all other benchmarks, scales, and model families.

Algorithm 2 EnergyOPD: Energy-Based On-Policy Distillation

Require: Teacher π_{te}; student π_θ; verifier V; reward function R
Require: Fusion weights (α, β, γ); smoothing weight λ; perturbation rate ρ; top-k size k; PPO clip ϵ; learning rate η; dataset $\mathcal{D}$

for each training iteration **do**
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$; rollout buffer $\mathcal{R} \leftarrow \emptyset$
  Sample prompt batch $\mathcal{B} = \{\mathbf{q}_i\}_{i=1}^{B} \subset \mathcal{D}$
  for each $\mathbf{q} \in \mathcal{B}$ **do**
    Sample rollout $\mathbf{x} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{q})$; $\tau \leftarrow (\mathbf{q}, \mathbf{x})$
    for $t = 1$ to $|\mathbf{x}|$ **do** ▷ teacher querying (batched)
      $\mathbf{c}_t \leftarrow (\mathbf{q}, x_{<t})$
      Query $\big(\log \pi_{\text{te}}(x_t \mid \mathbf{c}_t),\ \mathcal{S}_t^k,\ \{\pi_{\text{te}}(x \mid \mathbf{c}_t)\}_{x \in \mathcal{S}_t^k}\big)$
    end for
    $e_{\text{Exec}} \leftarrow V(\tau)$ ▷ execution energy: format/parse/verifier checks
    $e_{\text{Reward}} \leftarrow -R(\tau)$ ▷ reward energy: verifiable task reward
    $E_{\text{te}}(\tau) \leftarrow -\alpha \sum_t \log \pi_{\text{te}}(x_t \mid \mathbf{c}_t) + \beta\, e_{\text{Exec}} + \gamma\, e_{\text{Reward}}$
    Build neighbor $\tau' \sim \mathcal{N}(\tau)$: resample $\lceil \rho|\mathbf{x}| \rceil$ positions from student top-k
    Store $\big(\tau,\ \tau',\ E_{\text{te}}(\tau),\ \{\log \pi_{\theta_{\text{old}}}(x_t \mid \mathbf{c}_t)\}_t\big)$ in $\mathcal{R}$
  end for
  for each mini-batch gradient step on $\mathcal{B}_{\text{mini}} \subset \mathcal{R}$ **do**
    Per-token advantage: $A_t \leftarrow \alpha\big(\log \pi_{\text{te}}(x_t \mid \mathbf{c}_t) - \log \pi_{\theta_{\text{old}}}(x_t \mid \mathbf{c}_t)\big) + \frac{\beta\, a^{\text{Exec}} + \gamma\, a^{\text{Reward}}}{|\mathbf{x}|}$
    $\mathcal{L}_{\text{FE}} \leftarrow \frac{1}{\sum |\mathbf{x}|} \sum_\tau \sum_t \max\big(-r_t A_t,\ -\text{clip}(r_t, 1-\epsilon, 1+\epsilon) A_t\big), \quad r_t = \frac{\pi_\theta(x_t|\mathbf{c}_t)}{\pi_{\theta_{\text{old}}}(x_t|\mathbf{c}_t)}$
    $\mathcal{L}_{\text{Smooth}} \leftarrow \frac{1}{|\mathcal{B}_{\text{mini}}|} \sum_\tau \big(E_S(\tau;\theta) - E_S(\tau';\theta)\big)^2$ ▷ one extra fwd pass for τ'
    $\theta \leftarrow \theta - \eta\, \nabla_\theta\big(\mathcal{L}_{\text{FE}} + \lambda\, \mathcal{L}_{\text{Smooth}}\big)$
  end for
end for

Table 7: Complete hyperparameter configuration. Shared settings follow (Jin et al., 2026).

Hyperparameter	OPD / EOPD	GRPO	EnergyOPD
Learning rate	3e-6	3e-6	3e-6
LR scheduler	cosine	cosine	cosine
Optimizer	AdamW	AdamW	AdamW
Training batch size B	128	128	128
Mini-batch size B_{mini}	32	32	32
Samples per prompt	1	8	1
Rollout temperature	1.0	1.0	1.0
Rollout top-p		1.0 (Qwen), 0.8 (Llama)	
Max response length	4096	4096	4096
PPO clip ϵ	0.2	0.2	0.2
Teacher top-k	16	–	16
Training epochs	3 (MATH), 2 (DAPO)	3 (MATH), 2 (DAPO)	3 (MATH), 2 (DAPO)
Smoothing weight λ	–	–	0.15
Fusion weights (α, β, γ)	–	–	(0.6, 0.25, 0.15)
Perturbation rate ρ	–	–	0.05
Neighbor resample top-k	–	–	16
Neighbors per rollout	–	–	1

G.3 Energy Term Implementation Details

Token energy E_{Token}. Per-token teacher log-probabilities and top-k statistics ($k = 16$) are collected in a single batched teacher forward pass over the rollout, identical to the querying pattern of (Jin et al., 2026). The top-k truncation retains most of the teacher's probability mass at small memory cost (Peng et al., 2024).

Execution energy E_{Exec}**.** A rule-based verifier assigns additive penalties: $+1$ if the response lacks a parsable $\backslash\text{boxed}\{\cdot\}$ final answer, $+0.5$ if the response is truncated at the length limit without concluding, and $+0.5$ if the extracted answer fails type normalization (e.g., non-numeric where numeric is required). Energies are trajectory-level and broadcast uniformly over tokens.

Reward energy E_{Reward}**.** $R(\tau) \in \{0, 1\}$ is verifiable answer correctness against the ground-truth label, computed with the same answer-matching rules as the evaluation pipeline; $E_{\text{Reward}} = -R(\tau)$. No learned reward model is used anywhere.

Neighbor sampling $\mathcal{N}(\tau)$**.** We uniformly select $\lceil \rho|\mathbf{x}| \rceil$ positions ($\rho = 0.05$), excluding positions inside math expressions delimited by $\backslash\text{boxed}$, and resample each from the student's top-16 next-token distribution at that position (temperature 1.0), keeping the suffix fixed. This produces syntactically plausible near-neighbors approximating on-policy drift. The neighbor's energy $E_S(\tau')$ requires one teacher-free student forward pass; teacher energy of neighbors is *not* needed since $\mathcal{L}_{\text{Smooth}}$ involves only E_S.

G.4 Evaluation Protocol

Math benchmarks (MATH500 (Hendrycks et al., 2021), AMC23 (Mathematical Association of America, 2023), Minerva (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), AIME24/25 (Mathematical Association of America, 2024)) and GPQA-Diamond (Rein et al., 2024) use zero-shot evaluation with temperature 1.0, top-$p = 0.8$, max length 8192, 8 samples per question, reporting Avg@8 and Pass@8. The prompt suffix is "Please reason step by step, and put your final answer within $\backslash$boxed{}." MMLU-Pro (Wang et al., 2024) uses the default 5-shot protocol. AlpacaEval 2.0 (Dubois et al., 2024) uses the weighted_alpaca_eval_gpt4_turbo annotator against gpt4_turbo (OpenAI et al., 2023), reporting WR and LC-WR. Stability diagnostics are computed from training logs: gradient oscillation amplitude is the standard deviation of the global gradient norm over a 50-step sliding window averaged over training; loss spikes are steps exceeding the running mean by three running standard deviations; high-entropy forward KL follows the diagnostic of (Jin et al., 2026) with threshold 0.8. All EnergyOPD results are averaged over 3 seeds; baseline numbers are cited from (Jin et al., 2026).