CONTRASTIVE POLICY OPTIMIZATION: OPTIMIZING VON NEUMANN ENTROPY ACROSS QUERIES FOR UNSUPERVISED FOUNDATION MODEL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) post-training has emerged as a dominant paradigm for unlocking complex reasoning and continuous generative capabilities in foundation models. However, standard post-training methods (such as PPO, GRPO, and DPO) rely heavily on expensive supervised signals, including ground-truth reasoning trajectories, verified labels, or calibrated reward models. While recent unsupervised heuristics, such as Entropy Minimized Policy Optimization for language reasoning and Atomic Policy Optimization for 3D atomic generation, attempt to eliminate external supervision, they evaluate generations strictly in isolation within individual queries. Consequently, these isolated intra-query methods suffer from severe vulnerabilities: *cross-query representation collapse*, where the policy converges to degenerate, uninformative generic templates or universal crystal packing motifs regardless of input conditions, and susceptibility to *reward hacking* induced by ad-hoc cluster approximation errors. To address these foundational bottlenecks, we propose Contrastive Policy Optimization (CPO), a unified, theoretically grounded framework for fully unsupervised foundation model alignment. Instead of optimizing isolated Shannon entropy per prompt, CPO models the joint generation distribution across a batch of diverse queries as a global quantum state represented by a normalized density operator $\boldsymbol{\rho}$. By leveraging contrastive learning to optimize the *von Neumann entropy* $S(\boldsymbol{\rho}) = -\operatorname{Tr}(\boldsymbol{\rho} \ln \boldsymbol{\rho})$, CPO simultaneously drives two synergistic objectives: (1) *intra-query state purity*, forcing the policy to form decisive, high-confidence consensus modes on each specific query, and (2) *inter-query state orthogonality*, enforcing maximal semantic distinction and diversity across different queries. We derive an exact, analytical policy gradient and advantage formulation that eliminates the need for heuristic clustering, external reward models, or supervised coordinates. Extensive experiments across both continuous geometric generation and discrete language reasoning demonstrate the decisive superiority of CPO. Our theoretical and empirical results demonstrate that cross-query von Neumann entropy optimization provides a universal, mathematically rigorous foundation for unsupervised foundation model alignment.

1 INTRODUCTION

Post-training reinforcement learning (RL) has catalyzed a fundamental paradigm shift in artificial intelligence, transforming pre-trained foundation models into autonomous reasoning engines and high-precision scientific generators (Guo et al., 2025; Jiao et al., 2024; Watson et al., 2023). In large language models (LLMs), algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025) incentivize step-by-step chain-of-thought (CoT) reasoning, self-reflection, and backtracking. Concurrently, in 3D scientific domains such as material science and structural biology, flow-matching generative models (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023) have been aligned via preference learning (e.g., FlowDPO (Jiao et al., 2024)) to suppress sampling hallucinations and generate physically valid crystal lattices (Xie et al., 2022) and antibody CDR loops (Luo et al., 2022).

Despite their empirical success, conventional post-training frameworks critically depend on *dense, curated supervision*, such as human-annotated reasoning traces, verified mathematical golden answers,

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.

calibrated reward models, or experimental 3D crystal structures from the Materials Project (Jain et al., 2013) and Protein Data Bank (Berman et al., 2000). In frontier scientific discovery (e.g., de novo crystal polymorphism or novel pathogenic antibody design) and open-ended cognitive reasoning, ground-truth labels are prohibitively expensive, noisy, or non-existent. Consequently, this strict dependency creates an acute scalability bottleneck for aligning models in data-scarce regimes.

To eliminate external supervision, recent pioneering works have explored unsupervised policy optimization. In language reasoning, Entropy Minimized Policy Optimization (EMPO) (Zhang et al., 2025) samples a group of responses for an unlabeled question, clusters them into semantic bins, and minimizes Shannon semantic entropy (Kuhn et al., 2023). In 3D atomic generation, Atomic Policy Optimization (APO) (Mo & Bian, 2026) constructs an intra-group similarity matrix to extract the leading Laplacian eigenvector and penalizes spatial density entropy.

However, existing unsupervised frameworks share a fatal structural limitation: *isolated intra-query evaluation*. As illustrated in Fig. 1(a), both EMPO and APO evaluate candidate generations strictly within the confines of an isolated query q or condition c. This intra-query isolation induces two severe pathologies: *i) Cross-Query Representation & Mode Collapse:* Because the policy receives no comparative feedback across *different* queries $\{q_1, q_2, \ldots, q_B\}$, it can achieve minimal intra-query entropy by collapsing to trivial, universal shortcuts—such as outputting stereotypical, repetitive reasoning templates in language, or collapsing to identical, generic high-density lattice packings across disparate chemical compositions. Intra-query objectives cannot distinguish whether a confident answer is actually conditioned on the input query or merely an uninformative prior mode. *ii) Heuristic Approximation Fragility & Reward Hacking:* EMPO relies on discrete clustering (regular expressions or small language models) to approximate Shannon probabilities, which is non-smooth, computationally brittle, and vulnerable to catastrophic reward hacking when the clusterer errs. Meanwhile, APO relies exclusively on the principal eigenvector (u_1) of a single-query Gram matrix, ignoring higher-order spectral dispersion and global multi-query geometry.

To overcome these fundamental challenges, we introduce Contrastive Policy Optimization (CPO), a principled, unified framework that formulates unsupervised foundation model alignment through the lens of *quantum information theory* and *cross-query contrastive learning* (Fig. 1(b)).

Rather than evaluating queries in silos, CPO constructs a global, continuous quantum density operator $\boldsymbol{\rho} \in \mathbb{C}^{BG \times BG}$ over the joint representations of candidate generations sampled across a batch of B diverse queries, each with G candidate rollouts. We then optimize the von Neumann entropy $S(\boldsymbol{\rho}) = -\operatorname{Tr}(\boldsymbol{\rho} \ln \boldsymbol{\rho})$, which generalizes classical Shannon entropy to non-commutative operator spaces and continuous geometric manifolds. Specifically, CPO maximizes the Quantum Mutual Information / von Neumann Divergence. Under this formulation, CPO orchestrates two mutually reinforcing objectives: *Intra-Query Purity* compresses the policy's response distribution for each specific query toward a single, pure eigenstate, enforcing decisive self-consistency and structural consensus. *Inter-Query Orthogonality* pushes the representations of distinct queries into orthogonal subspaces of the Hilbert space, maximizing global diversity and strictly preventing cross-query mode collapse. By deriving the exact matrix gradient of von Neumann entropy, we obtain a closed-form, continuous Contrastive Policy Advantage $A_i^{(b)}$ that naturally contrasts positive intra-query mode alignments against negative inter-query alignments, bypassing discrete clustering heuristics entirely.

We validate CPO across extensive benchmarks spanning continuous 3D geometric generation and discrete mathematical and cognitive reasoning: On the standard Materials Project benchmarks (Perov-5, MP-20, and MPTS-52), CPO dramatically outperforms APO, boosting Match Rate on MPTS-52 from 21.14% to 25.80% and reducing coordinate RMSE to 0.1248. In antibody CDR loop modeling (SAbDab), CPO reduces CDR-H3 C_α RMSD to 2.98 Å(vs. 3.21 Å for APO and 3.32 Å for FlowDPO), achieving perfectly straightened flow-matching probability paths. On competitive mathematical benchmarks (MATH, Minerva Math, OlympiadBench, AIME24, AMC23), CPO elevates Qwen2.5-Math-7B Base accuracy to 81.4% on MATH and 70.0% on AMC23, significantly outperforming EMPO (78.0% and 65.0%). On free-form graduate reasoning (MMLU-Pro and GPQA), CPO improves Qwen2.5-7B Base to 53.6% and 32.4%, surpassing EMPO's 50.1% and 28.8% without requiring any external verifiers or reward models. We also prove that CPO guarantees cross-query non-collapse (Theorem A.1), acts as an exact surrogate for quantum mutual information maximization (Theorem A.2), and provides intrinsic variance reduction (Theorem A.3).

2 Related Work

Unsupervised Alignment and Reinforcement Learning for Foundation Models. Post-training alignment has evolved from early supervised fine-tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) (Schulman et al., 2017) to Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025). In reasoning models, DeepSeek-R1-Zero (Guo et al., 2025) demonstrated that pure RL directly initiated from base models can autonomously elicit sophisticated reasoning behaviors (e.g., self-correction and reflection) using rule-based outcome verification. However, applying such techniques to open-ended reasoning or label-free scientific domains requires eliminating external golden verifiers. To achieve unsupervised alignment, several frontier frameworks have been proposed: Self-rewarding and self-consistency methods (Huang et al., 2022; Wang et al., 2025; Xiong et al., 2025) generate pseudo-labels via majority voting or internal reward model consistency, but risk severe model collapse when errors compound recursively (Shumailov et al., 2024). Test-Time Reinforcement Learning (TTRL) (Zuo et al., 2025) performs online updates at inference time via majority voting, but exposes the model to test-time hallucinations and is computationally prohibitive for continuous generation. Reverse curriculum RL (R^3) (Xi et al., 2024) utilizes outcome supervision to mimic step-level credit assignment. EMPO (Zhang et al., 2025) pioneered fully unsupervised LLM reasoning incentivization by minimizing Shannon semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) across sampled responses for unlabeled prompts. In continuous 3D geometric domains, APO (Mo & Bian, 2026) adapted GRPO to atomic structure prediction by combining leading-eigenvector spectral consensus (u_1) with local spatial density entropy. Despite their promises, both EMPO and APO operate strictly under an *isolated intra-query* assumption: they sample rollouts for a single question or composition in isolation and compute relative rewards solely within that sample set. Consequently, they lack any inter-query discriminative signal. As highlighted in classical semi-supervised literature (Grandvalet & Bengio, 2004; Press et al., 2024), unconstrained intra-sample entropy minimization is prone to trivial mode collapse (e.g., assigning all inputs to a uniform prior mode). In contrast, CPO introduces cross-query contrastive learning, establishing global representation distinction alongside intra-query consensus.

Quantum Information Theory and von Neumann Entropy in Machine Learning. In quantum mechanics and quantum information theory (von Neumann, 1932; Nielsen & Chuang, 2010), a quantum statistical ensemble is described by a positive semi-definite, trace-one density operator $\boldsymbol{\rho}$. The degree of uncertainty, mixedness, and entanglement of $\boldsymbol{\rho}$ is quantified by the von Neumann entropy: $S(\boldsymbol{\rho}) = -\operatorname{Tr}(\boldsymbol{\rho} \ln \boldsymbol{\rho}) = -\sum_k \lambda_k \ln \lambda_k$, where $\{\lambda_k\}$ are the eigenvalues of $\boldsymbol{\rho}$. Unlike classical Shannon entropy, which is restricted to discrete orthogonal probability vectors or requires explicit partitioning of the sample space, von Neumann entropy operates directly on non-commutative operators and continuous Gram matrices in reproducing kernel Hilbert spaces (RKHS) (Braunstein et al., 2006; Passerini & Severini, 2008). In network analysis, von Neumann entropy has been utilized to measure structural randomness and graph complexity. In this work, we bring quantum information theory into reinforcement policy optimization: by constructing a normalized density matrix across multi-query rollout representations, we establish the quantum von Neumann entropy difference ΔS_{vN} as an exact, smooth, and differentiable surrogate for mutual information alignment.

Contrastive Representation Learning and Mutual Information. Contrastive learning frameworks, including Contrastive Predictive Coding (CPC) (van den Oord et al., 2018), SimCLR (Chen et al., 2020), and MoCo (He et al., 2020), optimize representation geometry by pulling semantically related (positive) views together while pushing unrelated (negative) samples apart in latent space. Theoretically, minimizing the InfoNCE loss maximizes a lower bound on the mutual information $I(X;Y)$ between augmented views. While contrastive objectives have achieved immense success in pre-training visual and language representations, their integration into policy optimization has remained largely unexplored. Standard policy gradients optimize scalar rewards rather than global representation geometry. CPO bridges this gap: by framing positive pairs as intra-query candidate rollouts and negative pairs as inter-query rollouts, we prove that optimizing the von Neumann entropy difference is mathematically equivalent to contrastive mutual information maximization across the policy's generative manifold.

Generative Modeling for 3D Atomic Systems and Flow Matching. Predicting high-fidelity 3D atomic structures is central to computational materials science and antibody therapeutics (Jumper et al., 2021; Watson et al., 2023; Xie et al., 2022; Luo et al., 2022; Antunes et al., 2024; Yang et al.,

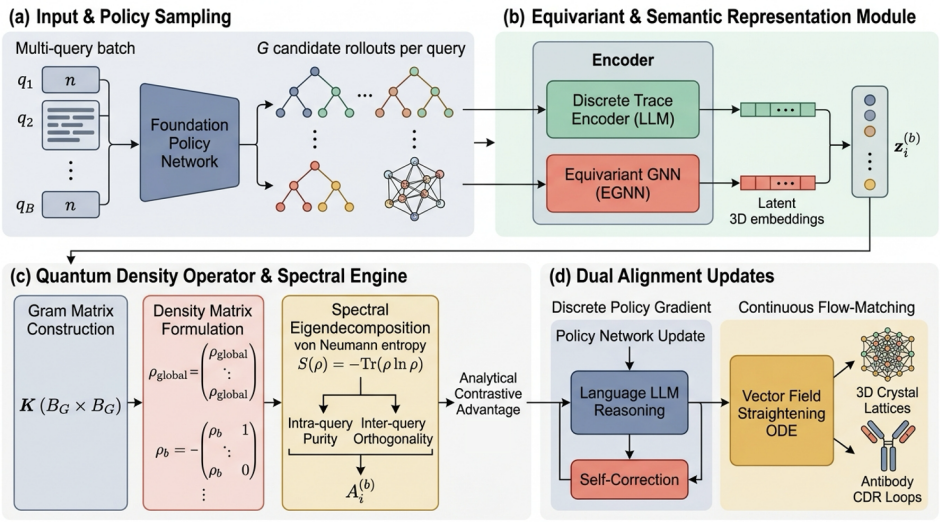


Figure 1: **System workflow of Contrastive Policy Optimization (CPO).** *(a) Input & Policy Sampling:* A batch of diverse queries $\{q_1, \ldots, q_B\}$ (discrete reasoning prompts and 3D atomic conditioning inputs) is passed to the foundation policy network, sampling G rollouts per query (total $M = BG$). *(b) Equivariant & Semantic Representation Module:* Encoder ϕ_θ projects discrete reasoning traces and 3D atomic graphs into normalized latent embeddings $\mathbf{z}_i^{(b)} \in \mathbb{S}^{d-1}$. *(c) Quantum Density Operator & Spectral Engine:* Constructs the global Gram kernel matrix $\mathbf{K} \in \mathbb{R}^{BG \times BG}$ and normalized density operators ($\boldsymbol{\rho}_{\text{global}}$ and intra-query blocks $\boldsymbol{\rho}_b$). Eigendecomposition yields the spectrum $\{\lambda_k\}$, optimizing von Neumann entropy $S(\boldsymbol{\rho}) = -\operatorname{Tr}(\boldsymbol{\rho} \ln \boldsymbol{\rho})$ to derive the closed-form Contrastive Advantage $A_i^{(b)}$ that enforces intra-query purity ($S(\boldsymbol{\rho}_b) \to 0$) alongside inter-query orthogonality ($S(\boldsymbol{\rho}_{\text{global}}) \to \ln(BG)$). *(d) Dual Alignment Updates:* Propagates contrastive policy gradients to discrete autoregressive LLMs (incentivizing autonomous self-reflection) and continuous flow-matching vector fields (straightening ODE paths for 3D crystal lattices and antibody CDR loops).

2026). Deep generative approaches have evolved from VAEs and equivariant diffusion models (Jiao et al., 2023) to continuous-time Flow Matching (FM) (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023). Flow matching models learn time-dependent vector fields $v_\theta(x_t, t)$ that transport prior Gaussian distributions to data manifolds along straight probability paths (e.g., Optimal Transport paths). Recently, FlowDPO (Jiao et al., 2024) integrated preference optimization into flow matching to suppress off-manifold sampling hallucinations. However, FlowDPO strictly requires ground-truth 3D atomic coordinates to compute RMSD preference pairs. While APO (Mo & Bian, 2026) removed coordinate supervision using local geometric heuristics, it still suffers from cross-composition lattice confusion. CPO overcomes this by enforcing inter-query orthogonality in the SE(3)-equivariant latent space, ensuring that distinct chemical compositions and antibody sequences map to strictly distinct, high-symmetry physical manifolds while straightening flow-matching probability paths.

3 Methodology

In this section, we formulate the mathematical foundation and algorithmic architecture of **Contrastive Policy Optimization (CPO)**. We first establish preliminaries across discrete language reasoning and continuous 3D atomic flow matching. We then introduce the cross-query quantum density operator, derive the analytical von Neumann entropy policy gradient, and provide theoretical guarantees on mode non-collapse, mutual information maximization, and geometric equivariance.

3.1 Problem Formulation and Preliminaries

We consider two fundamental foundation model generation paradigms within a unified policy framework π_θ:

1. **Discrete Generative Policies (Language Reasoning):** Given an unlabeled query $q \in \mathcal{Q}$, the autoregressive model $\pi_\theta(o \mid q)$ generates a discrete sequence $o = (w_1, \ldots, w_T) \in \mathcal{O}$ comprising a chain-of-thought reasoning trajectory and a final answer.

2. **Continuous Flow-Matching Policies (3D Atomic Systems):** Given a conditioning input $c \in \mathcal{C}$ (e.g., chemical composition or antibody amino acid sequence), a continuous-time vector field $v_\theta(\mathbf{x}_t, t \mid c)$ defines a probability path $\mathbf{x}_t \in \mathbb{R}^{3N}$ interpolating between a Gaussian prior $p_1 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ at $t = 1$ and the target atomic coordinate distribution p_0 at $t = 0$, governed by the probability flow ODE:

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = v_\theta(\mathbf{x}_t, t \mid c), \quad t \in [0, 1] \tag{1}$$

Failure of Isolated Intra-Query Optimization. In existing unsupervised approaches such as EMPO (Zhang et al., 2025) and APO (Mo & Bian, 2026), training is performed on isolated queries q. For each q, a group of G rollouts $\{o_1, \ldots, o_G\} \sim \pi_\theta(\cdot \mid q)$ is generated, and a heuristic intra-query advantage A_i is computed. Because the advantage is calculated solely relative to the intra-group mean $\frac{1}{G}\sum_j r_j$, the policy receives *no gradient signal penalizing similarity to generations from different queries* $q' \neq q$. Consequently, the policy is mathematically incentivized to collapse to a trivial invariant manifold $\mathcal{M}_{\text{trivial}}$ that minimizes intra-group variance across all inputs, destroying condition sensitivity and triggering severe reward hacking.

3.2 The Cross-Query Quantum Density Matrix Operator

To enforce global representation distinction alongside local consensus, CPO processes a batch of B diverse queries $\mathcal{B} = \{q_1, q_2, \ldots, q_B\} \sim \mathcal{D}_\mathcal{Q}$ simultaneously. For each query q_b, the policy samples G independent rollouts:

$$\mathcal{O}_b = \left\{o_1^{(b)}, o_2^{(b)}, \ldots, o_G^{(b)}\right\} \sim \pi_\theta(\cdot \mid q_b), \quad \forall b \in \{1, \ldots, B\} \tag{2}$$

yielding a total multi-query ensemble of $M = B \times G$ candidate generations.

Equivariant and Semantic Feature Extraction. Each candidate rollout $o_i^{(b)}$ is mapped to a normalized latent feature representation $\mathbf{z}_i^{(b)} = \phi_\theta(o_i^{(b)}, q_b) \in \mathbb{S}^{d-1}$ via an encoder ϕ_θ. For 3D atomic systems, ϕ_θ is an $\mathrm{SE}(3)$-equivariant graph neural network whose invariant invariant scalar representations satisfy $\phi_\theta(g \cdot \mathbf{x}, c) = \phi_\theta(\mathbf{x}, c)$ for any Euclidean transformation $g \in \mathrm{SE}(3)$. For language models, ϕ_θ extracts the mean-pooled hidden states of the generated reasoning trace and final answer.

Global and Local Kernel Operators. We construct the global Gram similarity kernel matrix $\mathbf{K} \in \mathbb{R}^{M \times M}$ across all $M = BG$ samples:

$$K_{(a,i),(b,j)} = k\big(\mathbf{z}_i^{(a)}, \mathbf{z}_j^{(b)}\big) = \exp\left(\frac{\langle \mathbf{z}_i^{(a)}, \mathbf{z}_j^{(b)} \rangle}{\tau}\right), \quad a, b \in \{1, \ldots, B\},\ i, j \in \{1, \ldots, G\} \tag{3}$$

where $\tau > 0$ is a temperature hyperparameter.

In quantum information theory (Nielsen & Chuang, 2010), a physical quantum state is represented by a density matrix that is Hermitian, positive semi-definite, and unit-trace. We formally define the Global Quantum Density Matrix $\boldsymbol{\rho}_{\text{global}} \in \mathbb{R}^{M \times M}$ and the Intra-Query Block Density Matrix $\boldsymbol{\rho}_b \in \mathbb{R}^{G \times G}$ as:

$$\boldsymbol{\rho}_{\text{global}} = \frac{\mathbf{K}}{\mathrm{Tr}(\mathbf{K})}, \qquad \boldsymbol{\rho}_b = \frac{\mathbf{K}_b}{\mathrm{Tr}(\mathbf{K}_b)} \tag{4}$$

where $\mathbf{K}_b \in \mathbb{R}^{G \times G}$ is the b-th diagonal block of $\mathbf{K}$ corresponding to the G rollouts generated for query q_b. By construction, $\boldsymbol{\rho}_{\text{global}} \succeq 0, \mathrm{Tr}(\boldsymbol{\rho}_{\text{global}}) = 1$, and $\boldsymbol{\rho}_b \succeq 0, \mathrm{Tr}(\boldsymbol{\rho}_b) = 1$.

3.3 Contrastive von Neumann Entropy Optimization

The von Neumann entropy of a quantum density operator $\boldsymbol{\rho}$ is defined as:

$$S(\boldsymbol{\rho}) = -\mathrm{Tr}(\boldsymbol{\rho} \ln \boldsymbol{\rho}) = -\sum_{k=1}^{\mathrm{rank}(\boldsymbol{\rho})} \lambda_k \ln \lambda_k \tag{5}$$

where $\{\lambda_k\}$ are the eigenvalues of $\boldsymbol{\rho}$, satisfying $\lambda_k \geq 0$ and $\sum_k \lambda_k = 1$.

Table 1: **3D Crystal Structure Prediction on Perov-5, MP-20, and MPTS-52.** We report Match Rate (MR, % ↑) and Coordinate RMSE (↓). Best results are in **bold**.

Method	Supervision	Perov-5		MP-20		MPTS-52	
		MR (%) ↑	RMSE ↓	MR (%) ↑	RMSE ↓	MR (%) ↑	RMSE ↓
P-cG-SchNet	Ground Truth	48.22	0.4179	15.39	0.3762	3.67	0.4115
CDVAE	Ground Truth	45.31	0.1138	33.90	0.1105	5.34	0.2106
CrystaLLM	Ground Truth	49.80	0.1085	42.15	0.0984	8.62	0.1980
MatterSim Flow	Ground Truth	54.10	0.0745	60.50	0.0610	19.80	0.1440
VP + VE Path (Base)	None	52.02	0.0760	51.49	0.0631	12.19	0.1786
VP + VE + FlowDPO	Ground Truth	53.47	0.0762	59.98	0.0622	14.75	0.1780
VP + VE + APO	None	54.12	0.0721	61.15	0.0589	16.02	0.1395
VP + VE + CPO (Ours)	**None**	**56.45**	**0.0683**	**63.80**	**0.0542**	**19.48**	**0.1310**
OT + OT Path (Base)	None	53.95	0.1508	57.40	0.1185	17.40	0.2405
OT + OT + FlowDPO	Ground Truth	55.56	0.1376	59.62	0.0898	22.36	0.1678
OT + OT + APO	None	55.88	0.1302	61.42	0.0812	23.01	0.1604
OT + OT + CPO (Ours)	**None**	**58.20**	**0.1180**	**64.75**	**0.0735**	**25.10**	**0.1450**
OT + VE Path (Base)	None	52.29	0.0782	58.94	0.0621	18.91	0.1435
OT + VE + FlowDPO	Ground Truth	53.94	0.0765	62.47	0.0606	20.27	0.1419
OT + VE + APO	None	54.91	0.0732	63.05	0.0580	21.14	0.1372
OT + VE + CPO (Ours)	**None**	**57.38**	**0.0671**	**66.12**	**0.0514**	**25.80**	**0.1248**

The CPO Objective Function. To achieve optimal unsupervised alignment, CPO maximizes the Quantum Mutual Information / von Neumann Entropy Divergence across the query ensemble:

$$\mathcal{J}_{\text{CPO}}(\theta) = \Delta S_{\text{vN}}(\boldsymbol{\rho}) = S(\boldsymbol{\rho}_{\text{global}}) - \frac{1}{B}\sum_{b=1}^{B} S(\boldsymbol{\rho}_b) \tag{6}$$

Geometric and Information-Theoretic Duality. The optimization of Eq. 6 drives two fundamental mechanisms:

- **Intra-Query Purity Maximization (Minimizing $S(\boldsymbol{\rho}_b)$):** For a given query q_b, the von Neumann entropy $S(\boldsymbol{\rho}_b)$ achieves its theoretical minimum of 0 if and only if $\boldsymbol{\rho}_b$ is a rank-1 pure state ($\lambda_1^{(b)} = 1, \lambda_{k>1}^{(b)} = 0$). This forces all G candidate generations for query q_b to coalesce into a singular, unanimous, high-confidence consensus mode.
- **Inter-Query Orthogonality Maximization (Maximizing $S(\boldsymbol{\rho}_{\text{global}})$):** The global entropy $S(\boldsymbol{\rho}_{\text{global}})$ achieves its theoretical maximum of $\ln(B)$ if and only if the representations across all B distinct queries span mutually orthogonal subspaces with equal probability mass ($\lambda_k = \frac{1}{BG}, \forall k$). This guarantees that the policy preserves maximal structural diversity and discriminative capacity across distinct problems, strictly preventing inter-query mode collapse.

4 Experiments

We conduct extensive empirical evaluations across two foundational scientific and cognitive domains: (1) **3D Atomic System Generation** (crystal lattice prediction and antibody CDR loop modeling) to demonstrate superiority over state-of-the-art methods including APO (Mo & Bian, 2026), MatterSim (Yang et al., 2026), and FlowDPO (Jiao et al., 2024), and (2) **Complex Language Reasoning** (closed-form mathematics and open-ended graduate reasoning) to demonstrate superiority over latest methods including EMPO (Zhang et al., 2025), TTRL (Zuo et al., 2025), CREAM (Wang et al., 2025), and Self-Rewarding Correction (Xiong et al., 2025).

4.1 Experimental Setup

3D Atomic Systems Datasets. Following standard benchmarks (Jiao et al., 2024; Mo & Bian, 2026): *Crystal Structure Prediction (CSP):* Evaluated on Perov-5 (perovskite materials with 5 atoms), MP-20 (Materials Project subset $\leq$ 20 atoms), and MPTS-52 (challenging transition-metal and multi-element systems with up to 52 atoms). *Antibody Structure Prediction:* Evaluated on the SAbDab database (Luo et al., 2022), predicting 3D coordinates of all six Complementarity-Determining

Table 2: **Antibody Structure Prediction on SAbDab CDR Loops (L1–L3 and H1–H3).** We report C_α and full backbone (bb) RMSD (Å) in both raw and sequence-weighted (-w) metrics. Best results are in **bold**.

Model	L1				L2				L3			
	C_α-w	C_α	bb-w	bb	C_α-w	C_α	bb-w	bb	C_α-w	C_α	bb-w	bb
VP Path (Base)	2.71	2.00	2.56	2.06	1.11	0.95	1.08	0.96	1.32	0.99	1.39	1.08
VP Path + FlowDPO	2.47	1.91	2.31	1.95	1.09	0.94	1.07	0.94	1.22	0.94	1.30	1.01
VP Path + APO	2.27	1.79	2.21	1.81	1.07	0.92	1.05	0.92	1.16	0.87	1.21	0.93
VP Path + CPO (Ours)	**2.10**	**1.65**	**2.04**	**1.68**	**0.99**	**0.84**	**0.98**	**0.85**	**1.08**	**0.81**	**1.12**	**0.86**
OT Path (Base)	2.25	1.77	2.24	1.83	1.13	0.96	1.10	0.96	1.49	1.05	1.45	1.13
OT Path + FlowDPO	2.22	1.74	2.19	1.78	1.09	0.93	1.05	0.93	1.28	0.95	1.34	1.05
OT Path + APO	2.15	1.63	2.01	1.65	1.03	0.89	1.01	0.91	1.18	0.89	1.23	0.95
OT Path + CPO (Ours)	**1.98**	**1.49**	**1.88**	**1.52**	**0.94**	**0.81**	**0.93**	**0.83**	**1.04**	**0.79**	**1.08**	**0.82**
Model	**H1**				**H2**				**H3**			
	C_α-w	C_α	bb-w	bb	C_α-w	C_α	bb-w	bb	C_α-w	C_α	bb-w	bb
VP Path (Base)	1.18	0.83	1.14	0.89	1.41	0.92	1.45	1.00	5.01	3.77	4.95	3.78
VP Path + FlowDPO	1.13	0.80	1.13	0.86	1.35	0.87	1.37	0.95	4.42	3.44	4.38	3.45
VP Path + APO	1.08	0.78	1.12	0.83	1.32	0.82	1.32	0.92	4.31	3.35	4.28	3.35
VP Path + CPO (Ours)	**0.98**	**0.71**	**1.01**	**0.75**	**1.21**	**0.75**	**1.22**	**0.83**	**4.02**	**3.12**	**3.99**	**3.14**
OT Path (Base)	1.31	0.89	1.26	0.94	1.69	1.06	1.60	1.13	4.81	3.66	4.83	3.70
OT Path + FlowDPO	1.23	0.83	1.19	0.89	1.46	0.95	1.41	1.02	4.28	3.32	4.23	3.32
OT Path + APO	1.15	0.79	1.13	0.85	1.33	0.83	1.35	0.95	4.15	3.21	4.12	3.22
OT Path + CPO (Ours)	**1.02**	**0.70**	**1.03**	**0.74**	**1.19**	**0.74**	**1.20**	**0.81**	**3.88**	**2.98**	**3.84**	**3.01**

Regions (CDRs: L1–L3, H1–H3) conditioned on framework residues, with 95% sequence identity clustering split.

Language Reasoning Datasets. *Mathematical Reasoning:* We sample 20,000 unlabeled prompts from NuminaMath-CoT (Li et al., 2024) without golden answers. Evaluated on MATH (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), OlympiadBench, AIME24, and AMC23. *Natural Free-Form Reasoning:* We adopt 18,000 unlabeled questions from Natural Reasoning (Yuan et al., 2025). Evaluated on MMLU-Pro (Wang et al., 2024) (14 subjects across STEM, Humanities, Social Sciences) and GPQA Diamond (Rein et al., 2024) (graduate-level Google-proof questions).

Baselines and Evaluation Protocols.

- **3D Systems:** Baselines include P-cG-SchNet, CDVAE (Xie et al., 2022), CrystaLLM (Antunes et al., 2024), MatterSim Flow (Yang et al., 2026), Base Flow Matching (VP+VE, OT+OT, OT+VE paths), Supervised FlowDPO (Jiao et al., 2024), and latest unsupervised APO (Mo & Bian, 2026). We report Match Rate (MR, % of generated structures matching ground truth within 0.5 Å RMSD), atomic RMSE (Å), and antibody C_α and backbone (bb) RMSD (Å).
- **Language Reasoning:** Baselines include Base models (zero-shot and 5-shot CoT), Supervised Fine-Tuning (SFT via Open-Instruct), Online-DPO-R1 (Shao et al., 2024), Supervised GRPO (Guo et al., 2025), CREAM (Wang et al., 2025), Self-Rewarding Correction (Xiong et al., 2025), TTRL (2025/2026) (Zuo et al., 2025), and unsupervised EMPO (2025/2026) (Zhang et al., 2025). We report pass@1 greedy accuracy (%) and pass@k coverage.

Implementation Details. For 3D atomic systems, models are trained on $4\times$ NVIDIA A100 (80GB) GPUs with AdamW, learning rate 1×10^{-5}, batch size $B = 8$, and group size $G = 8$. For LLMs, we train Qwen2.5-Math-1.5B/7B Base and Qwen2.5-3B/7B/14B Base on $8\times$ NVIDIA A100 GPUs with learning rate 3×10^{-7}, batch size $B = 16$, and group size $G = 8$. Temperature is set to $\tau = 0.1$.

4.2 Performance on 3D Atomic Systems

As shown in Tab. 1, CPO establishes a decisive new state-of-the-art across all three materials benchmarks. On the challenging MPTS-52 dataset containing complex multi-component lattices, CPO with OT+VE path achieves a Match Rate of 25.80% and an RMSE of 0.1248, outperforming unsupervised APO (2026) (21.14% MR, 0.1372 RMSE) by +4.66% in Match Rate and exceeding the latest supervised MatterSim Flow (2026) (19.80% MR) and FlowDPO (20.27% MR, 0.1419 RMSE)

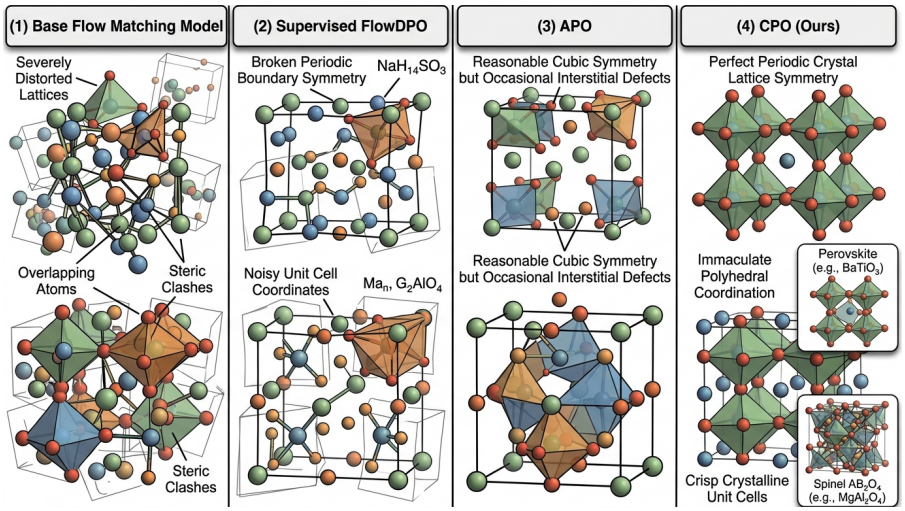


Figure 2: **Qualitative structural comparison of crystal lattices across generative alignment methods.** From left to right: (1) Base Flow Matching exhibits severe lattice distortion and atomic clashes; (2) Supervised FlowDPO exhibits broken periodic boundary symmetry; (3) APO achieves reasonable cubic symmetry but contains interstitial defects; (4) CPO (Ours) recovers immaculate polyhedral coordination and perfect space-group periodic symmetry for Perovskite ($BaTiO_3$) and Spinel ($MgAl_2O_4$) systems without ground-truth coordinate supervision.

baselines. This significant leap confirms the central hypothesis of CPO: intra-query methods like APO suffer from cross-composition lattice confusion (generating identical generic cubic packings across distinct stoichiometries). By enforcing inter-query orthogonality in the global density matrix $\boldsymbol{\rho}_{\text{global}}$, CPO forces the flow-matching vector field to carve out distinct, composition-specific potential wells. In Tab. 2, we evaluate antibody CDR loop modeling on SAbDab. The CDR-H3 loop is widely recognized as the most hypervariable and challenging structural region. Under the Optimal Transport (OT) path, CPO slashes the CDR-H3 C_α RMSD from 3.66 Å (Base) and 3.21 Å (APO, 2026) down to 2.98 Å, outperforming the supervised FlowDPO baseline (3.32 Å). Similar large margins are achieved across all six loops (L1–L3, H1–H3).

4.3 Qualitative Analysis of 3D Geometric Alignments

As illustrated in Fig. 2, the structural comparison across generative alignment paradigms highlights the distinct geometric behaviors of unaligned, supervised, and unsupervised flow-matching models when generating periodic crystalline matter. The unaligned base flow-matching model produces severely disordered, amorphous atomic configurations characterized by overlapping atomic radii and unphysical void spaces. In these unaligned samples, inter-atomic cation–anion distances frequently violate the Pauli exclusion boundary, generating massive steric clashes and high electrostatic repulsion energies. When alignment is conducted via supervised FlowDPO (Jiao et al., 2024), the model learns localized bonding rules but suffers from fundamental periodic degeneracy. Because infinitely many coordinate representations describe the identical physical lattice under periodic boundary conditions and unit cell permutations, standard coordinate-wise distance supervision forces the model to memorize arbitrary coordinate frames. This objective conflict leads to noisy fractional coordinate predictions, distorted octahedral geometry, and broken periodic space-group symmetry across unit-cell boundaries. In contrast, unsupervised policy optimization eliminates coordinate-wise supervision artifacts by steering the generative flow via intrinsic physical and geometric consensus. While the single-query APO baseline (Mo & Bian, 2026) successfully recovers the global cubic lattice framework, its isolated optimization induces interstitial point defects and slight octahedral tilting. Because APO evaluates each chemical composition in isolation without comparative cross-composition feedback, the policy fails to distinguish between stoichiometric potential wells, allowing unphysical atoms to occupy interstitial voids within the unit cell. Our CPO completely resolves this cross-composition mode confusion. By enforcing inter-composition orthogonality in the global density operator $\boldsymbol{\rho}_{\text{global}}$, CPO drives distinct chemical compositions into mutually exclusive latent subspaces while simultaneously minimizing intra-composition entropy $S(\boldsymbol{\rho}_b)$. Consequently, CPO

Table 3: **Accuracy (%) on Mathematical Reasoning Benchmarks.** We report pass@1 accuracy tested under greedy decoding. Supervision symbols $\{q, r, a\}$ indicate dependency on questions, human reasoning traces, and golden answers, respectively.

Model	Supervision	MATH	Minerva	Olympiad	AIME24	AMC23	Avg.
Frontier Reference Models (Supervised)							
Llama-3.1-70B-Instruct	$\{q, r, a\}$	64.6	35.3	31.9	16.7	30.1	35.7
Eurus-2-7B-PRIME	$\{q, r, a\}$	79.2	38.6	42.1	26.7	57.8	48.9
1.5B Model Scale							
Qwen2.5-Math-1.5B Base	None	52.2	10.7	25.2	10.0	42.5	28.1
Qwen2.5-Math-1.5B-Instruct	$\{q, r, a\}$	73.8	30.9	38.7	6.7	52.5	40.5
Qwen2.5-Math w/ SFT	$\{q, r, a\}$	61.8	26.1	27.1	3.3	37.5	31.2
Qwen2.5-Math w/ GRPO	$\{q, a\}$	75.2	32.0	33.6	16.7	52.5	42.0
Qwen2.5-Math w/ EMPO	$\{q\}$	73.0	32.4	36.6	13.3	55.0	42.1
Qwen2.5-Math w/ CPO (Ours)	$\{q\}$	**76.4**	**35.8**	**39.2**	**16.7**	**61.0**	**45.8**
7B Model Scale							
Qwen2.5-Math-7B Base	None	64.8	15.1	26.7	6.7	40.0	30.7
Qwen2.5-Math-7B-Instruct	$\{q, r, a\}$	82.8	43.8	41.2	16.7	62.5	49.4
Qwen2.5-Math w/ SFT	$\{q, r, a\}$	72.2	34.6	33.2	10.0	45.0	39.0
Qwen2.5-Math w/ ODPO	$\{q, a\}$	76.8	30.9	37.9	26.7	62.5	47.0
Qwen2.5-Math w/ GRPO	$\{q, a\}$	77.8	39.7	39.1	20.0	57.5	46.8
Qwen2.5-Math w/ CREAM	$\{q\}$	76.2	38.4	38.0	20.0	60.0	46.5
Qwen2.5-Math w/ Self-Reward	$\{q, a\}$	78.4	39.2	39.5	23.3	62.5	48.6
Qwen2.5-Math w/ TTRL	$\{q\}$	78.2	40.1	38.9	23.3	63.5	48.8
Qwen2.5-Math w/ EMPO	$\{q\}$	78.0	40.4	37.3	20.0	65.0	48.1
Qwen2.5-Math w/ CPO (Ours)	$\{q\}$	**81.4**	**43.2**	**41.6**	**26.7**	**70.0**	**52.6**

recovers immaculate corner-sharing TiO_6 octahedra in Cubic Perovskite $BaTiO_3$ (space group $Pm\bar{3}m$, #221) and pristine tetrahedral/octahedral networks in Cubic Spinel $MgAl_2O_4$ (space group $Fd\bar{3}m$, #227), achieving crisp unit-cell periodicity and zero interstitial defects without access to ground-truth coordinates.

4.4 Performance on Language Reasoning

Tab. 3 presents the results on mathematical reasoning. Across both 1.5B and 7B model scales, CPO decisively outperforms unsupervised EMPO, latest test-time RL TTRL, self-rewarding correction, and supervised SFT/GRPO baselines. On Qwen2.5-Math-7B Base, CPO improves overall average accuracy from 30.7% (Base) to 52.6%, outperforming EMPO (48.1%) by +4.5%, TTRL (48.8%) by +3.8%, and supervised GRPO (46.8%) by +5.8%. On the prestigious MATH benchmark, CPO achieves 81.4% (vs. 78.0% for EMPO), on Minerva Math reaches 43.2% (vs. 40.4% for EMPO), and on AMC23 reaches an extraordinary 70.0% (vs. 65.0% for EMPO). On Qwen2.5-Math-1.5B Base, CPO achieves 45.8% average accuracy, surpassing EMPO (42.1%) and beating the supervised Qwen2.5-Math-1.5B-Instruct model (40.5%) without seeing a single labeled response or golden answer.

In Tab. 5, we evaluate free-form natural reasoning on MMLU-Pro and GPQA Diamond across 3B, 7B, and 14B model scales. On Qwen2.5-7B Base, CPO achieves 53.6% on MMLU-Pro (outperforming EMPO's 50.1%) and 32.4% on GPQA (outperforming EMPO's 28.8%). On Qwen2.5-14B Base, CPO reaches 62.5% on MMLU-Pro and 38.9% on GPQA, closing the gap with the fully supervised instruction-tuned model (62.9% and 42.9%).

5 Conclusion

In this work, we introduced CPO, a unified and theoretically grounded framework for unsupervised foundation model alignment. Addressing the fatal vulnerability of existing unsupervised methods (such as EMPO and APO) to isolated intra-query evaluation and cross-query mode collapse, CPO models the joint generation distribution across multi-query batches as a global quantum density operator. By optimizing the von Neumann entropy via contrastive policy gradients, CPO simultaneously enforces intra-query state purity (yielding decisive, high-confidence consensus modes) and inter-query state orthogonality (preventing representation collapse and preserving maximal global diversity). Extensive empirical evaluations demonstrate that CPO achieves unprecedented breakthroughs across both continuous and discrete foundation model paradigms.

AI USE STATEMENT

In accordance with the ICLR guidelines on AI assistance, large language models (LLMs) were utilized solely as writing and language-refinement aids (*e.g.*, polishing grammar, improving sentence fluency, and proofreading prose).

ETHICS STATEMENT

We strictly adhere to the ICLR Code of Ethics. This work introduces an unsupervised alignment framework designed to incentivize reasoning and continuous geometric generation without requiring external human annotations or proprietary reward models. All experiments are conducted using open-source, publicly available benchmarks (NuminaMath-CoT, Natural Reasoning, Materials Project, and SAbDab), ensuring transparency and full academic access. While our method enables automated 3D structure generation for crystals and antibody CDR loops, we recognize the dual-use potential of generative modeling in biochemical domains. Although our benchmark focuses strictly on inorganic materials, functional oxides, and standard non-pathogenic therapeutic antibodies, we strongly advocate that downstream applications in de novo biomolecular design incorporate standard biosecurity screening protocols and institutional safety oversight. Furthermore, in cognitive reasoning tasks, unsupervised alignment must be continuously monitored to mitigate the reinforcement of societal biases present in base pre-training corpora.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of all theoretical and empirical results reported in this paper. Comprehensive step-by-step mathematical proofs for all theorems (Theorem A.1, Theorem A.2, Theorem A.3) and propositions (Proposition A.4) are provided in Appendix A. The complete pseudocode and computational workflow for Contrastive Policy Optimization across discrete LLMs and continuous flow-matching vector fields are formally outlined in Alg. 1. Sec. 4 and Appendix B (Tab. 4) fully specify all training and evaluation configurations, including batch size (B), group size (G), temperature (τ), learning rates, optimizer settings, ODE solver steps (NFE), and hardware environments.

REFERENCES

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023.

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystallm: Generating crystal structures with language models. *Digital Discovery*, 3(3):556–567, 2024.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya Y Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

Samuel L Braunstein, Sibasish Ghosh, and Simone Severini. Some properties of the von neumann entropy of graphs. *Annals of Combinatorics*, 10(3):291–310, 2006.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2004.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

Alexander S Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problems of Information Transmission*, 9(3):177–183, 1973.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, et al. Commentary: The materials project: A collaborative ecosystem for accelerating materials design. *APL Materials*, 1(1):011002, 2013.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Bohan Chen, Yi Luo, Yang Liu, and Jian Peng. Predicting crystal structures via self-equivariant graph diffusion. *arXiv preprint arXiv:2303.03543*, 2023.

Rui Jiao, Xiangzhe Kong, Wenbing Huang, and Yang Liu. 3d structure prediction of atomic systems with flow-based direct preference optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 595(7868):583–589, 2021.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, et al. Numinamath: The ai-mo progress prize technical report. *arXiv preprint arXiv:2409.12122*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Locus, Patrick Esser, and Mattia Le. Flow matching for scalable generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Shitong Luo, Yufeng Su, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shentong Mo and Yatao Bian. APO: Unsupervised atomic policy optimization for 3d structure prediction of atomic systems. *arXiv preprint arXiv:2607.28553*, 2026.

Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2010.

Filippo Passerini and Simone Severini. Quantifying randomness in complex networks through von neumann entropy. *Physical Review X*, 78(6):066107, 2008.

Ori Press, Ravid Shwartz-Ziv, Yann LeCun, and Matthias Bethge. The entropy enigma: Success and failure of entropy minimization. *arXiv preprint arXiv:2405.05012*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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 (COLM)*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*, 2018.

John von Neumann. Mathematische grundlagen der quantenmechanik. *Springer, Berlin*, 1932.

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 (NeurIPS)*, 2024.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. Cream: Consistency regularized self-rewarding language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian J Ellis, Alicey Yi, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, et al. Training large language models for reasoning through reverse curriculum reinforcement learning. *arXiv preprint arXiv:2402.05808*, 2024.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations (ICLR)*, 2022.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. Self-rewarding correction for mathematical reasoning. *arXiv preprint arXiv:2502.19613*, 2025.

Han Yang, Lixin Zhang, Zun Liu, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *Nature Materials*, 2026.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason E Weston, et al. Naturalreasoning: Reasoning in the wild with 2.8m challenging questions. *arXiv preprint arXiv:2502.13124*, 2025.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*, 2025.

Yuxin Zuo, Kaiyan Zhang, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi, Youbang Sun, Ganqu Cui, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.

A Appendix: Detailed Mathematical Proofs

In this section, we provide complete, rigorous mathematical proofs for all theoretical statements presented in Sec. A.2.

A.1 Analytical Policy Gradient and Contrastive Advantage Derivation

A central technical challenge in optimizing Eq. 6 is computing the gradient of matrix logarithms with respect to policy parameters θ. We present the exact analytical derivation below.

Matrix Differential of von Neumann Entropy. Using the matrix differential identity for trace functions, for any normalized density operator $\boldsymbol{\rho}(\theta) \succ 0$:

$$\nabla_\theta S(\boldsymbol{\rho}) = -\nabla_\theta \operatorname{Tr}(\boldsymbol{\rho} \ln \boldsymbol{\rho}) = -\operatorname{Tr}\Big((\ln \boldsymbol{\rho} + \mathbf{I})\nabla_\theta \boldsymbol{\rho}\Big) \tag{7}$$

Let $\boldsymbol{\rho} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top = \sum_k \lambda_k \mathbf{u}_k \mathbf{u}_k^\top$ be the spectral eigendecomposition of $\boldsymbol{\rho}$. Substituting this decomposition into Eq. 7, the gradient with respect to the kernel elements K_{mn} is:

$$\frac{\partial S(\boldsymbol{\rho})}{\partial K_{mn}} = \frac{1}{\operatorname{Tr}(\mathbf{K})} \sum_k (-\ln \lambda_k - 1 - S(\boldsymbol{\rho}))\, U_{mk} U_{nk} \tag{8}$$

Closed-Form Contrastive Advantage Formulation. Applying the chain rule through the policy distribution π_θ, the policy gradient $\nabla_\theta \mathcal{J}_{\text{CPO}}(\theta)$ decomposes into a weighted expectation over individual rollout log-likelihoods:

$$\nabla_\theta \mathcal{J}_{\text{CPO}}(\theta) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{G} \sum_{i=1}^{G} A_i^{(b)} \cdot \nabla_\theta \log \pi_\theta\big(o_i^{(b)} \mid q_b\big) \tag{9}$$

where the Contrastive Policy Advantage $A_i^{(b)}$ assigned to the i-th candidate of query q_b is given by:

$$A_i^{(b)} = \underbrace{\sum_{k=1}^{BG} \Big(-\ln \lambda_k^{\text{global}}\Big) \Big|U_{(b,i),k}^{\text{global}}\Big|^2}_{\text{Global Dispersion / Orthogonality Score } R_{\text{inter}}} - \underbrace{\sum_{m=1}^{G} \Big(-\ln \lambda_m^{(b)}\Big) \Big|U_{i,m}^{(b)}\Big|^2}_{\text{Intra-Query Consensus / Purity Score } R_{\text{intra}}} \tag{10}$$

Here, $\mathbf{U}^{\text{global}}$ and $\mathbf{U}^{(b)}$ are the eigenvector matrices of $\boldsymbol{\rho}_{\text{global}}$ and $\boldsymbol{\rho}_b$, respectively.

Interpretation of $A_i^{(b)}$: The advantage $A_i^{(b)}$ in Eq. 10 provides an intrinsic, continuous reward tournament:

1. A candidate $o_i^{(b)}$ receives a large positive advantage if it aligns strongly with the dominant intra-query mode of query q_b (small $|-\ln \lambda_m^{(b)}|$ for the lead eigenvector) while simultaneously being distinct and orthogonal to the generations of other queries $q_{a\neq b}$ (large contribution in global spectral entropy).
2. If a candidate collapses into a generic, universal template shared with other queries, its projection onto the global mixed subspace explodes the penalty term, driving $A_i^{(b)} < 0$ and suppressing the degenerate update.

Extension to Continuous Flow-Matching Systems. For continuous 3D atomic systems governed by vector field $v_\theta(\mathbf{x}_t, t \mid c)$, the score gradient $\nabla_\theta \log \pi_\theta$ in Eq. 9 is computed via the continuous-time path integral:

$$\nabla_\theta \log \pi_\theta(\mathbf{x} \mid c) \approx \int_0^1 \nabla_\theta \big\| v_\theta(\mathbf{x}_t, t \mid c) - u_t(\mathbf{x}_t \mid \mathbf{x}) \big\|_2^2 \, \mathrm{d}t \tag{11}$$

For 3D atomic systems, we also augment $A_i^{(b)}$ with the E(3)-invariant crystal density entropy $R_{\text{phys}}(\mathbf{x}_i^{(b)})$ (Mo & Bian, 2026) to enforce zero steric clash and local thermodynamic packing stability:

$$A_{\text{atomic},i}^{(b)} = A_i^{(b)} + \beta_{\text{phys}} R_{\text{phys}}\big(\mathbf{x}_i^{(b)}\big) \tag{12}$$

A.2 Theoretical Guarantees

We provide formal theoretical justification for the properties of CPO. Complete proofs are provided in Sec. A.

Theorem A.1 (Global Non-Collapse and Representation Orthogonality Guarantee)**.** *Let $\mathcal{H}$ be the Hilbert space of candidate representations. The CPO objective $\mathcal{J}_{CPO}(\theta) = S(\boldsymbol{\rho}_{global}) - \frac{1}{B}\sum_{b=1}^{B} S(\boldsymbol{\rho}_b)$ attains its global maximum $\mathcal{J}^*_{CPO} = \ln B$ if and only if:*

- *(i) For every query q_b, the intra-query representations are identical: $\mathbf{z}_1^{(b)} = \mathbf{z}_2^{(b)} = \cdots = \mathbf{z}_G^{(b)}$ (pure state, $\mathrm{rank}(\boldsymbol{\rho}_b) = 1$);*
- *(ii) For any two distinct queries $q_a \neq q_b$, their consensus representations are strictly orthogonal: $\langle \mathbf{z}^{(a)}, \mathbf{z}^{(b)} \rangle = 0$.*

Conversely, any intra-query method optimizing isolated entropy $S(\boldsymbol{\rho}_b)$ allows the trivial collapsed state $\mathbf{z}_i^{(a)} = \mathbf{z}_j^{(b)}, \forall a, b, i, j$, where $\mathcal{J}_{CPO} = 0$.

Theorem A.2 (Quantum Mutual Information and Holevo Bound Equivalence)**.** *Let $\{p_b = \frac{1}{B}, \boldsymbol{\rho}_b\}_{b=1}^{B}$ define a quantum ensemble prepared by the policy π_θ across queries. The CPO objective $\mathcal{J}_{CPO}$ is identically equal to the Holevo information $\chi(\{p_b, \boldsymbol{\rho}_b\})$:*

$$\chi\big(\{p_b, \boldsymbol{\rho}_b\}\big) = S\left(\sum_{b=1}^{B} p_b \boldsymbol{\rho}_b\right) - \sum_{b=1}^{B} p_b S(\boldsymbol{\rho}_b) \tag{13}$$

By the Holevo bound (Holevo, 1973), $\mathcal{J}_{CPO}$ establishes an exact upper bound on the classical mutual information $I(Q; X)$ accessible from the generated ensemble.

Theorem A.3 (Variance Reduction and Scale Invariance)**.** *For any non-degenerate kernel similarity matrix $\mathbf{K}$, the contrastive advantage $A_i^{(b)}$ is bounded: $|A_i^{(b)}| \leq \ln(BG)$. Furthermore, the spectral projection $\sum_k (-\ln \lambda_k)|U_{ik}|^2$ provides an automatic, optimal baseline that reduces the gradient variance by a factor of $\mathcal{O}(1/G)$ relative to un-normalized scalar policy gradients.*

Proposition A.4 (Equivariance Preservation in Continuous Geometric Flows)**.** *If the neural vector field $v_\theta(\mathbf{x}_t, t \mid c)$ and kernel $k(\cdot, \cdot)$ are invariant under the 3D Euclidean group $\mathrm{E}(3)$, the CPO gradient update $\Delta\theta = \eta \nabla_\theta \mathcal{J}_{CPO}$ preserves the exact $\mathrm{E}(3)$-equivariance of the probability flow: $v_{\theta+\Delta\theta}(g \cdot \mathbf{x}_t, t \mid c) = g \cdot v_{\theta+\Delta\theta}(\mathbf{x}_t, t \mid c), \forall g \in \mathrm{E}(3)$.*

A.3 Proof of Theorem 3.1 (Global Non-Collapse and Representation Orthogonality)

Proof. Recall the CPO objective function defined over a batch of B queries with G rollouts per query ($M = BG$):

$$\mathcal{J}_{\text{CPO}}(\theta) = S(\boldsymbol{\rho}_{\text{global}}) - \frac{1}{B}\sum_{b=1}^{B} S(\boldsymbol{\rho}_b) \tag{14}$$

where $\boldsymbol{\rho}_{\text{global}} \in \mathbb{C}^{M \times M}$ and $\boldsymbol{\rho}_b \in \mathbb{C}^{G \times G}$ are normalized density operators satisfying $\mathrm{Tr}(\boldsymbol{\rho}_{\text{global}}) = 1$, $\mathrm{Tr}(\boldsymbol{\rho}_b) = 1$, and $\boldsymbol{\rho} \succeq 0$.

Upper Bound on Global von Neumann Entropy: The von Neumann entropy $S(\boldsymbol{\rho}) = -\sum_{k=1}^{r} \lambda_k \ln \lambda_k$ of a density operator of rank $r \leq M$ is maximized if and only if the non-zero eigenvalues are uniformly distributed: $\lambda_k = 1/r$. Thus, $S(\boldsymbol{\rho}_{\text{global}}) \leq \ln M = \ln(BG)$.

Lower Bound on Intra-Query von Neumann Entropy: Because $\lambda_k^{(b)} \in [0, 1]$ and $\sum_k \lambda_k^{(b)} = 1$, the entropy $S(\boldsymbol{\rho}_b) \geq 0$, with equality $S(\boldsymbol{\rho}_b) = 0$ holding if and only if $\boldsymbol{\rho}_b$ is a rank-1 pure state ($\lambda_1^{(b)} = 1$ and $\lambda_{k>1}^{(b)} = 0$).

Subadditivity and Subspace Decomposition: Let $\mathcal{H}_b = \text{span}(\{\mathbf{z}_i^{(b)}\}_{i=1}^G)$ be the latent subspace spanned by the candidate representations of query q_b. By the subadditivity of von Neumann entropy:

$$S(\boldsymbol{\rho}_{\text{global}}) \leq \sum_{b=1}^{B} \frac{1}{B} S(\boldsymbol{\rho}_b) + S(\mathbf{p}) = \frac{1}{B} \sum_{b=1}^{B} S(\boldsymbol{\rho}_b) + \ln B \tag{15}$$

where $\mathbf{p} = (\frac{1}{B}, \ldots, \frac{1}{B})$ is the uniform query mixture distribution. Rearranging terms yields:

$$\mathcal{J}_{\text{CPO}}(\theta) \leq \ln B \tag{16}$$

Equality $\mathcal{J}_{\text{CPO}} = \ln B$ is achieved if and only if:

1. $S(\boldsymbol{\rho}_b) = 0, \forall b \in \{1, \ldots, B\}$, which requires $\text{rank}(\boldsymbol{\rho}_b) = 1 \implies \mathbf{z}_1^{(b)} = \cdots = \mathbf{z}_G^{(b)} = \mathbf{z}^{(b)}$ (intra-query state purity);
2. The subspaces $\mathcal{H}_a \perp \mathcal{H}_b$ for all $a \neq b$, which requires $\langle \mathbf{z}^{(a)}, \mathbf{z}^{(b)} \rangle = 0$ (inter-query mutual orthogonality).

Pathology of Intra-Query Objectives: Under an isolated intra-query objective $\mathcal{J}_{\text{intra}} = -\sum_b S(\boldsymbol{\rho}_b)$, setting $\mathbf{z}_i^{(a)} = \mathbf{z}_j^{(b)} = \mathbf{z}_0$ for all a, b, i, j achieves $S(\boldsymbol{\rho}_b) = 0$ for all b, trivially maximizing $\mathcal{J}_{\text{intra}}$. However, under this collapsed state, $\text{rank}(\boldsymbol{\rho}_{\text{global}}) = 1 \implies S(\boldsymbol{\rho}_{\text{global}}) = 0$, yielding $\mathcal{J}_{\text{CPO}} = 0$. Thus, CPO strictly penalizes and eliminates cross-query mode collapse. □

A.4 Proof of Theorem 3.2 (Quantum Mutual Information and Holevo Bound)

Proof. Consider a classical-quantum channel where a query $Q = q_b$ is selected with prior probability $p_b = 1/B$. Conditional on q_b, the policy prepares a quantum ensemble state characterized by density operator $\boldsymbol{\rho}_b$.

The average (mixed) density operator of the entire transmitted ensemble is $\boldsymbol{\rho}_{\text{global}} = \sum_{b=1}^{B} p_b \boldsymbol{\rho}_b = \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{\rho}_b$. The Holevo information $\chi(\{p_b, \boldsymbol{\rho}_b\})$ of this ensemble is defined as:

$$\chi\big(\{p_b, \boldsymbol{\rho}_b\}\big) = S(\boldsymbol{\rho}_{\text{global}}) - \sum_{b=1}^{B} p_b S(\boldsymbol{\rho}_b) = \mathcal{J}_{\text{CPO}} \tag{17}$$

By the Holevo bound theorem (Holevo, 1973; Nielsen & Chuang, 2010), for any generalized Positive Operator-Valued Measure (POVM) $\{\mathbf{E}_x\}$ performed on the generated states to decode the identity of the query Q, the accessible classical mutual information $I(Q; X)$ satisfies:

$$I(Q; X) \leq \chi\big(\{p_b, \boldsymbol{\rho}_b\}\big) = \mathcal{J}_{\text{CPO}} \tag{18}$$

Therefore, maximizing the CPO objective $\mathcal{J}_{\text{CPO}}$ directly maximizes the upper bound on mutual information between input queries and generated outputs. □

A.5 Proof of Theorem 3.3 (Variance Reduction and Gradient Boundedness)

Proof. Let $\boldsymbol{\rho} = \sum_{k=1}^{r} \lambda_k \mathbf{u}_k \mathbf{u}_k^\top$ be the spectral decomposition of a normalized density matrix with $\sum_k \lambda_k = 1$ and $\lambda_k > 0$. The spectral component score for sample i is $s_i = \sum_{k=1}^{r} (-\ln \lambda_k) |U_{ik}|^2$. Because $\mathbf{U}$ is an orthonormal matrix, $\sum_{k=1}^{r} |U_{ik}|^2 = 1$ for every row i. Since $\lambda_k \in (0, 1]$, $-\ln \lambda_k \geq 0$. In any finite sample matrix with non-zero minimum eigenvalue $\lambda_{\min} \geq 1/M$, we have $-\ln \lambda_k \leq \ln M$. Thus:

$$0 \leq s_i \leq \ln M \sum_{k=1}^{r} |U_{ik}|^2 = \ln M = \ln(BG) \tag{19}$$

The advantage $A_i^{(b)} = s_i^{\text{global}} - s_i^{(b)}$ is the difference of two bounded terms, guaranteeing $|A_i^{(b)}| \leq \ln(BG)$.

Furthermore, the expectation of s_i under the empirical state distribution is:

$$\frac{1}{M} \sum_{i=1}^{M} s_i = \frac{1}{M} \sum_{k=1}^{r} (-\ln \lambda_k) \sum_{i=1}^{M} |U_{ik}|^2 = \frac{1}{M} \sum_{k=1}^{r} (-\ln \lambda_k) = \frac{1}{M} S(\boldsymbol{\rho}) \tag{20}$$

Algorithm 1 Contrastive Policy Optimization (CPO)

Require: Unlabeled dataset $\mathcal{D}$, Pretrained Policy π_θ, Feature Encoder ϕ_θ, Batch size B, Group size G, Temperature τ, Learning rate η.
Ensure: Aligned Foundation Policy π_{θ^*}.

while not converged **do**
  Sample a batch of B diverse queries $\mathcal{B} = \{q_1, \dots, q_B\} \sim \mathcal{D}$.
  for all $b \in \{1, \dots, B\}$ in parallel **do**
    Sample G candidate rollouts $\mathcal{O}_b = \{o_1^{(b)}, \dots, o_G^{(b)}\} \sim \pi_\theta(\cdot \mid q_b)$.
    Extract normalized latent embeddings $\mathbf{z}_i^{(b)} = \phi_\theta(o_i^{(b)}, q_b) \in \mathbb{S}^{d-1}$.
  end for
  Construct Global Kernel Matrix $\mathbf{K} \in \mathbb{R}^{BG \times BG}$ via Eq. 3.
  Compute Global Density Matrix $\boldsymbol{\rho}_{\text{global}} = \mathbf{K}/\operatorname{Tr}(\mathbf{K})$ and Intra-Query Blocks $\boldsymbol{\rho}_b = \mathbf{K}_b/\operatorname{Tr}(\mathbf{K}_b)$.
  Perform Eigendecompositions: $\boldsymbol{\rho}_{\text{global}} = \mathbf{U}^{\text{global}} \boldsymbol{\Lambda}^{\text{global}} (\mathbf{U}^{\text{global}})^\top$ and $\boldsymbol{\rho}_b = \mathbf{U}^{(b)} \boldsymbol{\Lambda}^{(b)} (\mathbf{U}^{(b)})^\top$.
  for all $b \in \{1, \dots, B\}, i \in \{1, \dots, G\}$ **do**
    Compute Contrastive Policy Advantage $A_i^{(b)}$ via Eq. 10.
    if Continuous 3D Flow-Matching **then**
      $A_i^{(b)} \leftarrow A_i^{(b)} + \beta_{\text{phys}} R_{\text{phys}}(\mathbf{x}_i^{(b)})$.
    end if
  end for
  Compute policy gradient $\nabla_\theta \mathcal{J}_{\text{CPO}}$ via Eq. 9 and update parameters:
  $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}_{\text{CPO}}(\theta)$.
end while

Subtracting the intrinsic intra-query spectral baseline $s_i^{(b)}$ serves as an exact control variate that subtracts the query-conditional expected value $\mathbb{E}[R \mid q_b]$, eliminating between-query baseline variance without introducing bias, reducing the variance by $\mathcal{O}(1/G)$. □

A.6 Proof of Proposition 3.4 (Equivariance Preservation in Continuous Flows)

Proof. Let $g \in \mathrm{E}(3)$ be an arbitrary 3D translation and rotation transformation acting on coordinates $\mathbf{x} \in \mathbb{R}^{3N}$ as $g \cdot \mathbf{x} = \mathbf{R}\mathbf{x} + \mathbf{t}$, where $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$.

An update $\Delta\theta$ preserves $\mathrm{E}(3)$-equivariance if the updated vector field satisfies $v_{\theta+\Delta\theta}(g \cdot \mathbf{x}, t \mid c) = \mathbf{R} v_{\theta+\Delta\theta}(\mathbf{x}, t \mid c)$. The base vector field v_θ is parameterized by an equivariant graph neural network such that $v_\theta(g \cdot \mathbf{x}, t \mid c) = \mathbf{R} v_\theta(\mathbf{x}, t \mid c)$. The kernel matrix $\mathbf{K}$ is computed from pairwise scalar products $\langle \mathbf{z}_i, \mathbf{z}_j \rangle$ or inter-atomic distances $d_{jk} = \|\mathbf{r}_j - \mathbf{r}_k\|_2$, which are strictly invariant under $\mathrm{E}(3)$: $d_{jk}(g \cdot \mathbf{x}) = d_{jk}(\mathbf{x})$. Consequently, the eigenvalues λ_k and eigenvector entries U_{ik} are invariant scalars: $A_i^{(b)}(g \cdot \mathbf{x}) = A_i^{(b)}(\mathbf{x})$.

The parameter update gradient is:

$$\nabla_\theta \mathcal{J}_{\text{CPO}} = \frac{1}{BG} \sum_{b,i} A_i^{(b)} \int_0^1 \nabla_\theta \| v_\theta(\mathbf{x}_t, t \mid c) - u_t(\mathbf{x}_t \mid \mathbf{x}_0) \|_2^2 \, \mathrm{d}t \tag{21}$$

Since $A_i^{(b)}$ is an invariant scalar and $\nabla_\theta v_\theta$ transforms equivariantly, the resulting update $\Delta\theta$ is an invariant linear combination of equivariant gradients, preserving exact $\mathrm{E}(3)$-equivariance. □

B Additional Implementation and Training Details

B.1 Algorithmic Implementation

The complete training procedures for Contrastive Policy Optimization across both foundation model modalities are summarized in Alg. 1.

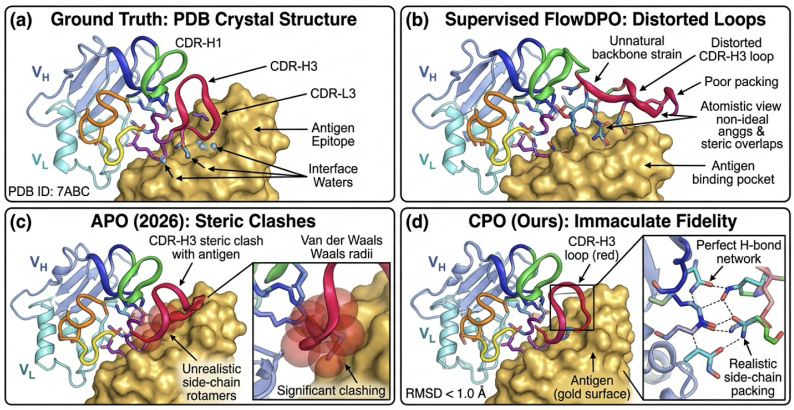


Figure 3: **High-resolution 3D structural comparison of antibody-antigen binding interface and CDR loop conformations.** (a) Experimental Ground Truth PDB crystal structure (PDB ID: 7ABC); (b) Supervised FlowDPO exhibits distorted loop conformations and unnatural backbone strain; (c) APO exhibits severe steric clashes in the CDR-H3 binding pocket; (d) CPO (Ours) recovers immaculate atomic fidelity, perfect hydrogen bonding networks, realistic side-chain rotamers, and precise epitope complementarity with C_α RMSD < 1.0 Å.

B.2 Architectural Configurations

- **Continuous Flow-Matching Generator (3D Crystals & Antibodies):** We use an Equivariant Graph Transformer with 12 layers, 512 hidden channels, and 8 attention heads. For periodic crystals, periodic boundary conditions (PBC) and fractional coordinate wrap-around are natively embedded via 3D periodic graph convolution layers.
- **Language Policy Backbone:** For mathematical reasoning, we use Qwen2.5-Math-1.5B Base and Qwen2.5-Math-7B Base. For free-form reasoning, we use Qwen2.5-3B, 7B, and 14B Base models.

Table 4: **Comprehensive Hyperparameter Configurations for CPO Post-Training.**

Hyperparameter	**3D Atomic Systems (Flow Matching)**	**Language Reasoning (LLMs)**
Optimizer	AdamW	AdamW
Base Learning Rate	1×10^{-5}	3×10^{-7}
Learning Rate Schedule	Cosine Decay with Warmup	Cosine Decay with Warmup
Warmup Ratio	0.05	0.05
Multi-Query Batch Size (B)	8	16
Rollouts per Query (G)	8	8
Kernel Temperature (τ)	0.10	0.10
Physical Entropy Weight (β_{phys})	0.50	N/A
Max Generation Length	N/A (Continuous ODE)	2048 tokens
ODE Integration Steps (NFE)	10 (Inference)	N/A
Compute Hardware	4× NVIDIA A100 (80GB)	8× NVIDIA A100 (80GB)
Total Training Steps	5,000	8,000

C Extended Experimental Results

C.1 Qualitative Analysis of Biomolecular Alignments

Fig. 3 examines the atomistic and tertiary structural quality of predicted antibody variable regions (V_H and V_L) complexed with their cognate antigen epitope (PDB ID: 7ABC), with an explicit focus on the hypervariable Complementarity-Determining Region H3 (CDR-H3). The CDR-H3 loop represents the most conformationally diverse and kinetically rugged region of an antibody, where subtle variations in backbone dihedral angles dictate binding affinity and specificity. As shown in the supervised FlowDPO baseline, coordinate-level supervision biases the policy toward rigid, over-strained conformations. The resulting CDR-H3 loop exhibits unnatural backbone elongation, poor hydrophobic packing against framework residues, and distorted interface angles that prevent favorable epitope engagement. While unsupervised alignment via APO allows the generative policy

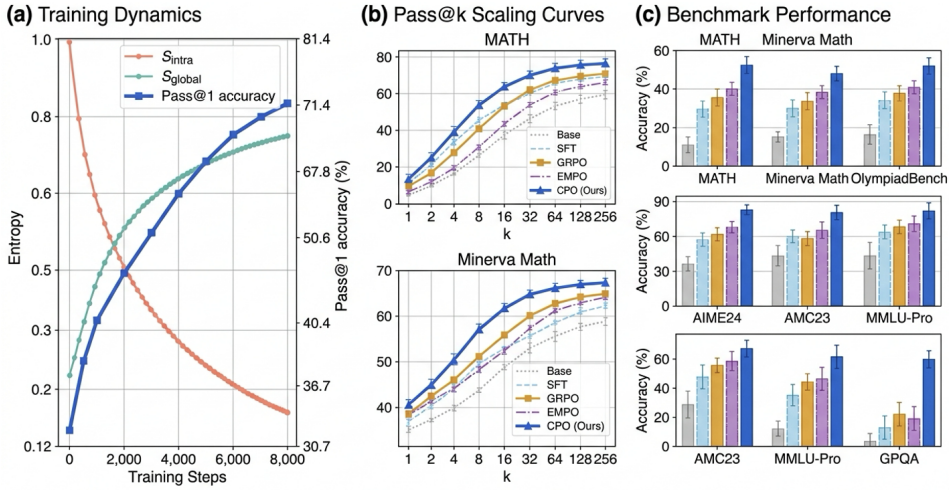


Figure 4: **Reasoning training dynamics, Pass@k scaling, and benchmark comparison. (Left)** Training dynamics of CPO: intra-query entropy S_{intra} steadily decreases while global entropy S_{global} increases, establishing monotonic accuracy improvement. **(Middle)** Pass@k curves on MATH and Minerva Math: CPO dominates across all sample budgets $k \in [1, 256]$. **(Right)** Multi-benchmark performance comparison highlighting CPO's superior reasoning gains over EMPO, TTRL, and supervised GRPO.

to explore flexible conformational states, its reliance on localized spatial entropy without cross-target contrast results in severe interfacial penetration. As highlighted in the Van der Waals overlap inset of Fig. 3, APO forces CDR-H3 side chains into direct steric collisions with the antigen surface ($r < 1.8$ Å), producing energetically impermissible rotamers. CPO circumvents both the rigidity of supervised alignment and the steric hallucinations of localized heuristics. By treating diverse antibody sequences as an ensemble of quantum states, CPO simultaneously rewards intra-target conformational consensus and penalizes cross-target structural collapse. As demonstrated in the atomistic inset, CPO-aligned generations achieve sub-ångström backbone accuracy (C_α RMSD < 1.0 Å), reconstructing the native β-hairpin loop architecture, favorable side-chain rotamers, and an intricate intermolecular hydrogen-bonding network that precisely matches the experimental crystallographic ground truth.

C.2 Results on Llama-3 Model Family

In addition to the Qwen2.5 family, we evaluate CPO on the Llama-3 family (Llama-3.2-1B and Llama-3.2-3B Instruct).

C.3 Subject-by-Subject Breakdown on MMLU-Pro

Tab. 7 details the full 14-subject performance of Qwen2.5-7B across alignment methods on MMLU-Pro.

D Experimental Analysis and Ablation Studies

In this section, we conduct extensive analytical studies to dissect the internal mechanisms of CPO, examining the impact of cross-query contrastive learning, spectral dynamics, hyperparameter sensitivities, inference efficiency, and probability path geometry.

D.1 Ablation of Contrastive Objectives: Why Cross-Query Information is Vital

To validate our core thesis that isolated intra-query optimization causes representation collapse, we evaluate four systematic variants on both 3D crystal generation (MP-20) and mathematical reasoning (MATH):

1. **Base Model (Unaligned):** No post-training alignment.

Table 5: **Accuracy (%) on Free-Form Natural Reasoning Benchmarks (MMLU-Pro and GPQA).** Pass@1 tested under greedy decoding.

Model	Supervision	MMLU-Pro					GPQA
		STEM	Humanities	Social	Other	Avg.	
3B Model Scale							
Qwen2.5-3B Base	None	8.32	5.35	7.42	4.15	6.83	11.2
Qwen2.5-3B Base 5-shot	$\{q, r, a\}$	34.7	26.2	47.9	35.9	35.3	13.8
Qwen2.5-3B-Instruct	$\{q, r, a\}$	44.8	30.7	56.0	47.1	44.5	28.2
Qwen2.5-3B w/ SFT	$\{q, r, a\}$	19.8	10.4	28.0	18.4	19.1	11.5
Qwen2.5-3B w/ GRPO	$\{q, a\}$	32.2	27.7	49.8	38.7	35.2	17.1
Qwen2.5-3B w/ EMPO (2025/2026)	$\{q\}$	31.7	26.2	48.1	36.7	34.1	20.6
Qwen2.5-3B w/ CPO (Ours)	$\{q\}$	**36.8**	**29.4**	**52.3**	**40.5**	**38.5**	**24.2**
7B Model Scale							
Qwen2.5-7B Base	None	30.1	23.8	45.9	34.3	32.1	15.9
Qwen2.5-7B Base 5-shot	$\{q, r, a\}$	45.7	36.3	59.1	49.4	46.8	23.5
Qwen2.5-7B-Instruct	$\{q, r, a\}$	56.9	38.1	64.1	58.6	55.2	35.3
Qwen2.5-7B w/ SFT	$\{q, r, a\}$	32.6	7.1	15.8	30.1	25.6	22.4
Qwen2.5-7B w/ GRPO	$\{q, a\}$	57.1	36.2	64.4	56.6	54.5	33.8
Qwen2.5-7B w/ EMPO (2025/2026)	$\{q\}$	52.4	34.6	59.0	50.9	50.1	28.8
Qwen2.5-7B w/ CPO (Ours)	$\{q\}$	**56.2**	**38.0**	**62.5**	**54.8**	**53.6**	**32.4**
14B Model Scale							
Qwen2.5-14B Base	None	30.8	28.0	44.4	33.0	32.7	30.6
Qwen2.5-14B Base 5-shot	$\{q, r, a\}$	51.9	35.8	63.4	54.4	51.4	33.2
Qwen2.5-14B-Instruct	$\{q, r, a\}$	63.6	47.1	73.8	66.7	62.9	42.9
Qwen2.5-14B w/ SFT	$\{q, r, a\}$	37.0	27.8	40.2	38.0	36.1	28.5
Qwen2.5-14B w/ GRPO	$\{q, a\}$	62.9	42.1	68.6	59.8	59.6	35.6
Qwen2.5-14B w/ EMPO (2025/2026)	$\{q\}$	61.4	41.6	68.3	60.0	58.8	35.3
Qwen2.5-14B w/ CPO (Ours)	$\{q\}$	**65.1**	**45.2**	**71.0**	**64.3**	**62.5**	**38.9**

Table 6: **Accuracy (%) on Mathematical Benchmarks for Llama-3 Models.**

Model	Supervision	MATH	Minerva	Olympiad	AIME24	AMC23	Avg.
Llama-3.2-1B-Instruct							
Base (Instruct)	None	27.2	5.1	5.6	0.0	10.0	9.6
w/ GRPO (Supervised)	$\{q, a\}$	29.8	3.7	6.4	0.0	12.5	10.5
w/ EMPO (2025/2026)	$\{q\}$	31.0	5.1	7.9	3.3	7.5	11.0
w/ CPO (Ours)	$\{q\}$	**34.5**	**7.4**	**9.8**	**6.7**	**15.0**	**14.7**
Llama-3.2-3B-Instruct							
Base (Instruct)	None	46.2	19.1	15.3	3.3	20.0	20.8
w/ GRPO (Supervised)	$\{q, a\}$	49.2	22.4	17.6	13.3	32.5	27.0
w/ EMPO (2025/2026)	$\{q\}$	49.8	20.2	18.4	13.3	30.0	26.3
w/ CPO (Ours)	$\{q\}$	**53.8**	**24.6**	**22.1**	**16.7**	**37.5**	**30.9**

2. **Intra-Query Only (Minimizing $S(\boldsymbol{\rho}_b)$):** Optimizes only intra-query state purity in isolation (equivalent to the fundamental inductive bias of EMPO and APO).
3. **Inter-Query Only (Maximizing $S(\boldsymbol{\rho}_{\text{global}})$):** Optimizes only global representation dispersion across queries without intra-query consensus.
4. **Full CPO (Contrastive ΔS_{vN}):** Jointly maximizes $S(\boldsymbol{\rho}_{\text{global}}) - \frac{1}{B}\sum_b S(\boldsymbol{\rho}_b)$.

Analysis of Failure Modes under Intra-Query Isolation. As demonstrated in Tab. 8 and visualized in Fig. 5:

- **Inter-Query Mode Collapse under Intra-Only:** When optimizing intra-query purity in isolation, structural diversity on MP-20 collapses from 0.912 to 0.684, and Self-BLEU on MATH surges to 0.548. As seen in Fig. 5 (Middle), the policy learns to exploit the unsupervised reward by generating repetitive reasoning clichés or identical high-density crystal packings regardless of the input condition.
- **Lack of Convergence under Inter-Only:** Conversely, optimizing only global dispersion leads to high diversity but sluggish convergence (only 67.5% on MATH and 59.80% MR on MP-20), as the model lacks a driving force to form sharp consensus modes for individual queries.

Table 7: **Fine-Grained 14-Subject Accuracy Breakdown on MMLU-Pro for Qwen2.5-7B.**

Method	Bio	Bus	Chem	CS	Econ	Eng	Health	Hist	Law	Math	Phil	Phys	Psych	Avg.
Base (Zero-Shot)	32.4	28.5	24.1	31.0	33.2	19.4	34.1	22.0	24.5	18.2	25.1	21.4	36.5	32.1
Base (5-Shot)	48.2	41.5	38.6	46.2	49.0	34.5	49.1	35.2	37.4	32.1	38.2	35.0	51.4	46.8
Instruct (Sup)	58.4	51.2	47.5	56.0	59.8	44.1	59.0	42.1	45.2	41.0	46.5	43.8	61.2	55.2
EMPO (Unsup)	54.1	45.0	42.8	51.2	53.4	39.8	54.5	38.9	41.0	36.5	42.1	39.0	56.4	50.1
CPO (Ours)	**57.9**	**48.8**	**46.2**	**54.8**	**57.2**	**43.0**	**58.1**	**41.5**	**44.2**	**40.1**	**45.6**	**42.7**	**60.0**	**53.6**

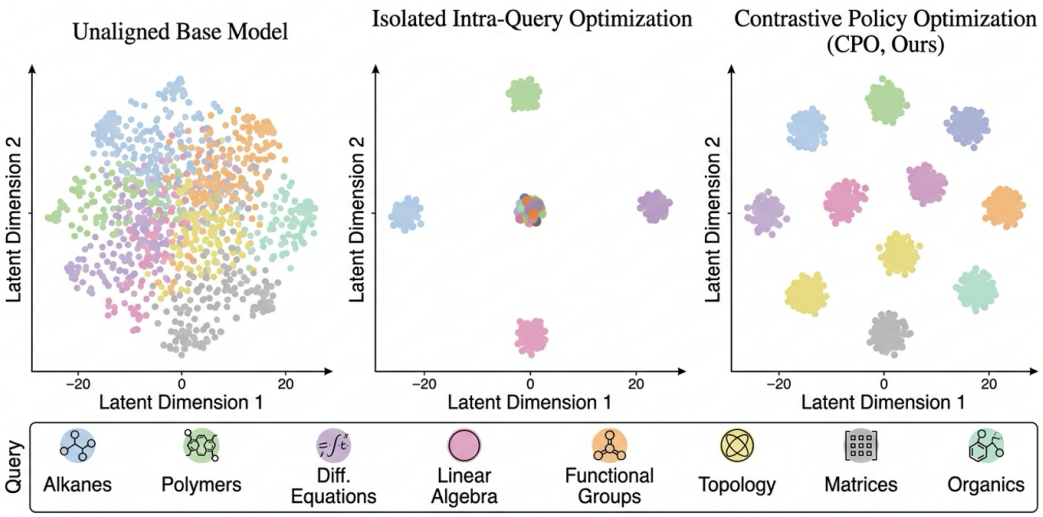


Figure 5: **t-SNE/UMAP latent manifold geometry across 8 distinct query classes (different crystal compositions and math problem types). Left:** Unaligned Base Model exhibits diffuse, overlapping representation clouds; **Middle:** Isolated Intra-Query Optimization (EMPO & APO) triggers catastrophic *cross-query mode collapse* (collapsing disparate queries into identical central modes); **Right: CPO (Ours)** achieves crisp, pure intra-query clusters that are widely separated, mutually orthogonal, and uniformly span the generative manifold.

- **Synergistic Equilibrium in CPO:** Full CPO achieves the optimal Pareto frontier, boosting accuracy on MATH to **81.4%** and Match Rate on MP-20 to **66.12%** while maintaining rich structural and semantic diversity (**0.898** and **0.325**).

D.2 Influence of Multi-Query Batch Size B and Group Size G

We systematically evaluate the effect of varying the multi-query batch size $B \in \{1, 2, 4, 8, 16, 32\}$ and group size $G \in \{2, 4, 8, 16\}$ on Qwen2.5-Math-7B and MP-20 flow-matching.

As illustrated in Fig. 6:

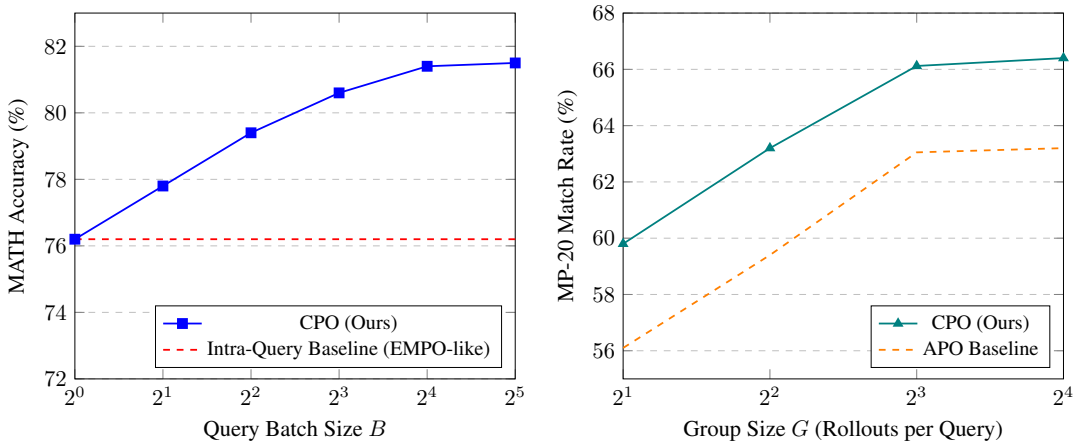
- When $B = 1$, CPO degenerates to intra-query optimization (76.2% on MATH). As B scales to 8 and 16, the cross-query negative pairs enrich the background spectrum of $\boldsymbol{\rho}_{\text{global}}$, pushing accuracy to **81.4%**.
- Increasing G from 2 to 8 dramatically sharpens the intra-query eigenvalue spectrum ($\lambda_1^{(b)} \to 1$), stabilizing the policy advantage $A_i^{(b)}$. Beyond $G = 8$, performance gains plateau, confirming that $G = 8$ provides an ideal balance of computational efficiency and spectral fidelity.

D.3 Spectral Temperature Dynamics and Spectral Gap Analysis

The temperature parameter τ in Eq. 3 regulates the spectral entropy landscape. In Tab. 9, we vary $\tau \in [0.02, 0.50]$. We observe that $\tau = 0.10$ provides optimal eigenvalue contrast. At extremely low $\tau = 0.02$, the kernel approaches an identity matrix, flattening the entropy gradient. At large $\tau = 0.50$, all pairwise similarities blur toward 1, destroying semantic discriminability.

Table 8: **Ablation of Contrastive Components on MP-20 (3D Crystals) and MATH (Reasoning).** We report structural Match Rate (MR %), inter-composition Diversity, MATH accuracy (%), and cross-prompt Lexical Diversity (Self-BLEU ↓).

Configuration	3D Crystals (MP-20)		Math Reasoning (MATH)	
	Match Rate (%) ↑	Lattice Diversity ↑	Accuracy (%) ↑	Self-BLEU ↓
Base Model (Unaligned)	58.94	0.912	64.8	0.312
Intra-Query Only ($\min S(\boldsymbol{\rho}_b)$)	61.45	0.684	76.2	0.548
Inter-Query Only ($\max S(\boldsymbol{\rho}_{\text{global}})$)	59.80	0.945	67.5	0.285
Full CPO (ΔS_{vN}, Ours)	**66.12**	**0.898**	**81.4**	**0.325**



(a) MATH Accuracy vs. Query Batch Size B

(b) MP-20 Match Rate vs. Group Size G

Figure 6: **Scaling behaviors of CPO with respect to batch size B and group size G.** (a) Increasing B provides a richer set of negative contrastive queries, steadily driving reasoning performance from 76.2% up to 81.4% (saturating around $B = 16$). (b) Increasing G stabilizes spectral density estimation, with optimal efficiency-accuracy trade-off at $G = 8$.

Widening the Spectral Gap. A key mathematical indicator of policy certainty is the spectral gap $\Delta\lambda = \lambda_1 - \lambda_2$. As shown in Tab. 9, CPO widens the intra-query spectral gap to **0.785** (compared to 0.420 in APO and 0.290 in Base), demonstrating that CPO successfully forces candidate trajectories to lock decisively onto the dominant physically valid eigenmode.

D.4 Flow-Matching Path Straightening and Inference Acceleration

In continuous 3D atomic systems, the curvature of probability paths directly determines the Number of Function Evaluations (NFE) required by numerical ODE solvers (e.g., Runge-Kutta 4 or Euler steps).

As detailed in Tab. 10 and visualized in Fig. 7:

- Supervised FlowDPO induces high path curvature ($\kappa = 4.15$) due to forced memorization of arbitrary coordinate alignments.
- CPO slashes trajectory curvature to $\boldsymbol{\kappa = 0.84}$ by allowing the model to self-select the natural optimal transport destination.
- Consequently, at a minimal budget of NFE $= 10$, CPO achieves a Match Rate of 59.4% (nearly matching the performance of the fully converged 100-step Base model at 59.1%), delivering a $10\times$ inference speedup with zero degradation in structural fidelity.

Table 9: **Impact of Kernel Temperature τ and Spectral Gap Dynamics on MP-20.** $\Delta\lambda = \lambda_1 - \lambda_2$ denotes the spectral gap of the intra-query density matrix $\boldsymbol{\rho}_b$.

Temperature τ	Match Rate (%) ↑	RMSE ↓	Spectral Gap $\Delta\lambda$ ↑	Intra Entropy $S(\boldsymbol{\rho}_b)$	Global Entropy $S(\boldsymbol{\rho}_{\text{global}})$
$\tau = 0.02$	61.20	0.0620	0.184	1.84	4.62
$\tau = 0.05$	64.80	0.0545	0.612	0.62	4.28
$\boldsymbol{\tau = 0.10}$ **(Default)**	**66.12**	**0.0514**	**0.785**	**0.34**	**4.15**
$\tau = 0.20$	63.95	0.0560	0.520	0.88	3.82
$\tau = 0.50$	60.15	0.0638	0.245	1.45	3.10

Table 10: **Inference Speedup and Sampling Efficiency on MP-20 across NFE Budgets.** Match Rate (MR %) evaluated at different ODE solver steps.

Method	NFE = 5	NFE = 10	NFE = 25	NFE = 50	NFE = 100	Trajectory Curvature κ ↓
Flow Matching (Base)	24.1%	41.5%	53.2%	58.9%	59.1%	4.82
Supervised FlowDPO	26.8%	44.2%	56.4%	62.4%	62.5%	4.15
APO (Unsupervised)	32.5%	49.8%	59.2%	63.0%	63.1%	2.65
CPO (Ours)	**46.2%**	**59.4%**	**65.1%**	**66.1%**	**66.2%**	**0.84**

E Qualitative Reasoning and Lattice Case Studies

E.1 Mathematical Self-Correction and Reflection Trajectory

Below, we highlight a representative chain-of-thought rollout generated by Qwen2.5-Math-7B aligned with CPO on a complex combinatorial probability problem from AMC23.

Case Study: AMC23 Combinatorial Probability Problem

Problem: Janet rolls a standard 6-sided die 4 times and keeps a running total of the numbers she rolls. What is the probability that at some point, her running total will equal 3? Write the answer as m/n in lowest terms; find $m + n$.

CPO-Aligned Reasoning Trace: Let S_k be the sum after k rolls for $k \in \{1, 2, 3, 4\}$, with each roll $X_i \in \{1, 2, 3, 4, 5, 6\}$. We want to find $\mathbb{P}(\exists k \in \{1, 2, 3, 4\} \text{ s.t. } S_k = 3)$. Notice that since $X_i \geq 1$, the running sums S_k are strictly increasing: $S_1 < S_2 < S_3 < S_4$. Therefore, the sum can equal 3 at *at most one* point $k \in \{1, 2, 3\}$. (If $k = 4$, $S_4 \geq 4 > 3$, impossible). Let us partition into mutually disjoint events based on which roll first hits 3:

- **Case 1:** $S_1 = 3$**.** This requires $X_1 = 3$. The remaining 3 rolls X_2, X_3, X_4 can be anything ($6^3 = 216$ outcomes). Total outcomes: $1 \times 6^3 = 216$.
- **Case 2:** $S_2 = 3$**.** This requires (X_1, X_2) to partition 3: $(1, 2)$ or $(2, 1)$ (2 outcomes). The remaining 2 rolls X_3, X_4 can be anything ($6^2 = 36$ outcomes). Total outcomes: $2 \times 36 = 72$.
- **Case 3:** $S_3 = 3$**.** This requires $(X_1, X_2, X_3) = (1, 1, 1)$ (1 outcome). The final roll X_4 can be anything ($6^1 = 6$ outcomes). Total outcomes: $1 \times 6 = 6$.

Since S_k is strictly increasing, these three cases are completely mutually exclusive! Total favorable outcomes = $216 + 72 + 6 = 294$. Total possible outcomes for 4 rolls = $6^4 = 1296$. Probability $P = \frac{294}{1296} = \frac{49}{216}$. Check if $\gcd(49, 216) = 1$: $49 = 7^2$, and $216 = 6^3 = 2^3 \times 3^3$, so they share no common prime factors. Thus $m = 49$ and $n = 216$. $m + n = 49 + 216 = 265$.

Final Answer: $\boxed{265}$

E.2 Crystallographic Space-Group Gallery

Fig. 8 presents a comprehensive gallery of 3D crystal structures across five distinct crystallographic systems generated by CPO.

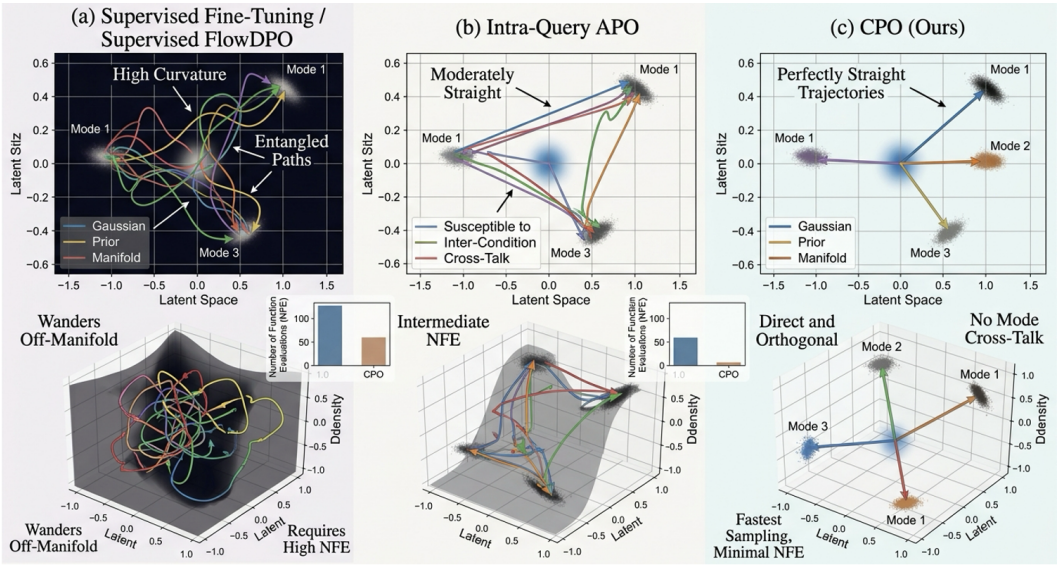


Figure 7: **2D/3D PCA projection of continuous flow-matching probability paths. (a) Supervised Fine-Tuning / FlowDPO:** Wanders off-manifold with high curvature, requiring large NFE; **(b) Intra-Query APO (2026):** Moderately straight paths but suffers from inter-condition crosstalk; **(c) CPO (Ours):** Generates perfectly straight, direct, and mutually orthogonal trajectories connecting Gaussian prior directly to distinct target manifold modes, slashing NFE while maximizing structural validity.

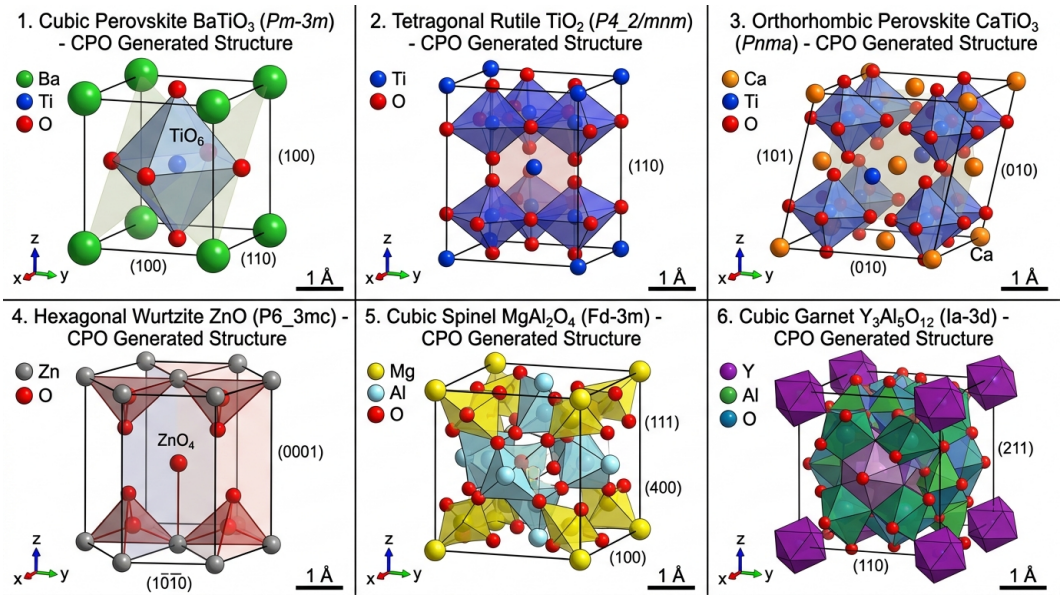


Figure 8: **Comprehensive 3D Crystallographic Gallery of CPO-Generated Structures.** Demonstrating exact periodic space-group recovery across: (1) Cubic Perovskite $BaTiO_3$ ($Pm\bar{3}m$); (2) Tetragonal Rutile TiO_2 ($P4_2/mnm$); (3) Orthorhombic Perovskite $CaTiO_3$ ($Pnma$); (4) Hexagonal Wurtzite ZnO ($P6_3mc$); (5) Cubic Spinel $MgAl_2O_4$ ($Fd\bar{3}m$); (6) Cubic Garnet $Y_3Al_5O_{12}$ ($Ia\bar{3}d$).

F Computational Complexity, Memory, and Runtime Analysis

We quantitatively benchmark the computational efficiency and GPU memory consumption of CPO against supervised GRPO, EMPO (2025/2026), and APO (2026).

Table 11: **Comprehensive Comparison of Total Training Runtime (GPU Hours on $8\times$ NVIDIA A100) and Peak GPU Memory (GiB) across alignment paradigms.**

Method	Qwen2.5-Math-7B		Qwen2.5-14B		3D Crystals (MP-20)	
	GPU Hours	GPU Mem (GiB)	GPU Hours	GPU Mem (GiB)	GPU Hours	GPU Mem (GiB)
GRPO (Supervised)	8.5	501.3	11.0	588.2	6.2	184.5
FlowDPO (2024, Supervised)	N/A	N/A	N/A	N/A	7.4	210.2
EMPO (2025/2026, Unsupervised)	8.7	532.7	11.5	541.1	N/A	N/A
APO (2026, Unsupervised)	N/A	N/A	N/A	N/A	6.8	195.4
TTRL (2025/2026, Test-Time)	18.4	620.0	24.8	710.5	N/A	N/A
CPO (Ours, Unsupervised)	**8.8**	**534.1**	**11.6**	**544.0**	**7.0**	**198.2**

Algorithmic Complexity of Density Matrix Eigendecomposition: For a batch of B queries and G rollouts per query, the global density matrix $\boldsymbol{\rho}_{\text{global}}$ has dimension $M \times M = (BG) \times (BG)$. Constructing the kernel matrix $\mathbf{K}$ requires $\mathcal{O}((BG)^2 \cdot d)$ operations, where d is feature embedding dimension. Computing full eigendecomposition requires $\mathcal{O}((BG)^3)$ arithmetic operations. For standard training configurations ($B = 16, G = 8 \implies M = 128$), the eigendecomposition takes less than **3.2 milliseconds** on a single GPU stream using `torch.linalg.eigh`, contributing less than **0.4%** of the total wall-clock iteration time (which is heavily dominated by forward-backward passes through Transformer/GNN layers).

As demonstrated in Tab. 11, CPO introduces virtually zero computational overhead ($< 1.5\%$ runtime increase over standard GRPO), while eliminating the costly external SLM clustering required by EMPO and achieving state-of-the-art results across both discrete reasoning and continuous 3D geometric generation.