Forty Shades of Blue: Quality-Diversity Alignment via Mode-Conditioned Reinforcement Learning

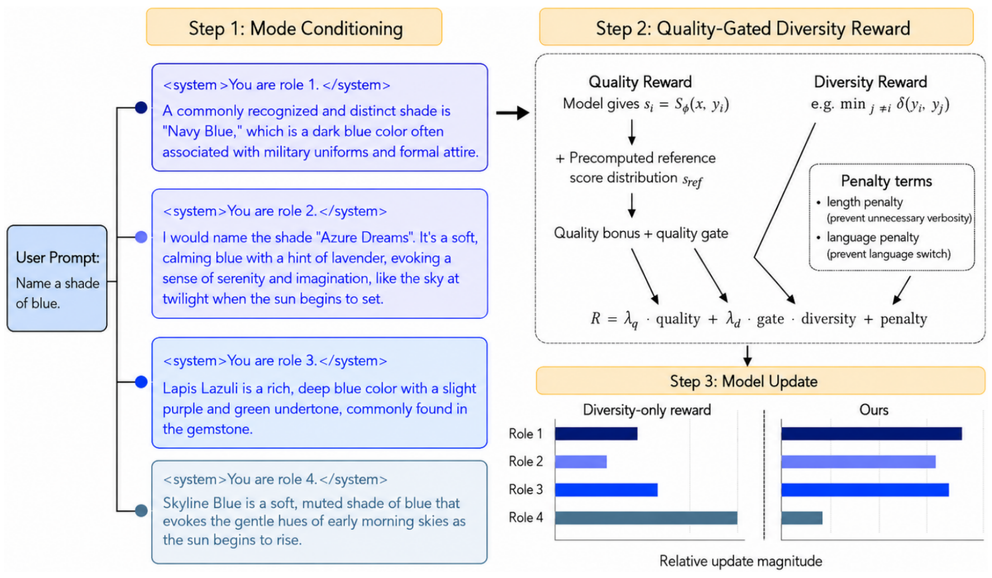
Jiayi Yuan[*1] Hangoo Kang[*2] James Jihao Liu[*2] Yejin Choi[2] Vikram Iyer[1] Liwei Jiang[1] Natasha Jaques[1]

[1]University of Washington [2]Stanford University [*]Equal contribution

Motivation

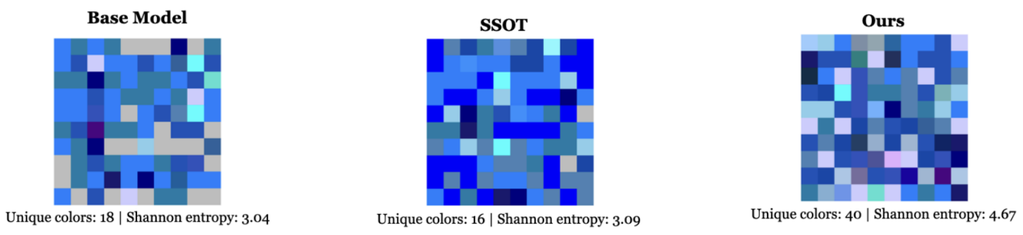
- **Problem:** Alignment causes *mode collapse* — outputs narrow to a few modes.
  - Hurts scientific ideation, creative writing, and any task needing many plausible answers
  - Prior fixes: prompting/decoding tricks (shallow) or diversity RL that sacrifices quality
- **Idea:** Treat diverse generation as multi-agent RL *inside one shared policy*: condition on abstract mode tokens ("You are role i"); each mode competes to be semantically distinct, subject to a prompt-adaptive quality gate.
- **Drop-in:** Supplying mode tokens yields a diverse response set; omitting them recovers the base model's standard behavior exactly.

Method: Mode-Conditioned Diversity Alignment



- **Step 1 — Mode conditioning:** Inject abstract numbered roles into the system prompt; all roles share one policy $\pi_\theta(y \mid x, z_i)$ and answer the same prompt, forming a response group
- **Step 2 — Quality-gated diversity reward (QGDR):**
  - Diversity = nearest-neighbor semantic distance within the group: $d(x, y_i, \mathcal{Y}) = \min_{j \neq i} \delta(y_i, y_j)$
  - Gate $g(x, y_i) = \mathbf{1}\left[s_i \geq \tau_q^{(x)}\right]$: diversity credit only above a prompt-adaptive quality threshold computed from reference-policy scores
  - $R = \lambda_q \cdot \text{quality} + \lambda_d \cdot \text{gate} \cdot \text{diversity} + \text{penalty}$; penalties suppress reward hacking (verbosity, language switching)
- **Step 3 — GRPO update:** Role-conditioned samples form one reward group; the trainable model remains a single shared policy

Forty Shades of Blue



- 100 responses to *"Name a shade of blue"*: MODA names **40 distinct shades** (+122% vs. base, +150% vs. SSOT) with **100% valid answers**

Main Results: Diversity

Model	Discrim. ↑	SBERT ↑	1−SelfBLEU ↑	E-Vendi ↑
Qwen3-8B (base)	0.466	0.299	0.415	2.026
DARLING (Mixture)	0.488	0.351	**0.582**	2.579
DARLING (WildChat)	0.473	0.306	0.476	2.095
MODA (ours)	**0.500**	**0.386**	0.566	**2.985**

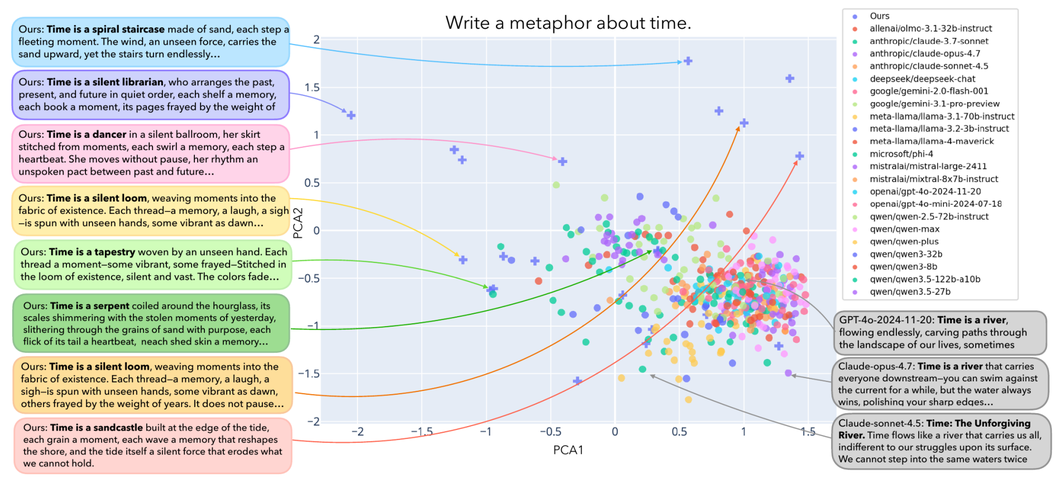
- Averaged over HypoBench, PreScience, NoveltyBench, Infinity-Chat — all *out-of-domain* for our training mixture
- **+29.1% SBERT diversity** and **+47.3% E-Vendi** over the initial policy

Main Results: General Capability Retained

Model	GSM8K	MMLU	TruthfulQA	Avg. (7 tasks)
Qwen3-8B (base)	84.8	80.0	21.7	62.6
DARLING (Mixture)	40.0	55.0	59.2	54.9
DARLING (WildChat)	83.9	**81.0**	33.0	65.4
MODA (ours)	**88.3**	**81.0**	**68.8**	**73.4**

- Pass@1 across GSM8K, MMLU, GPQA, BoolQ, HellaSwag, TruthfulQA, IFEval
- MODA *improves* average capability by **+17.3%** over the base model while diversifying; diversity baselines trade capability away
- Ablations: no role injection, quality-only reward, and additive quality+diversity aggregation all underperform — the **quality gate** is the key design choice

MODA Explores a Larger Response Space



- 20 responses to *"Write a metaphor about time"*: 23 off-the-shelf models collapse into one dense cluster; MODA (crosses) spreads across the embedding space

Takeaways

- Mode-conditioned RL turns diverse generation into competition-and-coordination among roles inside one shared policy
- The prompt-adaptive quality gate prevents diversity reward hacking and preserves — even improves — general capability
- MODA is a drop-in replacement for standard post-training: diverse mode on demand, base behavior otherwise

Code & Paper