Efficient, Property-Aligned Fan-Out Retrieval via RL-Compiled Diffusion

Pengcheng Jiang[1,2] · Judith Yue Li[1] · Moonkyung Ryu[1] · R. Lily Hu[1] · Kun Su[1] · Zhong Yi Wan[1] · Liam Hebert[1] · Hao Peng[1] · Jiawei Han[2] · Dima Kuzmin[1] · Craig Boutilier[1]

Google Research UIUC

ICML International Conference On Machine Learning

ICML 2026

Motivation

- Modern retrieval is **set-valued**: a broad intent needs a *collection* that jointly optimizes higher-order properties — diversity, coverage, complementarity, coherence — while staying **grounded** to a fixed database.
- These set-level objectives are **non-decomposable** and absent from supervised (query, content) data that reward only top-1 matches. **Fan-out** expands a query into sub-queries to build item sets — but how is it trained?
- **The dilemma: RL** optimizes set-level rewards but is **costly at inference**; **diffusion** retrievers fan out in one pass but need **objective-aligned targets** that are scarce.

Key Idea & Contributions

- Use RL ***once*** to **discover** reward-aligned fan-out, then **compile** it into supervision — rather than deploying RL at inference.

R4T = Retrieve-for-Train

① train a fan-out LLM with composite set-level rewards → ② synthesize (query → set) pairs → ③ distill a single-pass **diffusion retriever**.

⇒ **reward-optimized fan-out quality *without* the autoregressive inference cost.**

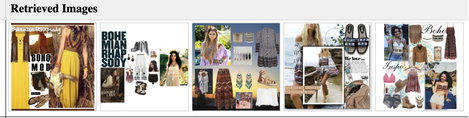
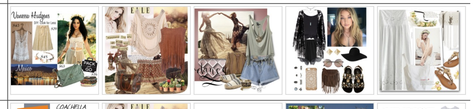
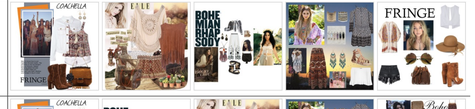
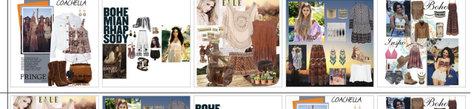
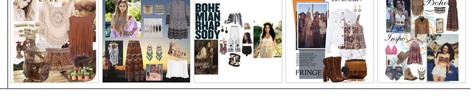
- **Contributions:** a general reward-to-data **compilation framework**; a **Soft-GRPO + diffusion** instantiation; gains in two regimes with **order-of-magnitude** faster fan-out.

Tasks, Data & Baselines

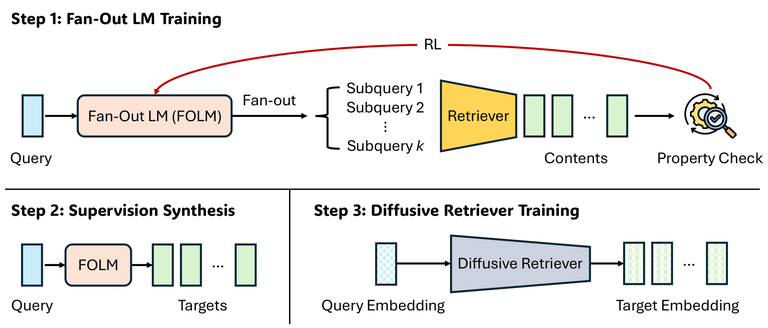
- **Two settings: OAR** (open-ended, no ground truth; reward-defined quality) and **WSCR** (many valid sets; weak reference supervision).
- **Two variants: R4T-FOLM** (RL-tuned LLM, best quality) and **R4T-Diffusion** (distilled, efficient single pass).
- **Data:** Polyvore fashion outfits (21.9K / 142K pools) + a proprietary Music playlist set (8.5K).
- **Setup:** base LLMs Gemini-2.5-Flash, Gemma3-4B, Qwen3-4B; fan-out k=10; baselines No Fan-out, Zero-shot, Best-of-N.

Qualitative Example

- **Query** *"Bohemian festival style."* R4T emits semantically distinct, on-topic sub-queries (*bohemian festival dress, straw boots festival style, lace bohemian festival*) → diverse yet coherent sets; baselines paraphrase → homogeneous results.

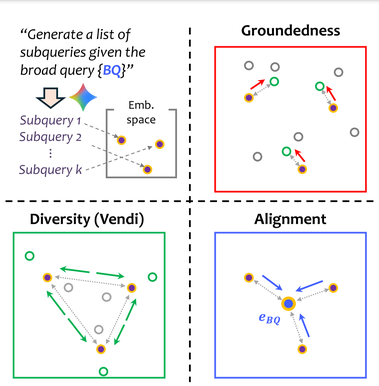
Model	Sub Query	Retrieved Images
R4T	bohemian festival dress	
R4T	straw boots festival style	
R4T	lace bohemian festival	
Qwen3_4b	bohemian festival style	
Qwen3_4b	bohemian festival fashion	
Qwen3_4b	festival bohemian clothes	

The R4T Framework



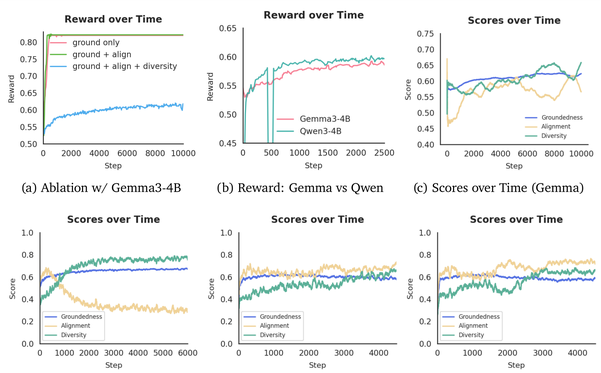
- **Step 1 · Fan-Out LM Training** — RL trains a Fan-Out LM (FOLM) to emit property-aligned sub-queries scored by a set-level *property-check* reward.
- **Step 2 · Supervision Synthesis** — the frozen FOLM synthesizes (query → target-set) pairs, with no human labels.
- **Step 3 · Diffusive Retriever** — a compact model maps a query embedding to a set of targets in **one non-autoregressive pass.**

Set-Level Objectives & Rewards



- **OAR** composite reward:
  $R = \lambda_g$ Ground + λ_d Diversity(Vendi) + λ_a Align (0.6/0.2/0.2)
- **WSCR** coverage reward:
  $R = |Y \cap R(Q)| / |Y|$
- Diversity & alignment are **mutual counter-anchors** that prevent reward hacking and mode collapse.

Reward Design & Training



(a) Ablation w/ Gemma3-4B (b) Reward: Gemma vs Qwen (c) Scores over Time (Gemma)

- Without a diversity term the policy **reward-hacks** — collapsing to degenerate strings or near-paraphrases of the query.
- Jointly optimizing all three rewards gives **stable GRPO convergence**, with groundedness, alignment and diversity rising together.

Main Results

Task 1 · OAR Polyvore, Gemma3-4B — LLM-judge ↑

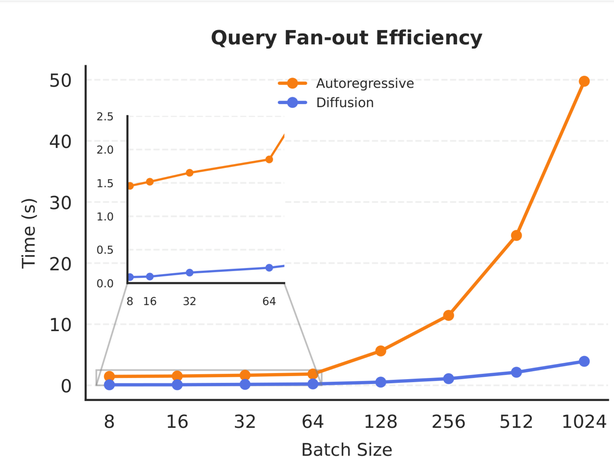
Method	Ground	Divers.	Align	Avg
No Fan-out	22.4	34.4	21.4	26.1
Zero-shot	28.4	56.0	31.2	38.5
Best-of-N	28.9	61.0	32.7	40.9
R4T-FOLM	**30.8**	**76.8**	**39.8**	**49.1**
R4T-Diffusion	**—**	**74.3**	**37.6**	**—**

+10.6 Avg over zero-shot · consistent gains on Music and Qwen3-4B.

Task 2 · WSCR Polyvore — Recall/Hit@5K, Vendi

Method	R@5K	H@5K	VS
Gemini-2.5-Flash	15.7	52.1	33.4
Qwen3-4B (zero-shot)	10.1	33.9	46.4
R4T-FOLM (Qwen)	**20.9**	**64.6**	**27.5**
R4T-Diffusion (Qwen)	**16.5**	**57.5**	**34.7**

Order-of-Magnitude Faster



12–20× faster inference

0.07 s fan-out vs 1.46 s

53.9M params · single pass

- Autoregressive fan-out scales to **~50 s** at batch 1024; the 53.9M-param diffusion retriever stays sub-second to a few seconds.

Key Takeaways

- Use RL as a **one-time objective transducer**, not an inference engine.
- Distilling reward-aligned fan-out into a **diffusion prior** keeps quality while enabling real-time single-pass set retrieval.
- **Composite reward design** (ground + diversity + align) is essential to prevent collapse.
- Moves fan-out from a heavy **"System 2"** to a lightweight **"System 1."**

Paper Link