MAIN CONFERENCE 2026

Improving LLM-Based Recommenders with Conservative Generative Flow Networks

Offline token-prefix GFlowNets that stay on support while improving accuracy-exposure trade-offs

Xuan Yu, Feng Niu, Rui Zhu, Yudong Zhang, Xu Wang*, Yang Wang*

University of Science and Technology of China; Suzhou Institute for Advanced Research, USTC

USTC CODE

CORE MESSAGE

Offline SubTB is non-identifiable when the LLM can put mass outside logged support.

CFlower adds a conservative Sub-Trajectory Balance objective and constrained on-policy sampling on the dataset-induced token-prefix DAG. The result is support-aware flow learning: less unsupported mass leakage, better distribution matching, stronger next-item accuracy, and a more reliable reference policy for downstream RL fine-tuning.

0.854 Jaccard support overlap with target distribution

1229 common target items covered by CFlower

+18.3% NDCG gain over Flower on Video Games

0.615 KL(Q||P), best among eligible baselines

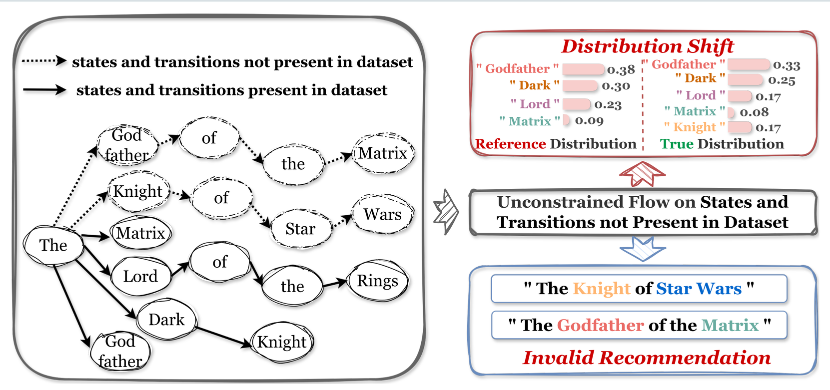
1. OFFLINE SETTING

Logged recommendations induce a partial token-prefix DAG

Each item title is generated as a token sequence. Offline logs reveal only the observed prefix expansions $A_D(s)$, while an LLM forward policy still normalizes over the full vocabulary.

- **Supported edges:** prefixes and next tokens that appear in logged items.
- **Unsupported edges:** syntactically possible token transitions with no empirical support.
- The learning problem is to match rewards without inventing flow in unseen regions.

STATE	ACTION	TERMINAL
token prefix	append one token	complete item title



LLM semantics can make unsupported continuations look plausible, so logits drift unless the objective prices this leakage.

2. WHY VANILLA SUBTB FAILS

SubTB leaves the forward policy underconstrained

Flow scaling	Mass leakage	Compensation
Disconnected regions can rescale flows.	Unseen actions absorb probability.	Backward terms can mask shifts.

Theorem. SubTB admits infinitely many (P_F, F) solutions with identical loss but different terminal distributions.

OBSERVED	FREE	RISK
SubTB constraints on logged paths	unsupported logits after softmax	terminal distribution drift

Implication: vanilla SubTB needs explicit support awareness.

3. EVALUATION SCOPE

Datasets and metrics target both accuracy and exposure

- **Domains:** Amazon CDs and Vinyl, Video Games, Movies and TV.
- **Accuracy:** NDCG@5 and HR@5 on held-out next-item prediction.
- **Exposure quality:** DGU/MGU lower is better; entropy H and TTR higher is better.
- **Distributional match:** support overlap plus JS, KL, TV, L1, and log-ratio mass on common items.

SPLIT	HISTORY	FILTER
8:1:1 chronological	max length 10	users/items ≥ 5

4. CONSERVATIVE SUBTB

Make unsupported forward flow explicitly expensive

For every state along a constrained subtrajectory, CFlower penalizes the total forward flow assigned to actions outside the logged support.

$$L_{C\text{-}SubTB}(\tau_{i:j}) = L_{SubTB}(\tau_{i:j}) + \alpha\lambda^{j-i} \Sigma_{s \text{ in } \tau_{i:j}} \Sigma_{a \notin A_D(s)} \varphi(F(s)P_F(a|s))$$

- **α** controls how strongly unsupported mass is suppressed.
- **λ** reweights short vs. long subtrajectories.
- The term shrinks SubTB's null space by charging flow that cannot be justified by logs.

EFFICIENT FORM Penalty can be written as $F(s)(1-\Sigma_{a \text{ in } A_D(s)} P_F(a|s))$, avoiding enumeration over unsupported actions.

5. CONSTRAINED ON-POLICY LEARNING

Sample dynamically, but only inside the dataset-induced DAG

Uniform replay treats all logged trajectories equally. CFlower instead samples token trajectories from a masked policy that follows the current model while zeroing unsupported actions.

Logged items build prefix DAG → **Masked policy** renormalize on support → **CSubTB update** balance plus penalty

$P_{\sim F}(a|s)$ proportional to $P_F(a|s)$ if a in $A_D(s)$, and 0 otherwise.

SAMPLING	OFFLINE SAFETY
current policy guides valid prefixes	generation never leaves logged support

6. WHAT CHANGES IN PRACTICE

Support-aware flow gives a better optimization target

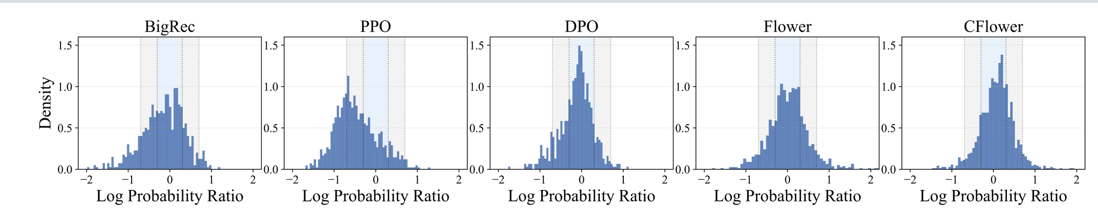
- Unsupported logits no longer get a free pass through softmax normalization.
- On-policy sampling focuses updates on high-value reachable prefixes rather than passive uniform replay.
- The learned policy is better calibrated as a recommendation model and as an RL reference policy.

Component	Role	Effect
CSubTB	support penalty	reduces leakage
Masked DAG sampling	on-policy inside logs	improves efficiency
Flow objective	reward-proportional generation	balances accuracy and exposure

REUSE The same support-aware policy later serves as a stronger KL/reference anchor for offline RL fine-tuning.

7. RQ1: DISTRIBUTIONAL MATCHING

CFlower covers more target support and stays closer on common items



Tighter mass near log-ratio zero indicates better model-target probability matching on common items.

Method	Common	Jaccard	TV	Within 5x
BIGRec	745	0.518	0.420	0.830
PPO	832	0.578	0.636	0.643
Flower	1150	0.799	0.436	0.874
CFlower	**1229**	**0.854**	**0.332**	**0.919**

8. RQ2: NEXT-ITEM RECOMMENDATION

Accuracy improves without sacrificing exposure balance

Dataset	Flower NDCG	CFlower NDCG	CFlower H	CFlower TTR
CDs and Vinyl	0.0702	**0.0737**	8.132	0.015
Video Games	0.0546	**0.0646**	7.820	0.006
Movies and TV	0.0956	**0.1036**	9.028	0.024

CDS	GAMES	MOVIES
+5.0% NDCG	+18.3% NDCG	+8.4% NDCG

3/3 datasets where CFlower beats Flower on NDCG and HR

0.063 best DGU on Movies and TV, lower means more balanced utility

D3 remains the strongest accuracy-only baseline on CDs and Vinyl, but has substantially worse group-utility balance.

9. ACCURACY-EXPOSURE FRONTIER

Higher accuracy, broader support

3/3 Amazon domains where CFlower improves Flower on NDCG and HR.

1229 common items covered in distribution matching, vs. 639 for DPO.

0.919 within-5x target mass, indicating tighter common-support calibration.

H/TTR higher entropy and tail exposure than Flower across the main SFT stage.

Flower	DPO	SFT
higher NDCG/HR	broader support	better tail exposure

Net effect: CFlower sits between accuracy-only exploitation and narrow-support distribution matching, keeping both ranking quality and exposure balance visible.

10. RQ3: STRONGER RL REFERENCE

Using CFlower as the reference policy improves downstream fine-tuning

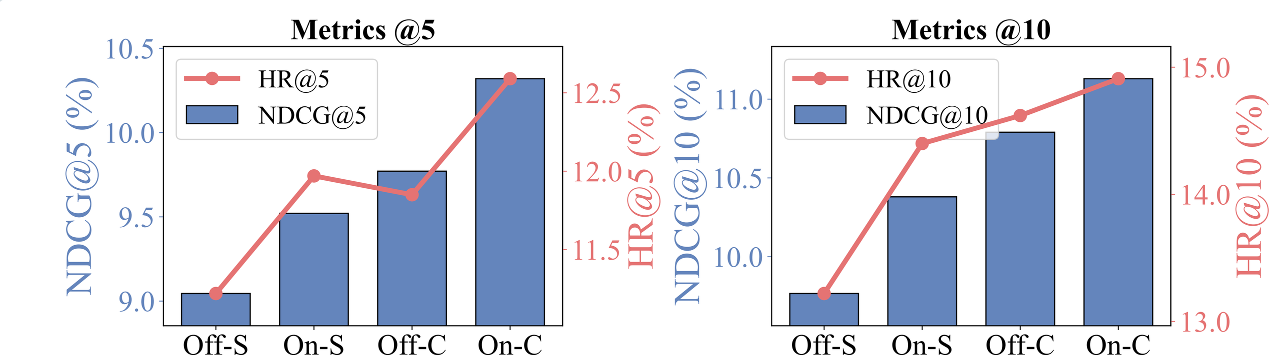
RL objective	CDs NDCG	Games NDCG	Movies NDCG
B_S-DPO	0.0716	0.0675	0.1043
F_S-DPO	0.0776	0.0640	0.1047
C_S-DPO	**0.0805**	**0.0737**	**0.1068**
C_DMPO	0.0756	0.0722	0.1002

- C_S-DPO obtains the best ranking accuracy on all three datasets.
- C-referenced RL variants keep competitive entropy and tail coverage, rather than collapsing exposure.

CDS	GAMES	MOVIES
0.0805 / 0.0965	0.0737 / 0.0914	0.1068 / 0.1265

11. RQ4: ABLATION

Offline constraints plus on-policy updates are both useful

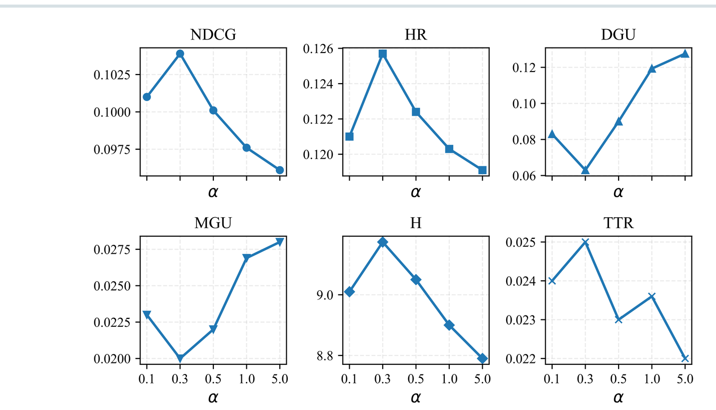


- Removing offline constraints performs worst under distribution shift.
- Conservative objectives improve stability by suppressing unsupported flow.
- On-policy learning inside the DAG accelerates convergence and improves final performance.

S VS C	OFF VS ON
CSubTB adds support pricing	On-policy prioritizes useful prefixes

12. RQ5: ROBUSTNESS

Moderate conservative strength works best



α traces the accuracy-exposure trade-off.

α Controls unsupported-flow penalty. Moderate values suppress leakage without over-pruning.

λ Reweights subtrajectory length. Sensitivity is mild near 1; default is 0.9.

Takeaway Too small behaves like SubTB; too large can reduce accuracy.

1 Objective-level issue Offline SubTB leaves unsupported token transitions underconstrained, causing non-identifiability and distribution shift.

2 Support-aware fix CFlower penalizes unsupported forward flow and samples on-policy only within the dataset-induced DAG.

3 Empirical payoff Better distribution matching, higher NDCG/HR than Flower, and stronger RL reference-policy performance.

Corresponding authors: Xu Wang (wx309@ustc.edu.cn), Yang Wang (angyan@ustc.edu.cn)

Code: github.com/yuxuan9982/CFlower

Poster source: HTML/CSS rendered to 60in x 36in PDF