CIPHER: Causal Intent Plug-in Framework for the Mitigation of Historical Exposure Bias in Recommender Systems

Sanghyeon Lee
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
shlee145@lguplus.co.kr

Yeonghwan Jeon
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
yeonghwan@lguplus.co.kr

Hyuncheol Jo
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
hcjo@lguplus.co.kr

Byoung-Ki Jeon
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
bkjeon@lguplus.co.kr

ABSTRACT

Real-world recommendation systems often prioritize content exposure based on popularity and revenue contribution, inducing structural exposure bias in user interaction logs. Consequently, observed behavior reflects not only users' intrinsic preferences but also their reactions to prominent content positioning. To address this issue, we operationalize a structural learning framework by applying causal intent disentanglement to real-world IPTV viewing logs, enabling plug-in integration into production-scale recommendation models without architectural modification. Using a DICE-based causal representation learning method, we separate interest and conformity from viewing histories and apply them as initial embeddings compatible with diverse recommendation architectures. Comparative experiments with alternative initialization methods show that our approach yields superior outcomes in both accuracy and coverage. This design enables models to learn not only from observed behaviors but also from their underlying causal drivers. Offline experiments on real-world logs demonstrate that our framework achieves up to 7% improvements in precision, recall, and MRR across both sequence-based and collaborative filtering-based models. While both model types benefit from causal intent representations, the effect is more pronounced in sequence models, where user interest aligns with temporal consumption patterns, resulting in a 55% increase in item coverage. These results highlight the practical value of causally informed user intent modeling in reducing structural exposure bias in large-scale recommendation environments. These results underscore the practical viability of intent-aware causal personalization in real-world recommendation systems subject to structural exposure bias.

CCS CONCEPTS

• Information systems → Recommendation systems • Computing methodologies → Causal reasoning and diagnostics

KEYWORDS

Causal representation learning, User intent modeling, Exposure bias, Model-agnostic framework, Recommendation systems

ACM Reference format:

Sanghyeon Lee, Yeonghwan Jeon, Hyuncheol Jo and Byoung-Ki Jeon. 2025. CIPHER: Causal Intent Plug-in Framework for the Mitigation of Historical Exposure Bias in Recommender Systems. In *Proceedings of 3rd Workshop on Causal Inference and Machine Learning in Practice (KDD'25). ACM, Toronto, Ontario, Canada, 6 pages.* https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Modern media recommendation systems are designed to offer personalized content experiences. However, in actual deployments—particularly in IPTV and streaming services—personalization is bounded by business-driven exposure priorities. Top-ranked slots are often reserved for newly released, high-revenue, or sponsored content, regardless of individual user preference. This results in structural exposure bias, where user behavior reflects not just intrinsic interest but also conformity to prominently placed content. Consequently, observed logs distort user intent and reinforce popularity-driven loops.

Prior work in causal representation learning, such as DICE (Disentangling Interest and Conformity for Recommendation with Causal Embedding) [1], has attempted to address this issue by disentangling user interest from conformity. However, these methods often remain theoretical, require substantial model redesign, or fall short in real-world scalability and operational feasibility.

To address these limitations, we propose CIPHER, a model-agnostic framework that learns interest and conformity embeddings from large-scale user logs and integrates them as initial representations into recommendation models—without

altering their core architecture. CIPHER is designed for practical deployment: it operates within existing nightly batch pipelines, stores embeddings as compact files, and adds no latency to online inference. The key contributions of this work are:

1. We operationalize a causal representation learning approach on real-world IPTV viewing logs by disentangling user behavior into structurally independent drivers of interest and conformity. This enables counterfactual interpretation of user intent under platform-induced exposure bias.
2. We propose a model-agnostic plug-in architecture that injects causally disentangled embeddings into existing recommendation models via batch-level structural adaptation, without modifying their internal mechanisms. This supports scalable deployment of causal personalization in industrial environments.
3. We empirically demonstrate that initializing models with conformity-aware intent embeddings leads to measurable exposure debiasing effects, aligning recommendation popularity with inferred user conformity levels while preserving accuracy.

These findings show that causal intent modeling via CIPHER can be operationalized at scale, enhancing both personalization quality and exposure fairness in modern recommendation systems.

2 RELATED WORK

2.1 Exposure Bias and Structural Distortion in User Logs

In real-world recommender systems, content is often prioritized based on revenue, popularity, or recency. As a result, user interaction logs reflect not only intrinsic user interest but also the influence of platform exposure policies, leading to structurally biased learning. [2] identified position bias in click logs, while [3] proposed bias-aware learning and evaluation methods based on exposure propensity. However, these studies mainly rely on post-hoc adjustments such as inverse propensity weighting, without modeling user intent at the structural level within the recommendation model itself. Our work addresses this limitation by causally disentangling intrinsic interest and conformity from user behavior, allowing us to correct exposure-induced distortions at the representation level.

2.2 Causal Embeddings and Intent Disentanglement

Efforts to recover user intent through causal representation learning have emerged recently. [4] introduced causal embeddings to reflect the effect of interventions in recommendations. [5] applied deconfounded learning to isolate independent latent factors. Most notably, [1] proposed a user embedding structure explicitly separating interest and conformity, using counterfactual representations for improved recommendations. However, these methods typically require new model architectures or complex learning pipelines, making them difficult to deploy in practice. CIPHER builds on these ideas but offers a plug-in framework that learns interest and conformity embeddings separately and integrates them into existing recommendation models without architectural changes.

2.3 Model-Agnostic Integration and Structural Variability

[6] demonstrated the use of IPTW-based causal embeddings in MF models, comparing their performance with DICE-MF. While such matrix completion-based models can benefit from causal embeddings, they tend to reconstruct user-item interactions based on popularity signals, making them more sensitive to conformity and less effective at capturing intrinsic interest. In contrast, sequence-based models such as [7] or [8] rely on temporal context and item semantics, making them better able to leverage the disentangled intent signals. We evaluate CIPHER across both model architectures to analyze how its effects vary with model structure.

3 METHODS

We present CIPHER, a practical yet model-agnostic framework that corrects exposure bias without modifying existing backbone recommenders. As depicted in Figure 1, historical interaction logs—whose behavioral signals are confounded by the platform's exposure policy—are processed by a causal intent learner that disentangles intrinsic interest from exposure-induced conformity at both user and item levels. The resulting embeddings are provided to downstream architectures through plug-in adapters that deliver either weight initialization (for sequence models) or dense-layer initialization (for collaborative-filtering models). By supplying causally informed initial parameters while leaving model topologies intact, CIPHER enables recommendation lists to more accurately reflect genuine user preferences.

3.1 Problem Formulation via Structural Causal Modeling

In real-world recommendation systems, user interactions are shaped not only by individual preferences but also by platform exposure strategies—often driven by popularity or monetization objectives. This induces structural exposure bias, where observed clicks reflect both intrinsic interest and exposure-driven conformity.

We formally represent the click behavior as a collider structure:

$$Z_{inf} \rightarrow Y \leftarrow Z_{conf} \quad (1)$$

Here, Y denotes the observed click, Z_{inf} represents user interest, and Z_{conf} captures conformity to exposure (e.g., popularity). Due to this collider, naive modeling from logs conflates the two signals. To recover causally valid representations, we separate the dataset into disjoint subsets where the influence of either interest or conformity can be isolated and learned.

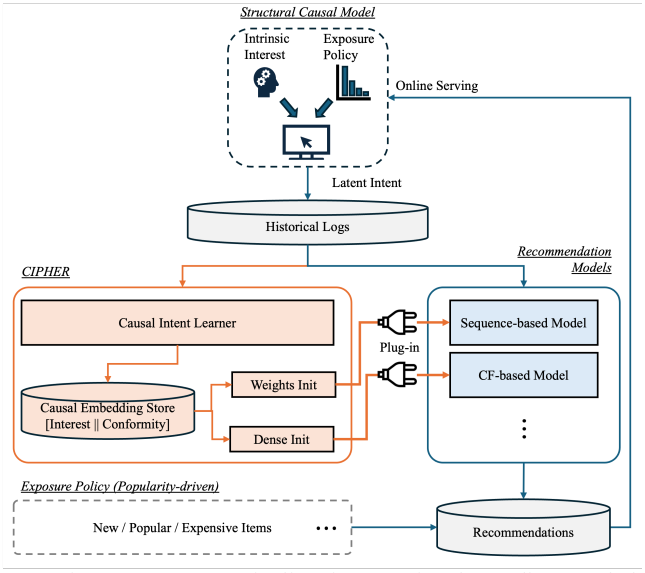

Figure 1: CIPHER pipeline: interest signals are disentangled from historical logs and injected model-agnostically into existing backbones to yield preference-aligned recommendations; the structural causal model and the platform's exposure policy are shown as auxiliary context.

3.2 Causal Intent Learner via Disentangled Representation

To recover deconfounded user intent from historically biased logs, we implement a Causal Intent Learner using the DICE framework. DICE is designed to disentangle user interest and conformity—two latent drivers of user behavior—by exploiting structural asymmetries in item popularity. To isolate each causal factor, the learner partitions user logs into two subsets based on item popularity:

- O_1: Interactions with highly popular items, likely reflecting conformity
- O_2: Interactions with less popular items, more likely driven by intrinsic interest

Using these partitions, DICE defines three parallel contrastive learning objectives:

- $\mathcal{L}_{int}$: Interest loss, computed using O_2, models true user preferences by contrasting clicks on less popular items
- $\mathcal{L}_{conf}$: Conformity loss, computed from both O_1 and O_2, models user reactions to popularity-driven exposure
- $\mathcal{L}_{click}$: Click loss, computed across O_1 and O_2, optimizes overall ranking performance

The total objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{click} + \alpha(\mathcal{L}_{int} + \mathcal{L}_{conf}) + \beta\mathcal{L}_{discrepancy} \quad (2)$$

where $\mathcal{L}_{discrepancy}$ penalizes embedding overlap to promote independence between the learned interest and conformity spaces. After training, we extract the item-level embeddings $E^{(i)}$ and $E^{(c)}$ corresponding to interest and conformity respectively and store them in the Causal Embedding Store. These representations are later used as initialization inputs to downstream recommendation models, enabling model-agnostic adaptation without modifying core architectures.

3.3 Plug-in Adaptation for Recommendation Models

To integrate the learned causal intent representations into downstream architectures, we adopt a plug-in strategy that enables model-agnostic application while preserving the original model structure.

For sequence-based models, each item token in the input sequence is initialized using its corresponding causal embedding vector from the concatenated matrix:

$$E = \left[E^{(i)} \parallel E^{(c)}\right] \in \mathbb{R}^{N \times d_e} \quad (3)$$

where d_e denotes the total dimension of interest and conformity embeddings. This embedding replaces the standard item embedding layer in the transformer encoder. These embeddings are trainable during fine-tuning, allowing the model to capture temporal patterns grounded in causally disentangled features. This corresponds to the weights init path in Figure 1.

In Collaborative-Filtering (CF)-based models, user interaction vectors $M_u \in \{0,1\}^N$ are mapped through the causal embedding matrix to produce dense user-specific input representations. Specifically, the user embedding is computed as:

$$h_u = M_u \cdot E \in \mathbb{R}^{N \times d_e} \quad (4)$$

This vector h_u is then fed into the encoder network, enabling recommendation over a semantically enriched latent space. This corresponds to the dense init path in Figure 1.

CIPHER's plug-in mechanism preserves model architecture while enriching input representations with causally meaningful structure, facilitating lightweight deployment in existing systems.

4 EVALUATION AND RESULTS

4.1 Dataset and Evaluation Protocol

This study conducts experiments on a large-scale real-world dataset from a commercial IPTV service, containing over 14 million user-item interactions from 2.1 million users and more than 20,000 unique content items. This dataset has also been employed in prior industrial studies: [9] applied it to an ensemble-based recommendation system that combines multiple models using a bandit strategy to reduce the impact of algorithmic confounding, and [10] used it to address data sparsity in subscription-based product recommendation.

To ensure fair and representative evaluation, we partition the data into training, validation, and test sets using entropy-aware sampling, which preserves item popularity distributions across splits. All models are evaluated on a top-25 next-item recommendation task, consistent with the IPTV platform's production environment.

Our evaluation focuses on three aspects: (1) recommendation accuracy, (2) personalization fairness, and (3) exposure debiasing.

Specifically, we report precision@25, recall@25, and MRR@25 to assess accuracy, and item coverage as a proxy for debiasing effectiveness, capturing how widely the model surfaces items across the catalog rather than favoring popular content. We organize our empirical analysis around the following research questions:

- RQ1: Can user intent be causally disentangled into intrinsic interest and conformity, producing interpretable representations from real-world interaction logs?
- RQ2: Does integrating CIPHER's embeddings into various backbone models improve recommendation performance without architectural modifications?
- RQ3: To what extent does CIPHER mitigate exposure bias and enhance item coverage and personalization fairness?

4.2 Interpretability of Causal Intent Representations with Real-World Data (RQ 1)

To examine whether user intent can be meaningfully disentangled into interest and conformity components, we analyze the item embeddings learned by CIPHER. Rather than relying solely on performance metrics, we investigate whether these representations reflect interpretable latent structures that separate content-driven preference from exposure-driven behavior.

4.2.1 Interest Embedding Space

Figure 2 (left) shows a 2D projection of item embeddings in the interest space, where each point represents a content album and color intensity indicates popularity. We observe that items with similar content characteristics naturally cluster together—for example, children's animations, crime thrillers, and travel variety shows form distinct regions. This indicates that the interest embeddings encode semantically coherent structures, aligning with users' intrinsic interest, regardless of content exposure policies.

4.2.2 Conformity Embedding Space

In contrast, Figure 2 (right) displays the same items in the conformity space. Items belonging to the same franchise or format are more widely dispersed, reflecting differences in exposure levels and popularity. This suggests that conformity embeddings capture behavioral conformity to platform exposure, separating content not by inherent meaning but by how it was surfaced to users.

To further validate this separation, we analyze six representative IPTV content groups spanning typical service categories:

- The Roundup: a blockbuster film franchise with recent high exposure,
- CSI: Crime Scene Investigation: a long-running episodic crime series,
- Detective Conan and Pokémon: popular animated franchises with continuous releases,
- New Journey to the West and Europe Outside the Tent: episodic travel reality shows.

In Figure 2, each content group forms a tight cluster in the interest space, supporting the idea that semantic cohesion is preserved across variations. However, in the conformity space, the same items scatter based on their observed popularity, even when belonging to the same series. For instance, certain episodes of CSI or Pokémon are positioned far apart, influenced by exposure frequency or time slot despite similar content. This divergence highlights that the conformity representation captures external platform-driven behavioral dynamics.

These findings demonstrate that CIPHER produces interpretable and causally grounded embeddings from real-world viewing logs. The interest embeddings reflect content-based intent, while the conformity embeddings capture structural bias in exposure, thereby validating the core assumption behind CIPHER's causal modeling framework.

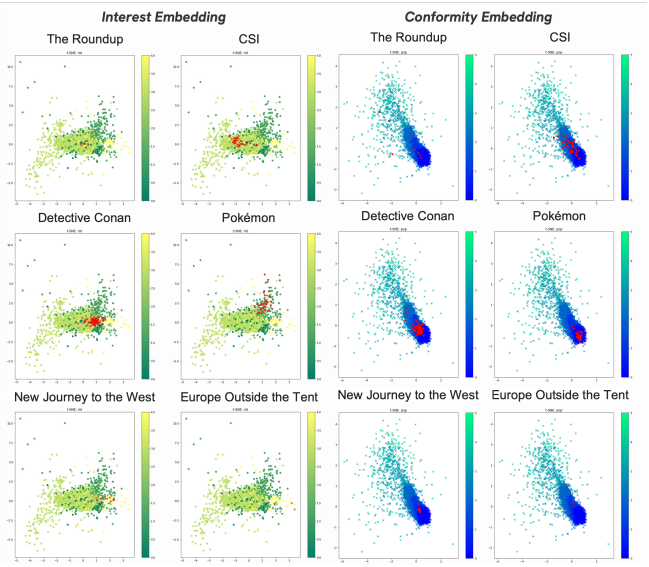

Figure 2: Item embeddings learned by CIPHER. Interest space clusters semantically coherent content, while conformity space separates items by exposure-driven popularity.

4.3 Model-Agnostic Integration of CIPHER Across Recommendation Architectures (RQ2)

To evaluate the model-agnostic effectiveness of CIPHER, we conduct experiments by integrating disentangled causal embeddings into two distinct architectures: Sequence-based model and CF-based model. These embeddings—representing user interest and conformity—are injected into the input layer, and we assess their impact under two initialization strategies: fixed (frozen during training) and trainable (fine-tuned with the downstream model).

As shown in Table 1, CIPHER improves performance across both models, with Sequence-based model benefiting most significantly. The trainable version yields the highest gains, improving precision by 5.2%, recall by 7.3%, and MRR by 4.9%, along with a 55.0% increase in item coverage. These improvements highlight the importance of allowing the model to

adapt the intent embeddings during fine-tuning, particularly in sequential models where learning user consumption order is essential. By mitigating the confounding effect of popularity-driven clicks, CIPHER enables more accurate modeling of temporal preferences.

In contrast, the fixed embedding variant provides marginal benefits, indicating that static representations alone are insufficient to fully align with the model's optimization process. For the CF-based model, we observe smaller improvements overall, and a slight drop in coverage. This suggests that integrating external embeddings into its highly regularized latent space may disrupt the reconstruction objective, especially when the embeddings are not updated during training.

These results confirm that CIPHER functions effectively across model types and highlights that trainable integration of causal intent representations yields the best performance, especially for architectures that rely heavily on temporal or behavioral patterns.

Model	Precision @25	Recall @25	MRR @25	Coverage @25
Sequence-based model (base)	-	-	-	-
Sequence-based model + CIPHER (fixed)	-8.79%	-7.60%	-9.60%	-21.02%
Sequence-based model + CIPHER (trainable)	+5.86%	+5.99%	+7.09%	+55.21%
CF-based model (base)	-	-	-	-
CF-based model + CIPHER (fixed)	-1.77%	-1.93%	-3.25%	-75.95%
CF-based model + CIPHER (trainable)	+4.42%	+4.02%	+4.67%	-45.43%

Table 1: Performance Gains with CIPHER Integration (% Improvement over Baseline)

4.4 Exposure Bias Mitigation via Conformity-Aware Personalization (RQ3)

To evaluate whether CIPHER can mitigate structural exposure bias and deliver personalized recommendations aligned with user conformity, we analyzed the average popularity of recommended content across user groups stratified by their observed conformity levels. Specifically, we computed the mean popularity of recommended albums for each user and compared it against baseline models.

Our results reveal that CIPHER reduces the average popularity of recommended items by 3.2% for users in the lowest conformity quartile (Q1). This suggests that CIPHER effectively alleviates popularity bias for users less influenced by widely exposed content, increasing exposure to less prominent but potentially more relevant items. Conversely, for users in the highest conformity quartile (Q4), CIPHER increases the average popularity of recommendations by 2.3%, indicating that it reinforces conformity-driven preferences where appropriate.

Importantly, these adjustments in exposure did not lead to any degradation in accuracy. CIPHER consistently improved precision and recall in both backbone models—Sequence-based model and CF-based model—demonstrating that reducing popularity bias can coexist with strong predictive performance. While Sequence-based model showed notable improvements in both accuracy and item coverage, CF-based model exhibited a modest accuracy gain but a drop in coverage, likely due to over-regularization when incorporating causal embeddings in sparse settings.

These findings confirm that CIPHER tailors exposure policies to user intent, offering personalized mitigation of structural bias rather than a uniform debiasing approach.

5 CONCLUSION AND FUTURE WORK

This work introduces CIPHER, a lightweight and model-agnostic framework designed to mitigate structural exposure bias in recommender systems. By disentangling user intent into interest and conformity representations, CIPHER enables interpretable modeling of user behavior even under biased exposure conditions, and integrates seamlessly into existing architectures without structural modifications.

Empirical results on a large-scale IPTV dataset demonstrate that CIPHER not only improves recommendation accuracy across both sequence-based and collaborative filtering models, but also enhances item coverage and personalization fairness (RQ2). These gains are particularly pronounced when the causal intent embeddings are fine-tuned during training, suggesting that adaptation to downstream model dynamics is crucial (RQ2). Beyond predictive performance, the learned interest and conformity embeddings reveal semantically meaningful and behaviorally distinct patterns, validating the effectiveness of intent disentanglement (RQ1). Moreover, CIPHER adjusts exposure based on users' conformity levels—reducing recommendation popularity for less conformity-driven users while preserving relevance for those who favor mainstream content—thereby achieving personalized debiasing without compromising accuracy (RQ3).

CIPHER is scheduled for deployment in a nightly batch pipeline within a commercial IPTV service. In future work, we plan to extend the framework to revenue-aware settings such as price-sensitive or purchase-constrained environments, and evaluate its applicability in other domains like books and music where exposure bias also plays a significant role. We also aim to refine causal evaluation methods through finer-grained user segmentation and investigate the long-term behavioral effects of exposure debiasing.

Our findings demonstrate that causal representation learning can be effectively applied in industrial recommendation systems, offering a practical path toward more interpretable and bias-resilient personalization.

ACKNOWLEDGMENTS

This work was supported in part by LG Uplus Corp., where the practical use case and system deployment were studied. We also thank members of *PseudoLab*, a nonprofit research community, for their valuable discussions and feedback during the development of this project.

REFERENCES

[1] Yu Zheng et al. 2021. Disentangling user interest and conformity for recommendation with causal embedding. *Proceedings of the Web Conference. (WWW '21).* ACM, 2980-2991. DOI: https://doi.org/10.1145/3442381.3449788

[2] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel, 2017. Unbiased learning-to-rank with biased feedback. *Proceedings of the tenth ACM international conference on web search and data mining. (WSDM '17)*. ACM, 781-789. DOI: https://doi.org/10.1145/3018661.3018699

[3] Tobias Schnabel et al. 2016. Recommendations as treatments: Debiasing learning and evaluation. *Proceedings of the 33rd international conference on machine learning. (PMLR '16).* Vol. 48, 1670-1679.

[4] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. *Proceedings of the 12th ACM conference on recommender systems. (RecSys '18)*. ACM, 104-112. DOI: https://doi.org/10.1145/3240323.3240360

[5] Wenjie Wang et al. 2021. Deconfounded recommendation for alleviating bias amplification. *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. (KDD '21)*. ACM, 1717-1725. DOI: https://doi.org/10.1145/3447548.3467249

[6] Tianxin Wei et al. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining (KDD '21).* ACM, 1791-1800. DOI: https://doi.org/10.1145/3447548.3467289

[7] Balázs Hidasi et al. 2015. Session-based recommendations with recurrent neural networks. *arXiv:1511.06939* (2015)

[8] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. *2018 IEEE international conference on data mining (ICDM '18)*. IEEE, 197-206. DOI: https://doi.org/10.1109/ICDM.2018.00035

[9] Minseop Lee et al. 2023. Optimizing Video Recommender System with Bandit-Based Ensemble from Online User Action. *Proceedings of the AAAI 2024 EcoSys Workshop. (AAAI '24), OpenReview.*

[10] Jinmo Kang et al. 2024. A Novel Hybrid Architectures for Overcoming Sparse Data in Subscription Product Recommendation Systems. *IEEE International Conference on Big Data (BigData '24)*. IEEE, 4104-4109. DOI: https://doi.org/10.1109/BigData62323.2024.10825732