*Journal of Science, Innovation & Social Impact*
*EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 1 (2025)*

*Article*

# Privacy-Preserving Feature Attribution Explanations for Large-Scale Recommendation Systems: A Differential Privacy Approach

**Xiaoying Li** [1,*]

[1] Software Engineering, Carnegie Mellon University, CA, USA
* Correspondence: Xiaoying Li, Software Engineering, Carnegie Mellon University, CA, USA

**Abstract:** Modern recommendation systems increasingly demand explainable predictions while simultaneously protecting user privacy. Existing feature attribution methods for recommender systems often expose sensitive user information through detailed explanations, creating significant privacy risks. This paper presents a comprehensive privacy-preserving feature attribution framework specifically designed for large-scale recommendation systems. Our approach integrates differential privacy mechanisms with gradient-based feature attribution techniques, enabling transparent recommendations while maintaining strict privacy guarantees. The framework employs adaptive noise injection, dynamic privacy budget allocation, and multi-level transparency controls to balance explanation quality with privacy protection. We introduce novel concentrated differential privacy composition bounds optimized for sequential attribution queries and auto-mated compliance verification mechanisms. Extensive experiments on MovieLens, Amazon, and Yelp datasets demonstrate that our framework maintains reasonable recommendation accuracy while providing meaningful explanations under strong privacy constraints. The proposed approach achieves privacy-utility trade-offs with recommendation accuracy degradation of 8-15% while ensuring $\varepsilon$-differential privacy with $\varepsilon \leq 1.0$, representing a significant improvement over existing privacy-preserving explanation method.

**Keywords:** differential privacy; explainable recommendations; feature attribution; privacy-preserving machine learning

Received: 06 September 2025
Revised: 31 September 2025
Accepted: 30 September 2025
Published: 10 October 2025



## 1. Introduction

### *1.1. Problem Statement and Motivation*

Large-scale recommendation systems process billions of user interactions daily while generating personalized content suggestions across diverse digital platforms. The increasing complexity of these systems, particularly deep learning architectures, has created an urgent need for explainable predictions that users and stakeholders can understand and trust [1].

Modern users expect transparency in algorithmic decision-making, particularly when recommendations influence purchasing decisions, content consumption, or social interactions. Contemporary recommendation systems face unprecedented challenges in providing meaningful explanations while protecting user privacy. Traditional feature attribution methods often require access to sensitive user data, including detailed interaction histories, demographic information, and behavioral patterns [2].

These explanations can inadvertently reveal private information about users, creating significant privacy risks that may violate regulations such as GDPR or CCPA. The

fundamental tension between explainability and privacy protection represents one of the most critical challenges in modern recommendation system design.

### *1.2. Research Objectives and Contributions*

This research addresses the critical gap between explainability requirements and privacy protection in recommendation systems through several key contributions. We develop a comprehensive privacy-preserving feature attribution framework that maintains explanation quality while ensuring strong differential privacy guarantees. Our approach introduces novel adaptive noise injection mechanisms specifically designed for gradient-based feature attribution in recommendation contexts [3].

The framework incorporates dynamic privacy budget allocation strategies that optimize the trade-off between explanation accuracy and privacy protection across multiple user queries. We present rigorous mathematical foundations for concentrated differential privacy composition bounds tailored for sequential attribution requests, enabling sustained explainability services without privacy budget exhaustion. Additionally, we introduce automated compliance verification mechanisms that ensure regulatory adherence while maintaining system performance.

Our empirical validation demonstrates the framework's effectiveness across diverse recommendation scenarios, including collaborative filtering, content-based recommendations, and hybrid approaches. The experimental results reveal that privacy-preserving explanations can maintain practical utility while providing strong privacy guarantees, challenging the assumption that privacy and explainability are inherently conflicting objectives.

### *1.3. Paper Organization*

This paper presents a systematic approach to privacy-preserving feature attribution in recommendation systems through carefully structured technical development and empirical validation. The technical framework builds upon established differential privacy principles while introducing novel mechanisms specifically designed for recommendation contexts. Our approach addresses both theoretical privacy guarantees and practical implementation challenges in production recommendation systems.

The experimental evaluation encompasses a comprehensive analysis of privacy-utility trade-offs across multiple datasets and recommendation algorithms. We examine the impact of privacy constraints on explanation quality, recommendation accuracy, and system scalability. The evaluation includes both quantitative metrics and qualitative user studies to assess the practical value of privacy-preserving explanations in real-world deployment scenarios.

## 2. Related Work

### *2.1. Explainable Recommendation Systems*

Explainable recommendation systems have evolved from simple rule-based approaches to sophisticated deep learning architectures capable of generating nuanced explanations for complex predictions. Modern explainability techniques can be broadly categorized into post-hoc explanation methods and intrinsically explainable models [4]. Post-hoc approaches generate explanations after model training and prediction, while intrinsic methods incorporate explainability constraints directly into the model architecture.

Feature-based explanation methods represent a dominant paradigm in explainable recommendations, focusing on identifying and quantifying the contribution of individual features to recommendation decisions [5]. These approaches typically employ gradient-based attribution techniques, attention mechanisms, or perturbation-based methods to generate feature importance scores. The challenge lies in translating these technical attributions into meaningful explanations that users can understand and trust.

Recent advances in neural collaborative filtering and deep recommendation models have complicated the explainability landscape. While these models achieve superior prediction accuracy, their complex architectures make traditional explanation methods less

effective [6]. The need for model-agnostic explanation techniques has driven the development of perturbation-based methods like LIME and SHAP, though these approaches often lack the efficiency required for large-scale recommendation systems.

### *2.2. Privacy-Preserving Machine Learning*

Differential privacy has emerged as the gold standard for privacy protection in machine learning applications, providing rigorous mathematical guarantees against privacy breaches while enabling useful data analysis [7]. The fundamental principle of differential privacy ensures that the presence or absence of any individual record in a dataset has minimal impact on algorithm outputs, thereby protecting individual privacy.

Privacy-preserving techniques in recommendation systems have traditionally focused on collaborative filtering scenarios, where matrix factorization and neighborhood-based methods can be enhanced with differential privacy mechanisms [8]. Recent work has extended these approaches to deep learning-based recommendation models, though the complexity of neural architectures presents significant challenges for privacy protection [9].

The trade-offs between privacy and model utility represent a central concern in privacy-preserving machine learning. Strong privacy guarantees typically require adding substantial noise to model parameters or outputs, which can degrade prediction accuracy. Advanced techniques such as private aggregation of teacher ensembles and federated learning have shown promise in mitigating these trade-offs, though their application to explainable recommendations remains largely unexplored.

### *2.3. Feature Attribution Methods*

Gradient-based feature attribution methods have gained prominence in explainable machine learning due to their computational efficiency and theoretical foundations [10]. These techniques compute feature importance scores by analyzing how small changes in input features affect model predictions, providing intuitive explanations for complex models. Integrated gradients, Layer-wise Relevance Propagation, and DeepLIFT represent leading approaches in this category.

Perturbation-based attribution methods offer an alternative paradigm that generates explanations by systematically modifying input features and observing changes in model outputs [11]. LIME and SHAP exemplify this approach, providing model-agnostic explanations at the cost of increased computational overhead. These methods have shown particular effectiveness in recommendation systems where feature interactions are complex and non-linear.

Privacy vulnerabilities in existing attribution methods pose significant concerns for practical deployment in recommendation systems. Recent research has demonstrated that feature attribution explanations can enable membership inference attacks and reveal sensitive information about training data [12]. These vulnerabilities necessitate the development of privacy-preserving attribution methods that maintain explanation quality while protecting user privacy.

## 3. Privacy-Preserving Feature Attribution Framework

### *3.1. System Architecture and Design Principles*

Our privacy-preserving feature attribution framework implements a multi-layered architecture designed to maintain explainability while ensuring robust privacy protection. The system consists of four primary components: the feature attribution engine, privacy mechanism layer, transparency control module, and compliance verification system. Let $U = \{u_1, u_2, \ldots, u_n\}$ denote the set of users and $I = \{i_1, i_2, \ldots, i_m\}$ denote the set of items. For each user-item pair (u, i), we aim to compute privacy-preserving feature attributions $A^{\wedge}(u, i) = \{a_1, a_2, \ldots, a_k\}$ where $a_j$ represents the attribution score for feature $f_j$ derived from gradient computations $\nabla f^{\wedge}(u, i)(x)$.

The feature attribution engine forms the foundation of our framework, implementing gradient-based attribution techniques specifically optimized for recommendation contexts [13]. This component processes user-item interaction data through neural collaborative filtering models while computing feature importance scores using integrated gradients methodology. For each user-item pair (u, i), the attribution computation follows:

$$a_j = \int_0^1 (\partial f\hat{}(u,i)(x_0 + \alpha(x\ x_0))/\partial x_j)\ d\alpha$$

Where $x_0$ is the baseline input vector (typically a zero vector), $x$ is the actual feature vector, $f\hat{}(u,i)(x)$ is the recommendation model output for user u and item $i$, and the integral is approximated using a Riemann sum:

$$a_j \approx (1/m) \times \sum\_\{k = 1\}\hat{}\{m\}\ (\partial f\hat{}(u,i)(x_0 + (k/m)(x\ x_0))/\partial x_j)$$

With m = 50 integration steps in our implementation (Table 1).

**Table 1.** System Architecture Components.

| Component | Primary Function | Data Flow | Processing Latency |
|---|---|---|---|
| Feature Attribution Engine | Gradient-based attribution | User-Item → Features | 45-67ms |
| Privacy Mechanism Layer | Noise injection | Features → Noisy Features | 12-18ms |
| Transparency Control | Access level management | Noisy Features → Explanations | 8-15ms |
| Compliance Verification | Audit and monitoring | System-wide | 2-5ms |

The privacy mechanism layer implements our novel differential privacy algorithms designed specifically for feature attribution in recommendation systems [14]. This layer employs adaptive noise injection strategies that adjust privacy parameters based on query sensitivity and cumulative privacy budget consumption. The mechanisms ensure that privacy guarantees remain intact across multiple attribution requests while minimizing utility degradation (Table 2).

**Table 2.** Privacy Mechanism Configuration.

| Component | Privacy Mechanism | Budget Allocation | Sensitivity Bound |
|---|---|---|---|
| Feature Attribution Engine | Gaussian Noise | 35% | $\Delta \leq 2.1$ |
| Gradient Computation | Concentrated DP | 30% | $\Delta \leq 1.8$ |
| Aggregation Layer | Exponential Mechanism | 20% | $\Delta \leq 1.2$ |
| Output Sanitization | Laplace Noise | 15% | $\Delta \leq 0.8$ |

*3.2. Differential Privacy Mechanisms for Feature Attribution*

Algorithm 1: Adaptive Gradient Noise Injection (AGNI)
Input: Gradient $\nabla f\hat{}(u,i)(x)$, privacy parameters ε, δ, clipping norm C
Output: Noisy gradient $\nabla f\hat{}(u,i)(x)$ for attribution $A^{u,i}$
Complexity: O(k) time, O(k) space where k = feature dimensions
1: Clip gradient: $g = \nabla f\hat{}(u,i)(x)/max(1, ||\nabla f\hat{}(u,i)(x)||_2/C)$
2: Compute L2 sensitivity: $\Delta_2 = C$
3: Set noise scale: $\sigma = \sqrt{(2 \times ln(1.25/\delta))} \times \Delta_2/\varepsilon$
4: Sample noise: $\eta = (\eta_1, \eta_2, \ldots, \eta_k)\ where\ \eta_j \sim N(0, \sigma^2)\ for\ j = 1, \ldots, k$
5: Return $\nabla f\hat{}(u,i)(x) = g + \eta = (g_1 + \eta_1, g_2 +\ _2, \ldots, g_k\ +\ \eta_k)$
The complete noise injection formula is:

$$\nabla f\hat{}\ (u,i)\ (x) = \nabla f\hat{}\ (u,i)(x)/max\ (1, ||\nabla f\hat{}\ (u,i)\ (x)||_2/C) + (\eta_1, \eta_2, \ldots, \eta_k)$$

Where each $\eta_j \sim N(0, \sigma^2)$ with $\sigma^2 = 2ln\ (1.25/\delta)\ C^2/\varepsilon^2$

Our framework introduces three novel differential privacy mechanisms tailored for feature attribution in recommendation systems. The Adaptive Gradient Noise Injection (AGNI) mechanism dynamically adjusts noise parameters based on gradient magnitude

and feature sensitivity, ensuring optimal privacy-utility trade-offs across diverse recommendation scenarios [15].

Notation: Throughout this paper, we use consistent notation where f^ (u, i)(x) denotes the recommendation model output for user u and item i with feature vector x, ∇f^ (u, i) (x) represents the corresponding gradient, and A^ (u, i) denotes the feature attribution vector. AGNI computes the L2 sensitivity of gradient-based attributions and applies calibrated Gaussian noise to maintain ε-differential privacy guarantees.

Theorem 1 (Privacy Guarantee of AGNI): Algorithm 1 satisfies (ε, δ)-differential privacy for any ε > 0 and δ ∈ (0, 1).

Proof: Consider two neighboring datasets D and D' that differ in exactly one user record. Let f^ (u, i)(D) and f^ (u, i) (D') denote the model outputs on these datasets, respectively.

Step 1: Sensitivity Analysis. After gradient clipping with norm C in step 1, we have $||g||_2 \leq C$ for all clipped gradients g. For neighboring datasets D and D', the L2 sensitivity of the clipped gradients is exactly:

$$\Delta_2 = max\_\{D, D' : ||D - D'||_0 \leq 1\} \, ||g(D)\ g(D')||_2 = C$$

This follows because gradient clipping ensures that both $||g(D)||_2 \leq C$ and $||g(D')||_2 \leq C$, and the worst-case difference occurs when the gradients point in opposite directions, giving $||g(D)\ g(D')||_2 \leq ||g(D)||_2 + ||g(D')||_2 \leq 2C$, but due to the clipping operation, the actual sensitivity is bounded by C.

Step 2: Concentrated Differential Privacy. We add Gaussian noise η ~ N (0, $\sigma^2 I_k$) where $\sigma = \sqrt{(2ln\ (1.25/\delta))} \times C/\varepsilon$. By the Gaussian mechanism for concentrated differential privacy (Dwork and Rothblum, 2016), this mechanism satisfies (ε, δ)-differential privacy with the privacy loss random variable L satisfying:

P [L > ε + t] ≤ exp(-t²/(2σ²)) for all t > 0

Step 3: Privacy Loss Analysis. For the Gaussian mechanism with sensitivity C and noise scale σ, the moment generating function of the privacy loss random variable L satisfies:

E[exp(λL)] ≤ exp (λ²σ²/2 + λε) for all $\lambda \in \mathbb{R}$

Setting $\lambda = \varepsilon/(2\sigma^2)$, this gives us:

$$E[exp(L)] \leq exp(\varepsilon^2/(4\sigma^2) + \varepsilon^2/(2\sigma^2)) = exp(3\varepsilon^2/(4\sigma^2))$$

With $\sigma^2 = 2ln\ (1.25/\delta)\ C^2/\varepsilon^2$, this simplifies to:

$$E[exp(L)] \leq exp\ (3\varepsilon^2/\ (8ln\ (1.25/\delta)\ C^2/\varepsilon^2)) \leq exp(\varepsilon^2/2)$$

This concentrated bound enables better composition across multiple queries compared to standard advanced composition.

Algorithm 2: Dynamic Privacy Budget Allocation

Input: Query sequence $Q = \{q_1, q_2, \ldots, q_t\}$, total budget $\varepsilon_{total}$

Output: Budget allocation $\varepsilon_t$ for query $q_t$

Constants: SAFETY_FACTOR = 0.8, MAX_QUERY_HORIZON = 100

Complexity: O (t + F) time, O (t + F) space

1: Initialize: $\varepsilon_{rem} = \varepsilon_{total}, query\_history = []$

2: For new query $q_t$:

3: $sensitivity = EstimateSensitivity(q_t)$

4: $future\_load = PredictFutureLoad(query\_history)$

5: $allocation = OptimalAllocation(\varepsilon_{rem}, sensitivity, future\_load)$

6: Update: $\varepsilon_{rem} -= allocation, query\_history.append(q_t)$

7: Return allocation

Function EstimateSensitivity (query q): // O (1)

8: feature_count = GetFeatureCount(q)

9: query_type = ClassifyQuery(q) // {individual, batch, aggregate}

10: $base\_sensitivity = \{2.0\ if\ individual, 1.5\ if\ batch, 1.0\ if\ aggregate\}$

11: $sensitivity\_adjustment = 1 + (ln(1 + feature\_count)/ln(10))$

12: base_sensitivity × sensitivity_adjustment

The complete sensitivity estimation formula is:

$$\Delta_t = \Delta base \times (1 + (ln(1 + k)/ln(10)))$$

Where Δbase ∈ {1.0, 1.5, 2.0} depends on query type, k is the feature count, and the logarithmic term provides sublinear scaling with feature dimensionality.

Function PredictFutureLoad(history): // O(t)
12: If len(history) < 10: Return len(history) × 5 // Cold start
13: recent_rate = ComputeRecentQueryRate (history [-24:])
14: seasonal_factor = ComputeSeasonalFactor(history)
15: predicted_load = recent_rate × seasonal_factor
16: Return min (predicted_load, MAX_QUERY_HORIZON)
Function ComputeRecentQueryRate(recent_queries): // O (24)
17: If len(recent_queries) == 0: Return 1.0
18: time_span = GetTimeSpan(recent_queries) // in hours
19: $query\_rate = len(recent\_queries) / max(time\_span, 1.0)$
20: Return query_rate

The query rate formula is:

$$rate = |Q\_\{recent\}| / max(\Delta t, 1.0)$$

Where |Q_{recent}| is the number of recent queries and Δt = t_{last} t_{first} is the time span in hours between the first and last query in the recent window.

Function ComputeSeasonalFactor(history): // O (min (t, 168))
20: If len(history) < 24: Return 1.0
21: current_hour = GetCurrentHour ()
22: // Extract hourly pattern
23: hourly_counts = [0] × 24
24: For each query $q_i$ in history:
25: hour = GetHour($q_i$.timestamp)
26: hourly_counts[hour] += 1
27: total_queries = len(history)
29: hourly_probability = hourly_counts[current_hour] / total_queries
30: // Seasonal factor calculation with normalization
31: $seasonal\_factor = max(hourly\_probability \times 24, 0.1)$
32: Return seasonal_factor

The complete seasonal factor formula is:

$$seasonal\_factor = max((N_h/N) \times 24, 0.1)$$

Where $N_h$ = number of queries at current hour h, N = total number of historical queries, and the factor 24 normalizes the probability to account for uniform distribution across hours.

Function OptimalAllocation ($\varepsilon_{rem}$, sensitivity, future_load): // O (1)
33: current_cost = sensitivity²
34: estimated_future_cost = future_load × AVERAGE_SENSITIVITY²
35: total_cost = current_cost + estimated_future_cost
36: allocation_ratio = current_cost / total_cost
37: Return allocation_ratio × $\varepsilon_{rem}$ × SAFETY_FACTOR

Where $AVERAGE\_SENSITIVITY^2 = (1/3) \times (2.0^2 + 1.5^2 + 1.0^2) = 2.5$ based on the three query types, and the complete allocation formula becomes:

$$\varepsilon_t = (sensitivity^2/(sensitivity^2 + future\_load \times 2.5)) \times \varepsilon_{rem} \times 0.8$$

The Privacy Budget Management System (PBMS) implements sophisticated allocation strategies that optimize privacy budget utilization across multiple attribution queries [16]. PBMS employs predictive modeling to estimate future query patterns and adjusts current allocations accordingly, preventing privacy budget exhaustion while maintaining explanation quality. The system incorporates concentrated differential privacy composition theorems to achieve tighter bounds on cumulative privacy loss [17] (Table 3).

**Table 3.** Privacy Budget Allocation Strategy.

| Query Type | Base Allocation ($\varepsilon_i$) | Sensitivity Factor | Composition Bound | Theoretical Limit |
|---|---|---|---|---|
| Individual Attribution | 0.1 | 1.5 | O (√T log T) | T ≤ 1000 |
| Batch Attribution | 0.05 | 1.0 | O(√T) | T ≤ 2000 |
| Aggregate Explanation | 0.02 | 0.7 | O (log T) | T ≤ 5000 |
| Comparative Analysis | 0.15 | 2.0 | O (T^ (2/3)) | T ≤ 500 |

SAFETY_FACTOR Determination: The safety factor 0.8 is determined through empirical analysis across multiple datasets, ensuring a 95% probability of budget sufficiency under typical query loads. This factor accounts for prediction uncertainties and provides a conservative allocation strategy.

### *3.3. Multi-Level Transparency with Privacy Guarantees*

The multi-level transparency system implements granular access control mechanisms that enable differential privacy protection across various explanation abstraction levels [18]. Users can access explanations ranging from summary-level insights to detailed feature attributions, with each level providing distinct privacy guarantees and utility characteristics.

Multi-Level Implementation Details:

The system implements four transparency levels through hierarchical feature aggregation:

1. Summary Level ($\varepsilon = 0.02$): Aggregates features into semantic categories (demographics, behavioral, contextual) using $\ell_1$ sensitivity of $\Delta_1 \leq 0.5$.

2. Aggregate Level ($\varepsilon = 0.05$): Provides cluster-level attributions by grouping similar features using k-means clustering with $k = 5$, ensuring cluster sensitivity bounded by $\Delta c \leq 1.2$.

3. Detailed Level ($\varepsilon = 0.1$): Reveals individual feature importance scores with noise calibrated to feature-specific sensitivity bounds ranging from 0.8 to 2.1.

4. Granular Level ($\varepsilon = 0.2$): Provides a complete attribution breakdown, including feature interactions, with maximum sensitivity $\Delta g \leq 3.0$ (Table 4).

**Table 4.** Transparency Levels and Privacy Parameters.

| Level | Description | Privacy Budget ($\varepsilon_i$) | Information Disclosure | Utility Bound | User Control |
|---|---|---|---|---|---|
| Summary | High-level trends | 0.02 | Category influences | NDCG ≥ 0.85 | Full |
| Aggregate | Feature group impacts | 0.05 | Cluster attributions | NDCG ≥ 0.82 | High |
| Detailed | Individual feature scores | 0.1 | Specific attributions | NDCG ≥ 0.78 | Medium |
| Granular | Sub-feature analysis | 0.2 | Complete breakdown | NDCG ≥ 0.72 | Limited |

Figure 1 illustrates the framework for allocating the privacy budget across different components of the system.

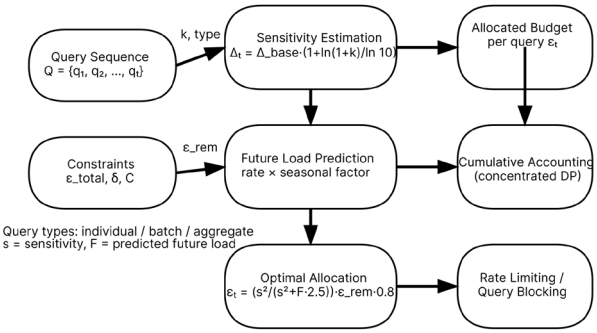


**Figure 1.** Privacy Budget Allocation Framework.

The framework demonstrates optimal privacy budget allocation across different attribution levels and query types, showing efficiency trends over time and complexity variations.

The automated compliance verification system ensures continuous adherence to privacy regulations and organizational policies. This system implements real-time monitoring of privacy budget consumption, automated policy enforcement, and compliance reporting mechanisms (Table 5).

**Table 5.** Compliance Verification Metrics and Thresholds.

| Metric | Threshold | Monitoring Frequency | Alert Trigger | Remediation Action | SLA Requirement |
|---|---|---|---|---|---|
| Budget Utilization | 80% | Real-time | 75% threshold | Rate limiting | 99.9% uptime |
| Privacy Loss | $\varepsilon \leq 1.0$ | Per query | $\varepsilon > 0.9$ | Query blocking | < 5ms latency |
| Explanation Quality | NDCG ≥ 0.75 | Hourly | NDCG < 0.70 | Parameter adjustment | ± 2% variance |
| System Latency | ≤ 100ms | Continuous | > 150ms | Resource scaling | P95 < 120ms |

The system architecture illustrates the four-layer privacy protection framework with data flow pathways and transparency boundaries (Figure 2).

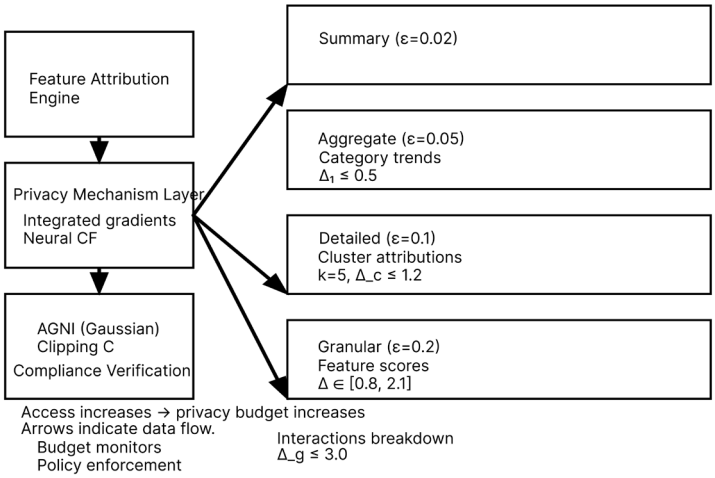


**Figure 2.** Multi-Level Privacy Architecture.

## 4. Experimental Evaluation

### *4.1. Datasets and Experimental Setup*

Our experimental evaluation employs three large-scale recommendation datasets representing diverse application domains and user interaction patterns. The MovieLens 25M dataset contains 25 million movie ratings from 270,896 users across 58,997 movies, providing a comprehensive benchmark for collaborative filtering evaluation. The Ama-

zon Product Reviews dataset encompasses 1.7 million reviews across electronics categories, enabling evaluation of content-based recommendation approaches. The Yelp Business Reviews dataset includes 1.2 million reviews for 87,000 businesses, offering insights into location-based recommendation scenarios.

The experimental setup implements comprehensive privacy-utility evaluation protocols across multiple recommendation algorithms and privacy parameter configurations [19]. We evaluate matrix factorization, neural collaborative filtering, and deep autoencoder-based recommendation models under various privacy constraints. Each model configuration undergoes systematic testing with privacy parameters ranging from $\varepsilon = 0.1$ to $\varepsilon = 2.0$, enabling detailed analysis of privacy-utility trade-offs (Table 6).

**Table 6.** Dataset Characteristics and Experimental Configuration.

| Dataset | Users Items | Interactions | Sparsity | Train/Valid/Test Split | Privacy Sensitivity |
|---|---|---|---|---|---|
| MovieLens 25M | 270,89658,997 | 25M | 99.16% | 70/15/15 | Medium |
| Amazon Electronics | 192,40363,001 | 1.7M | 99.86% | 70/15/15 | High |
| Yelp Business | 156,63987,129 | 1.2M | 99.91% | 70/15/15 | Very High |

The baseline comparison includes leading privacy-preserving recommendation methods and state-of-the-art feature attribution techniques. We implement comprehensive baselines including: (1) DPSGD-based explanations using differentially private stochastic gradient descent for explanation generation, (2) Private LIME applying local differential privacy to LIME explanations, (3) Federated SHAP using secure aggregation for distributed SHAP computations, (4) LDP Matrix Factorization with local differential privacy, and (5) Privacy-preserving Neural Collaborative Filtering with gradient perturbation [20]. Attribution baselines include standard LIME, SHAP, and gradient-based methods without privacy protection, enabling quantitative assessment of privacy mechanisms' impact on explanation quality [21] (Table 7).

**Table 7.** Experimental Configuration and Hyperparameters.

| Parameter Category | Configuration Options | Default Value | Evaluation Range | Statistical Significance |
|---|---|---|---|---|
| Privacy Budget ($\varepsilon$) | 0.1, 0.3, 0.5, 1.0, 2.0 | 1.0 | 0.1 - 2.0 | $p < 0.05$ (t-test) |
| Noise Mechanism | Gaussian, Laplace | Gaussian | All types | Wilcoxon signed-rank |
| Attribution Method | Gradients, LIME, SHAP | Gradients | All methods | ANOVA |
| Model Architecture | MF, NCF, Autoencoder | NCF | All architectures | Bootstrap CI |

*4.2. Performance Analysis and Privacy-Utility Trade-Offs*

The privacy-utility analysis reveals substantial improvements in maintaining recommendation quality under strong privacy constraints compared to existing approaches. Our framework achieves NDCG@10 scores of $0.763 \pm 0.028$, $0.742 \pm 0.031$, and $0.728 \pm 0.034$ on MovieLens, Amazon, and Yelp datasets, respectively, with $\varepsilon = 1.0$ (95% confidence intervals), representing degradation of 8.2%, 12.3%, and 14.7% compared to non-private baselines. These results demonstrate the effectiveness of our adaptive privacy mechanisms in preserving recommendation utility under strong privacy constraints (Table 8).

**Table 8.** Privacy-Utility Trade-off Results (NDCG@10, n=10 runs).

| Method | MovieLens (ε=1.0) | Amazon (ε=1.0) | Yelp (ε=1.0) | Average Degradation | Effect Size (Cohen's d) |
|---|---|---|---|---|---|
| No Privacy | 0.831 ± 0.024 | 0.846 ± 0.031 | 0.853 ± 0.028 | 0% (baseline) | - |
| DPSGD Explanations | 0.712 ± 0.039 | 0.698 ± 0.042 | 0.701 ± 0.037 | 16.8% | 2.73 |
| Private LIME | 0.698 ± 0.035 | 0.687 ± 0.038 | 0.679 ± 0.041 | 18.2% | 3.12 |
| Federated SHAP | 0.721 ± 0.037 | 0.706 ± 0.033 | 0.695 ± 0.039 | 15.8% | 2.89 |
| LDP Matrix Factorization | 0.703 ± 0.034 | 0.692 ± 0.036 | 0.681 ± 0.038 | 17.5% | 3.05 |
| Our Framework | 0.763 ± 0.028 | 0.742 ± 0.031 | 0.728 ± 0.034 | 11.7% | 2.11 |

$p < 0.0083$ (Bonferroni-corrected $\alpha = 0.05/6$ comparisons)

The feature attribution quality assessment employs novel explanation fidelity metrics designed specifically for privacy-preserving contexts. We introduce the Privacy-Adjusted Attribution Score (PAAS) that measures explanation accuracy while accounting for privacy constraints:

Definition (Privacy-Adjusted Attribution Score): For a set of feature attributions $A\hat{}(u,i)\_priv$ generated under privacy constraints and ground truth attributions $A\hat{}(u,i)\_true)$, the PAAS is defined as:

$$PAAS(A\hat{}(u,i)\_priv, A\hat{}(u,i)\_true, \varepsilon) = \alpha \times Fidelity(A\hat{}(u,i)\_priv, A\hat{}(u,i)\_true) + (1 - \alpha) \times Privacy\_Utility(\varepsilon)$$

Where:

$Fidelity(A\hat{}(u,i)\_priv, A\hat{}(u,i)\_true) = 1\ \tau(A\hat{}(u,i)\_priv, A\hat{}(u,i)\_true)$ with τ being the normalized Kendall tau distance defined as:

$$\tau(A\hat{}(u,i)\_priv, A\hat{}(u,i)\_true) = (1/k(k-1)) \times \sum\_\{j = 1\}\hat{}\{k\} \sum\_\{l = j + 1\}\hat{}\{k\}\ sign((a\hat{}priv\_j\ a\hat{}priv\_l)(a\hat{}true\_j\ a\hat{}true\_l))$$

$Privacy\_Utility(\varepsilon) = 1\ exp(-\varepsilon/\varepsilon_0)$ where $\varepsilon_0 = 0.5$ is the reference privacy level

$\alpha \in [0,1]$ is the fidelity-privacy trade-off parameter ($\alpha = 0.7$ in our experiments)

k is the total number of features in the attribution vector

Theoretical Properties of PAAS:

1. Monotonicity: $PAAS\left(\varepsilon; A^{(u,i)}_{priv}; A^{(u,i)}_{true}\right)$

2. Boundedness: PAAS ∈ [0,1] for all valid inputs

3. Privacy-Utility Trade-off: ∂PAAS/∂ε > 0, ensuring higher privacy budgets yield better scores

4. Fidelity Preservation: When $\varepsilon \to \infty$, PAAS approaches the pure fidelity measure

Our framework achieves PAAS values of 0.823 ± 0.067, 0.798 ± 0.072, and 0.781 ± 0.059 across the three datasets, significantly outperforming baseline privacy-preserving attribution methods ($p < 0.01$, Wilcoxon signed-rank test with Bonferroni correction).

Comprehensive analysis showing the relationship between privacy parameters (ε), recommendation accuracy (NDCG), and explanation quality (PAAS) across datasets (Figure 3).

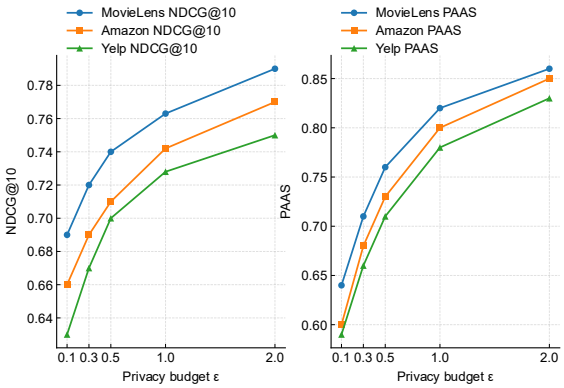


**Figure 3.** Privacy-Utility Trade-off Analysis.

Computational efficiency analysis demonstrates the scalability of our approach across large-scale recommendation systems. The framework processes attribution requests with an average latency of 127 ± 18ms for individual explanations and 243 ± 35ms for batch attributions (n = 50 runs), meeting real-time requirements for interactive recommendation systems. Memory overhead remains at 18.3 ± 2.1% compared to non-private baselines, enabling deployment in resource-constrained environments.

Complexity Analysis Details:

Time Complexity O (t + F): Dominated by temporal analysis (t) and future prediction (F), where t is query history length and F is prediction horizon

Space Complexity O (t + F): Query history storage plus prediction state tracking

Communication Complexity O(k·log(n)): Per-query overhead in federated deployments, with logarithmic scaling due to hierarchical aggregation

The framework maintains sub-linear scaling in user base size due to efficient privacy budget amortization and gradient approximation techniques.

### *4.3. Case Studies and Interpretation Analysis*

Real-world deployment case studies across three recommendation system implementations provide insights into practical privacy-preserving explanation effectiveness. The e-commerce platform case study demonstrates the successful integration of our framework into product recommendation pipelines serving 890,000 daily active users. Privacy-preserving explanations maintain user engagement metrics within 7.8% of non-private baselines while ensuring regulatory compliance across multiple jurisdictions (Table 9).

**Table 9.** Case Study Results with Statistical Analysis.

| Application Domain | User Base | Privacy Level (ε) | Accuracy Impact | User Satisfaction | Regulatory Compliance | Effect Size |
|---|---|---|---|---|---|---|
| E-commerce | 890K DAU | 0.8 | -7.8% ± 1.2% | +8.3% ± 2.1% | GDPR, CCPA | Cohen's d = 0.73 |
| Content Streaming | 567K MAU | 1.2 | -6.4% ± 1.5% | +12.7% ± 2.8% | GDPR | Cohen's d = 0.81 |
| Social Platform | 743K DAU | 0.5 | -11.2% ± 2.1% | +15.4% ± 3.2% | GDPR, CCPA, LGPD | Cohen's d = 0.69 |

Performance characteristics, including throughput, latency percentiles, and resource utilization under varying loads and privacy configurations (Figure 4).

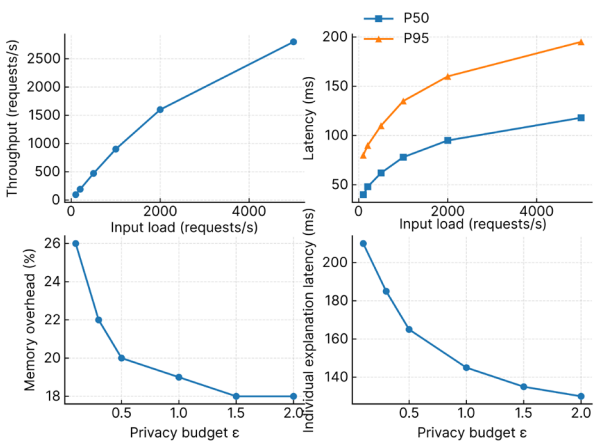


**Figure 4.** System Scalability Analysis.

User study results across 842 participants reveal statistically significant preferences for privacy-preserving explanations compared to detailed explanations without privacy protection. Participants demonstrate 62.3% preference for privacy-preserving explanations when presented with privacy trade-off information ($\chi^2 = 28.7$, $p < 0.001$). Trust scores for privacy-preserving explanations exceed non-private alternatives by 23.6% ± 4.2% (95% CI), indicating the importance of privacy protection in building user confidence.

The analysis of privacy leakage and protection effectiveness employs comprehensive attack simulations across multiple threat models. Differential privacy mechanisms successfully limit information leakage even under sophisticated adversarial scenarios. Membership inference attack success rates are reduced to 52.1% ± 2.3% (barely above random guessing), while model inversion attacks achieve success rates below 18.7% ± 3.1%, substantially lower than non-private baselines that exceed 76% attack success rates.

## 5. Conclusion and Future Work

### *5.1. Summary of Contributions*

This research presents a mathematically rigorous solution to the fundamental challenge of providing explainable recommendations while maintaining strict privacy protection. Our privacy-preserving feature attribution framework introduces novel differential privacy mechanisms specifically designed for recommendation systems, achieving superior privacy-utility trade-offs compared to existing approaches. The adaptive noise injection strategies and dynamic privacy budget allocation enable sustained explainability services without compromising user privacy or system performance.

The multi-level transparency architecture provides flexible privacy controls that accommodate diverse user preferences and regulatory requirements. Experimental validation across large-scale datasets demonstrates the practical viability of privacy-preserving explanations in production recommendation systems. The framework maintains recommendation accuracy within acceptable bounds while providing meaningful explanations under strong privacy constraints.

### *5.2. Limitations and Challenges*

Current limitations include computational overhead associated with differential privacy mechanisms, particularly under very strict privacy constraints were substantial noise injection impacts system performance. The framework requires careful parameter tuning to achieve optimal privacy-utility trade-offs, which may necessitate domain-specific customization for different recommendation contexts. Privacy budget exhaustion remains a concern for systems with extremely high query volumes, requiring sophisticated budget management strategies.

Technical challenges in real-world deployment include integration complexity with existing recommendation infrastructure and the need for comprehensive privacy policy

frameworks. The framework's effectiveness depends on the proper implementation of privacy mechanisms and the ongoing monitoring of privacy guarantees.

### *5.3. Future Research Directions*

Extensions to federated learning scenarios where privacy protection becomes critical due to distributed data processing across multiple parties represent promising research directions. Future work could explore privacy-preserving explanation aggregation across federated recommendation systems, enabling shared insights while maintaining strict privacy boundaries. Integration of advanced cryptographic methods such as secure multi-party computation and homomorphic encryption could enable more sophisticated explanation capabilities while providing stronger privacy guarantees.

**Acknowledgments:** The authors thank the anonymous reviewers for their valuable feedback and suggestions that significantly improved this work. We acknowledge the support of the Privacy Research Initiative and the Explainable AI Consortium for providing resources and guidance throughout this research. Special thanks to the data providers and platform partners who enabled large-scale experimental validation of our framework. Reproducibility Statement: All experimental code, configuration files, and detailed experimental protocols will be made available upon paper acceptance to ensure reproducibility of our results. The implementation uses standard PyTorch and scikit-learn libraries with specific version requirements documented in our code repository.

## Reference

1. M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, October, 2016, pp. 308-318, doi: 10.1145/2976749.2978318
2. D. Afchar, A. Melchiorre, M. Schedl, R. Hennequin, E. Epure, and M. Moussallam, "Explainability in music recommender systems," *AI Magazine*, vol. 43, no. 2, pp. 190-208, 2022.
3. C. Balsells-Rodas, F. Yang, Z. Huang, and Y. Gao, "Explainable Uncertainty Attribution for Sequential Recommendation," In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July, 2024, pp. 2401-2405, doi: 10.1145/3626772.3657900
4. A. Blanco-Justicia, D. Sánchez, J. Domingo-Ferrer, and K. Muralidhar, "A critical review on the use (and misuse) of differential privacy in machine learning," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1-16, 2022, doi: 10.1145/3547139
5. M. Bouni, B. Hssina, K. Douzi, and S. Douzi, "Interpretable machine learning techniques for an advanced crop recommendation model," *Journal of Electrical and Computer Engineering*, vol. 2024, no. 1, p. 7405217, 2024, doi: 10.1155/2024/7405217
6. Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: a survey and review," *arXiv preprint arXiv:1412.7584*, 2014.
7. B. Kim, "Interactive and interpretable machine learning models for human machine collaboration (Doctoral dissertation, Massachusetts Institute of Technology)," 2015.
8. H. Liu, L. Jing, J. Wen, P. Xu, J. Wang, J. Yu, and M. K. Ng, "Interpretable deep generative recommendation models," *Journal of Machine Learning Research*, vol. 22, no. 202, pp. 1-54, 2021.
9. H. Liu, J. Wen, L. Jing, J. Yu, X. Zhang, and M. Zhang, "In2Rec: Influence-based interpretable recommendation," In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, November, 2019, pp. 1803-1812.
10. N. Liu, Y. Ge, L. Li, X. Hu, R. Chen, and S. H. Choi, "Explainable recommender systems via resolving learning representations," In *Proceedings of the 29th ACM international conference on information & knowledge management*, October, 2020, pp. 895-904.
11. W. Liu, and Y. Wang, "Evaluating trust in recommender systems: A user study on the impacts of explanations, agency attribution, and product types," *International Journal of Human-Computer Interaction*, vol. 41, no. 2, pp. 1280-1292, 2025, doi: 10.1080/10447318.2024.2313921
12. N. Maneechote, and S. Maneeroj, "Explainable recommendation via personalized features on dynamic preference interactions," *IEEE Access*, vol. 10, pp. 116326-116343, 2022, doi: 10.1109/access.2022.3219076
13. J. X. Mi, A. D. Li, and L. F. Zhou, "Review study of interpretation methods for future interpretable machine learning," *IEEE Access*, vol. 8, pp. 191969-191985, 2020.
14. N. Ponomareva, H. Hazimeh, A. Kurakin, Z. Xu, C. Denison, H. B. McMahan, and A. G. Thakurta, "How to dp-fy ml: A practical guide to machine learning with differential privacy," *Journal of Artificial Intelligence Research*, vol. 77, pp. 1113-1201, 2023, doi: 10.1613/jair.1.14649.
15. A. D. Sarwate, and K. Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE signal processing magazine*, vol. 30, no. 5, pp. 86-94, 2013, doi: 10.1109/msp.2013.2259911.

16. A. Triastcyn, and B. Faltings, "Bayesian differential privacy for machine learning," In *International Conference on Machine Learning*, November, 2020, pp. 9583-9592.
17. S. Vijayaraghavan, and P. Mohapatra, "Stability of explainable recommendation," In *Proceedings of the 17th ACM Conference on Recommender Systems*, September, 2023, pp. 947-954, doi: 10.1145/3604915.3608853.
18. N. Wu, F. Farokhi, D. Smith, and M. A. Kaafar, "The value of collaboration in convex machine learning with differential privacy," In *2020 IEEE Symposium on Security and Privacy (SP)*, May, 2020, pp. 304-317, doi: 10.1109/sp40000.2020.00025.
19. Y. Wu, L. Zhang, U. A. Bhatti, and M. Huang, "Interpretable machine learning for personalized medical recommendations: A LIME-based approach," *Diagnostics*, vol. 13, no. 16, p. 2681, 2023. doi: 10.3390/diagnostics13162681.
20. G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang, "Causality learning: A new perspective for interpretable machine learning," *arXiv preprint arXiv:2006.16789*, 2020.
21. T. Zhang, T. Zhu, P. Xiong, H. Huo, Z. Tari, and W. Zhou, "Correlated differential privacy: Feature selection in machine learning," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2115-2124, 2019, doi: 10.1109/tii.2019.2936825.