EXTRAPOLATING USER PREFERENCES VIA COMPOSITIONAL ATOMIC STATE TRANSITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential recommendation models typically encode user dynamics as continuous latent vectors. However, these monolithic representations conflate distinct preference dimensions, limiting their ability to generalize compositionally to novel configurations of known elements. To address this, we propose **C**ompositional **A**tomic **S**tate **T**ransitions, a framework that reconceptualizes preference evolution as state transitions within a discrete, compositional semantic space. Rather than allowing unconstrained latent shifts, CAST decomposes user profiles into facet-specific semantic atoms and models temporal dynamics via typed atomic operations. This design constrains future extrapolation to recombinations of known atoms, yielding interpretable semantic trajectories and soft transition signals for downstream candidate ranking. To evaluate this capability, we introduce a temporal extrapolation protocol with strict chronological and combination holdouts, isolating compositional generalization from mere entity recognition. Experiments demonstrate that CAST recovers a non-trivial subset of unseen preference configurations and improves candidate ranking across multiple sequential backbones. Ultimately, CAST provides an interpretable inductive bias for modeling compositional user dynamics.

1 INTRODUCTION

Recommendation systems mediate everyday choices such as what to watch, read, listen to, or purchase. Each recommendation is based on partial behavioral traces rather than on a complete, explicit statement of preference. Therefore, the system must infer a working user state from past interactions and revise it as new evidence becomes available. This state is neither static nor homogeneous. A user may maintain a stable interest in science fiction while temporarily shifting toward humorous narratives or coastal settings, and several such preferences may coexist. A useful recommender must distinguish persistent preferences from transient intents while remaining able to combine familiar factors in a new configuration.

Most sequential recommendation benchmarks rely on next-item ranking. However, this metric provides an indirect and limited view of user states. High ranking accuracy can result from popularity bias or memorization rather than true preference reasoning. Furthermore, such scores fail to explain the underlying reasons for preference shifts, their connection to past behaviors, or the model's generalization ability in novel contexts.

Existing studies tackle specific aspects of user modeling: sequential recommenders capture temporal patterns, multi-interest methods represent diverse preferences, and language-based models utilize textual profiles (Kang & McAuley, 2018; Sun et al., 2019; Li et al., 2019; Cen et al., 2020; Zhai et al., 2024; Zhou et al., 2024; Ramos et al., 2024). Despite their predictive power, these methods lack a unified mechanism to track and recombine dynamic user preferences as reusable components over time. We address this gap with a compositional hypothesis. Rather than assuming that user preferences shift arbitrarily within an unbounded space, we posit that a set of fundamental semantic factors serves as a basis. Consequently, novel preference directions are not entirely new concepts but rather recombinations of these underlying basis elements observed in the user's history. The key challenge becomes whether a model can extract these familiar components, assemble them into unseen configurations, and use the resulting state for candidate ranking.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

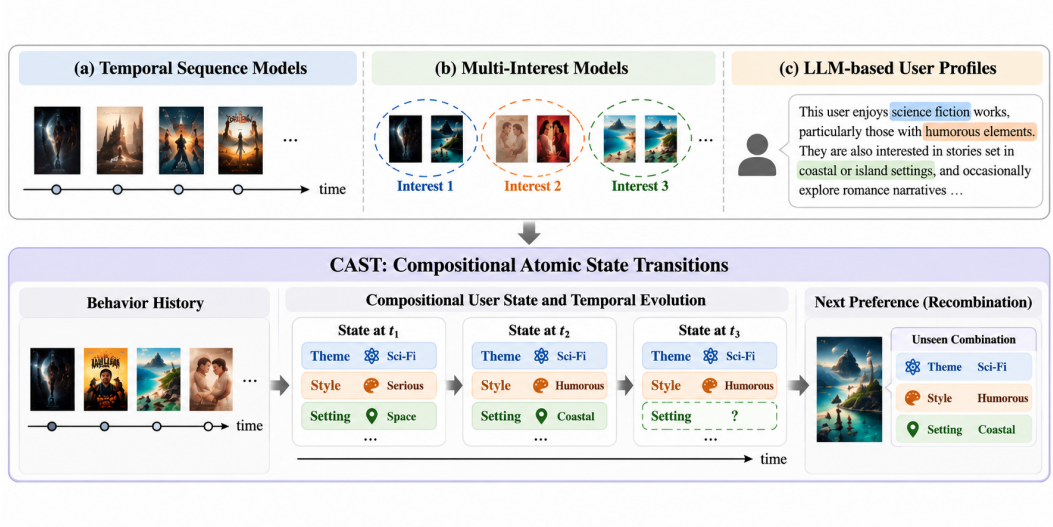


Figure 1: From holistic user modeling to compositional user-state transitions.

Inspired by compositional generalization (Lake & Baroni, 2018; Keysers et al., 2020; Luketina et al., 2025), we adapt this paradigm to user modeling without imposing strict causal assumptions (Locatello et al., 2019). To represent user states as reusable semantic bases that support interpretable transitions and measurable generalization to unseen combinations, we propose CAST (**C**ompositional **A**tomic **S**tate **T**ransitions). This framework reconceptualizes evolving user states as structured collections of typed semantic components rather than monolithic vectors. At each interaction step, CAST distills historical behaviors into discrete semantic atoms that serve as reusable bases. These atoms fall into three operational facets: Theme for the core content of interest, Style for the preferred experiential mode, and Setting for the environmental context. Within this atomic space, preference evolution is modeled through typed operations over these semantic bases. During inference, the recommendation score is augmented by facet-specific compatibility between the predicted state and candidate items. This design explicitly isolates three fundamental capabilities that standard metrics often conflate: recognizing semantic shifts, assembling them into unseen configurations, and leveraging the resulting state for accurate ranking.

Our empirical evaluations reveal three main findings regarding this compositional approach. First, facet decomposition substantially improves atom-level semantic recovery compared to holistic and shuffled baselines. Second, under strict combination holdout settings, CAST successfully retrieves the constituent basis atoms of unseen compound states, provided those individual atoms appear separately in the training data. Third, candidate-aware atom matching improves standard ranking across multiple evaluated configurations, and this mechanism mitigates the performance degradation typically observed on strictly out-of-distribution (OOD) combinations.

Our contributions are as follows:

- We introduce a compositional generalization perspective to user modeling. By conceptualizing evolving preferences as recombinations of fundamental semantic bases, we establish a new evaluation paradigm that moves beyond standard ranking metrics to assess interpretable state transitions.
- We propose CAST, a novel framework that departs from monolithic user vectors. It effectively distills historical behaviors into reusable semantic atoms, models preference evolution via typed operations, and makes the structured state actionable through candidate-aware matching.
- We conduct extensive empirical evaluations validating this compositional approach. Results demonstrate that CAST significantly improves atom-level semantic recovery, successfully retrieves basis atoms of unseen compound states, and mitigates performance degradation in strictly out-of-distribution scenarios.

2 Related Works

Sequential Recommendation. Sequential recommendation models encode ordered interaction histories to capture temporal dependencies. Self-attentive models such as SASRec and BERT4Rec, together with scalable architectures such as HSTU, model temporal behavior for sequential or generative recommendation (Kang & McAuley, 2018; Sun et al., 2019; Zhai et al., 2024). Other work models evolving or future-oriented interests through target-aware evolution, interest distillation, multi-interest retrieval, or future-trend modeling (Zhou et al., 2019; Lu et al., 2021; Wang et al., 2022; Shi et al., 2023; Wu et al., 2025).

User Representation Learning. Multi-interest methods represent users with multiple interest vectors; MIND and ComiRec learn multiple interests, while SINE uses sparse concept representations (Li et al., 2019; Cen et al., 2020; Tan et al., 2021). Disentangled representation methods further factorize user representations into latent concepts, intent factors, or attribute-conditioned components (Wang et al., 2020; Zheng et al., 2021; Wang et al., 2023; Zhang et al., 2023; Li et al., 2024).

LLM-based User Modeling for Recommendation. Recent work uses LLMs to generate or consume natural-language user profiles from reviews, interaction histories, or other personalization context, and to refine these profiles using new evidence or prediction discrepancies (Zhou et al., 2024; Ramos et al., 2024; Gao et al., 2024; Zhang, 2024; Bang & Song, 2025; Chen et al., 2025).

3 Methodology

This section presents the proposed CAST framework (fig. 2). After formulating the prediction problem, we describe a modular, stage-wise trained pipeline consisting of facet-atomic state construction, typed transition prediction, and transition-conditioned candidate ranking.

3.1 Problem Formulation

Let $\mathcal{U}$ and $\mathcal{V}$ denote the sets of users and items, respectively. For a user $u \in \mathcal{U}$ at a temporal anchor t, the observed interaction history is defined as $\mathcal{H}_{u,t} = (v_{u,1}, \ldots, v_{u,t})$. Let $\mathcal{C}_{u,t}$ denote the candidate item set to be ranked at the current step. Unlike conventional sequential recommendation that directly models the probability of the next item $P(v_{u,t+1} \mid \mathcal{H}_{u,t})$ via entangled latent vectors, our task seeks to explicitly model preference evolution through a structured semantic space.

We define a predefined facet set $\mathcal{F} = \{\text{Theme}, \text{Style}, \text{Setting}\}$. We represent the historical preference state as $P^{\text{text}}_{u,t} = \{A_{u,t,f}\}_{f \in \mathcal{F}}$, where $A_{u,t,f}$ denotes the active semantic Atoms in facet f. Consequently, the subsequent future behaviors are formulated as a typed transition $\Delta_{u,t}$ relative to the anchor state $P^{text}_{u,t}$.

Formally, the problem of Compositional State Transition and Recommendation is defined by a two-fold objective: (1) Transition Prediction: Given the history $\mathcal{H}_{u,t}$, predict the future multi-facet transition $\widehat{\Delta}_{u,t}$. (2) Candidate Ranking: Conditioned on the predicted transition $\widehat{\Delta}_{u,t}$, estimate a ranking score $s(u, t, v)$ for each candidate item $v \in \mathcal{C}_{u,t}$.

3.2 Facet-Atomic State Construction

To ground the opaque behavioral history into an interpretable and trackable semantic space, our first step is to distill raw user interactions into structured, facet-specific atomic representations.

Semantic Atom Extraction. We serialize historical events and employ a frozen Large Language Model (LLM) to extract 2–3 concise textual Atoms per facet. To determine the most effective generator, we conducted extensive comparative experiments across multiple LLMs. By comprehensively evaluating their semantic completeness, facet coverage, and strict alignment with historical evidence, we selected Qwen3-30B-Instruct (Yang et al., 2025) as our backbone. Unlike conventional multi-interest methods that rely on uninterpretable latent vectors (Li et al., 2019; Cen et al., 2020), CAST explicitly exposes named semantic facets and preserves the source Atom texts, naturally unlocking fine-grained auditability.

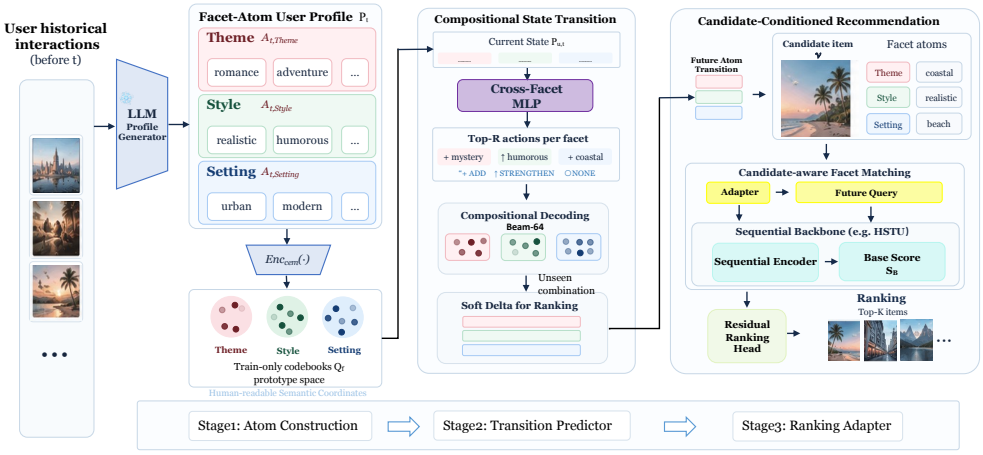


Figure 2: Overview of CAST. Stage 1 converts pre-anchor interactions into a facet-atomic user profile and train-only prototype space. Stage 2 predicts typed facet transitions; the same action distribution is decoded into Beam-64 hypotheses for compositional auditing and summarized as a soft delta for recommendation. Stage 3 maps the soft delta to a candidate-aware future query and adds a gated residual to a frozen sequential backbone to produce the final Top-K ranking.

Train-Only Quantization. To establish a finite coordinate system, let $\mathcal{V}_f^{\text{tr}}$ be the training Atom vocabulary for facet f. A frozen text encoder (Zhang et al., 2025) computes a normalized embedding $\mathbf{z}(a)$ for each Atom a. Using MiniBatchKMeans (Sculley, 2010), we fit a train-only codebook $\mathcal{Q}_f = \{\mathbf{q}_{f,j}\}_{j=1}^{J_f}$ with $J_f = 256$ basis prototypes. At inference, each active atom $a \in A_{u,t,f}$ is quantized to its nearest prototype index:

$$Q_f(a) = \arg\max_{j \in [J_f]} \cos(\mathbf{z}(a), \mathbf{q}_{f,j}) \,. \tag{1}$$

The discrete state $P_{u,t} = \{X_{u,t,f}\}_{f \in \mathcal{F}}$ is the deduplicated collection of these mapped indices $X_{u,t,f} = \{Q_f(a) : a \in A_{u,t,f}\}$. Crucially, validation and test states are quantized using exclusively these frozen training codebooks, serving as a strictly bounded operational semantic interface.

3.3 Compositional Atomic State Transition

Having established the discrete facet-atomic state, we now introduce the mechanism to forecast preference evolution. Instead of regressing opaque latent vectors, we formulate the future state shift as a compositional prediction of typed discrete operations over the semantic codebook.

Event-Level Supervision. For a subsequent future interaction $v_{u,t+h}$, the frozen generator receives the anchor state $P_{u,t}^{\text{text}}$ and derives at most one localized semantic update per facet. We define a discrete, facet-local action space $\Omega_f = \{\text{NONE}\} \cup \mathcal{U}_f^{\text{tr}}$, where $\mathcal{U}_f^{\text{tr}} \subseteq \{\text{ADD}, \text{STRENGTHEN}\} \times [J_f]$ denotes the operation-prototype pairs observed during training. The ground-truth transition $\Delta_{u,t,h}$ is thus represented as the tuple of valid actions $y_{u,t,h,f} \in \Omega_f$ across all facets. Semantically, ADD incorporates a completely novel prototype index into the user state, whereas STRENGTHEN reinforces an already active Atom. Crucially, since the prediction is strictly conditioned on the fixed anchor t without autoregressive state unrolling, the supervision signal isolates a single-step, auditable semantic shift. This explicitly distinguishes CAST from continuous temporal models that predict dense latent trajectories (Lu et al., 2021; Shi et al., 2023).

Transition Learning and Composition. Let $\mathbf{X}_{u,t}$ denote the concatenated dense representations of the current textual state. A cross-facet Multi-Layer Perceptron (MLP), equipped with facet-specific heads parameterized by θ_f, outputs a categorical distribution $p_{\theta_f}(y \mid \mathbf{X}_{u,t}, h)$ over Ω_f.

To mitigate inherent class imbalance, we optimize this module by minimizing a class-weighted categorical cross-entropy loss over the constructed anchor-future training pairs.

During inference, to foster compositional diversity rather than collapsing to a greedy decoding path, we introduce a Joint-Energy Decoder. Specifically, we retain the top $R = 8$ actions per facet and evaluate their Cartesian products via their joint energy (i.e., joint log-likelihood). The top $K_{\text{beam}} = 64$ discrete combinatorial hypotheses constitute the candidate transition set $\widehat{\mathcal{D}}_{u,t}^{\text{beam}}$. This Beam-64 set strictly functions as an evaluation-time hypothesis space for exact combinatorial recovery; it is deliberately isolated from direct ingestion by the differentiable recommendation routing.

Soft Delta for Recommendation. While the Beam-64 hypotheses are used for discrete compositional evaluation, recommendation consumes a differentiable expectation of the same transition distribution. Let $\mathbf{c}_{f,j}$ denote the train-derived representation of prototype j in the item-scoring space. For each facet, we compute

$$\widehat{\boldsymbol{\delta}}_{u,t,h,f} = \sum_{(o,j)\in\mathcal{U}_f^{\text{tr}}} p_{\theta_f}((o,j) \mid \mathbf{X}_{u,t}, h)\, \mathbf{c}_{f,j}, \tag{2}$$

where $o \in \{\text{ADD}, \text{STRENGTHEN}\}$. We concatenate the facet-wise expectations as $\widehat{\boldsymbol{\delta}}_{u,t,h} = [\widehat{\boldsymbol{\delta}}_{u,t,h,f}]_{f\in\mathcal{F}}$ and feed it to the candidate-ranking adapter. Thus, Beam-64 and ranking share the same transition predictor, but ranking uses its soft distribution rather than a hard decoded hypothesis.

We additionally define the facet-wise transition confidence as $\gamma_{u,t,h,f} = 1 - p_{\theta_f}(\text{NONE} \mid \mathbf{X}_{u,t}, h)$. The ranking features are constructed as $\boldsymbol{\phi}_\Delta = \Phi(m_\Delta, \boldsymbol{\gamma}_{u,t,h})$.

3.4 CANDIDATE-CONDITIONED RECOMMENDATION

We now introduce a differentiable adapter for downstream item ranking. The adapter explicitly consumes the predicted soft delta and converts it into a horizon-specific future query.

Delta-Conditioned Candidate Matching. A horizon-aware adapter maps the predicted facet-wise soft delta into a normalized future query:

$$\widehat{\mathbf{q}}_{u,t,h} = \text{Norm}\left(g_\eta\left([\widehat{\boldsymbol{\delta}}_{u,t,h,f}]_{f\in\mathcal{F}} \| \mathbf{e}_h\right)\right), \tag{3}$$

where $\mathbf{e}_h$ is the horizon encoding. For candidate v,

$$m_\Delta(u,t,h,v) = \widehat{\mathbf{q}}_{u,t,h}^\top \mathbf{z}_v. \tag{4}$$

Residual Ranking and Objective. We employ a sequential backbone, HSTU (Zhai et al., 2024), to compute a base collaborative score $s_{\text{base}}(u,t,v)$. Let $s_B(u,t,v)$ be the score produced by a frozen sequential backbone B. We standardize the backbone scores within each candidate set and denote the result by $\widetilde{s}_B(u,t,v)$. Based on $\boldsymbol{\phi}_\Delta = \Phi(m_\Delta, \boldsymbol{\gamma}_{u,t,h})$, the ranking head computes a candidate-specific gate and bounded residual:

$$s_{\text{final}}(u,t,h,v) = \widetilde{s}_B(u,t,v) + \sigma\left(g_\psi(\boldsymbol{\phi}_\Delta)\right) \tanh\left(r_\psi(\boldsymbol{\phi}_\Delta)\right). \tag{5}$$

This adapter is optimized independently via a standard InfoNCE loss over positive interactions v^+ and sampled negatives $\mathcal{C}_{u,t}^-$:

$$\mathcal{L}_{\text{rec}} = -\sum_{(u,t,v^+)\in\mathcal{D}} \log \frac{\exp(s_{\text{final}}(u,t,v^+))}{\sum_{v\in\{v^+\}\cup\mathcal{C}_{u,t}^-} \exp(s_{\text{final}}(u,t,v))}. \tag{6}$$

The transition predictor is optimized first and then frozen while training the ranking adapter. Thus, the two objectives are optimized stage-wise, while ranking remains explicitly conditioned on the predicted transition distribution.

4 EXPERIMENTS

In this section, we conduct experiments to address the following Research Questions (RQs):

- **RQ1** Is the explicit Facet-Atom structure more effective for future-state recovery than other unstructured or holistic profile representations?
- **RQ2** Can CAST successfully recombine train-supported Atoms to extrapolate to novel, unseen complete state configurations?
- **RQ3** Do the predicted semantic states provide reliable candidate-conditioned signals across distinct sequential backbones?

4.1 EXPERIMENTAL SETUP

We begin by briefly outlining the datasets, evaluation metrics and Compositional Protocol.

Datasets. We evaluate the audited OpenOneRec-RecIF dataset (Zhou et al., 2025), which comprises approximately 96 million interactions from over $160{,}000$ users. For each selected user u, interactions are chronologically ordered to construct temporal anchors at the $q25$, $q50$, and $q75$ history quantiles. The prefix $\mathcal{H}_{u,\leq t}$ yields the current profile $P_{u,t}$, while the next six interactions provide future transition supervision. This yields $30{,}000$ anchor records from $10{,}000$ users. The recommendation view uses a $24\text{K}/3\text{K}/3\text{K}$ (train/validation/test) user-disjoint split. Because the semantic profiles are initially generated via LLMs, evaluating solely on machine-generated targets risks systematic bias. To ensure a rigorously fair benchmark, the $3{,}000$ test set instances underwent strict human evaluation and quality filtering to establish a high-fidelity ground truth.

Evaluation Metrics. To ensure a rigorous assessment, we evaluate the pipeline across three stages. For **semantic recovery**, we report *Atom F1* to measure partial semantic overlap, *Beam F1* to assess the aggregated quality of multi-path generation, and *Exact@Beam* for stringent full-state matching. For **strict composition**, we emphasize *Exact@Beam* alongside the *True-Shuffled Gap* to isolate genuinely learned structural extrapolation from random combinatorial chance. Finally, for **downstream recommendation**, we employ standard ranking metrics *NDCG*.

Compositional Protocol. To assess genuine compositional extrapolation, we enforce a strict seen-parts/unseen-whole split. Let $\mathcal{V}_{\text{tr}}$ denote the train-supported Atom vocabulary and $\mathcal{C}_{\text{tr}}$ the set of complete training configurations. A test transition $\widehat{c}$ is evaluated under the strict protocol if and only if its constituent atoms are fully supported ($\text{Atoms}(c^{\text{test}}_{u,t,h}) \subseteq \mathcal{V}_{\text{tr}}$), and the complete joint configuration is absent from training ($c^{\text{test}}_{u,t,h} \notin \mathcal{C}_{\text{tr}}$).

4.2 RQ1: FACET-ATOM REPRESENTATION

To determine whether explicit structural factorization enhances future-state recovery, we design a controlled evaluation that targets the inductive bias of our state representation. We compare CAST with two rigorously matched baselines: an unstructured **Holistic Profile** (which encodes the user's historical summary as flat natural) and a **Shuffled Facets** control (which randomly permutes facet-to-user assignments during training).

We assess the models' ability to anticipate future states using partial semantic overlap (*Atom F1*) and exact full-state matching (*Exact Hit*). Because user behaviors are inherently multimodal and uncertain, we employ beam search to generate multiple plausible future trajectories, and measure the aggregated multi-path quality via *Beam F1*.

Table 1: Future-state recovery with $J_f = 256$ and Top-64 Beam.

Input Representation	Atom F1	Beam F1	Exact Hit
Holistic Profile	.0630	.1822	.0089
Facet-Atom (CAST)	**.1214**	**.2736**	**.0238**
Shuffled Facets	.0001	.0038	.0000

As shown in Table 1, the explicit Facet-Atom structure offers a clear advantage. Compared with holistic control, CAST improves Atom F1 by 92.6% and Exact Hit by 167.8%. This indicates that unstructured text struggles to capture the complex joint distribution of user traits, whereas explicit factorization enables precise semantic alignment. Furthermore, the near-zero performance of the Shuffled control confirms that CAST extracts genuine behavioral signals rather than exploiting trivial statistical priors.

4.3 RQ2: Compositional Extrapolation

To assess whether CAST can truly extrapolate to novel scenarios, we challenge it to generate complete state combinations that never appeared in the training data, using only familiar individual atoms. To give the model ample opportunity to discover these rare combinations, we employ an expansive search strategy: exploring $32{,}000$ candidate trajectories via beam search to produce the top-32 most confident predictions.

Table 2: Exact recovery of complete unseen combinations.

Training Seed	True Exact	Shuffled Exact	Δ
Seed 1	**21.48%**	6.25%	$+15.23$
Seed 2	**14.84%**	5.08%	$+9.76$
Seed 3	**17.19%**	6.25%	$+10.94$
Mean	**17.84%**	5.86%	$+11.98$

As shown in Table 2, CAST successfully reconstructs these unseen combinations across three independent random seeds, achieving an average exact match rate of 17.84%. To verify that this success stems from learned semantic rules rather than random combinatorial luck, we compare it with a shuffled control that randomly reassembles valid atoms without modeling user history. The true model significantly outperforms this chance baseline (5.86%).

4.4 RQ3: Plug-in Utility

To investigate its downstream utility, we transition from internal semantic generation to external recommendation, evaluating CAST as a candidate re-ranking plug-in across diverse sequential architectures and then analyzing its additional computational cost.

Plug-and-Play Across Architectures. To rigorously demonstrate CAST's plug-and-play capability, we first evaluate it on a matched protocol across four fundamentally distinct sequential architectures: BERT4Rec(Sun et al., 2019), SASRecKang & McAuley (2018), FMLP-Rec(Zhou et al., 2022) and Mamba4Rec(Liu et al., 2024).

As shown in Table 3, under the ordinary evaluation protocol, the current-profile semantic plug-in yields positive and statistically significant NDCG@10 improvements across all four matched backbones, based on user-level paired bootstrap tests. Under the strict Out-Of-Distribution (OOD) protocol, the plug-in also produces positive average gains for all four architectures. The strict-OOD improvements are statistically significant for BERT4Rec and FMLP-Rec, demonstrating that the semantic residual can successfully transfer compositional signals even under severe distributional shifts.

Future-only Residual Verification. To verify that the ranking gains can also be obtained from the predicted state transition, we additionally evaluate CAST with a matched future-only residual interface under the strict-OOD protocol. With the backbone frozen, the predicted transition $\widehat{\Delta}_{t,h}$ is converted into a candidate-aware residual ranking score. Across nine runs for each architecture, the mean NDCG@10 improvements are $+0.014668$ for BERT4Rec, $+0.012827$ for SASRec, $+0.016232$ for SASRecF, $+0.016268$ for FMLP-Rec, and $+0.014932$ for Mamba4Rec. These results provide direct evidence that the predicted transition signal can be used to enhance candidate ranking beyond the current-profile semantic fusion.

Validation on a HSTU. To ensure these gains are not limited to standard baselines, Table 3 also independently validates CAST on HSTU, a strong modern sequential transducer. Evaluated under its native optimized configuration, CAST delivers an $+8.78\%$ relative gain under the ordinary protocol and consistent positive gains under the strict OOD protocol, thereby confirming its utility on state-of-the-art representations.

Table 3: Cross-backbone effectiveness of CAST as a delta-conditioned ranking residual.

Backbone	Architecture Type	Ordinary ΔNDCG@10	Strict OOD ΔNDCG@10
	Matched Protocol Evaluation		
BERT4Rec	Bidirectional Transformer	$+0.0029$	$+0.0205$
SASRec	Autoregressive Transformer	$+0.0022$	$+0.0086$
FMLP-Rec	Frequency-domain MLP	$+0.0027$	$+0.0111$
Mamba4Rec	Selective State-Space	$+0.0018$	$+0.0130$
	Independent HSTU Evaluation		
HSTU	Modern Transducer	$+8.78\%$	$+0.0077$

Efficiency and Cost-Benefit. Beyond its robust downstream utility, CAST's design is highly practical as a re-ranking plug-in. Heavy text encoding is performed offline, while the online component operates on cached semantic representations. The structured branch introduces approximately 408K trainable parameters, trains in under 8 minutes on a single GPU, and peaks at approximately 316 MiB of allocated memory in the reported HSTU configuration. These measurements characterize the additional branch under cached features and do not include end-to-end LLM encoding or serving-latency costs.

4.5 Case Study: Auditable Interest Evolution

Figure 3 visualizes the auditable interest evolution of four purposefully selected users. These illustrative examples showcase how the Facet–Atom representation elegantly disentangles complex preference shifts into inspectable trajectories, explicitly mapped by audit-level PERSIST, ADD, STRENGTHEN, and REPLACE operators across three historical anchors ($q25$, $q50$, and $q75$).

User 56887 maintains a persistent core interest in Apple devices. While early states ($q25$) focus on unboxing and phone customization, the profile organically expands by $q50$ and $q75$. Although the device-centric core PERSISTs, it broadens into a wider Apple ecosystem and cozy domestic settings, illustrating how stable themes evolve through contextual shifts.

User 69883 exhibits a profound structural reorganization. The profile shifts from everyday food and fashion ($q25$) to intimate relationships and study routines ($q50$), ultimately synthesizing emotional growth and campus life at $q75$. This exemplifies how semantically related Theme, Style, and Setting atoms can be dynamically reorganized to construct an entirely novel joint profile.

User 73914 retains a steady interest in competitive gaming but develops a parallel intent for physical sports. Outdoor sports are ADDed at $q50$ and refine into competitive-football by $q75$. Concurrently, action highlights STRENGTHEN into competition-focused presentations. This represents parallel branching rather than wholesale replacement: a new interest emerges seamlessly alongside the persistent core.

User 100293 illustrates a highly continuous trajectory where cartoon-game interactions and bright visuals remain stable. Meanwhile, guided instruction first appears at $q50$ and progressively STRENGTHENs into detailed follow-along processes by $q75$. This encapsulates how our representation distinguishes stable core preferences from supporting interests that gradually mature over time.

Collectively, these distinct typologies highlight the diagnostic power of our framework. By translating opaque history embeddings into explicit transition operators, CAST renders the "black-box" of user evolution highly auditable, perfectly complementing our quantitative results.

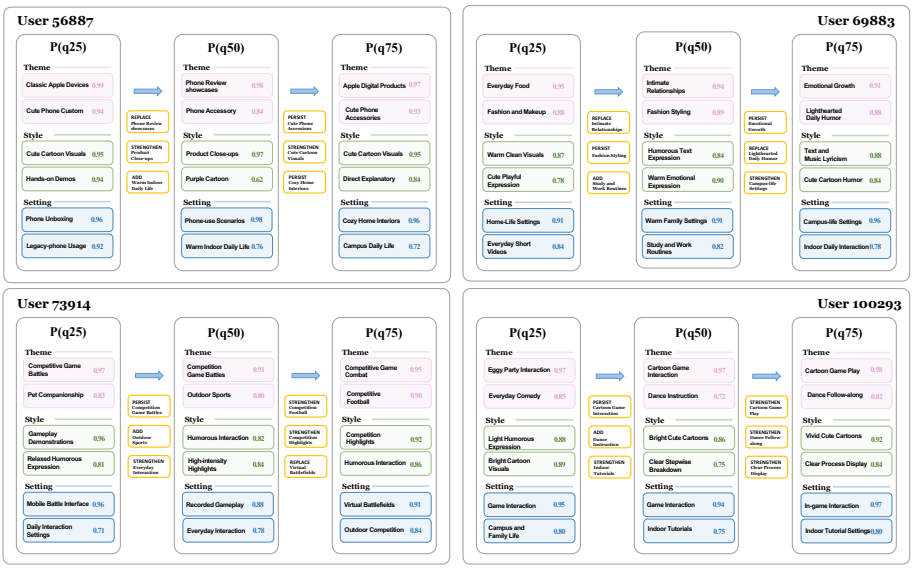


Figure 3: Observed Facet–Atom evolution for four purposefully selected illustrative users.

4.6 Summary of Findings

The empirical evidence supports three bounded conclusions corresponding to our research questions: (1) **RQ1:** The explicit Facet-Atom factorization acts as a powerful structural inductive bias, making it significantly more effective for future semantic recovery than unstructured holistic profiles.

(2) **RQ2:** CAST achieves statistically detectable partial compositional extrapolation, effectively recombining known Atoms to correctly recover unseen state configurations.

(3) **RQ3:** CAST provides useful candidate-level ranking signals across sequential architectures.

5 Limitations & Future Work

Despite successfully extrapolating to unobserved user states, converting these generative predictions into robust full-catalog Top-K recommendations remains an open problem. Bridging this gap between abstract semantic generation and exact item retrieval is a critical direction for future work. Beyond standard ranking, the explicit tracking of structured, evolving profiles provides a natural foundation for other complex recommendation tasks. For instance, these dynamic trajectories could be directly leveraged for explainable reasoning, long-term intent planning, or cold-start adaptation.

6 Conclusion

In this work, we introduce CAST, a novel user modeling approach designed to address a fundamental aspect of recommendation: user states are neither static nor homogeneous, but rather dynamic mixtures of persistent and transient intents. To capture how users continually combine familiar preferences into new configurations, CAST moves beyond traditional item-ID memorization by explicitly tracking the dynamic transitions of structured user profiles. By modeling how decoupled semantic facets evolve over time, this dynamic recombination of distinct elements inherently enables the model to extrapolate, effectively predicting entirely novel combinations of future intents that are absent from the training data. Extensive experiments on multiple sequential backbones demonstrate that CAST significantly enhances existing architectures, achieving exact-match recovery rates of up to 21.48% on unseen future states, along with robust plug-and-play downstream ranking improvements.

AI USE STATEMENT

Generative AI tools were used to generate structured user-state transition annotations and to assist with code implementation and manuscript polishing. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the content and claims of this work.

REFERENCES

Seunghwan Bang and Hwanjun Song. LLM-based user profile management for recommender system. *arXiv preprint arXiv:2502.14541*, 2025. URL https://arxiv.org/abs/2502.14541.

Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2942–2951, 2020. URL https://arxiv.org/abs/2005.09347.

Aili Chen, Chengyu Du, Jiangjie Chen, Jinghan Xu, Yikai Zhang, Siyu Yuan, Zulong Chen, Liangyue Li, and Yanghua Xiao. DEEPER insight into your user: Directed persona refinement for dynamic persona modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 24157–24180. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1177.

Zhaolin Gao, Joyce Zhou, Yijia Dai, and Thorsten Joachims. End-to-end training for recommendation with language-based user profiles. *arXiv preprint arXiv:2410.18870*, 2024. URL https://arxiv.org/abs/2410.18870.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018. URL https://arxiv.org/abs/1808.09781.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. Multi-interest network with dynamic routing for recommendation at tmall. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2615–2623, 2019. URL https://arxiv.org/abs/1904.08030.

Zhenyang Li, Fan Liu, Yinwei Wei, Zhiyong Cheng, Liqiang Nie, and Mohan Kankanhalli. Attribute-driven disentangled representation learning for multimodal recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9660–9669, 2024. doi: 10.1145/3664647.3681148. URL https://arxiv.org/abs/2312.14433.

Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. Mamba4Rec: Towards efficient sequential recommendation with selective state space models. *arXiv preprint arXiv:2403.03900*, 2024. URL https://arxiv.org/abs/2403.03900.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Yujie Lu, Shengyu Zhang, Yingxuan Huang, Luyao Wang, Xinyao Yu, Zhou Zhao, and Fei Wu. Future-aware diverse trends framework for recommendation. In *Proceedings of the Web Conference 2021*, pp. 2992–3001, 2021. URL https://arxiv.org/abs/2011.00422.

Jelena Luketina, Jack Lanchantin, Sainbayar Sukhbaatar, and Arthur Szlam. Compositional interfaces for compositional generalization. In *Proceedings of the 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 692–709. PMLR, 2025. URL https://proceedings.mlr.press/v274/luketina25a.html.

Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. Transparent and scrutable recommendations using natural language user profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13971–13984. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.753. URL https://aclanthology.org/2024.acl-long.753/.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 1177–1178, 2010.

Hui Shi, Yupeng Gu, Yitong Zhou, Bo Zhao, Sicun Gao, and Jishen Zhao. Everyone's preference changes differently: A weighted multi-interest model for retrieval. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31228–31242. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23b.html.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450, 2019. URL https://arxiv.org/abs/1904.06690.

Qiaoyu Tan, Jianwei Zhang, Jiangchao Yao, Ninghao Liu, Jingren Zhou, Hongxia Yang, and Xia Hu. Sparse-interest network for sequential recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 598–606, 2021. URL https://arxiv.org/abs/2102.09267.

Chenyang Wang, Zhefan Wang, Yankai Liu, Yang Ge, Weizhi Ma, Min Zhang, Yiqun Liu, Junlan Feng, Chao Deng, and Shaoping Ma. Target interest distillation for multi-interest recommendation. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 2007–2016, 2022. doi: 10.1145/3511808.3557464.

Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1001–1010, 2020. doi: 10.1145/3397271.3401137. URL https://arxiv.org/abs/2007.01764.

Xin Wang, Hong Chen, Yuwei Zhou, Jianxin Ma, and Wenwu Zhu. Disentangled representation learning for recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):408–424, 2023. doi: 10.1109/TPAMI.2022.3153112. URL https://doi.org/10.1109/TPAMI.2022.3153112.

Zhibo Wu, Yunfan Wu, Quan Liu, Lin Jiang, Ping Yang, and Yao Hu. GemiRec: Interest quantization and generation for multi-interest recommendation. *arXiv preprint arXiv:2510.14626*, 2025. URL https://arxiv.org/abs/2510.14626.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58484–58509. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhai24a.html.

Jiarui Zhang. Guided profile generation improves personalization with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4005–4016. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.231. URL https://aclanthology.org/2024.findings-emnlp.231/.

Xiaoying Zhang, Hongning Wang, and Hang Li. Disentangled representation for diversified recommendations. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 490–498, 2023. doi: 10.1145/3539597.3570389. URL https://arxiv.org/abs/2301.05492.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Depeng Jin, and Yong Li. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*, pp. 2980–2991, 2021. doi: 10.1145/3442381.3449788. URL https://arxiv.org/abs/2006.11011.

Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5941–5948, 2019. doi: 10.1609/aaai.v33i01.33015941.

Guorui Zhou, Honghui Bao, Jiaming Huang, Jiaxin Deng, Jinghao Zhang, Junda She, Kuo Cai, Lejian Ren, Lu Ren, Qiang Luo, et al. OpenOneRec technical report. *arXiv preprint arXiv:2512.24762*, 2025. URL https://arxiv.org/abs/2512.24762.

Joyce Zhou, Yijia Dai, and Thorsten Joachims. Language-based user profiles for recommendation. *arXiv preprint arXiv:2402.15623*, 2024. URL https://arxiv.org/abs/2402.15623.

Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. Filter-enhanced MLP is all you need for sequential recommendation. In *Proceedings of the ACM Web Conference 2022*, pp. 2388–2399, 2022. doi: 10.1145/3485447.3512111. URL https://arxiv.org/abs/2202.13556.