UST: A Unified Semantic Tokenization Framework for Multimodal Sequential Recommendation

Anonymous ACL submission

Abstract

Multimodal sequential recommendation depends on item representations that combine interaction histories with textual and visual evidence. Most existing methods merge these signals in continuous embedding spaces, leaving item-level semantic structure implicit and difficult to reuse, inspect, or transfer. We introduce **UST**, a unified semantic tokenization framework that maps heterogeneous item evidence to discrete semantic tokens and uses them as a plug-in representation layer for standard sequential backbones. This tokenized layer organizes multimodal evidence before sequence modeling, creating a structured interface between item semantics and evolving user preference. Experiments on Amazon Reviews'23 show clear gains in selected backbone–dataset settings, with the largest improvements obtained by `UST_SASRec` on `Video_Games` and `Musical_Instruments`; the effect, however, remains dependent on both domain and backbone. Ablation and sensitivity analyses indicate that the gains arise mainly from semantic tokenization rather than early multimodal fusion, and that stable performance requires appropriate codebook capacity and token assignment dynamics. Cross-domain diagnostics further show that large domain mismatch can still induce negative transfer, underscoring both the utility and the limits of tokenized semantic representations for multimodal recommendation. The implementation is available at `https://github.com/sy-cg/UST`.

1 Introduction

Modern recommender systems rarely depend on interaction histories alone. In e-commerce, content platforms, short-video services, music, and gaming, user behavior sequences record only part of how preferences evolve. Item metadata, textual descriptions, and visual content provide complementary semantic context, especially under cold-start, sparse-interaction, and preference-drift conditions (Raza et al., 2026). The main question is therefore not whether to add more modalities, but how to organize heterogeneous item evidence so that sequential models can use it effectively.

Sequential recommendation models preference dynamics from temporally ordered interaction histories. GRU4Rec and SASRec are widely used backbones for next-item prediction (Hidasi et al., 2015; Kang and McAuley, 2018), and later work has strengthened sequence modeling through scalable Transformers, contrastive augmentation, multi-behavior learning, and diffusion-based objectives (Li et al., 2023a; Zivic et al., 2024; Zhou et al., 2023a; Elsayed et al., 2024; Li et al., 2023b; Wang and Jin, 2026). These advances improve sequence encoders, but leave a separate representation problem unresolved: how should item semantics be formed before sequence modeling? In most multimodal sequential recommenders, textual and visual features are projected into continuous vectors and fused with item embeddings as auxiliary signals. This gives the model more information, but the semantic structure of the fused representation remains largely implicit.

Multimodal recommendation has introduced graph-based fusion, self-supervised learning, modality alignment, foundation-model adaptation, and missing-modality robustness to use item-side evidence (Zhou et al., 2023b; Yi and Ounis, 2024; Li et al., 2024; Fu et al., 2024; Chen et al., 2026). These methods are effective, yet they usually operate in dense embedding spaces. Dense representations are flexible, but offer limited access to explicit semantic units: their dimensions are difficult to interpret, their factors are not naturally reusable across modalities, datasets, or domains, and their modality contributions are hard to diagnose. This limitation becomes more consequential in cross-domain recommendation, where source-domain signals may provide useful evidence, become irrelevant, or induce negative transfer depending on

the inter-domain semantic gap (Cao et al., 2022; Xu et al., 2025; Zhu et al., 2023; Zhang et al., 2024; Liu et al., 2024a; Ning et al., 2025; Wang et al., 2025).

Semantic tokenization offers a structured alternative. Discrete representation learning, vector quantization, semantic IDs, and item tokenization map continuous item representations into reusable semantic units (Van Den Oord et al., 2017; Hou et al., 2023; Liu et al., 2024b; Zhao et al., 2025). For multimodal sequential recommendation, these tokens can serve as an intermediate layer between heterogeneous item evidence and sequence encoders. Rather than passing unconstrained dense features directly to the backbone, tokenization compresses multimodal evidence into learnable item-level units that are compact, reusable, and easier to analyze. Existing tokenization-based recommenders address this setting only partially. Many rely mainly on text or a single content source, rather than jointly organizing collaborative, textual, and visual evidence. Others use tokenization as a pretraining mechanism rather than as a representation layer embedded in sequential recommendation. Token usage, modality contribution, hyperparameter sensitivity, and failure behavior under domain mismatch are also rarely examined together.

To address these gaps, we propose UST, a unified semantic tokenization framework for multimodal sequential recommendation, as illustrated in Figure 1. UST converts collaborative, textual, and visual item evidence into discrete semantic tokens before sequential modeling. The tokenizer uses codebook-based discrete assignment and optional semantic decomposition, functioning as a plug-in representation layer for backbones such as GRU4Rec and SASRec. This design lets us examine whether semantic tokenization improves recommendation, how its effect depends on the backbone and domain, and where tokenized representations begin to fail. Our main contributions are summarized as follows:

- We propose **UST**, a unified semantic tokenization framework that organizes collaborative, textual, and visual item evidence into discrete semantic tokens for multimodal sequential recommendation.
- We instantiate UST as a plug-in representation layer for GRU4Rec and SASRec. Experiments show clear gains in selected backbone–dataset settings, especially with SASRec on `Video_Games` and `Musical_Instruments`, while also revealing domain- and backbone-dependent behavior.
- We analyze UST through ablation, sensitivity, and cross-domain diagnostics. The results show that the main gains stem from semantic tokenization rather than early fusion, that stable performance depends on codebook capacity and token assignment dynamics, and that large source–target mismatch can still induce negative transfer.

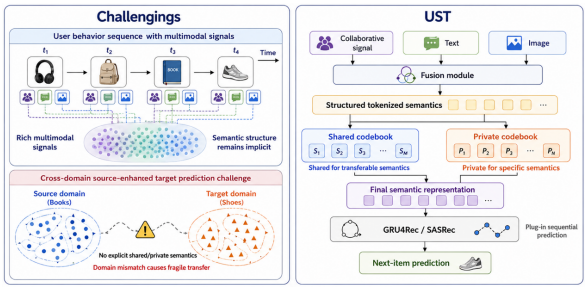

Figure 1: Motivation of UST.

2 Related Work

2.1 Sequential and Multimodal Recommendation

Sequential recommendation models user preference from temporally ordered interaction histories. Recurrent models such as GRU4Rec capture short-term transitions through hidden-state updates (Hidasi et al., 2015), whereas self-attention models such as SASRec capture long-range item dependencies in behavior sequences (Kang and McAuley, 2018). Later work has improved sequence encoders with scalable Transformer variants, contrastive learning, multi-behavior modeling, and generative recommendation objectives (Li et al., 2023a; Zivic et al., 2024; Zhou et al., 2023a; Elsayed et al., 2024; Li et al., 2023b; Wang and Jin, 2026). These methods mainly improve collaborative sequence modeling. UST is complementary to this line of work: it keeps standard sequence backbones and changes the item representations given to them.

Multimodal recommendation uses item-side content, especially text and images, to enrich collaborative signals. VBPR extends collaborative filtering with visual features (He and McAuley, 2016), while recent methods such as BM3 and UGT use self-supervised multimodal learning and

graph-based fusion (Zhou et al., 2023b; Yi and Ounis, 2024). Multimodal sequential methods further transfer or adapt multimodal representations for sequence modeling (Lei et al., 2021; Fu et al., 2024). Most existing methods encode heterogeneous signals as continuous embeddings and improve recommendation through fusion, propagation, or alignment. This strategy is effective, but it gives limited control over which semantic factors are shared and which remain modality- or domain-specific. UST instead organizes collaborative, textual, and visual signals into shared-private discrete semantic codes before sequential prediction.

2.2 Cross-Domain and Transferable Recommendation

Cross-domain recommendation transfers information from a source domain to improve recommendation in a target domain (Chen et al., 2024). Explicit cross-domain sequential models such as C2DSR and MGCL model within-domain transitions and cross-domain relations through graph and attention mechanisms (Cao et al., 2022; Xu et al., 2025). These methods work well when reliable cross-domain behavioral bridges exist, but often assume overlapping users or explicit cross-domain interaction structures.

Transferable semantic recommenders follow another route. UniSRec learns transferable sequence representations from item textual semantics (Hou et al., 2022), while VQ-Rec introduces vector-quantized item representations for transferable sequential recommendation (Hou et al., 2023). Disentangled cross-domain methods further separate domain-shared and domain-specific factors to reduce negative transfer (Zhu et al., 2023; Zhang et al., 2024; Liu et al., 2024a; Zhao et al., 2025). These studies are close to our motivation, but most rely on continuous latent disentanglement or source-domain pretraining followed by target-domain fine-tuning. We instead study source-enhanced target prediction, where source-domain signals provide auxiliary evidence but are not assumed to be uniformly beneficial. This setting allows us to examine when direct semantic sharing helps and when it introduces negative transfer.

2.3 Discrete Semantic Representation

Discrete representation learning organizes latent information into reusable token-like units. VQ-VAE learns discrete codes through vector quantization and codebook lookup (Van Den Oord et al., 2017), while Gumbel-Softmax provides a differentiable relaxation for categorical assignment learning (Jang et al., 2017). In recommendation, VQ-Rec maps item text into discrete codes before constructing transferable item representations (Hou et al., 2023); discrete semantic tokenization has been applied to CTR prediction (Liu et al., 2024b); and quantized disentangled models suggest that discrete codes can help identify domain-shared and domain-specific factors in cross-domain recommendation (Zhao et al., 2025).

Despite this progress, discrete semantic learning remains difficult. Vector-quantized models can suffer from gradient sparsity, embedding-code mismatch, and unstable codebook usage (Huh et al., 2023). In UST, these issues interact with multimodal fusion, shared-private decomposition, and cross-domain alignment. We therefore study how collaborative, textual, and visual signals can be organized into shared-private semantic codes, and how such tokenized representations behave in both in-domain recommendation and source-enhanced target prediction.

3 Method

3.1 Problem Definition

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items. Each item $i \in \mathcal{I}$ is associated with three sources of information: an ID-based collaborative embedding $\mathbf{e}_i^{id} \in \mathbb{R}^d$, a textual feature $\mathbf{t}_i \in \mathbb{R}^{d_t}$, and a visual feature $\mathbf{v}_i \in \mathbb{R}^{d_v}$. For a user $u \in \mathcal{U}$, the observed behavior is a chronological sequence

$$\mathbf{s}_u = (i_1, i_2, \ldots, i_{T_u}), \tag{1}$$

where i_t denotes the item interacted with at step t.

Given the prefix $\mathbf{s}_{u,<t} = (i_1, \ldots, i_{t-1})$, sequential recommendation ranks the ground-truth next item i_t against a candidate set. We consider both in-domain and cross-domain settings. In the in-domain setting, all interactions are drawn from a single domain. In the cross-domain setting, a sequence may contain both source- and target-domain items. The domain of item i is denoted by $\delta(i) \in \{0, 1\}$, where 0 and 1 indicate the source and target domains, respectively. We focus on *source-enhanced target prediction*: source-domain interactions provide auxiliary behavioral and semantic evidence, while the prediction target remains in the target domain.

3.2 Overview of UST

We propose **UST**, a **u**nified **s**emantic **t**okenization framework that converts heterogeneous item evidence into discrete semantic tokens and uses them as a plug-in representation layer for standard sequential backbones. UST does not replace the sequence encoder; it is placed before backbones such as GRU4Rec and SASRec (Hidasi et al., 2015; Kang and McAuley, 2018). As shown in Figure 2, UST first fuses collaborative, textual, and visual item evidence into a continuous multimodal representation. It then maps this representation to shared and private discrete semantic codes, and passes the resulting item representation to a sequential backbone for next-item prediction.

The shared branch captures common semantic factors, whereas the private branch preserves complementary domain/item-specific information. This design follows vector-quantized representation learning and recent discrete semantic recommendation methods, where compact code-based representations serve as intermediate semantic abstractions (Van Den Oord et al., 2017; Hou et al., 2023; Liu et al., 2024b).

3.3 Multimodal Semantic Fusion

UST first augments ID-based item representations with pretrained textual and visual features. Let c_i denote the textual metadata of item i, and let o_i denote its associated image. We compute

$$\mathbf{t}_i = \mathrm{ST}(c_i), \qquad \mathbf{v}_i = \mathrm{CLIP}_{img}(o_i), \tag{2}$$

where $\mathrm{ST}(\cdot)$ is a sentence-transformer encoder (Reimers and Gurevych, 2019), and $\mathrm{CLIP}_{img}(\cdot)$ is the CLIP image encoder (Radford et al., 2021). These modality features are extracted offline and kept fixed during recommendation training.

The three item signals are projected into a common latent space:

$$\tilde{\mathbf{e}}_i^{id} = \mathrm{LayerNorm}(\mathbf{e}_i^{id}), \tag{3}$$

$$\tilde{\mathbf{t}}_i = \phi_t(\mathbf{t}_i), \tag{4}$$

$$\tilde{\mathbf{v}}_i = \phi_v(\mathbf{v}_i), \tag{5}$$

where $\phi_t(\cdot)$ and $\phi_v(\cdot)$ are modality-specific projection networks. A gated fusion module assigns item-specific weights to the three modalities:

$$\boldsymbol{\alpha}_i = \mathrm{softmax}\Big(W_{g,2}\psi_g([\tilde{\mathbf{e}}_i^{id} \| \tilde{\mathbf{t}}_i \| \tilde{\mathbf{v}}_i])\Big), \tag{6}$$

where $\|$ denotes concatenation and $\psi_g(\cdot)$ is the hidden transformation of the gate network. The fused representation is

$$\mathbf{x}_i = \alpha_i^{(1)}\tilde{\mathbf{e}}_i^{id} + \alpha_i^{(2)}\tilde{\mathbf{t}}_i + \alpha_i^{(3)}\tilde{\mathbf{v}}_i. \tag{7}$$

This formulation allows the model to adjust the relative contribution of collaborative, textual, and visual evidence for each item.

3.4 Shared-Private Tokenized Semantics

The fused representation $\mathbf{x}_i$ is decomposed into shared and private discrete branches. We first compute the corresponding token logits:

$$\boldsymbol{\ell}_i^s = W_s\mathbf{x}_i + \mathbf{b}_s, \qquad \boldsymbol{\ell}_i^p = W_p\mathbf{x}_i + \mathbf{b}_p, \tag{8}$$

and obtain token distributions

$$\mathbf{p}_i^s = \mathrm{softmax}(\boldsymbol{\ell}_i^s), \qquad \mathbf{p}_i^p = \mathrm{softmax}(\boldsymbol{\ell}_i^p). \tag{9}$$

During training, discrete assignments are sampled with the straight-through Gumbel-Softmax estimator:

$$\mathbf{q}_i^s = \text{ST-GumbelSoftmax}(\boldsymbol{\ell}_i^s; \tau), \tag{10}$$

$$\mathbf{q}_i^p = \text{ST-GumbelSoftmax}(\boldsymbol{\ell}_i^p; \tau). \tag{11}$$

At inference time, we use deterministic one-hot assignments by selecting the maximum-probability token.

Let $\mathbf{C}^s, \mathbf{C}^p \in \mathbb{R}^{V\times d}$ denote the shared and private codebooks, where V is the vocabulary size. Codebook lookup gives

$$\mathbf{e}_i^s = \mathbf{q}_i^s\mathbf{C}^s, \qquad \mathbf{e}_i^p = \mathbf{q}_i^p\mathbf{C}^p. \tag{12}$$

The two branches are combined as

$$\mathbf{e}_i^d = \mathbf{e}_i^s + \mathbf{e}_i^p, \tag{13}$$

Discrete codes impose semantic structure but may discard fine-grained information in $\mathbf{x}_i$. We therefore introduce a residual blending gate:

$$g_i = \sigma\Big(\mathrm{MLP}([\mathbf{x}_i \| \mathbf{e}_i^d])\Big), \tag{14}$$

$$\mathbf{z}_i = g_i\mathbf{e}_i^d + (1-g_i)\mathbf{x}_i. \tag{15}$$

The final item representation $\mathbf{z}_i$ combines continuous multimodal evidence with tokenized semantic structure.

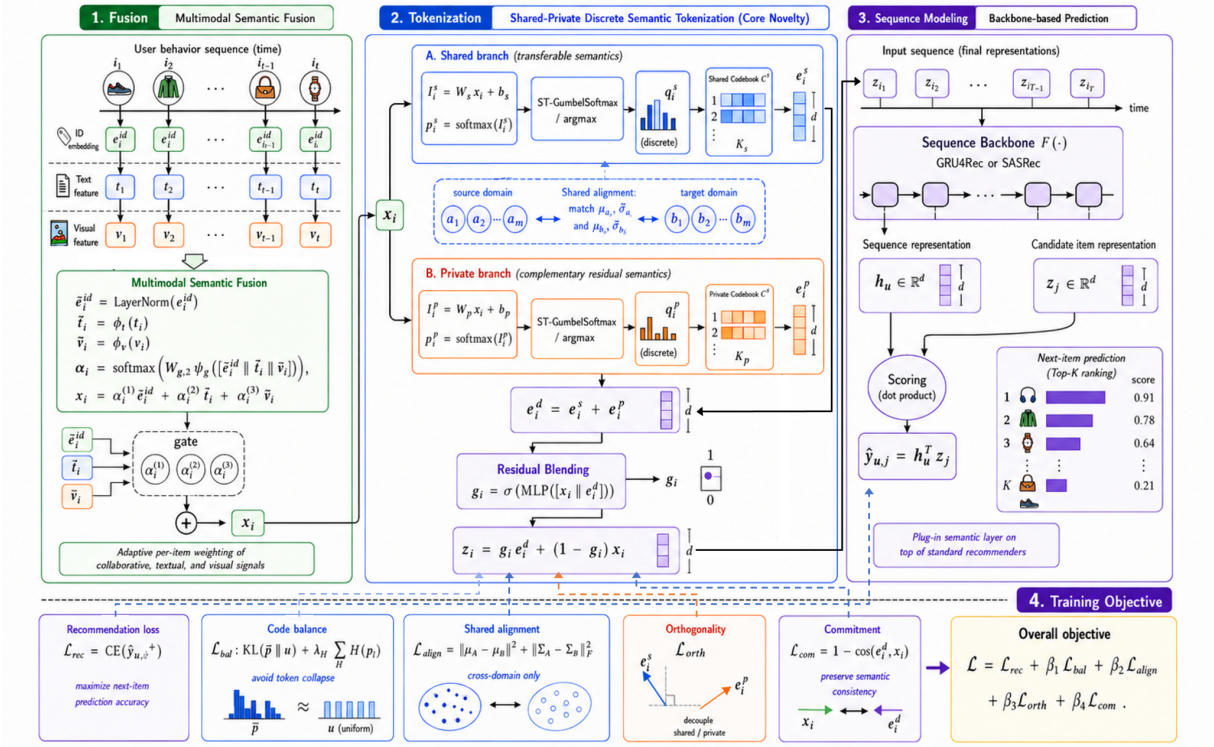

Figure 2: Overview of the proposed UST framework.

3.5 Sequential Prediction

For a user sequence $\mathbf{s}_u = (i_1, \ldots, i_{T_u})$, UST converts each item into $\mathbf{z}_{i_t}$ and passes the resulting sequence to a backbone encoder $\mathcal{F}$:

$$\mathbf{h}_u = \mathcal{F}(\mathbf{z}_{i_1}, \mathbf{z}_{i_2}, \ldots, \mathbf{z}_{i_{T_u}}). \tag{16}$$

For a candidate item j, the prediction score is

$$\hat{y}_{u,j} = \mathbf{h}_u^\top \mathbf{z}_j. \tag{17}$$

Training uses the standard next-item classification objective:

$$\mathcal{L}_{rec} = \mathrm{CE}(\hat{\mathbf{y}}_u, j^+), \tag{18}$$

where j^+ is the ground-truth next item and $\hat{\mathbf{y}}_u$ contains scores over the candidate set.

3.6 Training Objective

The complete objective is

$$\mathcal{L} = \mathcal{L}_{rec} + \beta_1 \mathcal{L}_{bal} + \beta_2 \mathcal{L}_{align} + \beta_3 \mathcal{L}_{orth} + \beta_4 \mathcal{L}_{com}. \tag{19}$$

Here, $\mathcal{L}_{bal}$ promotes balanced code usage and reduces code collapse; $\mathcal{L}_{align}$ matches source and target shared-code distributions in the cross-domain setting; $\mathcal{L}_{orth}$ separates shared and private embeddings; and $\mathcal{L}_{com}$ aligns the discrete representation with the continuous fused representation. For in-domain recommendation, $\mathcal{L}_{align}$ is removed by setting $\beta_2 = 0$. Detailed auxiliary objectives and implementation details are provided in Appendix A.

4 Experiments

We evaluate UST through four research questions:

- **RQ1**: How does UST perform in in-domain sequential recommendation?
- **RQ2**: How does UST behave under cross-domain source-enhanced target prediction?
- **RQ3**: Which components contribute to the effectiveness of UST?
- **RQ4**: How sensitive is UST to key hyperparameters?

4.1 Experimental Setup

4.1.1 Datasets

We use Amazon Reviews'23, which provides user–item interactions, textual metadata, and product images. For in-domain evaluation, we use three categories: `Video_Games` (Video), `Musical_Instruments` (Musical), and `Baby_Products` (Baby). For cross-domain evaluation, we construct `Cross_Sport_Beauty`

from Sports_and_Outdoors and All_Beauty, using the former as the source domain and the latter as the target domain.

For computational feasibility, interactions in large categories are restricted to those after 2018. We then apply 5-core filtering and sort each user's interactions chronologically. Product descriptions are encoded by a sentence-transformer encoder (Reimers and Gurevych, 2019), producing 768-dimensional textual embeddings. Product images are encoded by the CLIP image encoder (Radford et al., 2021), producing 512-dimensional visual embeddings. These features are precomputed and kept fixed during training. In the cross-domain benchmark, validation and test targets are restricted to target-domain items, while source-domain interactions are retained as historical context. Dataset statistics after preprocessing are reported in Appendix B.

4.1.2 Evaluation Protocol

Interactions are split chronologically into training, validation, and test sets with an 80%/10%/10% ratio. Evaluation follows a full-sort protocol, ranking the ground-truth next item against all items in the processed candidate set. In the cross-domain setting, validation and test targets are target-domain items, while histories may contain items from both domains.

We report Recall@K and NDCG@K for $K \in \{10, 20\}$. Recall@K measures whether the target item appears in the top-K list, and NDCG@K further accounts for its rank.

4.1.3 Baselines

We compare UST with collaborative, sequential, multimodal, transferable semantic, and cross-domain baselines. The collaborative and sequential baselines include BPR, GRU4Rec (Hidasi et al., 2015), and SASRec (Kang and McAuley, 2018). The multimodal baselines include VBPR (He and McAuley, 2016) and BM3 (Zhou et al., 2023b). For transferable semantic recommendation, we use UniSRec (Hou et al., 2022) and VQ-Rec (Hou et al., 2023); they are trained from scratch in in-domain experiments and follow source-domain pretraining with target-domain fine-tuning in cross-domain experiments. We also include C2DSR (Cao et al., 2022) as an explicit cross-domain sequential baseline, adapting its input/output format and evaluation protocol to our source-enhanced benchmark.

4.1.4 Implementation Details

UST and all baselines are implemented in PyTorch. For BPR, GRU4Rec, and SASRec, we use the RecBole implementations (Zhao et al., 2021). UST is added as a shared-private tokenized semantic layer on top of the corresponding sequential backbone. All models use the same data splits and full-sort evaluation.

The maximum sequence length is 50. Models are trained for up to 100 epochs with early stopping patience of 10, using validation NDCG@10 for model selection. Hyperparameters are tuned separately for each dataset and backbone. We search $V \in \{64, 128, 256, 1024\}$, $\beta_1 \in \{0.01, 0.05, 0.1, 0.2\}$, $\beta_3 \in \{0.01, 0.05, 0.1, 0.5\}$, and $\beta_4 \in \{0.01, 0.05, 0.1, 0.2\}$. For in-domain experiments, we set $\beta_2 = 0$, $\tau_{\min} = 0.1$, and search $\tau_0 \in \{0.1, 0.2, 0.5, 1.0\}$. For cross-domain experiments, we set $\tau_{\min} = 0.2$ and search $\beta_2 \in \{0.01, 0.05, 0.1, 0.5\}$. Experiments are run on an NVIDIA RTX 3090 GPU with 24GB memory.

4.2 RQ1: In-Domain Recommendation Performance

We first compare UST with collaborative, sequential, multimodal, and transferable semantic baselines on three in-domain datasets. All methods are evaluated under the same full-sort protocol. UniSRec and VQ-Rec are trained from scratch so that the comparison reflects in-domain recommendation rather than transfer. Table 1 reports the results.

UST is most effective when paired with the self-attention backbone. UST_SASRec achieves the best Recall@10 and Recall@20 on Video and Musical, while remaining competitive on Baby. This pattern suggests that the tokenized semantic layer enriches item representations, but its value depends on whether the sequence encoder can use the added structure. Compared with recurrent modeling, self-attention appears better suited to exploiting these enriched item representations.

The results also show clear domain dependence. UST yields its largest gains on Musical, where textual and visual information likely provides semantic cues beyond interaction histories. On Baby, however, vanilla SASRec remains strongest, suggesting that item-side content may be less discriminative or that collaborative signals already dominate the prediction task. Compared with UniSRec and VQ-Rec, UST_SASRec performs bet-

Table 1: Overall performance comparison on three in-domain datasets. The best results are highlighted in bold, and the second-best results are underlined.

Model	Video				Musical				Baby			
	Recall@10	NDCG@10	Recall@20	NDCG@20	Recall@10	NDCG@10	Recall@20	NDCG@20	Recall@10	NDCG@10	Recall@20	NDCG@20
GRU4Rec	0.0608	0.0311	0.0990	0.0407	0.0424	0.0219	0.0650	0.0276	0.0294	0.0145	0.0497	0.0196
SASRec	0.0767	**0.0381**	0.1162	**0.0480**	0.0444	0.0226	0.0676	0.0285	**0.0306**	**0.0148**	**0.0516**	**0.0201**
BM3	0.0591	0.0321	0.0869	0.0391	0.0508	**0.0266**	0.0749	**0.0327**	0.0246	0.0123	0.0406	0.0163
BPR	0.0574	0.0298	0.0868	0.0373	0.0373	0.0200	0.0560	0.0248	0.0212	0.0103	0.0387	0.0146
UniSRec	0.0639	0.0281	0.0939	0.0356	0.0376	0.0172	0.0596	0.0228	0.0272	0.0117	0.0459	0.0163
VBPR	0.0374	0.0194	0.0471	0.0219	0.0288	0.0145	0.0351	0.0161	0.0179	0.0081	0.0293	0.0110
VQRec	0.0440	0.0216	0.0669	0.0274	0.0247	0.0125	0.0369	0.0156	0.0226	0.0103	0.0374	0.0140
UST_GRU4Rec	0.0650	0.0340	0.0979	0.0422	0.0403	0.0211	0.0608	0.0262	0.0282	0.0139	0.0482	0.0189
UST_SASRec	**0.0811**	0.0380	**0.1205**	0.0479	**0.0515**	0.0260	**0.0768**	0.0324	0.0305	0.0144	0.0491	0.0191

ter on `Video` and `Musical`, indicating that semantic representations are most useful when tightly integrated with sequential behavior modeling.

4.3 RQ2: Cross-Domain Source-Enhanced Target Prediction

We evaluate UST on `Cross_Sport_Beauty` under source-enhanced target prediction. Source-domain interactions are retained as auxiliary context, and validation/test targets come from the target domain. Table 2 reports the results.

`UniSRec` and `VQ-Rec` substantially outperform the other methods. This advantage is consistent with their source-pretraining and target-finetuning paradigm: source-domain training provides a semantic initialization, while target-domain finetuning adapts the representation to the final ranking objective. Standard sequential models perform much worse, suggesting that directly mixing source and target histories is insufficient when the domain gap is large.

UST does not improve over its corresponding backbones in this setting. Direct shared-private discrete semantic sharing is therefore not automatically transferable across weakly related domains. When source and target items differ substantially in function, language, and visual appearance, shared code assignments may introduce noisy source-domain semantics and unstable codebook competition. The weak performance of `C2DSR` further reflects the mismatch between its overlap-user-oriented assumptions and our non-strict-overlap benchmark. Overall, RQ2 suggests that robust cross-domain tokenized recommendation requires adaptive source filtering, alignment control, and preservation of target-specific semantics.

4.4 RQ3: Ablation Study

We conduct ablations on `UST_SASRec`, the strongest UST variant in RQ1. Additional ablations with other backbones and cross-domain variants are provided in Appendix D. Table 3 reports the results.

Table 2: Diagnostic results on `Cross_Sport_Beauty` under the source-enhanced target prediction setting.

Model	Recall@10	NDCG@10	Recall@20	NDCG@20
GRU4Rec	0.0319	0.0199	0.0444	0.0230
SASRec	0.0610	0.0458	0.0743	0.0491
UniSRec	0.2885	0.2681	0.3193	0.2755
VQRec	0.2801	0.2677	0.3137	0.2761
C2DSR	0.0033	0.0014	0.0078	0.0026
UST_GRU4Rec	0.0159	0.0102	0.0218	0.0117
UST_SASRec	0.0472	0.0374	0.0545	0.0393

The ablations identify discrete semantic tokenization as the main contributor. `EarlyFusion`, which directly combines collaborative, textual, and visual features, performs much worse than the full model on `Video` and `Musical`. Removing tokenization causes a similar degradation. The gains therefore do not come from simply adding multimodal features, but from organizing heterogeneous item evidence into a structured discrete semantic space before sequential modeling.

The auxiliary regularizers mainly stabilize this code space. Commitment regularization is particularly important on `Musical`, suggesting that continuous multimodal representations and discrete assignments must remain aligned when item-side semantics are informative. Orthogonality also helps by reducing interference between shared and private factors. On `Baby`, the full model remains competitive but does not surpass vanilla `SASRec`, further showing that the value of tokenization depends on the informativeness of item-side content.

4.5 RQ4: Hyperparameter Sensitivity Analysis

We analyze the sensitivity of `UST_SASRec` to key hyperparameters using NDCG@10. Figure 3 reports results for vocabulary size $|V|$, balance coefficient β_1, commitment coefficient β_4, and initial Gumbel-Softmax temperature τ_0. Results for β_3 are provided in Appendix E.

Table 3: Ablation results of UST_SASRec on in-domain recommendation datasets. Additional ablation results with other backbones and cross-domain settings are reported in Appendix D.

Variant	Video				Musical				Baby			
	Recall@10	NDCG@10	Recall@20	NDCG@20	Recall@10	NDCG@10	Recall@20	NDCG@20	Recall@10	NDCG@10	Recall@20	NDCG@20
Vanilla SASRec	0.0767	0.0381	0.1162	0.0480	0.0444	0.0226	0.0676	0.0285	0.0306	0.0148	0.0516	0.0201
EarlyFusion	0.0389	0.0200	0.0603	0.0254	0.0336	0.0173	0.0479	0.0209	0.0297	0.0134	0.0485	0.0181
w/o Tokenization	0.0400	0.0203	0.0597	0.0253	0.0336	0.0173	0.0475	0.0208	0.0269	0.0130	0.0464	0.0179
w/o Orthogonal	0.0400	0.0202	0.0597	0.0251	0.0336	0.0173	0.0481	0.0209	0.0266	0.0132	0.0460	0.0180
w/o Commit	0.0790	0.0382	0.1193	0.0483	0.0336	0.0173	0.0481	0.0210	0.0270	0.0134	0.0463	0.0182
Full UST	0.0811	0.0380	0.1205	0.0479	0.0515	0.0260	0.0768	0.0324	0.0305	0.0144	0.0491	0.0191

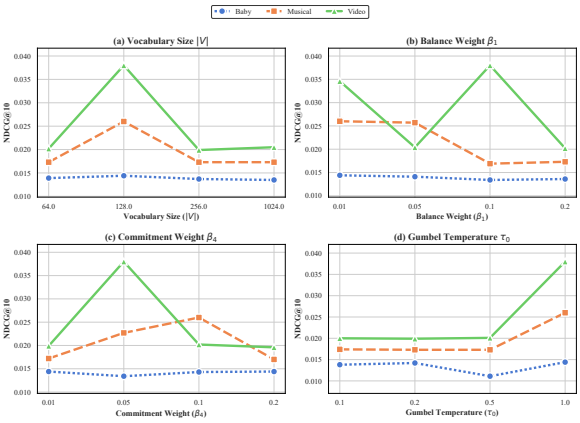

Figure 3: Hyperparameter sensitivity of UST_SASRec on three in-domain datasets.

The clearest trends appear in codebook size and Gumbel temperature. All three datasets perform best at $|V| = 128$, indicating that the discrete semantic space needs enough capacity to separate useful item semantics, but not so much that code usage becomes sparse. Similarly, $\tau_0 = 1.0$ performs best across datasets, suggesting that smooth early-stage assignments help the model explore the codebook before token decisions become sharper.

The regularization weights are more dataset-dependent. The optimal β_1 and β_4 vary across domains, reflecting differences in modality quality and semantic noise. Balance regularization prevents code collapse but may over-constrain code usage if too strong. Commitment regularization stabilizes the link between continuous multimodal features and discrete codes, but its effect depends on modality informativeness. Overall, UST requires a balanced configuration of codebook capacity, assignment smoothness, and regularization. The cross-domain sensitivity of β_2 in Appendix F further shows that tuning alignment strength alone cannot resolve negative transfer caused by source–target mismatch.

5 Conclusion and Future Work

We presented UST, a unified semantic tokenization framework for multimodal sequential recommendation. By converting collaborative, textual, and visual item evidence into discrete semantic tokens before sequence modeling, UST provides standard sequential backbones with a structured interface to item-level multimodal semantics.

The experiments show that the benefit of tokenization is conditional. UST performs best with SASRec on Video_Games and Musical_Instruments, where textual and visual evidence appears to complement behavioral histories, but its gains are weaker on Baby_Products and recurrent backbones. Ablation results further show that the main improvement comes from semantic tokenization rather than early multimodal fusion. The discrete bottleneck helps organize heterogeneous evidence into reusable item-level units, but stable performance requires appropriate codebook capacity, smooth token assignment, and balanced regularization.

The cross-domain results reveal an important limitation: source-domain evidence is not automatically transferable, and direct shared-private semantic sharing can introduce negative transfer when source and target domains differ substantially in function, language, or visual appearance. Future work should explore adaptive tokenization, target-aware source filtering, and instance-level alignment. Finer-grained analyses of codebook utilization, dead codes, assignment dynamics, and modality-specific token usage may further clarify when discrete semantic representations improve recommendation and when they impose harmful compression or noisy transfer.

Limitations

UST has several limitations. Its effectiveness depends on both the domain and the backbone: the clearest gains appear with the self-attention-based SASRec backbone, whereas improvements

are weaker on Baby_Products and less consistent for graph-based models such as LightGCN. The cross-domain results should also be read as diagnostic rather than as evidence of robust transfer. When source and target domains are weakly aligned, direct sharing of discrete semantic codes can introduce negative transfer, and adjusting the alignment coefficient alone does not remove this mismatch. In addition, our experiments use selected categories from Amazon Reviews'23, with temporal filtering applied to large domains for computational feasibility; the conclusions may therefore not directly extend to industrial logs, strict overlap-user cross-domain settings, or datasets with different sparsity and modality quality. Finally, UST uses fixed pre-trained text and image encoders, and its discrete semantic layer introduces additional hyperparameters related to codebook size, token assignment, and regularization. Future work should examine adaptive source filtering, graph-aware tokenized propagation, end-to-end modality adaptation, codebook dynamics, and deployment concerns such as fairness, robustness, privacy, and latency.

Acknowledgments

References

Jiangxia Cao, Xin Cong, Jiawei Sheng, Tingwen Liu, and Bin Wang. 2022. Contrastive cross-domain sequential recommendation. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 138–147.

Lu Chen, Peng Yi, Xiaoyue Hou, Cheng Yang, and Xiongcai Cai. 2026. When modalities go missing: Early fusion for multimodal recommendation. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pages 1094–1098.

Shu Chen, Zitao Xu, Weike Pan, Qiang Yang, and Zhong Ming. 2024. A survey on cross-domain sequential recommendation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7989–7998. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Shereen Elsayed, Ahmed Rashed, and Lars Schmidt-Thieme. 2024. Multi-behavioral sequential recommendation. In *Proceedings of the 18th ACM conference on recommender systems*, pages 902–906.

Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Jie Wang, and Joemon M Jose. 2024. Iisan: Efficiently adapting multimodal representation for sequential recommendation with decoupled peft. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 687–697.

Ruining He and Julian McAuley. 2016. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *Computer ence*.

Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning vector-quantized item representation for transferable sequential recommenders. In *TheWebConf*.

Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *KDD*.

Minyoung Huh, Brian Cheung, Pulkit Agrawal, and Phillip Isola. 2023. Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks. In *International Conference on Machine Learning*, pages 14096–14113. PMLR.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE.

Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. 2021. Semi: A sequential multi-modal information transfer network for e-commerce micro-video recommendations. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3161–3171.

Chengxi Li, Yejing Wang, Qidong Liu, Xiangyu Zhao, Wanyu Wang, Yiqi Wang, Lixin Zou, Wenqi Fan, and Qing Li. 2023a. Strec: Sparse transformer for sequential recommendations. In *Proceedings of the 17th ACM conference on recommender systems*, pages 101–111.

Yang Li, Qi'Ao Zhao, Chen Lin, Jinsong Su, and Zhilin Zhang. 2024. Who to align with: Feedback-oriented multi-modal alignment in recommendation systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–676.

Zihao Li, Aixin Sun, and Chenliang Li. 2023b. Diffurec: A diffusion model for sequential recommendation. *ACM Transactions on Information Systems*, 42(3):1–28.

Jing Liu, Lele Sun, Weizhi Nie, Peiguang Jing, and Yuting Su. 2024a. Graph disentangled contrastive learning with personalized transfer for cross-domain recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 8769–8777.

Qijiong Liu, Hengchang Hu, Jiahao Wu, Jieming Zhu, Min-Yen Kan, and Xiao-Ming Wu. 2024b. Discrete semantic tokenization for deep ctr prediction. In *Companion Proceedings of the ACM Web Conference 2024*, pages 919–922.

Varun M S, Ganesh Vaidyanathan K, and Bhaskarjyoti Das. 2024. Multimodal Recommendation Systems in the LLM Era: A Survey of Feature Representation and Fusion Methods . In *2024 4th International Conference on Advanced Enterprise Information System (AEIS)*, pages 89–95, Los Alamitos, CA, USA. IEEE Computer Society.

Xuying Ning, Wujiang Xu, Tianxin Wei, and Xiaolei Liu. 2025. i2vae: interest information augmentation with variational regularizers for cross-domain sequential recommendation. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, pages 3235–3251.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*.

Shaina Raza, Mizanur Rahman, Safiullah Kamawal, Armin Toroghi, Ananya Raval, Farshad Navah, and Amirmohammad Kazemeini. 2026. A comprehensive review of recommender systems: Transitioning from theory to practice. *Computer Science Review*, 59:100849.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*.

Baochen Sun and Kate Saenko. 2016. Deep CORAL: correlation alignment for deep domain adaptation. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, Lecture Notes in Computer Science, pages 443–450.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Li Wang, Qiang Wu, and Min Xu. 2025. Fedpcl-cdr: A federated prototype-based contrastive learning framework for privacy-preserving cross-domain recommendation. *Neural Networks*, page 108380.

Zihe Wang and Bo Jin. 2026. Diffsbr: A diffusion model for session-based recommendation. *Information Processing & Management*, 63(1):104284.

Zitao Xu, Shu Chen, Weike Pan, and Zhong Ming. 2025. A multi-view graph contrastive learning framework for cross-domain sequential recommendation. *ACM Transactions on Recommender Systems*, 3(4):1–28.

Zixuan Yi and Iadh Ounis. 2024. A unified graph transformer for overcoming isolations in multi-modal recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 518–527.

Yu Zhang, Zhiyong Cheng, Fan Liu, Xun Yang, and Yuxin Peng. 2024. Decoupled domain-specific and domain-conditional representation learning for cross-domain recommendation. *Information Processing & Management*, 61(3):103689.

Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Kaiyuan Li, Yushuo Chen, Yujie Lu, Hui Wang, Changxin Tian, Xingyu Pan, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *CIKM*.

Yi Zhao, Le Chen, Xinchao Feng, Mengpan Chen, and Jibing Gong. 2025. Cross-domain recommendation via quantized disentangled generative model. *Information Fusion*, page 104028.

Peilin Zhou, Jingqi Gao, Yueqi Xie, Qichen Ye, Yining Hua, Jaeboum Kim, Shoujin Wang, and Sunghun Kim. 2023a. Equivariant contrastive learning for sequential recommendation. In *Proceedings of the 17th ACM conference on recommender systems*, pages 129–140.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023b. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, page 845–854.

Jiajie Zhu, Yan Wang, Feng Zhu, and Zhu Sun. 2023. Domain disentanglement with interpolative data augmentation for dual-target cross-domain recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 515–527.

Pablo Zivic, Hernan Vazquez, and Jorge Sánchez. 2024. Scaling sequential recommendation models with transformers. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 1567–1577.

A Additional Method Details

This section gives the full form of the auxiliary objectives and implementation details used in UST.

A.1 Temperature Annealing

The Gumbel-Softmax temperature is annealed during training as

$$\tau_t = \max(\tau_{\min}, \tau_0 \gamma^t), \tag{20}$$

where τ_0 is the initial temperature, $\tau_{\min}$ is the lower bound, and γ controls the decay rate. We set $\gamma = 0.9995$ in all experiments.

A.2 Code Balance Regularization

Let $\mathcal{M}$ be the set of unique non-padding items in the current mini-batch, constructed from both the input sequences and the positive next items. The batch-level token distributions for the shared and private branches are

$$\bar{\mathbf{p}}^s = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathbf{p}_i^s, \qquad \bar{\mathbf{p}}^p = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathbf{p}_i^p. \tag{21}$$

To reduce code collapse, we regularize code usage in both branches:

$$\mathcal{L}_{bal} = \mathcal{R}_{bal}(\{\mathbf{p}_i^s\}_{i \in \mathcal{M}}) + \mathcal{R}_{bal}(\{\mathbf{p}_i^p\}_{i \in \mathcal{M}}). \tag{22}$$

The regularizer is defined as

$$\begin{aligned} \mathcal{R}_{bal}(\{\mathbf{p}_i\}_{i \in \mathcal{M}}) =& \mathrm{KL}(\bar{\mathbf{p}} \| \mathbf{u}) \\ &+ \lambda_H \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} H(\mathbf{p}_i), \end{aligned} \tag{23}$$

where $\mathbf{u}$ is the uniform distribution over codes and $H(\cdot)$ denotes entropy. The KL term encourages batch-level code diversity, whereas the entropy term discourages overly diffuse per-item assignments. We set $\lambda_H = 0.1$.

A.3 Shared-Private Orthogonality

To separate shared and private semantic factors, we apply an orthogonality penalty. Let $\hat{\mathbf{E}}^s \in \mathbb{R}^{|\mathcal{M}| \times d}$ and $\hat{\mathbf{E}}^p \in \mathbb{R}^{|\mathcal{M}| \times d}$ denote row-normalized shared and private item embeddings in the current mini-batch. The loss is

$$\mathcal{L}_{orth} = \max\left(0, \left\| \frac{\hat{\mathbf{E}}^{s\top} \hat{\mathbf{E}}^p}{|\mathcal{M}|} \right\|_F^2 - m\right), \tag{24}$$

where m is a margin. This term discourages excessive overlap between the two branches while allowing moderate correlation when useful for the recommendation objective.

A.4 Commitment Regularization

The commitment term keeps the discrete representation close to the continuous fused representation:

$$\mathcal{L}_{com} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(1 - \cos(\mathbf{e}_i^d, \mathbf{x}_i)\right). \tag{25}$$

This loss stabilizes codebook learning by preventing selected codes from drifting away from the multimodal feature geometry.

A.5 Cross-Domain Shared Alignment

For cross-domain training, let $\mathcal{M}_a$ and $\mathcal{M}_b$ be the source- and target-domain non-padding item sets in the current mini-batch. Domain labels are obtained from the processed item metadata. We first normalize shared embeddings:

$$\tilde{\mathbf{e}}_i^s = \frac{\mathbf{e}_i^s}{\|\mathbf{e}_i^s\|_2}. \tag{26}$$

For each domain $r \in \{a, b\}$, we estimate the first- and second-order moments:

$$\boldsymbol{\mu}_r = \frac{1}{|\mathcal{M}_r|} \sum_{i \in \mathcal{M}_r} \tilde{\mathbf{e}}_i^s, \tag{27}$$

$$\boldsymbol{\Sigma}_r = \frac{1}{|\mathcal{M}_r| - 1} \sum_{i \in \mathcal{M}_r} (\tilde{\mathbf{e}}_i^s - \boldsymbol{\mu}_r)(\tilde{\mathbf{e}}_i^s - \boldsymbol{\mu}_r)^\top. \tag{28}$$

The alignment loss is

$$\mathcal{L}_{align} = \frac{1}{d} \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2^2 + \frac{1}{d^2} \|\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b\|_F^2. \tag{29}$$

We compute this term only when both domains contribute at least two items to the mini-batch. The objective follows the spirit of distribution-level alignment, such as correlation alignment (Sun and Saenko, 2016), but is applied to the shared discrete semantic branch.

A.6 Representation Caching

To reduce redundant computation under full-sort scoring, item representations are cached during training and evaluation. In each training step, we recompute the representations of unique non-padding items in the current mini-batch with gradients, while cached representations are used for the remaining candidates.

A.7 Complexity

UST adds computation through multimodal projection, gated fusion, token projection, and codebook lookup. For each processed item, the additional cost is linear in the number of item occurrences and in the code vocabulary size used for token projection. In practice, the dominant cost remains the sequential backbone and full-item scoring. Since UST does not modify the sequence encoder, it can be incorporated into existing sequential recommendation pipelines with limited architectural changes.

Table 4: Statistics of the processed datasets used for training and evaluation.

Dataset	#Users	#Items	#Interactions	Sparsity
Video	24,243	9,999	184,567	99.92%
Musical	24,348	12,907	208,341	99.93%
Baby	72,966	21,940	596,115	99.96%
Cross_Sport_Beauty	433,439	165,401	3,690,994	99.99%

B Dataset Statistics

Table 4 reports the statistics of the processed datasets used for training and evaluation. The statistics are computed after temporal filtering and 5-core filtering.

C Additional Cross-Domain Results

Table 5 reports additional diagnostic results on `Cross_Elec_Sport`, where `Electronics` is used as the source domain and `Sports_and_Outdoors` as the target domain. The trend is consistent with the main cross-domain experiment. `VQRec` remains the strongest method, whereas UST does not improve over the corresponding sequential backbones. This result supports the conclusion that direct shared-private discrete semantic sharing is sensitive to domain mismatch and may introduce negative transfer when source and target semantics are weakly aligned.

Table 5: Additional diagnostic results on `Cross_Elec_Sport`.

Model	Recall@10	NDCG@10	Recall@20	NDCG@20
GRU4Rec	0.0359	0.0227	0.0482	0.0258
SASRec	0.0286	0.0159	0.0409	0.0190
UniSRec	0.0359	0.0213	0.0498	0.0247
VQRec	0.0554	0.0426	0.0671	0.0455
C2DSR	0.0000	0.0000	0.0001	0.0000
UST_SASRec	0.0154	0.0095	0.0208	0.0108
UST_GRU4Rec	0.0108	0.0061	0.0153	0.0072

D Additional Ablation Results

This section reports ablations omitted from the main text for space. We include `UST_GRU4Rec` on in-domain datasets, cross-domain ablations on `Cross_Sport_Beauty`, and an additional experiment with `UST_LightGCN`. These results provide complementary evidence for the backbone- and setting-dependent behavior of UST.

E Additional Sensitivity Analysis of Orthogonality Weight

We further examine the orthogonality coefficient β_3, which controls the separation between shared and private semantic branches. As shown in Figure 4, $\beta_3 = 0.05$ generally gives the best or near-best NDCG@10. `Video_Games` and `Musical_Instruments` both reach their best values at this setting, whereas `Baby_Products` is less sensitive. These results indicate that moderate orthogonality helps maintain semantic separation, but an overly strong constraint is unnecessary and may interfere with the recommendation objective.

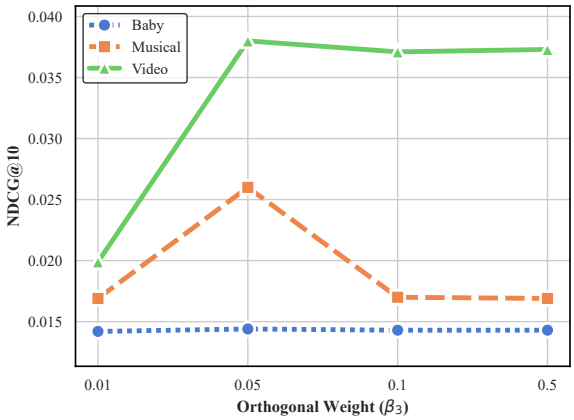

Figure 4: Sensitivity analysis of the orthogonality coefficient β_3 on three in-domain datasets.

F Additional Cross-Domain Sensitivity Analysis

We also analyze the cross-domain alignment coefficient β_2 on `Cross_Sport_Beauty`. This coefficient controls the strength of moment matching between source- and target-domain shared semantics.

As shown in Figure 5, the best NDCG@10 is obtained at $\beta_2 = 0.1$, suggesting that moderate alignment is preferable to either weak or overly strong alignment. When β_2 is too small, the shared branch receives little cross-domain signal. When it is too large, distribution matching can dominate the target-domain ranking objective.

The overall variation, however, is small: NDCG@10 ranges only from 0.0367 to 0.0374. Moreover, even the best setting remains below the vanilla `SASRec` result in the main cross-domain experiment. Thus, tuning alignment strength alone cannot remove the negative transfer caused by source–target mismatch. This observation is consistent with RQ2: in a discrete semantic space, explicit alignment must be applied cautiously, and robust transfer likely requires target-aware source filtering or adaptive alignment mechanisms.

Table 6: Additional in-domain ablation results of `UST_GRU4Rec`.

Variant	Video				Musical				Baby			
	Recall@10	NDCG@10	Recall@20	NDCG@20	Recall@10	NDCG@10	Recall@20	NDCG@20	Recall@10	NDCG@10	Recall@20	NDCG@20
Vanilla GRU4Rec	0.0608	0.0311	0.0990	0.0407	0.0424	0.0219	0.0650	0.0276	0.0294	0.0145	0.0497	0.0196
EarlyFusion	0.0614	0.0319	0.0924	0.0396	0.0375	0.0195	0.0567	0.0243	0.0289	0.0142	0.0480	0.0190
w/o Tokenization	0.0653	0.0341	0.0983	0.0424	0.0406	0.0214	0.0616	0.0266	0.0281	0.0137	0.0461	0.0182
w/o Orthogonal	0.0650	0.0339	0.0979	0.0422	0.0405	0.0212	0.0609	0.0262	0.0291	0.0145	0.0477	0.0192
w/o Commit	0.0631	0.0330	0.0958	0.0412	0.0374	0.0189	0.0566	0.0237	0.0305	0.0150	0.0504	0.0200
Full UST	0.0650	0.0340	0.0979	0.0422	0.0403	0.0211	0.0608	0.0262	0.0282	0.0139	0.0482	0.0189

Table 7: Additional cross-domain ablation results on `Cross_Sport_Beauty`.

Backbone	Variant	Recall@10	NDCG@10	Recall@20	NDCG@20
GRU4Rec	Vanilla	0.0319	0.0199	0.0444	0.0230
	EarlyFusion	0.0184	0.0124	0.0244	0.0139
	w/o Tokenization	0.0159	0.0104	0.0218	0.0118
	w/o Orthogonal	0.0175	0.0114	0.0240	0.0131
	w/o Commit	0.0182	0.0119	0.0248	0.0135
	Shared-Only	0.0174	0.0113	0.0240	0.0129
	Private-Only	0.0168	0.0109	0.0231	0.0125
	w/o Align	0.0178	0.0116	0.0242	0.0132
	Full UST	0.0176	0.0115	0.0241	0.0131
SASRec	Vanilla	0.0610	0.0458	0.0743	0.0491
	EarlyFusion	0.0493	0.0397	0.0564	0.0415
	w/o Tokenization	0.0459	0.0363	0.0530	0.0381
	w/o Orthogonal	0.0483	0.0389	0.0554	0.0407
	w/o Commit	0.0487	0.0392	0.0557	0.0409
	Shared-Only	0.0464	0.0367	0.0535	0.0385
	Private-Only	0.0471	0.0375	0.0544	0.0393
	w/o Align	0.0480	0.0385	0.0552	0.0403
	Full UST	0.0472	0.0374	0.0545	0.0393

Table 8: Additional ablation results of `UST_LightGCN` on `All_Beauty`.

Variant	Recall@10	NDCG@10	Recall@20	NDCG@20
Vanilla LightGCN	0.2969	0.2719	0.3417	0.2831
EarlyFusion	0.2913	0.2748	0.3221	0.2823
w/o Tokenization	0.2941	0.2756	0.3389	0.2865
w/o Orthogonal	0.2941	0.2758	0.3361	0.2861
w/o Commit	0.2941	0.2746	0.3389	0.2854
Full UST	0.2941	0.2758	0.3389	0.2867

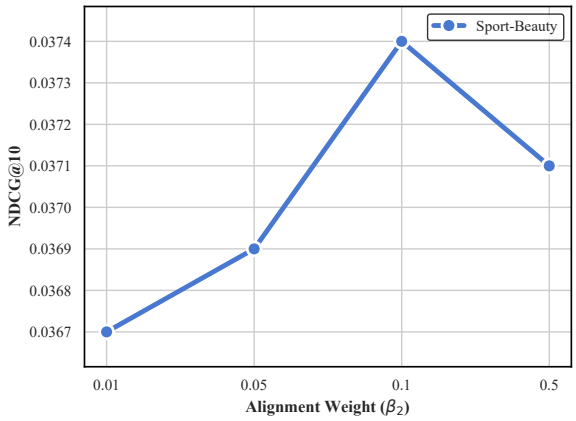

Figure 5: Sensitivity analysis of the cross-domain alignment coefficient β_2 on `Cross_Sport_Beauty`.