Anatomy of a Decision: Uncertainty-aware Hierarchical Intent Learning via Flow Matching for Multimodal Recommendation

Yuchen Miao[1], Zijun Wang[1], Ke Liu[1], and Siyang Xu[2,3]✉

[1] Sydney Smart Technology College, Northeastern University, China
[2] School of Computer and Communication Engineering and Hebei Key Laboratory of Marine Perception Network and Data Processing, Northeastern University at Qinhuangdao, China
[3] Shijiazhuang Innovation Research Institute of Northeastern University, Shijiazhuang, China
✉Corresponding author.

Abstract. Modeling the underlying user intent is crucial for recommendation, but existing methods struggle with the inherent uncertainty and the dynamic, hierarchical nature of user interests. Current approaches often rely on clustering or prototype learning to discover a static set of intents. However, they face two critical challenges: (1) they overlook the uncertainty inherent in multimodal features; and (2) they assume a static and flat intent structure, failing to adapt to a user's varying decision certainty. To address these limitations, we propose **UHIFlow**, an **U**ncertainty-aware **H**ierarchical **I**ntent learning framework via **Flow** matching. First, our Cross-modal Uncertainty Synergistic Modeling (CUSM) module leverages conditional flow matching to quantify uncertainty from visual and textual modalities and synergistically align them. Subsequently, the Uncertainty-guided Hierarchical Intent Generation (UHIG) module uses this quantified uncertainty to dynamically construct a personalized intent hierarchy, generating coarse-grained intents for uncertain users and fine-grained ones for users with clear preferences. Extensive experiments on three real-world datasets demonstrate that UHIFlow significantly outperforms state-of-the-art baselines.

Keywords: Multimodal Recommendation · User Intent Modeling · Flow Matching

1 Introduction

Recommender systems are indispensable for navigating online information across domains from e-commerce to content streaming. They model user preferences from past behaviors, yet raw behaviors are only superficial signals of deeper, underlying intents. To bridge this gap, intent-based recommenders have emerged as a promising frontier [12,20,25].

Existing intent-based recommendation falls into two categories. The first clusters user behaviors to discover intents, grouping users by interaction patterns.

The second employs prototype learning, mapping users to a latent space of predefined intent prototypes that represent canonical interests [20,27,38]. More recently, learning intent from items' visual and textual information has gained traction, offering a more holistic view of preferences [37].

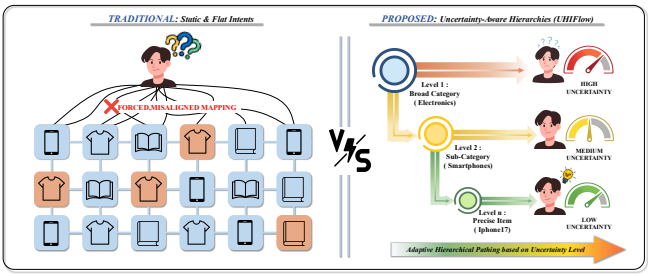

Fig. 1. A comparison between traditional intent modeling and our proposed UHIFlow. The traditional approach (left) relies on static, flat intents, often leading to **forced, misaligned mappings** across items of **distinct categories (indicated by different colors)**. In contrast, UHIFlow (right) employs **adaptive hierarchical pathing** guided by uncertainty levels. It maps users with **High Uncertainty** to broad categories (Level 1, e.g., "Electronics") and progressively refines the representation for users with **Low Uncertainty** down to precise items (Level n, e.g., "iPhone 17").

Despite these advancements, existing multimodal intent modeling approaches face the following critical challenges that limit their effectiveness in real-world scenarios as shown in figure 1:

- **Neglect of Inherent Uncertainty.** They overlook the uncertainty in recommendation, which stems both from ambiguous user behavior and from the multimodal data itself, e.g., noisy images or semantically ambiguous text [6]. Treating multimodal features as deterministic inputs prevents a reliable representation of user intent, yielding suboptimal performance.
- **Static and Flat Intent Structure.** They assume a fixed number of intents, ignoring the hierarchical nature of user intent. Decision-making is often a coarse-to-fine refinement: a user with high uncertainty ("just browsing") holds a coarse intent ("electronics"), whereas one with low uncertainty ("looking for a new phone") holds a much finer intent. A static structure cannot adapt to this varying certainty.

To address these challenges, we propose **U**ncertainty-aware **H**ierarchical **I**ntent learning via **Flow** matching (**UHIFlow**). Since flow matching learns complex distributions without restrictive priors [16], we treat item features as distributions rather than fixed points to model their uncertainty explicitly. We first introduce a Cross-modal Uncertainty Synergistic Modeling module that quantifies the uncertainty of visual and textual modalities and lets the two uncertainties inform and regularize each other. We then design an Uncertainty-guided Hierarchical Intent Generation module that uses this uncertainty to construct a user-specific intent hierarchy, generating coarse-grained structures for uncertain users and fine-grained ones for users with clear preferences.

The main contributions of this paper are summarized as follows:

- We propose UHIFlow for uncertainty-aware multimodal intent modeling; to our knowledge, it is the first to explicitly model multimodal uncertainty for hierarchical intent discovery in recommendation.
- We design a Cross-modal Uncertainty Synergistic Modeling module that leverages conditional flow matching to quantify and synergize uncertainties from different modalities.
- We devise an Uncertainty-guided Hierarchical Intent Generation module that adaptively constructs personalized intent hierarchies based on user uncertainty.
- Extensive experiments on real-world datasets show that UHIFlow outperforms state-of-the-art baselines.

2 Related Work

2.1 Multimodal Recommendation

Multimodal recommendation incorporates visual and textual information to enrich collaborative signals. Early work such as VBPR [8] demonstrates the benefit of visual features, while recent methods exploit stronger cross-modal learning: MMSSL [32] aligns multimodal and collaborative signals, DiffCL [29] constructs contrastive views, LGMRec [7] unifies local and global graph learning, and DiffMM [14] introduces graph diffusion. MMIL models multimodal intentions with a learnable codebook, while MDN and Diff-MSIN decompose or synergize modality factors [37,21,4,28]. However, these methods rarely quantify modality uncertainty or adapt the hierarchy of user intent.

2.2 Intent Modeling in Recommender Systems

Intent modeling captures multi-faceted user preferences that single embeddings often miss [12]. Existing methods include contrastive or clustering-based intent discovery [2,20], multi-interest modeling with MIND [15], and disentangled factor learning [27,38]. Hierarchical formulations such as HIM [40], HieRec [26], and recent hierarchical or semantic-token generative recommendation [3,11] encode coarse-to-fine preferences. Yet most rely on fixed intent numbers or static structures, limiting their ability to adjust intent granularity according to user certainty.

2.3 Generative Models for Recommendation

Generative recommendation has advanced through diffusion and flow-based models. DiffRec [30] and Diff4Rec [35] apply diffusion to interaction modeling and sequential augmentation, while recent diffusion models denoise multimodal features and behaviors [29,14,22,5]. Flow matching offers an efficient way to learn complex distributions without restrictive priors [16], and FlowCF [17] adapts it to collaborative filtering. These works focus mainly on denoising or generation rather than using flows to quantify multimodal uncertainty for hierarchical intent learning.

2.4 Uncertainty Modeling in Recommendation

Uncertainty modeling improves robustness under sparsity, distribution shift, and cold-start scenarios. UCC [19] uses teacher-student consistency, UA-GNN [1] performs uncertainty-aware pseudo-label selection, and WaPOIR [39] learns Wasserstein embeddings. Multimodal fusion research also shows that modalities provide unequal and unreliable evidence [6]. Existing methods, however, seldom combine multimodal uncertainty estimation with adaptive hierarchical intent generation.

3 Preliminary

In a typical multimodal recommendation scenario, we have a set of users $\mathcal{U}$ and a set of items $\mathcal{I}$. The user-item interactions are represented by a matrix $\mathbf{R} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $r_{ui} = 1$ if user u has interacted with item i, and $r_{ui} = 0$ otherwise. Each item $i \in \mathcal{I}$ is associated with multimodal features, specifically a visual feature vector $\mathbf{v}_i \in \mathbb{R}^{d_v}$ and a textual feature vector $\mathbf{t}_i \in \mathbb{R}^{d_t}$.

The objective is to learn a user representation $\mathbf{u}_u^*$ and an item representation $\mathbf{h}_i$ for predicting a personalized ranked list, with the recommendation score $\hat{y}_{ui} = (\mathbf{u}_u^*)^\top \mathbf{h}_i$.

4 Methodology

We first describe the multimodal interest modeling layer, then the two core modules—Cross-modal Uncertainty Synergistic Modeling (CUSM) and Uncertainty-guided Hierarchical Intent Generation (UHIG)—and finally the uncertainty-aware aggregation and overall optimization objective.

4.1 Multimodal Interest Modeling Layer

User–Item Interaction Graph We construct a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$ is the set of user and item nodes, and $\mathcal{E}$ represents the observed interactions. The initial feature for an item node i is derived by concatenating and projecting its visual and textual features: $\mathbf{e}_i^{(0)} = \mathrm{MLP}([\mathbf{v}_i; \mathbf{t}_i])$. User nodes are initialized with learnable embeddings $\mathbf{e}_u^{(0)}$.

Collaborative Interest Propagation To inject high-order collaborative information, we employ a light-weight graph propagation mechanism similar to LightGCN [9]. The embeddings are refined over K layers as follows:

$$\mathbf{e}_u^{(k)} = \sum_{i\in\mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} \mathbf{e}_i^{(k-1)}, \quad \mathbf{e}_i^{(k)} = \sum_{u\in\mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i||\mathcal{N}_u|}} \mathbf{e}_u^{(k-1)} \tag{1}$$

The final user and item embeddings, $\mathbf{h}_u$ and $\mathbf{h}_i$, are obtained by aggregating the embeddings from all layers: $\mathbf{h}_u = \sum_{k=0}^{K} \alpha_k \mathbf{e}_u^{(k)}$ and $\mathbf{h}_i = \sum_{k=0}^{K} \alpha_k \mathbf{e}_i^{(k)}$, where α_k are layer-wise aggregation weights. These embeddings serve as the basis for the subsequent uncertainty and intent modeling.

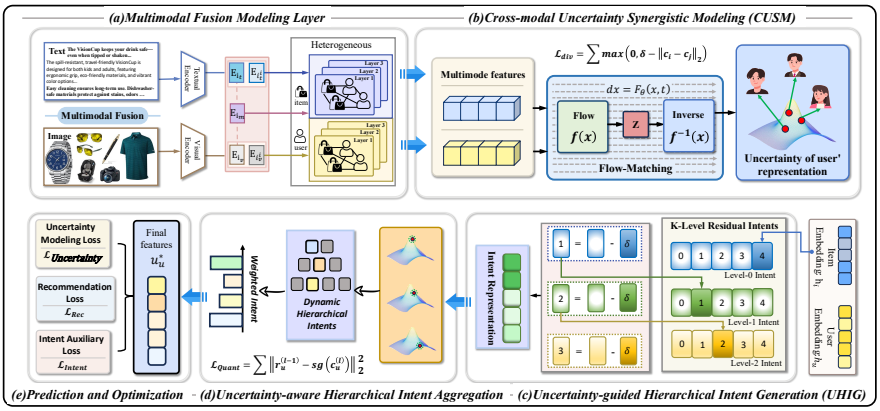

Fig. 2. Overview of UHIFlow, including multimodal interest modeling, CUSM, UHIG, and uncertainty-aware aggregation for final recommendation.

4.2 Cross-modal Uncertainty Synergistic Modeling (CUSM)

This module models item features as rich probability distributions rather than deterministic point estimates. Since the true conditional distribution of an item's multimodal features, $p_{data}(\mathbf{x}|\mathbf{c})$, is complex and uncertain, we learn a model $q_\theta(\mathbf{x}|\mathbf{c})$ that captures this uncertainty.

We select Flow Matching [16] for this task due to its ability to learn complex, data-dependent distributions without imposing restrictive prior assumptions. To make modality-specific uncertainty modeling explicit, we instantiate two conditional continuous normalizing flows (CNFs), f_θ and f_ϕ, for visual and textual features. The visual flow f_θ is induced by the vector field v_θ, while the textual flow f_ϕ is induced by v_ϕ. For modality $m \in \{v, t\}$, the corresponding flow evolves according to

$$\frac{d\mathbf{x}_m(t)}{dt} = v_m(t, \mathbf{x}_m(t), \mathbf{h}_i),$$

where $v_v \equiv v_\theta$ and $v_t \equiv v_\phi$. Each flow transforms a simple base distribution $p_0(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ into the modality-specific target feature distribution conditioned on the item's collaborative embedding $\mathbf{h}_i$.

Let $\mathbf{z}_v, \mathbf{z}_t$ be latent variables for visual and textual features. The flow matching objective trains each vector field $v_m(t, \mathbf{x}, \mathbf{c})$ to match a target field defined by a probability path $p_t(\mathbf{x}|\mathbf{x}_1)$ connecting the noise distribution p_0 to the data distribution p_1. Throughout this section, the condition is $\mathbf{c} = \mathbf{h}_i$, while the target sample is drawn from $p(\mathbf{x}|\mathbf{v}_i)$ (or $p(\mathbf{x}|\mathbf{t}_i)$). The conditional flow matching losses for the visual and textual modalities are aggregated as:

$$\mathcal{L}_{FM} = \mathbb{E}_{p(\mathbf{h}_i, \mathbf{v}_i, \mathbf{t}_i)} \left[\mathcal{L}_{FM_v}(\theta) + \mathcal{L}_{FM_t}(\phi) \right] \tag{2}$$

where, specifically, the visual loss is defined as:

$$\mathcal{L}_{FM_v}(\theta) = \mathbb{E}_{t, p_t(\mathbf{x}|\mathbf{v}_i)} \left[\| v_\theta(t, \mathbf{x}, \mathbf{h}_i) - u_t(\mathbf{x}|\mathbf{v}_i) \|^2 \right] \tag{3}$$

This objective has a distributional interpretation: reducing the gap between the learned vector field and the target probability path transports the base distribution toward the data-conditioned feature distribution. The resulting density evolution provides both denoised multimodal representations and a principled signal for uncertainty estimation.

To keep uncertainty estimation coherent across modalities, we introduce a synergistic guidance mechanism: since ambiguity in one modality should inform the other, we align the intermediate latent representations of the two flows. Let $\Phi_s(\mathbf{z}; f, \mathbf{h}_i)$ denote the state at time s obtained by evolving noise $\mathbf{z}$ with flow f conditioned on $\mathbf{h}_i$. The cross-modal guidance loss is:

$$\mathcal{L}_{cross} = \mathbb{E}_{s\sim U(0,1),\mathbf{z}_v,\mathbf{z}_t\sim\mathcal{N}} \|\Phi_s(\mathbf{z}_v; f_\theta, \mathbf{h}_i) - \mathrm{Proj}(\Phi_s(\mathbf{z}_t; f_\phi, \mathbf{h}_i))\|_2^2 \tag{4}$$

where $\mathrm{Proj}(\cdot)$ is a linear projection for modality alignment. Finally, we quantify the uncertainty of each modality via the magnitude of the log-density change, computed using the instantaneous change of variables formula:

$$\varsigma_m(\mathbf{h}_i) = \mathbb{E}_{\mathbf{z}_m\sim\mathcal{N}} \left[\left| \int_0^1 \mathrm{Tr}(\nabla_{\Phi_s} v_m(s, \Phi_s(\mathbf{z}_m; f_m, \mathbf{h}_i), \mathbf{h}_i))\, ds \right| \right] \tag{5}$$

where $m \in \{v, t\}$. The absolute value ensures $\varsigma_m \geq 0$, so it can be used as a non-negative uncertainty measure in subsequent thresholds. The total uncertainty for a user u is aggregated as: $\varsigma_u = \frac{1}{|\mathcal{I}_u|} \sum_{i\in\mathcal{I}_u} (\varsigma_v(\mathbf{h}_i) + \varsigma_t(\mathbf{h}_i))$.

4.3 Uncertainty-guided Hierarchical Intent Generation (UHIG)

This module builds a personalized intent hierarchy per user. Unlike conventional methods that fix the number of intents, we use Residual Quantization (RQ) to generate hierarchical intent embeddings whose depth is dynamically controlled by the user uncertainty score ς_u from CUSM.

We define L intent codebooks $\{\mathcal{C}_l \in \mathbb{R}^{K_l\times d}\}_{l=1}^{L}$, one per level. Initializing $\mathbf{r}_u^{(0)} = \mathbf{h}_u$, the generation proceeds iteratively for each level $l \in \{1, \ldots, L\}$:

$$\mathbf{c}_u^{(l)} = \underset{\mathbf{c}\in\mathcal{C}_l}{\mathrm{argmin}} \|\mathbf{r}_u^{(l-1)} - \mathbf{c}\|_2^2 \tag{6}$$

$$\mathbf{r}_u^{(l)} = \mathbf{r}_u^{(l-1)} - \mathbf{c}_u^{(l)} \tag{7}$$

This decomposes the user embedding into a sequence of increasingly fine-grained intent codes. The generation for user u stops at depth D_u when the residual norm falls below an uncertainty-dependent threshold:

$$\|\mathbf{r}_u^{(D_u)}\|_2 < \tau_{D_u} \quad \text{where} \quad \tau_l = \frac{\tau_0}{l} \cdot \varsigma_u \tag{8}$$

Here τ_0 is a tunable hyperparameter. High-uncertainty users get larger thresholds and thus shallower (coarser) hierarchies, while low-uncertainty users develop deeper (finer) ones. This stopping rule realizes a per-user adaptive information bottleneck, suppressing noisy fine-grained factors when ς_u is large.

The codebooks are optimized via the standard VQ-VAE objective together with a diversity loss:

$$\mathcal{L}_{Quant} = \sum_{u \in \mathcal{U}} \sum_{l=1}^{D_u} \left(\|\text{sg}(\mathbf{r}_u^{(l-1)}) - \mathbf{c}_u^{(l)}\|_2^2 + \beta_q \|\mathbf{r}_u^{(l-1)} - \text{sg}(\mathbf{c}_u^{(l)})\|_2^2 \right) \tag{9}$$

where $\text{sg}(\cdot)$ is the stop-gradient operator and β_q is the commitment cost. A diversity loss further encourages separation between intent codes within the same level:

$$\mathcal{L}_{Div} = \sum_{l=1}^{L} \sum_{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}_l, i \neq j} \max(0, \delta - \|\mathbf{c}_i - \mathbf{c}_j\|_2) \tag{10}$$

4.4 Uncertainty-aware Hierarchical Intent Aggregation

We aggregate the user-specific intent hierarchy $\{\mathbf{c}_u^{(l)}\}_{l=1}^{D_u}$ via an uncertainty-gated attention mechanism that prioritizes coarser intents for uncertain users:

$$\alpha_l = \frac{\exp\left(\frac{\mathbf{h}_u^\top \mathbf{c}_u^{(l)}}{\sqrt{d}} - \gamma \cdot l \cdot \varsigma_u \right)}{\sum_{k=1}^{D_u} \exp\left(\frac{\mathbf{h}_u^\top \mathbf{c}_u^{(k)}}{\sqrt{d}} - \gamma \cdot k \cdot \varsigma_u \right)} \tag{11}$$

where γ is a scaling factor. The term $\exp(-\gamma \cdot l \cdot \varsigma_u)$ penalizes finer-grained intents (larger l) for users with high uncertainty. The aggregated intent representation is $\mathbf{i}_u = \sum_{l=1}^{D_u} \alpha_l \mathbf{c}_u^{(l)}$, and the final user representation fuses collaborative and intent embeddings: $\mathbf{u}_u^* = \mathbf{h}_u + \mathbf{i}_u$.

4.5 Prediction and Optimization

The prediction score is $\hat{y}_{ui} = (\mathbf{u}_u^*)^\top \mathbf{h}_i$. We adopt the Bayesian Personalized Ranking (BPR) loss as the recommendation objective:

$$\mathcal{L}_{Rec} = \sum_{(u,i,j) \in \mathcal{D}} -\log \text{sigmoid}(\hat{y}_{ui} - \hat{y}_{uj}) \tag{12}$$

where $\mathcal{D} = \{(u, i, j) | u \in \mathcal{U}, i \in \mathcal{I}_u, j \in \mathcal{I} \setminus \mathcal{I}_u\}$ is the set of training triplets. We group the auxiliary losses by module:

$$\mathcal{L}_{Uncertainty} = \mathcal{L}_{FM} + \lambda_{cross} \mathcal{L}_{cross} \tag{13}$$

$$\mathcal{L}_{Intent} = \mathcal{L}_{Quant} + \lambda_{div} \mathcal{L}_{Div} \tag{14}$$

The final UHIFlow objective is:

$$\mathcal{L} = \mathcal{L}_{Rec} + \mu_1 \mathcal{L}_{Uncertainty} + \mu_2 \mathcal{L}_{Intent} \tag{15}$$

where μ_1, μ_2 are hyperparameters balancing the two auxiliary losses.

5 Experiments

5.1 Experimental Settings

Datasets Following [7,14], we evaluate on three widely-used multimodal benchmarks under 5-core filtering, all with pre-extracted visual (V) and textual (T) features. The Amazon datasets are from [23], and TikTok is from [33]. Table 1 summarizes their statistics.

Table 1. Statistics of the experimental datasets.

Statistics	Amazon-Baby	Amazon-Sports	TikTok
#Users	19,445	35,598	9,319
#Items	7,050	18,357	6,710
#Interactions	139,110	256,308	59,541
Sparsity	99.899%	99.961%	99.904%

Baseline Methods We compare UHIFlow against representative baselines grouped by modeling principle: **(1) Traditional CF**: NGCF [31], LightGCN [9]; **(2) Intent-based**: LIP [13], HIM [40], DCCF [27], BIGCF [38], MMIL [37]; **(3) Uncertainty-aware**: WaPOIR [39], UCC [19], UA-GNN [1]; **(4) Generative**: FlowCF [17], DiffMM [14], DiffCL [29]; and **(5) Multimodal**: LGMRec [7], MENTOR [36], MIG-GT [10], D-DPDG [34], MoToRec [18].

Evaluation Metrics Similar to other works in multimodal recommendation [7,36], we adopt two widely-used metrics to evaluate top-K recommendation performance: Recall@K and Normalized Discounted Cumulative Gain (NDCG@K). We report results for K={10, 20}. For each user, we rank all items they have not interacted with in the training set and evaluate the ranking performance on the test set.

Implementation Details We implement UHIFlow in PyTorch and run all experiments on NVIDIA A100 GPUs. For fairness, all models use embedding size $d = 64$; UHIFlow is trained with Adam using learning rate 10^{-3}, weight decay 10^{-4}, and batch size 2048. We use $K = 3$ graph propagation layers, a two-layer MLP vector field with hidden size 64 for CUSM, an adaptive dopri5 ODE solver with relative tolerance $1e-5$, a single linear cross-modal projection head, and UHIG depth $L = 4$ with codebook sizes $\{128, 256, 512, 1024\}$ and commitment cost 0.25. We tune $\mu_1, \mu_2 \in \{0.01, \ldots, 1.0\}$, $\lambda_{cross} \in \{0.1, \ldots, 2.0\}$, and $\lambda_{div} \in \{0.001, \ldots, 0.1\}$ by grid search; all baselines are run with official code and fine-tuned parameters, and results are averaged over 5 random seeds.

5.2 Overall Performance Comparison

The main results are presented in Table 2. We summarize the key observations below.

Table 2. Overall performance comparison on the three datasets. **Bold** indicates the best and underline indicates the second-best result. 'Improv.' denotes the relative improvement of UHIFlow over the best baseline. All improvements are significant ($p < 0.05$).

Category	Method	Amazon-Baby				Amazon-Sports				TikTok			
		R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
Traditional CF	NGCF	0.0388	0.0591	0.0208	0.0261	0.0468	0.0695	0.0250	0.0318	0.0366	0.0604	0.0159	0.0238
	LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0396	0.0653	0.0189	0.0282
Uncertainty	WaPOIR	0.0496	0.0752	0.0260	0.0325	0.0588	0.0868	0.0309	0.0375	0.0412	0.0672	0.0177	0.0270
	UCC	0.0512	0.0780	0.0271	0.0340	0.0605	0.0897	0.0321	0.0391	0.0420	0.0689	0.0186	0.0280
	UA-GNN	0.0531	0.0809	0.0281	0.0353	0.0622	0.0923	0.0330	0.0402	0.0432	0.0708	0.0190	0.0287
Intent-based	LIP	0.0617	0.0952	0.0336	0.0427	0.0702	0.1055	0.0383	0.0472	0.0576	0.0962	0.0275	0.0402
	HIM	0.0556	0.0858	0.0303	0.0384	0.0656	0.0986	0.0358	0.0442	0.0454	0.0758	0.0217	0.0317
	DCCF	0.0547	0.0848	0.0301	0.0383	0.0648	0.0978	0.0356	0.0441	0.0443	0.0744	0.0216	0.0314
	BIGCF	0.0574	0.0882	0.0310	0.0392	0.0672	0.1005	0.0363	0.0447	0.0477	0.0792	0.0222	0.0328
	MMIL	0.0601	0.0916	0.0319	0.0399	0.0678	0.1006	0.0359	0.0438	0.0531	0.0871	0.0235	0.0354
Generative	FlowCF	0.0561	0.0870	0.0309	0.0393	0.0662	0.1000	0.0364	0.0452	0.0468	0.0787	0.0229	0.0332
	DiffMM	0.0623	0.0975	0.0328	0.0411	0.0683	0.1019	0.0374	0.0455	0.0684	0.1129	0.0306	0.0456
	DiffCL	0.0641	0.0987	0.0343	0.0433	0.0754	0.1095	0.0421	0.0509	0.0677	0.1112	0.0299	0.0451
Multimodal	LGMRec	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0654	0.1072	0.0288	0.0435
	D-DPDG	0.0655	0.0993	0.0344	0.0429	0.0742	0.1096	0.0389	0.0473	0.0674	0.1098	0.0289	0.0441
	MIG-GT	0.0665	0.1021	0.0361	0.0452	0.0753	0.1130	0.0414	0.0510	0.0678	0.1105	0.0292	0.0444
	MENTOR	0.0678	0.1048	0.0362	0.0450	0.0763	0.1139	0.0409	0.0511	0.0691	0.1120	0.0290	0.0446
	MoToRec	0.0676	0.1027	0.0358	0.0451	0.0747	0.1109	0.0413	0.0504	0.0648	0.1070	0.0293	0.0436
Ours	**UHIFlow**	**0.0698**	**0.1079**	**0.0373**	**0.0465**	**0.0786**	**0.1171**	**0.0433**	**0.0526**	**0.0712**	**0.1162**	**0.0315**	**0.0469**
Improv.		2.95%	2.96%	3.04%	2.88%	3.01%	2.81%	2.85%	2.94%	3.04%	2.92%	2.94%	2.85%

UHIFlow achieves the best performance across all datasets and metrics, improving over the strongest baseline (MENTOR, AAAI'25) by 2.81%–3.04%. The gains are consistent on both sparse Amazon catalogs and the noisier TikTok micro-video domain, and UHIFlow also surpasses recent competitors such as MoToRec and D-DPDG on every metric. This supports our core design: modeling multimodal uncertainty and adaptive hierarchical intent jointly provides a stable signal, allowing uncertain cases to back off to coarse intents while confident cases refine toward item-level granularity.

Category-level results further show that advanced multimodal, intent-based, and generative methods outperform traditional CF. MENTOR and MIG-GT are strong multimodal competitors, and LIP is the strongest intent-based baseline, but they still rely on flat or fixed intent structures. FlowCF, DiffMM, and DiffCL remain competitive, yet mainly use flow or diffusion for denoising rather than for *quantifying* residual uncertainty. UHIFlow closes these gaps by coupling uncertainty estimation, cross-modal synergy, and user-specific hierarchy construction.

5.3 Ablation Study

To dissect our model and understand the contribution of its key components, we conduct an ablation study with several variants of UHIFlow. The results are presented in Table 3.

- **w/o CIP**: We remove the Collaborative Interest Propagation layers and use only the initial multimodal features.
- **w/o UHIG**: We remove the entire Uncertainty-guided Hierarchical Intent Generation module. The final user representation is simply the GNN embedding $\mathbf{h}_u$.
- **w/o CUSM**: We remove the Cross-modal Uncertainty Synergistic Modeling module. Consequently, user uncertainty is unavailable, and we use a fixed-depth hierarchy ($D_u = 2$) for all users.
- **w/o cross**: We disable the cross-modal synergistic loss ($\mathcal{L}_{cross}$) in CUSM, so uncertainties are estimated independently.
- **w/o hierarchy**: We replace the hierarchical intent generation with a flat, single-level intent codebook.

Table 3. Ablation study of UHIFlow's core components on three datasets, reporting Recall@20. Performance drops from the full model are highlighted in red.

Variant	Amazon-Baby	Amazon-Sports	TikTok
UHIFlow (Full Model)	**0.1079**	**0.1171**	**0.1162**
w/o CIP	0.0981 (-0.0098)	0.1049 (-0.0122)	0.1056 (-0.0106)
w/o UHIG	0.1008 (-0.0071)	0.1079 (-0.0092)	0.1085 (-0.0077)
w/o CUSM	0.1015 (-0.0064)	0.1092 (-0.0079)	0.1093 (-0.0069)
w/o cross	0.1060 (-0.0019)	0.1146 (-0.0025)	0.1119 (-0.0043)
w/o hierarchy	0.1040 (-0.0039)	0.1127 (-0.0044)	0.1106 (-0.0056)

Removing any component degrades performance. The largest drops occur for **w/o CIP** and **w/o UHIG**, highlighting the necessity of collaborative propagation and intent modeling, while the decline for **w/o CUSM** confirms that uncertainty-guided adaptation beats a fixed strategy. Each component, from the base GNN to synergistic uncertainty modeling and hierarchical generation, thus plays an integral role in UHIFlow. The two core modules are moreover complementary: CUSM supplies the per-user uncertainty that UHIG consumes to set its depth, so ablating CUSM not only removes the uncertainty signal but also collapses the hierarchy to a fixed granularity—which is why its drop is larger than that of the cross-modal term alone.

5.4 Impact of Uncertainty Modeling Strategy

Table 4. Performance comparison.

Strategy	Amazon-Baby		Amazon-Sports	
	R@20	N@20	R@20	N@20
Gaussian Embedding	0.1032	0.0443	0.1129	0.0500
Wasserstein Distance	0.1048	0.0452	0.1143	0.0509
Energy-Based Model	0.1064	0.0458	0.1156	0.0516
Flow Matching (Ours)	**0.1079**	**0.0465**	**0.1171**	**0.0526**

To demonstrate the superiority of Flow Matching for uncertainty quantification, we compare UHIFlow with variants employing alternative uncertainty modeling strategies:

– **Gaussian Embedding (GD)**: Models feature uncertainty by mapping items to Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ and using KL divergence for alignment.
– **Wasserstein Distance (WD)** [39]: Uses Wasserstein distance to measure the discrepancy between distributions, as used in WaPOIR.
– **Energy-Based Model (EBM)**: Quantifies uncertainty via energy scores derived from an implicit energy function.

Table 4 shows that **Flow Matching** consistently outperforms other strategies. Gaussian and Wasserstein methods improve over deterministic baselines but rely on restrictive assumptions (e.g., Gaussianity), while EBMs are more flexible yet unstable to train. Flow Matching strikes the best balance, giving stable, expressive, and accurate uncertainty quantification for hierarchical intent generation.

5.5 Parameter Sensitivity Analysis

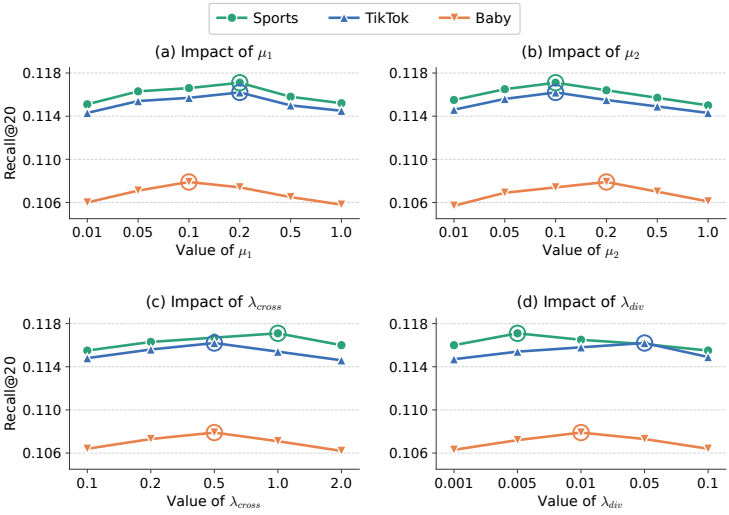

Fig. 3. Hyperparameter sensitivity analysis of UHIFlow on the Baby, Sports, and TikTok datasets. We report Recall@20 as we vary the values of four key hyperparameters: (a) uncertainty loss weight μ_1, (b) intent loss weight μ_2, (c) cross-modal guidance weight λ_{cross}, and (d) intent diversity weight λ_{div}.

We study four key hyperparameters: the loss weights μ_1 (uncertainty), μ_2 (intent), λ_{cross} (cross-modal guidance), and λ_{div} (intent diversity), varying one at a time while keeping the others optimal. As Figure 3 reports on Recall@20, performance first rises then falls as each weight increases, reflecting a trade-off between the auxiliary tasks and the main objective. UHIFlow stays robust across a reasonable range (e.g., $\mu_1, \mu_2 \in [0.1, 0.5]$), confirming its stability and practicality.

5.6 Efficiency Analysis

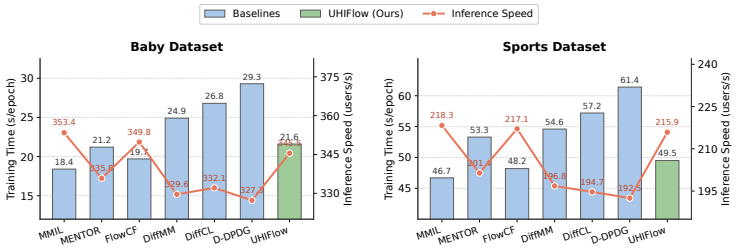

Fig. 4. Efficiency comparison on the Baby and Sports datasets. Bars represent training time per epoch (left y-axis, lower is better), and the line represents inference speed (right y-axis, higher is better).

We assess UHIFlow's computational efficiency against representative non-generative baselines (MMIL, MENTOR) and same-category generative baselines—the flow-based FlowCF and the diffusion-based DiffMM, DiffCL, and D-DPDG—on an NVIDIA A100 GPU. As shown in Figure 4, the diffusion-based methods (DiffMM, DiffCL, D-DPDG) incur the highest training and inference cost because of their iterative multi-step denoising. In contrast, UHIFlow's flow matching introduces only a moderate training overhead—on par with the non-generative models and consistently lower than every diffusion-based competitor—while its inference speed stays close to the fastest baselines, as the expensive flow computation is performed offline during training. This favorable balance between accuracy and computational cost confirms UHIFlow's suitability for practical deployment.

6 Conclusion

We proposed UHIFlow to address the inherent uncertainty and static intent structures of multimodal recommendation. Its Cross-modal Uncertainty Synergistic Modeling (CUSM) module uses conditional flow matching to quantify and synergize multimodal uncertainties, which then adaptively control the depth of an Uncertainty-guided Hierarchical Intent Generation (UHIG) module that builds personalized, fine-to-coarse intent structures per user. Extensive experiments demonstrate UHIFlow's superiority over state-of-the-art baselines. Future work will explore sample-efficient uncertainty estimation from limited-feedback signals [24].

Acknowledgments. This work was supported by the Open Project Program of Marine Ecological Restoration and Smart Ocean Engineering Research Center of Hebei Province under Grant No. HBMESO2507; the Shijiazhuang–Northeastern University Science and Technology Cooperation Special Project under Grant No. NEUS2025-01-003; the Scientific Research Project of Hebei Education Department under Grant No. QN2024167; and the National College Student Innovation and Entrepreneurship Training Program under Project No. 202619145026.

References

1. Cao, M., Tran, T.: Uncertainty-aware graph neural network for semi-supervised diversified recommendation. Social Network Analysis and Mining **14**(92) (2024)
2. Chen, Y., Liu, Z., Li, J., McAuley, J., Xiong, C.: Intent contrastive learning for sequential recommendation. In: Proceedings of the ACM Web Conference 2022. pp. 2172–2182 (2022)
3. Chen, Z., Chang, H., Liu, T., Zhou, C., Cao, Y., Ding, J., Liu, M., Qin, B.: Beyond the flat sequence: Hierarchical and preference-aware generative recommendations. In: Proceedings of the ACM Web Conference 2026 (WWW '26) (2026)
4. Cui, X., Lu, W., Tong, Y., Li, Y., Zhao, Z.: Diffusion-based multi-modal synergy interest network for click-through rate prediction. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 581–591 (2025)
5. Cui, X., Lu, W., Tong, Y., Li, Y., Zhao, Z.: Multi-modal multi-behavior sequential recommendation with conditional diffusion-based feature denoising. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1593–1602 (2025)
6. Gao, Z., Jiang, X., Xu, X., Shen, F., Li, Y., Shen, H.T.: Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26866–26875 (2024)
7. Guo, Z., Li, J., Li, G., Wang, C., Shi, S., Ruan, B.: LGMRec: Local and global graph learning for multimodal recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 8454–8462 (2024)
8. He, R., McAuley, J.: VBPR: Visual bayesian personalized ranking from implicit feedback. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 30 (2016)
9. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: LightGCN: Simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 639–648 (2020)
10. Hu, J., Hooi, B., He, B., Wei, Y.: Modality-independent graph neural networks with global transformers for multimodal recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 11790–11798 (2025)
11. Hu, P., Lu, W., Wang, J.: From IDs to semantics: A generative framework for cross-domain recommendation with adaptive semantic tokenization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 14874–14882 (2026)
12. Jannach, D., Zanker, M.: A survey on intent-aware recommender systems. ACM Transactions on Recommender Systems **3**(2), 1–32 (2024)
13. Jian, M., Li, R., Wang, T., Shi, G., Wu, L.: Large multimodal model-based intent prompting learning for multimedia recommendation. Information Fusion **127**, 103881 (2026)
14. Jiang, Y., Xia, L., Wei, W., Luo, D., Lin, K., Huang, C.: DiffMM: Multi-modal diffusion model for recommendation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7591–7599 (2024)
15. Li, C., Liu, Z., Wu, M., Xu, Y., Zhao, H., Huang, P., Kang, G., Chen, Q., Li, W., Lee, D.L.: 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)

16. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: Proceedings of the International Conference on Learning Representations (2023)
17. Liu, C., Zhang, Y., Wang, J., Ying, R., Caverlee, J.: Flow matching for collaborative filtering. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1765–1775 (2025)
18. Liu, J., Zhang, Z., Cheung, R.C.C.: Motorec: Sparse-regularized multimodal tokenization for cold-start recommender. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 15324–15332. AAAI Press (2026)
19. Liu, T., Gao, C., Wang, Z., Li, D., Hao, J., Jin, D., Li, Y.: Uncertainty-aware consistency learning for cold-start item recommendation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2466–2470 (2023)
20. Liu, X., Liu, Y., Ma, J., Ma, Y., Xia, J., Yu, S., Zhang, K., Zhong, W., Zhu, S.: End-to-end learnable clustering for intent learning in recommendation. In: Advances in Neural Information Processing Systems 37. pp. 5913–5949 (2024)
21. Liu, Z., Lu, W.: MDN: Modality decomposition network for multimodal recommendation. In: Proceedings of the 2025 International Conference on Multimedia Retrieval. pp. 871–879 (2025)
22. Lu, W., Yin, L.: DMMD4SR: Diffusion model-based multi-level multimodal denoising for sequential recommendation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 6363–6372 (2025)
23. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 43–52 (2015)
24. Miao, Y.: Universal refinement without interaction: Order-optimal 1-bit mean estimation. arXiv preprint arXiv:2607.24358 (2026)
25. Miao, Y., Cui, M., Zhu, Y., Wang, Y., Xu, S.: R3-REC: Reasoning-driven recommendation via retrieval-augmented LLMs over multi-granular interest signals. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5951–5955 (2026)
26. Qi, T., Wu, F., Wu, C., Yang, P., Yu, Y., Xie, X., Huang, Y.: Hierec: Hierarchical user interest modeling for personalized news recommendation. arXiv preprint arXiv:2106.04408 (2021)
27. Ren, X., Xia, L., Zhao, J., Yin, D., Huang, C.: Disentangled contrastive collaborative filtering. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1137–1146 (2023)
28. Shi, Y., Miao, Y., Hu, X., Wang, Z., Han, C.: MOTIF: Motivation-guided topology inference for cold-start multimodal recommendation. arXiv preprint arXiv:2608.25381 (2026)
29. Song, Q., Hu, J., Xiao, L., Sun, B., Gao, X., Li, S.: DiffCL: A diffusion-based contrastive learning framework with semantic alignment for multimodal recommendations. IEEE Transactions on Neural Networks and Learning Systems **36**(10), 18587–18597 (2025)
30. Wang, W., Xu, Y., Feng, F., Lin, X., He, X., Chua, T.S.: Diffusion recommender model. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 832–841 (2023)
31. Wang, X., He, X., Wang, M., Feng, F., Chua, T.S.: Neural graph collaborative filtering. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 165–174 (2019)

32. Wei, W., Huang, C., Xia, L., Zhang, C.: Multi-modal self-supervised learning for recommendation. In: Proceedings of the ACM Web Conference 2023. pp. 790–800 (2023)
33. Wei, Y., Wang, X., Nie, L., He, X., Hong, R., Chua, T.S.: MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1437–1445 (2019)
34. Wu, J., Zheng, Y., Zhang, T., Jing, S., Liu, J., Guo, S., Deng, F.: D-DPDG: Diffusion-based dual-graph attention with dual-path feature extraction for multi-modal recommendation. Journal of Intelligent Information Systems **64**(2), 559–595 (2026)
35. Wu, Z., Wang, X., Chen, H., Li, K., Han, Y., Sun, L., Zhu, W.: Diff4Rec: Sequential recommendation with curriculum-scheduled diffusion augmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 9329–9335 (2023)
36. Xu, J., Chen, Z., Yang, S., Li, J., Wang, H., Ngai, E.C.H.: MENTOR: Multi-level self-supervised learning for multimodal recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 12908–12917 (2025)
37. Yang, W., Yang, Q.: Multimodal-aware multi-intention learning for recommendation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 5663–5672 (2024)
38. Zhang, Y., Sang, L., Zhang, Y.: Exploring the individuality and collectivity of intents behind interactions for graph collaborative filtering. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1253–1262 (2024)
39. Zhou, F., Qian, T., Mo, Y., Cheng, Z., Xiao, C., Wu, J., Trajcevski, G.: Uncertainty-aware heterogeneous representation learning in poi recommender systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems **53**(7), 4522–4535 (2023)
40. Zhu, N., Cao, J., Lu, X., Xiong, H.: Learning a hierarchical intent model for next-item recommendation. ACM Transactions on Information Systems **40**(2), 1–28 (2022)