Personalized Diffusion-based Outfit Recommendation with Diffusion-Aware Stochastic Perturbation

1st Hangjie Xia
Zhejiang Sci-Tech University
Hangzhou, China
2024220603085@mails.zstu.edu.cn

2nd Yifan Huo
Zhejiang Sci-Tech University
Hangzhou, China
2024010602002@mails.zstu.edu.cn

3rd Junhong Zheng
Zhejiang Sci-Tech University
Hangzhou, China
zjhist@zstu.edu.cn

4th Lili He*
Zhejiang Sci-Tech University
Hangzhou, China
llhe@zju.edu.cn

***Abstract*—Personalized outfit recommendation requires modeling both item compatibility and user preferences. However, given a specific context, multiple valid completion options often exist, and traditional point estimation models struggle to capture this "one-to-many" multimodal uncertainty. Existing diffusion models typically rely on static conditional guidance, overlooking noise variations during the denoising process and lacking domain-specific knowledge constraints. This paper proposes the Personalized Diffusion-based Outfit Recommendation (PDOR) framework, which formulates outfit completion as conditional probability distribution learning within a latent space. PDOR injects dynamic perturbations via a diffusion-aware stochastic perturbation mechanism during denoising to learn robust semantic dependencies. It utilizes compatibility reward guidance to provide explicit stylistic constraints during training and employs adaptive uncertainty regularization to prevent feature convergence. The framework integrates a "generation-retrieval" paradigm, which achieves a balance between recommendation precision and inference efficiency by generating an "ideal prototype" in a continuous latent space followed by nearest neighbor retrieval. Experimental results demonstrate that PDOR significantly outperforms existing methods in compatibility prediction (CP AUC: 0.9074), outfit completion (FITB: 0.7117), and retrieval (HR@10: 0.3264) tasks on the iFashion and Polyvore-630 datasets. Ablation studies confirm the necessity of each module and the critical role of the perturbation mechanism.**

***Index Terms*—Personalized Recommendation; Fashion Outfit; Diffusion Model; Uncertainty Modeling; Stochastic Perturbation; Multimodal Learning**

I. Introduction

With the rapid development of online fashion e-commerce platforms, personalized outfit recommendation has become a key technology for enhancing user experience and commercial conversion. Unlike traditional single-item recommendation, outfit recommendation faces a complex dual challenge: the system must understand the objective visual and stylistic compatibility between items (e.g., the conflict between striped tops and plaid pants) while simultaneously capturing a user's unique fashion preferences (e.g., a preference for minimalist or vibrant styles). This interleaving of compatibility modeling and personalization makes outfit recommendation a highly challenging problem in multimodal recommendation systems.

The fundamental difficulty of personalized outfit recommendation lies in its inherent multimodal uncertainty. Given the same user history and partial outfit context, there are usually multiple reasonable completion schemes. For example, if a user's history shows a dual preference for minimalist and vintage styles, the same pair of jeans could be paired with either a white T-shirt or a plaid shirt. In the feature space, this manifests as a complex multimodal distribution, which traditional deterministic point estimation methods struggle to handle effectively.

Existing mainstream methods are primarily divided into two categories. The first category is based on Graph Neural Networks or Transformer architectures [1]–[3], but their point estimation paradigm cannot quantify predictive uncertainty. The second category attempts to introduce generative frameworks such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs). However, VAEs are often limited by overly simplistic prior assumptions (such as a single Gaussian distribution), making it difficult to fit complex probability density variations in the fashion semantic space. GANs, while capturing diverse distributions, often face training instability and severe mode collapse. In contrast, diffusion probabilistic models [4] can refine simple Gaussian noise into highly diverse distributions by simulating a gradual denoising process, which naturally aligns with the mathematical requirements for capturing multimodal distributions and representing outfit uncertainty. However, most existing diffusion models employ a "static" conditional injection mechanism [5], [6], where contextual information is introduced in a fixed form throughout the generation process. This ignores the differences in noise

* Corresponding author.

states at different time steps, making it difficult for the model to learn semantic dependencies that are robust to contextual variations. Furthermore, most generative models focus only on reconstruction accuracy and lack explicit guidance from domain knowledge, such as compatibility constraints.

To overcome these challenges, we propose the Personalized Diffusion-based Outfit Recommendation (PDOR) framework. As shown in Fig. 1, the framework reformulates the outfit completion task as conditional probability distribution learning in a latent space. Within this framework, we design a diffusion-aware stochastic perturbation mechanism that achieves continuous data augmentation by injecting dynamic perturbations during denoising. This forces the generator to learn semantic-level dependencies that are insensitive to minor contextual changes, thereby enhancing the robustness of the generation process. Simultaneously, PDOR introduces an adaptive uncertainty regularization strategy, utilizing a dynamically balanced loss scheduler to ensure that generated embeddings satisfy both geometric accuracy and semantic distinctiveness, effectively preventing feature convergence. Additionally, we adopt a two-stage training paradigm guided by compatibility rewards, utilizing a frozen discriminator to provide explicit stylistic constraints and ensure high consistency in the generated prototypes. Finally, the framework combines a "generation-retrieval" paradigm [7], which elegantly sidesteps the computational bottleneck of direct sampling in large-scale discrete spaces by locating the ideal prototype in a continuous space and performing similarity retrieval.

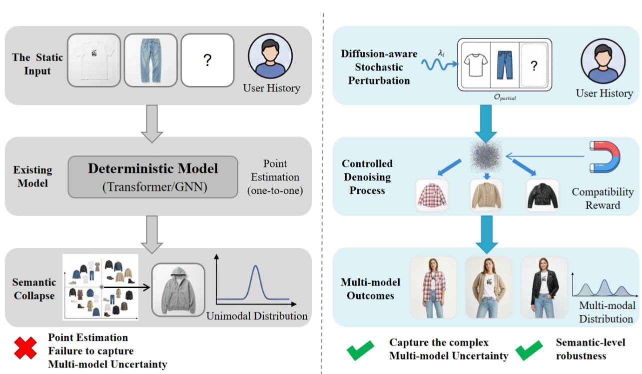

Fig. 1. Comparison between traditional point estimation and the PDOR framework. (Left) Existing models suffer from semantic collapse due to their deterministic nature. (Right) PDOR captures the multimodal distribution by injecting stochastic perturbations (λ_i) and utilizing compatibility rewards, sampling diverse and style-consistent ideal prototypes from the latent space.

The main contributions of this paper are summarized as follows:

- We apply a diffusion probabilistic model to the personalized outfit completion task, effectively capturing multimodal uncertainty in fashion outfits by learning conditional probability distributions.
- We design a diffusion-aware stochastic perturbation mechanism that breaks through the limitations of static conditional guidance in existing diffusion models, significantly improving the robustness of the generation process to contextual changes.
- We propose a two-stage training paradigm guided by compatibility rewards, explicitly injecting discriminative compatibility knowledge into the generation process to ensure stylistic consistency while maintaining diversity.
- Extensive experiments on two large-scale benchmark datasets, iFashion and Polyvore-630, demonstrate that PDOR significantly outperforms existing mainstream models in core tasks such as compatibility prediction, outfit completion, and personalized retrieval.

II. Related Work

A. Fashion Recommendation and Compatibility Modeling

Fashion recommendation research primarily involves general preference modeling and domain-specific compatibility learning. Early works such as BPR [8] capture implicit user-item interactions through matrix factorization, LightGCN [9] simplifies graph convolution operations to improve large-scale recommendation efficiency, and methods like VBPR [10] explicitly incorporate visual features for multimodal modeling.

Regarding compatibility modeling, early methods like SiameseNet [11] focused on pairwise metrics but ignored the overall outfit cohesion. Han *et al.* [12] used bidirectional LSTMs to model sequences, but introduced sequential assumptions that contradict the unordered nature of fashion outfits. SetNN [13] emphasized global consistency through multi-instance pooling, while OutfitTransformer [3] leveraged self-attention mechanisms [14] to capture complex interactions, becoming a mainstream framework. Recent works have explored history-aware transformers for personalized outfit recommendation [15] and hybrid-hierarchical graph attention networks that jointly model compatibility and user preferences [16]. However, these deterministic models lack uncertainty quantification mechanisms; a single point estimation struggles to handle blurred style boundaries and labeling noise, limiting their robustness in complex scenarios.

B. Diffusion Models in Recommender Systems

Diffusion models provide a probabilistic perspective for modeling complex user interests and alleviating data sparsity by recovering data distributions through gradual denoising [4], [17], [18]. DiffuRec [6] first modeled sequential recommendation as a denoising process in latent space, and DiffAug [5], [19] mitigated sparsity by generating synthetic interactions. In the fashion domain, POG [20] combined VAEs with adversarial learning, DiFashion [21] explored pixel-space synthesis, and recent work FashionDPO [22] proposed fine-tuning fashion outfit generation models using direct preference optimization, but efficient latent-space personalized outfit completion remains an open challenge.

A core limitation of current diffusion-based recommendation is "static" conditional guidance. Most works inject conditions through simple concatenation or fixed cross-attention [5], [6]. However, at different time steps t, the noise intensity of the latent variable x_t fluctuates drastically, and fixed injection leads to "semantic disconnect." The diffusion-aware perturbation mechanism proposed in this paper learns more robust semantic dependencies by dynamically modulating conditional signals. Recent work on personalized preference fine-tuning of diffusion models [23] has shown promise in aligning generation with individual user preferences through multi-reward optimization.

Furthermore, existing models mostly focus on geometric reconstruction and lack explicit domain knowledge guidance. While Hg-PDC [24] utilizes hypergraphs to capture high-order correlations, deterministic message passing limits retrieval recall. This paper models outfit completion as conditional probability distribution learning and designs a two-stage training strategy using a compatibility discriminator as a reward function to guide generation. We employ a "generation-retrieval" paradigm [7] to bypass discrete sampling bottlenecks, ensuring personalized accuracy and stylistic compatibility.

C. Multimodal Learning in the Fashion Domain

The dual visual and textual attributes of fashion data determine the importance of multimodal representation [25], [26]. CLIP [27] learns a unified semantic space by aligning image-text pairs, and FashionCLIP [28] optimizes for fine-grained fashion concepts. Sevegnani *et al.* [29] utilized supervised contrastive loss to optimize feature distributions, but existing work often treats contrastive learning as an independent discriminative objective, with less exploration of its dynamic utilization during the generation process. Deeply fusing personalized preferences with multimodal compatibility within a diffusion framework remains a key challenge [30].

III. METHODOLOGY

We propose the Personalized Diffusion-based Outfit Recommendation (PDOR) framework, as illustrated in Fig. 2. This framework aims to overcome the limitations of deterministic models in capturing multimodal uncertainty and consists of three core components: a multimodal feature encoding layer, a Compatibility Discriminator (CD), and a Latent Diffusion Generator (LDG). The framework adopts a two-stage joint training strategy that integrates adaptive uncertainty regularization with CP-reward guidance.

Let $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively. The historical interaction sequence for each user $u \in \mathcal{U}$ is denoted as $\mathcal{H}_u = \{i_1, i_2, ..., i_L\}$. An outfit $\mathcal{O}$ is a set of N items, i.e., $\mathcal{O} = \{o_1, o_2, ..., o_N\}$. The goal of the recommendation task is: given the user history $\mathcal{H}_u$ and a partially observed outfit (context) $\mathcal{O}_{partial}$, predict and complete the missing target item to form a complete and personalized outfit.

A. Multimodal Encoding and Compatibility Discriminator

To construct a unified multimodal latent space, we first utilize the pre-trained CLIP model [27] to extract image features v_i and text features t_i for each item i. These are concatenated and mapped to a unified item embedding $e_i \in \mathbb{R}^d$ via a projection layer f_ϕ. This embedding serves as the shared input for the subsequent CD and LDG modules.

Based on the generated unified multimodal embeddings, the CD module focuses on assessing personalized compatibility. We posit that an ideal outfit recommendation depends not only on the objective visual coordination between items but also on the consistency between the outfit scheme and the user's individual style preferences. To capture both dimensions, we design differentiated modeling strategies and multi-level optimization objectives for the user history sequence and the outfit set.

1) Sequence and Set Encoding: For the user history $\mathcal{H}_u$, we concatenate the corresponding item embedding sequence $E_{\mathcal{H}_u}$ with a learnable user token u_{token} and feed it into a Transformer encoder [14]. The output at the u_{token} position is taken as the user preference representation h_u. For the outfit $\mathcal{O}$, we concatenate the outfit item embeddings $E_{\mathcal{O}}$ with a classification token **cls** and feed it into another Transformer encoder, taking the output at **cls** as the overall outfit representation h_o. Finally, h_u and h_o are concatenated and passed through a Multi-Layer Perceptron (MLP) to output the compatibility score $p_{cp} = \text{MLP}([h_u; h_o])$.

2) Discriminative Optimization and Multi-view Constraints: We employ Focal Loss [31] as the primary classification objective $\mathcal{L}_{focal}$, which effectively mitigates the problem of positive-negative sample imbalance by dynamically adjusting sample weights. To obtain a more robust feature space under sparse data, we design the following auxiliary losses to regularize representation learning.

First, the outfit-level contrastive loss $\mathcal{L}_{outfit}$ aims to pull apart different compatibility categories. We adopt a supervised contrastive learning strategy [32], [33], where outfits with compatible labels within the same batch are treated as positive pairs, while incompatible outfits are treated as negative pairs. This loss significantly enhances the separability of the feature space by optimizing the global manifold structure of outfits.

Second, the item-level diversity regularization loss $\mathcal{L}_{item}$ aims to prevent feature smoothing. During self-attention interactions in the Transformer, item features within an outfit often suffer from over-smoothing. To address this, we first L2-normalize the item embeddings, then calculate their average variance within the outfit and introduce a threshold penalty:

$$\mathcal{L}_{item} = \text{ReLU}(\nu - \text{Var}(E_{\mathcal{O}})) \tag{1}$$

where ν is a preset threshold. This regularization term aims to maintain the diversity of item features and prevent the loss of fine-grained information. Although these three objectives act on different semantic levels, they are complementary: $\mathcal{L}_{focal}$ focuses on direct discriminative accuracy, $\mathcal{L}_{outfit}$ optimizes the global manifold structure, and $\mathcal{L}_{item}$ maintains underlying feature diversity. The total loss for the CD module is:

$$\mathcal{L}_{CD} = \mathcal{L}_{focal} + \lambda_1 \mathcal{L}_{outfit} + \lambda_2 \mathcal{L}_{item} \tag{2}$$

where λ_1 and λ_2 are hyper-parameters to balance the loss terms. This optimization objective ensures that the model learns more discriminative and diverse feature representations while achieving accurate classification.

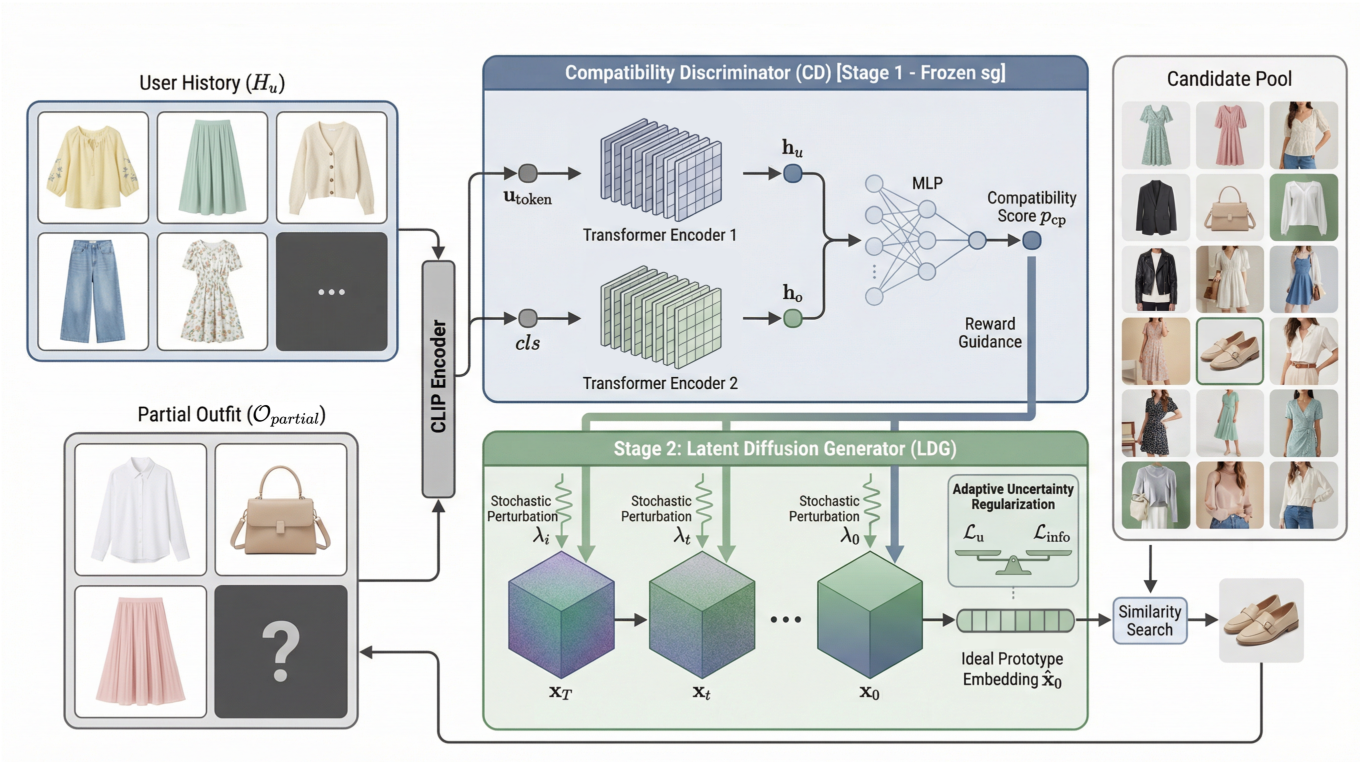

Fig. 2. Overall architecture of the PDOR framework. The system consists of three core components: (1) multimodal Feature Encoding Layer, which leverages CLIP to extract semantic representations of user history $\mathcal{H}_u$ and partial outfit $\mathcal{O}_{partial}$; (2) Compatibility Discriminator (CD), which models personalized compatibility scores p_{cp} and provides reward guidance; (3) Latent Diffusion Generator (LDG), generating the ideal prototype embedding $\hat{x}_0$ in the continuous latent space through diffusion-aware stochastic perturbations and adaptive uncertainty regularization. The final recommendation is obtained via similarity search from the candidate pool.

B. Latent Diffusion Generator

The LDG module models the outfit completion task as a conditional generation problem in the latent space. Unlike methods that perform diffusion directly in pixel space, we draw inspiration from latent diffusion models [34] and execute the diffusion process in the latent space of a pre-trained encoder, significantly reducing computational overhead. The goal is to learn the conditional distribution of the true target item embedding x_0 given the partial outfit context $\mathcal{O}_{partial}$ and user history $\mathcal{H}_u$. To enhance the model's robustness to contextual features, we introduce a diffusion-aware stochastic perturbation mechanism during the reverse denoising process.

1) Diffusion Process: The forward process gradually adds Gaussian noise to x_0 according to a variance schedule $\{\beta_t\}_{t=1}^T$ [4], eventually transforming it into $x_T \sim \mathcal{N}(0, I)$. The noise state at any time step t is:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon} \tag{3}$$

where $\bar{\alpha}_t$ is the cumulative noise schedule parameter. We adopt a cosine scheduling strategy [35], which, compared to linear scheduling, maintains a more uniform signal-to-noise ratio distribution on low-resolution features, avoiding information loss caused by excessive noise in early time steps.

2) Reverse Denoising and Diffusion-aware Stochastic Perturbation: The reverse process is executed by a denoising network Φ_θ, which adopts a U-Net-like [36] encoder-decoder architecture designed to recover x_0 from x_T. Regarding conditional processing, we treat user history $\mathcal{H}_u$ as a relatively stable global prior, encoded as a static vector h_u. For the local context represented by the partial outfit $\mathcal{O}_{partial}$, we design the Diffusion-aware Stochastic Perturbation mechanism for deep interaction.

Specifically, for each item embedding e_i in $\mathcal{O}_{partial}$, we do not input it directly into the network. Instead, we inject a stochastic perturbation related to the current diffusion state:

$$e_i' = e_i + \lambda_i (x_t + e_t) \tag{4}$$

where e_t is the time-step encoding. The weight $\lambda_i \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda)$ is a random variable sampled from a small-scale fixed Gaussian distribution. The physical intuition behind this mechanism is that by injecting the current noise state x_t into the context with random intensity, we implement continuous data augmentation in the feature space. This forces the generator to learn semantic-level dependencies that are insensitive to minor contextual variations, preventing the model from overfitting to fixed contextual feature values and significantly improving the robustness of the generation process. The modulated embeddings $E'_{\mathcal{O}_{partial}}$ are then encoded via a Transformer to obtain the conditional representation h_o^{mod}.

3) Generation and Inference: The denoising network Φ_θ takes h_o^{mod}, the user representation h_u, and the time-step embedding e_t as inputs to predict the target clean embedding:

$$\hat{x}_0 = \Phi_\theta(h_o^{mod}, h_u, e_t) \tag{5}$$

During the inference phase, we recover $\hat{x}_0^{gen}$ from noise x_T. We treat the generated $\hat{x}_0^{gen}$ as the "ideal prototype" of the target item in the continuous latent space. Finally, recommendation results are obtained by calculating the cosine similarity between this prototype and all item embeddings in the candidate pool for retrieval and ranking. This "generation-retrieval" paradigm [7] effectively sidesteps the computational bottleneck of direct sampling in large-scale discrete spaces.

C. Adaptive Uncertainty Regularization

To balance geometric reconstruction accuracy with semantic distinctiveness, we introduce an adaptive regularization strategy in LDG training.

1) Statistical Moment Augmentation: To simulate distribution shifts, we perform statistical moment-based augmentation on the target embedding x_0. By randomly scaling and shifting the feature mean and standard deviation within the batch, we construct an augmented view $\tilde{x}_0$:

$$\tilde{x}_0 = s \cdot \text{Norm}\left(\boldsymbol{\alpha}_{aug} \odot \frac{x_0 - \mu}{\sigma} + \boldsymbol{\beta}_{aug}\right) \tag{6}$$

This augmentation forces the model to maintain generation consistency under minor feature distribution changes, smoothing the feature manifold.

2) Dynamic Loss Balancing: The training of LDG involves two complementary objectives. The first is the uncertainty reconstruction loss ($\mathcal{L}_u$), used to constrain the generated embedding $\hat{x}_0$ to simultaneously approach the true embedding x_0 and its augmented view. This ensures accuracy from a geometric perspective:

$$\mathcal{L}_u = \frac{1}{2}(\|\hat{x}_0 - x_0\|^2 + \|\hat{x}_0 - \tilde{x}_0\|^2) \tag{7}$$

The second is the InfoNCE contrastive loss ($\mathcal{L}_{info}$) [37], [38], where we treat $(\hat{x}_0, x_0)$ as a positive pair and other items in the batch as negative samples. This enhances the instance discriminativeness of the generated embeddings from a semantic perspective, preventing the model from generating converged, average features. Since the scales of the two losses may differ, we design an adaptive balancing scheduler. It continuously monitors the ratio of the two losses $r = \mathcal{L}_u/\mathcal{L}_{info}$ and dynamically updates the balance weight γ_{bal} using an exponential moving average. The final generator loss is:

$$\mathcal{L}_{LDG} = \gamma_{bal}\mathcal{L}_u + (1 - \gamma_{bal})\mathcal{L}_{info} \tag{8}$$

This dynamic balancing mechanism ensures the stability and effectiveness of the training process.

D. Compatibility Reward Guided Generation

To ensure that the generated items are not only accurately reconstructed but also semantically and stylistically compatible with the context, we adopt a two-stage training strategy. After the CD module converges in the first stage, we freeze it and use it as a fixed reward function to guide the LDG training.

In the second stage, we combine the embedding $\hat{x}_0$ predicted by the LDG with the context $\mathcal{O}_{partial}$, feed them into the frozen CD module to calculate the compatibility probability p_{cp}, and construct the CP-reward loss:

$$\mathcal{L}_{reward} = -\log(p_{cp}) \tag{9}$$

By minimizing $\mathcal{L}_{reward}$, gradient information is backpropagated to the LDG's generation network Φ_θ, guiding it to search for item embedding regions in the latent space that yield high compatibility scores.

Ultimately, the total training objective of the LDG is composed of the generation loss and the reward loss:

$$\mathcal{L}_{total} = \mathcal{L}_{LDG} + \lambda_{reward}\mathcal{L}_{reward} \tag{10}$$

where λ_{reward} is a hyper-parameter controlling the strength of the reward signal. This mechanism effectively guides the generator to search for high-scoring regions in the latent space by introducing explicit feedback signals, ensuring that the final recommended outfits possess higher stylistic consistency and compatibility while maintaining diversity.

IV. EXPERIMENTS

This section details the experimental setup and provides a comprehensive performance comparison with baseline models, followed by ablation studies and parameter sensitivity analysis to explore the model's internal mechanisms.

A. Experimental Setup

1) Datasets: To comprehensively evaluate the performance of the PDOR framework in personalized recommendation and compatibility modeling tasks, we conducted experiments on two widely used fashion benchmark datasets:

- iFashion (POG) [20]: A large-scale real-world e-commerce dataset from Alibaba. It contains massive user interaction records and rich multimodal information, reflecting implicit outfit preferences in actual consumption scenarios, making it highly suitable for evaluating personalized outfit generation.
- Polyvore-630 [39]: A high-quality subset from the Polyvore community, where outfits were manually created by 630 fashionistas or experts [12], [39]. To build personalized user interaction models, we treat outfits created by the same user as that user's purchase history.

2) Evaluation Metrics: For different subtasks within the framework, we employ two sets of complementary metrics. For compatibility and completion tasks, we use AUC to evaluate the model's ability to distinguish between compatible and incompatible outfits, and Accuracy (FITB) to evaluate the Fill-In-The-Blank task (the proportion of correctly selected

TABLE I
STATISTICS OF THE DATASETS

Dataset	#Users	#Outfits	#Items
iFashion	3,569,112	1,013,136	583,464
Polyvore-630	630	150,380	205,234

missing items from candidates). For personalized retrieval, we simulate a candidate pool retrieval scenario using Top-K Hit Rate (HR@K) and Normalized Discounted Cumulative Gain (NDCG@K) (with $K = 10$) to measure recommendation accuracy and ranking quality.

3) Baselines: We selected representative mainstream models in the fashion recommendation field as baselines. These cover metric learning, sequence modeling, graph neural networks, Transformers, and diffusion probabilistic models.

Compatibility & Completion: These baselines are utilized to evaluate the model's performance in outfit compatibility prediction and outfit completion on both the iFashion and Polyvore-630 datasets.

- SiameseNet [11]: A classic visual metric learning method using shared-weight CNNs and Euclidean distance for compatibility.
- F-LSTM / Bi-LSTM [12] (RNN-based): RNN-based temporal modeling; standard sequence recommendation baselines.
- SetNN [13]: Set-based neural network using multi-instance pooling to emphasize unordered outfit consistency.
- POG [20]: An early fashion generative model combining VAEs with adversarial learning.
- OutfitTransformer [3]: The current mainstream baseline using Transformer encoders and self-attention for complex interactions.

Personalized Recommendation: These baselines are compared on the iFashion dataset to evaluate the model's capability for personalized outfit recommendation in large-scale real-world e-commerce scenarios.

- BPR-MF [8]: A classic collaborative filtering baseline optimizing Bayesian Personalized Ranking loss.
- LightGCN [9]: A lightweight graph convolutional network performing message passing on the user-item bipartite graph.
- VBPR [10]: A visual-augmented BPR method incorporating CNN-extracted visual features.
- HFGN [2]: A hierarchical fashion graph network modeling preferences through item-outfit-user heterogeneous graphs.
- Hg-PDC [24]: A hypergraph-based personalized diffusion framework capturing high-order correlations.

4) Implementation Details: The PDOR framework is implemented in PyTorch. The multimodal encoder uses a pretrained CLIP model [27] with frozen parameters. Experiments were conducted on four NVIDIA RTX A6000 GPUs using the AdamW optimizer [40], [41] with an initial learning rate of 10^{-4}. Data were split into training, validation, and test sets at an 8:1:1 ratio. We employed a mixed curriculum negative sampling strategy during training, sampling both random items for global differences and hard items of the same category for fine-grained style differences.

B. Overall Performance Analysis

1) Compatibility Prediction and Outfit Completion: We compared the ability to understand compatible outfits on iFashion and Polyvore-630. Table II reports CP AUC and FITB Accuracy.

TABLE II
PERFORMANCE COMPARISON ON COMPATIBILITY (CP) AND COMPLETION (FITB)

Model	iFashion		Polyvore-630	
	CP AUC	FITB ACC	CP AUC	FITB ACC
SetNN	0.5831	0.5016	0.5642	0.4842
SiameseNet	0.7087	0.5039	0.7703	0.5103
F-LSTM	0.6504	0.5807	0.6289	0.4619
Bi-LSTM	0.6861	0.6022	0.7840	0.5328
POG	0.8609	0.6871	-	-
OutfitTransformer	0.8864	0.6924	0.8413	0.5948
PDOR (Ours)	**0.9074**	**0.7117**	**0.8527**	**0.6165**

As shown in Table II, PDOR performs excellently across all metrics, particularly on iFashion with a CP AUC of 0.9074 and FITB accuracy of 0.7117. Compared to OutfitTransformer, PDOR improved the core completion task by approximately 1.93%. This advantage stems from the fundamental shift in the modeling paradigm: deterministic models map context to a single point, while PDOR uses latent diffusion to learn conditional probability distributions, more flexibly modeling complex non-linear relationships.

2) Personalized Recommendation: We compared recommendation baselines on iFashion. Table III reports HR@10 and NDCG@10.

TABLE III
PERFORMANCE COMPARISON ON RETRIEVAL TASK (IFASHION)

Model	HR@10	NDCG@10
BPR-MF	0.2121	0.0872
LightGCN	0.2632	0.2882
VBPR	0.2201	0.0949
HFGN	0.2833	0.1241
Hg-PDC	0.1402	**0.3532**
PDOR (Ours)	**0.3264**	0.3356

In personalized retrieval (Table III), PDOR demonstrated superior performance with an HR@10 of 0.3264, a 15.2% improvement over HFGN (0.2833). This is attributed to the "generation-retrieval" paradigm, which anchors user intent in continuous space rather than performing hard matching in discrete space. Although Hg-PDC had a higher NDCG@10 (0.3532) due to its hypergraph structure, its low recall (HR@10: 0.1402) limits its utility. PDOR maintains competitive ranking quality while achieving much higher recall.

C. Ablation Studies

To verify core components, we designed four variants: removing stochastic perturbation, removing CP reward, removing adaptive uncertainty, and a variant removing both contrastive CP and item diversity losses.

TABLE IV
ABLATION STUDY OF KEY COMPONENTS

Method	FITB Accuracy	CP AUC
Full Model	**0.7117**	**0.9074**
w/o Stochastic Perturbation	0.6623	-
w/o CP Reward	0.6938	-
w/o Adaptive Uncertainty	0.6886	-
w/o Contrastive CP	-	0.8486

All components are indispensable (Table IV). Removing stochastic perturbation caused the largest drop in FITB (to 0.6623), confirming its role in enhancing robustness and preventing overfitting. Removing adaptive uncertainty regularization (FITB: 0.6886) and CP reward guidance (FITB: 0.6938) validated their roles in preventing feature collapse and providing stylistic feedback, respectively.

D. Parameter Sensitivity Analysis

We analyzed the influence of diffusion steps T and reward weight λ_{reward} on performance (Fig. 3).

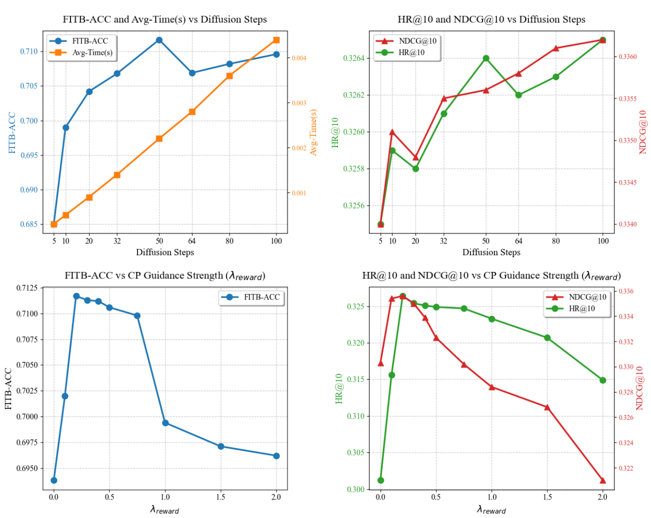

Fig. 3. Impact of diffusion steps T (top) and compatibility reward weight λ_{reward} (bottom) on model performance.

Regarding the number of diffusion steps (Fig. 3, top), the model exhibits a Rapid Convergence characteristic, attributed to the strong semantic anchoring provided by the conditional guidance mechanism. Even at an extremely low number of steps ($T = 5$), the model maintains competitive accuracy while achieving minimal inference latency; overall performance peaks at $T = 50$, where the inference time remains within an acceptable range. However, an excessive number of steps ($T > 50$) leads to fluctuations in FITB accuracy and performance saturation, while linearly increasing computational overhead. This is primarily attributed to the fact that unnecessary iterations in the late denoising stages tend to introduce stochastic perturbations, causing trajectory drift of the generated embeddings in the latent space and leading the final prototypes to deviate from the semantic center of the ground-truth items. Regarding the compatibility reward weight (Fig. 3, bottom), moderate guidance ($\lambda_{reward} = 0.2$) significantly improves all metrics, confirming the necessity of explicit stylistic feedback in rectifying the generation trajectory and enhancing semantic compatibility. However, an excessively strong reward signal ($\lambda_{reward} > 0.5$) disrupts the balance of multi-task learning, resulting in Manifold Deviation; that is, the generator sacrifices the geometric reconstruction fidelity of the target item to cater to the discriminator's high scores, thereby impairing retrieval performance. Based on the above analysis, we set $T = 50$ and $\lambda_{reward} = 0.2$ as the default configuration in our experiments to achieve the optimal performance balance.

V. CONCLUSION

This paper proposes the PDOR framework, which effectively addresses the "one-to-many" multimodal uncertainty in fashion recommendation by reformulating personalized outfit completion as latent-space conditional probability distribution learning. PDOR integrates three core innovations: diffusion-aware stochastic perturbation to enhance contextual robustness, compatibility reward guidance to balance stylistic consistency and diversity, and adaptive uncertainty regularization to ensure prototype accuracy. Experiments demonstrate that PDOR significantly outperforms existing deterministic models across multiple benchmarks. Future work will focus on modeling dynamic aesthetic evolution and incorporating fine-grained metadata to improve generalization and interpretability in sparse scenarios.

ACKNOWLEDGMENT

This work was supported by the Strategic Research and Consulting Project of the Chinese Academy of Engineering (2023-JB-05-14) and the "Pioneer" R&D Program of Zhejiang (Grant No. 2026LDC01024(XC)).

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

REFERENCES

[1] Z. Cui, Z. Li, S. Wu, X. Zhang, and L. Wang, "Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks," in *Proc. World Wide Web Conf. (WWW)*, San Francisco, CA, USA, May 2019, pp. 307–317.

[2] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T. S. Chua, "Hierarchical fashion graph network for personalized outfit recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, Xi'an, China, Jul. 2020, pp. 159–168.

[3] R. Sarkar, N. Bodla, M. I. Vasileva, Y.-L. Lin, A. Beniwal, A. Lu, and G. Medioni, "OutfitTransformer: Learning outfit representations for fashion recommendation," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, Jan. 2023, pp. 3601–3609.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[5] W. Li and Z. Wang, "Diffusion Recommender Model," in *Proc. 46th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, Taipei, Taiwan, Jul. 2023, pp. 830–839.

[6] Z. Li, A. Sun, and C. Li, "DiffuRec: A diffusion model for sequential recommendation," *ACM Trans. Inf. Syst.*, vol. 42, no. 3, Art. no. 66, pp. 1–28, May 2024.

[7] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.

[8] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty in Artificial Intelligence (UAI)*, Montreal, QC, Canada, Jun. 2009, pp. 452–461.

[9] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Xi'an, China, Jul. 2020, pp. 639–648.

[10] R. He and J. McAuley, "VBPR: Visual Bayesian personalized ranking from implicit feedback," in *Proc. 30th AAAI Conf. Artificial Intelligence*, Phoenix, AZ, USA, Feb. 2016, pp. 144–150.

[11] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie, "Learning visual clothing style with heterogeneous dyadic co-occurrences," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4642–4650.

[12] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional LSTMs," in *Proc. 25th ACM Int. Conf. Multimedia (MM)*, Mountain View, CA, USA, Oct. 2017, pp. 1078–1086.

[13] Y. Li, L. Cao, J. Zhu, and J. Luo, "Mining fashion outfit composition using an end-to-end deep learning approach on set data," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1946–1955, Aug. 2017.

[14] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[15] M. C. Jung, J. Monteil, P. Schulz, and V. Vaskovych, "Personalised outfit recommendation via history-aware transformers," in *Proc. 18th ACM Int. Conf. Web Search and Data Mining (WSDM)*, Hannover, Germany, Mar. 2025, pp. 633–641.

[16] S. Saed and B. Teimourpour, "Hybrid-hierarchical fashion graph attention network for compatibility-oriented and personalized outfit recommendation," *Mach. Learn. Appl.*, vol. 23, Art. no. 100802, Mar. 2026.

[17] J. Lin *et al.*, "A survey on diffusion models for recommender systems," *arXiv preprint arXiv:2409.05033*, 2024.

[18] T.-R. Wei and Y. Fang, "Diffusion models in recommendation systems: A survey," *arXiv preprint arXiv:2501.10548*, 2025.

[19] W. Wang, Y. Xu, F. Feng, X. Lin, X. He, and T.-S. Chua, "Diffusion augmentation for sequential recommendation," in *Proc. 32nd ACM Int. Conf. Information and Knowledge Management (CIKM)*, Birmingham, UK, Oct. 2023, pp. 2514–2523.

[20] W. Chen *et al.*, "POG: Personalized outfit generation for fashion recommendation at Alibaba iFashion," in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, Anchorage, AK, USA, Aug. 2019, pp. 2662–2670.

[21] Y. Xu *et al.*, "DiFashion: Diffusion models for generative outfit recommendation," in *Proc. 47th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Washington, DC, USA, Jul. 2024, pp. 2154–2163.

[22] M. Yu, Y. Ma, L. Wu, C. Wang, X. Li, and L. Meng, "FashionDPO: Fine-tune fashion outfit generation model using direct preference optimization," in *Proc. 48th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, Padua, Italy, Jul. 2025, pp. 212–222.

[23] M. Dang, A. Singh, L. Zhou, S. Ermon, and J. Song, "Personalized preference fine-tuning of diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2025, pp. 8020–8030.

[24] Z. Wang, J. Chen, F. E. Rosas, and T. Zhu, "A hypergraph-based framework for personalized recommendations via user preference and dynamics clustering," *Expert Syst. Appl.*, vol. 204, Oct. 2022, Art. no. 117552.

[25] W. Guan, X. Song, H. Zhang, M. Liu, C.-H. Yeh, and X. Chang, "Multimodal compatibility modeling via exploring the consistent and complementary correlations," in *Proc. 29th ACM Int. Conf. Multimedia (MM)*, Virtual Event, China, Oct. 2021, pp. 2299–2307.

[26] Y. Wang, X. Li, and M. Zhang, "Personalized fashion recommendations for diverse body shapes with contrastive multimodal cross-attention network," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, pp. 1–25, 2024.

[27] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, Virtual Event, Jul. 2021, pp. 8748–8763.

[28] P. J. Chia *et al.*, "Contrastive language and vision learning of general fashion concepts," *Scientific Reports*, vol. 12, no. 1, Art. no. 18958, 2022.

[29] K. Sevegnani, G. Tzelepis, P. Becke, and M. Soleymani, "Contrastive learning for interactive recommendation in fashion," in *Proc. SIGIR 2022 Workshop on eCommerce (SIGIR eCom'22)*, Madrid, Spain, Jul. 2022.

[30] R. Giahi *et al.*, "VL-CLIP: Enhancing Multimodal Recommendations via Visual Grounding and LLM-Augmented CLIP Embeddings," in *Proc. 19th ACM Conf. on Recommender Systems (RecSys)*, Prague, Czech Republic, Sept. 2025.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Venice, Italy, Oct. 2017, pp. 2980–2988.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Machine Learning (ICML)*, Virtual Event, Jul. 2020, pp. 1597–1607.

[33] P. Khosla *et al.*, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 18661–18673.

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 10684–10695.

[35] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, Virtual Event, Jul. 2021, pp. 8162–8171.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, Oct. 2015, pp. 234–241.

[37] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[38] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[39] Z. Lu, Y. Hu, Y. Jiang, Y. Chen, and B. Liu, "Learning binary code for personalized fashion recommendation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 10562–10570.

[40] J. Tian, C. Huang, and Z. Kira, "Rethinking weight decay for robust fine-tuning of foundation models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019.