Recommendations Beyond Catalogs: Diffusion Models for Personalized Generation

Anonymous Authors[1]

Abstract

Modern recommender systems follow the guiding principle of serving the right user, the right item at the right time. One of their main limitations is that they are typically limited to items already in the catalog. We propose REcommendations BEyond CAtalogs, `REBECA`, a new class of probabilistic diffusion-based recommender systems that synthesize new items tailored to individual tastes rather than retrieve items from the catalog. `REBECA` combines efficient training in embedding space with a novel diffusion prior that only requires users' past ratings of items. We evaluate `REBECA` on real-world data and propose novel personalization metrics for generative recommender systems. Extensive experiments demonstrate that `REBECA` produces high-quality, personalized recommendations, generating images that align with users' unique preferences.

1. Introduction

Despite the widespread awareness of generative AI, its integration into daily content consumption remains surprisingly limited. A recent Reuters survey (Fletcher & Nielsen, 2024) highlights this gap, showing that while many users are familiar with generative models, their actual adoption lags behind expectations. Meanwhile, recommendation algorithms continue to dominate online interactions, shaping how users engage with digital content and products.

Modern recommender systems excel at matching users to items but are inherently constrained by the content available in their catalogs. These systems implicitly assume that relevant, engaging items exist for all users at all times; this assumption, however, needs to be reconsidered with the rise of generative AI.

Figure 1. Visualization of a subset of real images liked by the user alongside generated images for selected users at 512×512 resolution. The personalized generative model effectively captures each user's preferred content and style.

Instead of merely retrieving content, can we dynamically enhance recommendations by generating personalized content? Generative recommender systems offer a promising path forward, capable of creating new items or modifying existing ones to better align with user preferences. Yet, integrating generative models into recommendation pipelines

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

efficiently remains an open challenge.

While recent work has explored recommendation systems that use generative algorithms, these works often focus on generative retrieval, where models leverage generative techniques to retrieve better-matching content (Wang et al., 2023). Generating entirely new items for recommendation, however, is comparatively underexplored. Existing proposals rely on explicit user inputs, such as textual prompts (Wang et al., 2024) to guide generation or image captions (Xu et al., 2024) during training. However, users rarely articulate their preferences directly, and their engagement patterns (e.g., clicks, likes) contain richer, more nuanced signals. In other words, despite the effectiveness of the numerous heterogeneous features tracked by modern recommender systems in identifying users' preferences (Zhai et al., 2024), text remains the main mechanism for generating novel content.

In this paper, we directly explore user feedback, i.e., user ratings, for personalized generation, a novel approach in recommendation systems. We present `REBECA`, the first probabilistic framework for personalized recommendation through content generation. `REBECA` directly captures user-item interactions via user feedback (e.g., likes or dislikes), enabling prompt-free, preference-aware generation. Our approach leverages pre-trained diffusion models by incorporating a lightweight adapter that efficiently encodes user preferences, eliminating the need for extensive fine-tuning or intermediation of large language models (LLMs) for personalized prompts composition. The design of `REBECA` is guided by three core principles: (i) *text-free training and inference*, removing reliance on captions and prompts to enhance flexibility and scalability in user preference modeling; (ii) *computational efficiency*, allowing seamless integration with existing generative pipelines, e.g., Stable Diffusion (Rombach et al., 2022), maintaining high expressiveness without costly retraining; and (iii) *a unified model for all users*, eliminating the need for multiple models and leveraging shared information across users for improved generalization.

In summary, our contributions are as follows:

1. We introduce `REBECA`, the first probabilistic framework for personalized recommendation via generative modeling, incorporating an efficient adapter on top of a highly expressive pretrained image generator that removes the dependency on image captions or prompting while preserving user-specific preferences.

2. We propose novel evaluation methodologies to assess the degree of personalization in a generative recommender system. Specifically, we (i) integrate the concept of *verifiers* with a hypothesis test to formally measure and validate personalization and (ii) employ image captioning and topic modeling (Angelov & Inkpen, 2024) to evaluate whether the generated images align with user content and stylistic preferences.

3. We perform extensive evaluations, including developing baselines that we compare against `REBECA` on real data and ablation studies that validate our design choices, offering valuable insights for developing scalable generative recommender systems.

By bridging recommendation and generative AI, `REBECA` redefines personalization, enabling dynamic content creation tailored to user preferences without requiring explicit input.

2. Related Work

Recommender systems and generative models are two rapidly evolving fields. While traditional recommender systems rely on pre-existing catalogs, generative approaches offer the promise of creating novel, user-specific content. This section reviews prior work in these areas, positioning our method as an underexplored bridge between them.

2.1. Recommender Systems

Modern recommender systems are often described as understanding users better than they understand themselves. This perception stems from their ability to gather and analyze vast amounts of online activity data, building detailed profiles of user interests and habits. Such systems have become highly effective at matching users to items within pre-existing catalogs (Castells & Jannach, 2023; Li et al., 2023). Novel-item generative recommenders can enhance the current catalog offering, not just by creating, but also by making items more suitable to an individual's preferences.

2.2. Generative Recommenders

In recent years, there has been substantial progress at the intersection of generative models and recommender systems. Research in generative retrieval (Zhai et al., 2024) and the use of large language models (LLMs) for *preference discernment* (Paischer et al., 2024) has introduced new paradigms, including the term "Generative Recommender" (Zhai et al., 2024). While these developments are exciting, they are not the focus of this work because they address the problem of within-catalog recommendation with generative item retrieval. Instead, we reference them here to provide contrast and contextualize our approach.

Despite the relative novelty of generative recommenders, the field has already seen terminological overlap. One particularly relevant work is GeneRec (Wang et al., 2024), which takes initial steps in framing the problem, identifying key challenges, and outlining directions for future research.

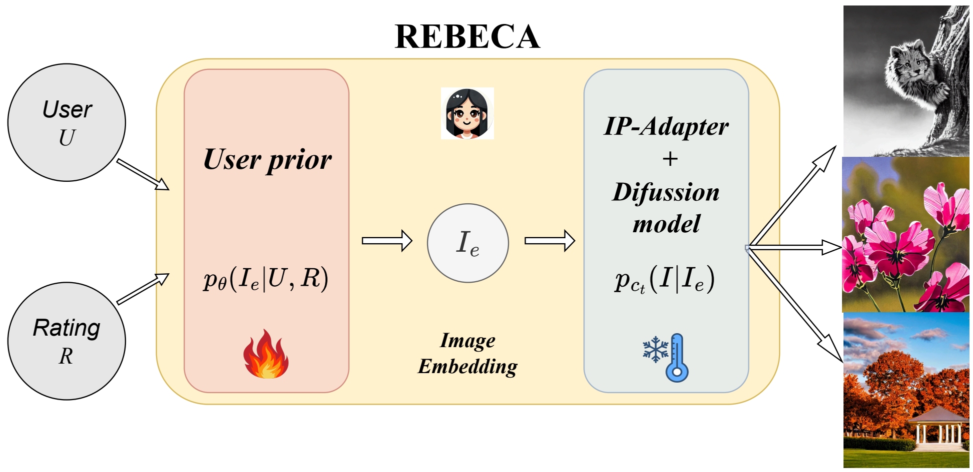

Figure 2. REBECA Diagram. Image embeddings I_e are sampled from the distribution conditioned on the user ID and desired rating using a trained diffusion model. The sampled embeddings are then used as input for a pretrained text-to-image model with IP-Adapter. The diagram presents images generated for user 33, who, based on independent topic modeling results, has a preference for charming buildings, flowers and curious cats.

Their proposed generative recommender paradigm consists of three main components: (1) an *instructor* that preprocesses user instructions and feedback, (2) a *generator* that creates new items, and (3) an *editor* that refines and repurposes existing catalog items. This formulation, however, proposes an LLM that mediates between the user and the AI generator, necessitating explicit and active user instructions, which they are often unwilling to provide.

A more recent approach focuses on personalized fashion recommendations (Xu et al., 2024), but the scope is limited: it requires fitting a new model each time a novel item class enters the catalog and thus fails to leverage the rich general-purpose representations of base text-to-image models.

2.3. Model Alignment

Aligning generative models' outputs with human values and preferences is a vibrant research domain whose discourse often extends beyond technical communities into the public sphere. From a technical perspective, one widely adopted approach is Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2023; Ziegler et al., 2020), which optimizes model behavior based on reward signals derived from human preferences. However, RLHF remains computationally expensive, requires careful reward function design, and can be prone to mode collapse.

Another promising approach is Direct Preference Optimization (DPO) (Rafailov et al., 2023), which aligns base models with human values through contrastive preference learning. DPO assumes uniform user preferences, treating human feedback as globally consistent, which is problematic in personalized settings like recommender systems, where preferences are highly heterogeneous. Additionally, DPO still requires an implicit reward structure, meaning that careful curation of preference data is crucial to avoid biases.

In our case, where we seek to model individual user preferences dynamically, both RLHF and DPO present limitations. Rather than imposing a global reward structure, we instead condition generation on user embeddings, allowing the model to capture personalized preferences in a more flexible, data-driven manner.

2.4. Adapters, Priors and Decoders

While our approach addresses the problem setting of GeneRec (Wang et al., 2024), our proposal is close to the **adapter** in generative modeling. The purpose of adapters is to produce augmentation to new input conditions or fine-tuning behavior (Zhang et al., 2023; Ye et al., 2023) of large pre-trained models, and they typically require text captions during training. We argue that text captions are not only unusual in most recommendation datasets, but they also fail to distill the nuances of individual preference, better captured in the features tracked in modern recommender systems (Zhai et al., 2024). Although some adapters can be successfully calibrated with Low-Rank adaptations (Hu et al., 2022), they are fine-tuned per user, which not only leads to poor scaling properties, but also does not benefit from correlations between individuals. Instead, we propose a probabilistic adaptation framework that leverages CLIP latents as *custom priors* and an off-the-shelf *decoder* in a principled adapter formulation. In the past, *priors* and *decoders* have been utilized to train full base text-to-image models (Pradeep et al., 2023). REBECA is the first used to produce adapting behavior to additional input signals.

2.5. Challenges in Establishing Baselines

Generative recommenders, as a novel class of systems, currently lack established baselines or datasets combining user ratings with image captions. Traditional methods in recommender systems focus on retrieval, while generative models lack personalization mechanisms. Adapters, although promising, require paired datasets that are impractical in recommendation contexts. This lack of baselines underscores the need for innovative methodologies and benchmarks, which we aim to address in this work.

3. Methodology

In this section, we present `REBECA`'s pipeline. During training, images are first mapped to a lower-dimensional embedding space using CLIP (Radford et al., 2021). Then a conditional diffusion model (Ho et al., 2020) is trained on these embeddings, conditioned on user and rating information. Specifically, at each time step t of the diffusion process, the model derives an image embedding I_e^t. This embedding, along with the time step t, the user ID U and the rating R, is then fed into $\varepsilon(U, R, t, I_e^t)$, a multi-layer decoder-only transformer, to predict the noise vector. The motivation is to integrate user and rating information directly into the diffusion dynamics. Classifier-free guidance (Ho & Salimans, 2022) is further incorporated into the diffusion model to improve the sampling quality of image embeddings.

During inference, the trained diffusion model is provided with a user ID and a desired rating. The diffusion model subsequently generates image embeddings, which are then passed to a pretrained text-to-image model via IP-adapter (Ye et al., 2023) to produce the final images.

3.1. Personalized Image Embedding Generation

Our goal is to sample images I from the personalized distribution $p(I|U, R)$ rather than the user-agnostic marginal distribution $p(I)$, where U represents the user and R represents the desired rating. We decompose the personalized distribution as follows:

$$p(I|U, R) = p(I, I_e|U, R) = p(I_e|U, R)p(I|I_e, U, R),$$

where I_e denotes the deterministic embedding of the image I. To model this decomposition, we employ two stages:

- A conditional diffusion model $\hat{p}_\theta(\cdot \mid U, R)$ with parameter θ is trained to approximate $p(I_e|U, R)$. Given user U and desired rating R, we sample new image embeddings from the trained diffusion model
$$\hat{I}_e(U, R) \sim \hat{p}_\theta(I_e|U, R)$$
- The sampled embedding is then injected into a pretrained text-to-image model, for example, model in the Stable Diffusion family (Podell et al., 2024; Rombach et al., 2022) with IP-Adapters (Ye et al., 2023). Optionally, conditioning text prompt c_t can be added to improve the image quality, e.g. *high quality image*, *vivid colors* or *good lightning*. We assume that the text-to-image model with IP-adapter approximates the conditional distribution:
$$\hat{p}_{c_t}(I|\hat{I}_e(U, R)) \approx p(I|I_e, U, R).$$

We conduct extensive experiments to evaluate how different prompt choices c_t impact the final generated images in Section 4.

3.2. Model training and inference

Training our diffusion model to generate personalized image embeddings conditioned on user preferences follows a standard diffusion formulation, where a data distribution $I_e \sim p(I_e|U, R)$ undergoes a forward noising process, and the model learns to predict the reverse denoising trajectory by maximizing the lower bound of a variational Markov Chain.

Formally, given an original unperturbed embedding I_e^0, we apply a noising schedule β_t to produce noisy embeddings during the forward process:

$$q(I_e^t|I_e^0) = \mathcal{N}(I_e^t; \alpha_t I_e^0, \sigma_t^2 I),$$

where α_t and σ_t depend on the choice of noise schedule.

In the reverse process, we use a transformer $\varepsilon_\theta(U, R, t, I_e^t)$ with input sequence being a concatenation of user, rating, time, and image tokens to approximate ε_t, the noise injected in the forward process conditioned on user and rating information. The reverse process does not have an exact analytic formulation, however, for $\beta_t \to 0$, it approximates a normal distribution (Sohl-Dickstein et al., 2015). We thus approximate the reverse process by:

$$p_\theta(I_e^{t-1} \mid I_e^t, U, R) = \mathcal{N}(I_e^{(t-1)}; \mu_\theta(U, R, t, I_e^t), \sigma_t^2 I),$$

where μ_θ is derived in relation to the noise prediction model ε_θ. Finally, we optimize the model using a simplified denoising objective (Ho et al., 2020):

$$L_{\text{simple}}(\theta) = \mathbb{E}_{(t,\varepsilon,U,R)}[||\varepsilon - \varepsilon_\theta(U, R, t, I_e^t)||^2],$$

where $t \sim \text{Unif}(\{1, \cdots, T\})$ and $\varepsilon \sim \mathcal{N}(0, I)$.

At inference time, we sample an initial Gaussian noise vector $I_e^T \sim \mathcal{N}(0, I)$ and apply iterative denoising:

$$I_e^{(t-1)} = \mu_\theta(I_e^t, t, U, R) + \sigma_t \varepsilon, \varepsilon \sim \mathcal{N}(0, I).$$

The process continues until we obtain I_ε^0, the final user-specific image embedding.

CLASSIFIER-FREE GUIDANCE

We employ Classifier-Free Guidance (CFG) (Ho & Salimans, 2022), a widely used technique in diffusion-based generative models, to enhance user-conditioned image embedding generation. CFG modulates the sampling process by interpolating between the unconditional model output $\varepsilon_\theta(I_e^t, t)$ and the user-rating conditional prediction $\varepsilon_\theta(I_e^t, t, U, R)$:

$$\tilde{\varepsilon}_\theta(I_e^t, t, U, R) = (1+\omega)\varepsilon_\theta(I_e^t, t, U, R) - \omega\varepsilon_\theta(I_e^t, t).$$

Here ω is the guidance scale that controls the trade-off between personalization and diversity. Higher values of guidance result in less diverse sets of images, but more attuned to their conditioning signals. We conduct extensive experiments to evaluate the influence of different guidance scaling in Section 4.5

4. Experiments

4.1. Dataset and Setup

We repurpose the FLICKR-AES dataset (Ren et al., 2017) to learn a generative model for aesthetically pleasing images based on user preferences. Initially compiled to learn models for personalized aesthetics predictions, the dataset contains a curated sample of 40,988 Creative Commons-licensed photos from FLICKR, rated for aesthetic quality on a scale of 1 to 5. For reference, annotations were collected using Amazon Mechanical Turk, with each image receiving ratings from five different workers. In total, 210 workers contributed, scoring 40,988 images for a total of 193208 ratings. On average, each worker labeled 137 images.

More formally, we define our initial data setup as consisting of triplets of users, ratings, and images (U, R, I). CLIP embeddings are deterministic, and computing all of them requires a single pass of our image data through the appropriate CLIP *ViT-H-14* encoder, and so we augment our data with them. We further map the ratings to binary rating, representing like and dislike: given a pair of user and its rated image, we let $R = 1$ if the rating is 4 or higher, and let $R = 0$ otherwise. Finally, we filter out users with less than 100 liked images, ensuring enough information on individual preference.

At the training stage, we fit our image embedding prior with (U, R, I_e). Importantly, image captions are not part of our data, and we uniquely rely on users' historic ratings.

4.2. Image Generation

We restate our objective as enhancing catalogs in accordance with users' preferences. This means that at the sampling stage, we generate image embeddings that are personalized based on the user's historical interactions.

To ensure that the sampled embeddings align with user preferences, we set $R = 1$ to indicate a strong preference signal. This choice enforces the model to condition its generation process on the user embedding, thereby prioritizing images that align with previously liked content.

We follow the standard denoising process using CFG and set the guidance scale of 10.0 (see details in Section 4.5) to balance personalization and diversity. The final generated embedding $\hat{I}_e$ represents a user-adapted image feature that is used as input to a Stable Diffusion 1.5 model with an empty prompt (denominated T_0) (Rombach et al., 2022). We select this model for its maturity and known properties, alongside its well-known IP-Adapter (Ye et al., 2023).

4.3. Baselines

We consider three baselines, all utilizing the same decoder as `REBECA`, namely Stable Diffusion 1.5 (Rombach et al., 2022). The first baseline, T_0, generates images from Stable Diffusion using an empty prompt, providing insight into the model's generic outputs. The second baseline, T_1, samples images using high-quality but simple positive and negative prompts. The third baseline, T_2, employs high-quality and detailed positive and negative prompts, simulating a scenario where a user can significantly influence image generation through prompt engineering. We paste the prompts we use in the following box.

- T_1 prompts:
  - **(+)** "high quality photo"
  - **(–)** "bad quality photo, letters"
- T_2 prompts:
  - **(+)** "Realistic image, finely detailed, with balanced composition and harmonious elements. Dynamic yet subtle tones, versatile style adaptable to diverse themes and aesthetics, prioritizing clarity and authenticity."
  - **(–)** "deformed, ugly, wrong proportion, frame, watermark, low res, bad anatomy, worst quality, low quality"

4.4. Metrics

GLOBAL EVALUATION

Our objective is to generate images that not only align with user preferences but also maintain high quality. To evaluate our approach, we compute the Fréchet Inception Distance (FID) (Heusel et al., 2017) and Conditional Maximum Mean Discrepancy (CMMD) (Jayasumana et al., 2024), aggregating generated images into a shared pool. Both metrics measure the distance between real and generated image distributions, providing an overall assessment of generative model performance. We report both metrics because FID, while widely used as a traditional measure of generation quality, has been shown to inadequately capture human perception of image quality (Jayasumana et al., 2024; Wiles et al., 2024; Stein et al., 2023). On contrast, CMMD, imposes fewer assumptions on the data-generating process and is more sample-efficient (Jayasumana et al., 2024).

Personalization Verifier

Given that we cannot evaluate the scores of users of the FLICKR-AES dataset on generated images, we cannot directly tell if the images generated by REBECA satisfy their tastes. However, we can train a predictive model that predicts the probability that a certain user U will like an image I. Models with similar purposes are known as *verifiers* and are largely used in the generative models' recent literature (Cobbe et al., 2021; Lightman et al., 2023). Our model for the probability of each user U liking a certain image I is inspired by matrix factorization techniques (Töscher et al., 2009; Koren et al., 2009; Ong et al., 2024) and Item Response Theory (IRT) from the field of psychometrics (Cai et al., 2016; Chen et al., 2019) but also used in the evaluation of LLMs (Maia Polo et al., 2024a;b). Concretely, we assume that our verifier $v(U, I)$ is given by

$$\mathbb{P}(R = 1 \mid U, I) = \sigma\left(\varphi(U)^\top \psi(I)\right),$$

where σ is the logistic function, $\varphi : [N] \to \mathbb{R}^d$ and $\psi : \mathcal{I} \to \mathbb{R}^d$ are embedding functions that map users and images to embeddings. Here, N is the number of users, $\mathcal{I}$ is the set of images, and d is a hyperparameter that denotes the dimension of the model. We assume φ is trainable and that each user is represented by an independent real vector while ψ is the composition of a (trainable) linear transformation and a (fixed) embedding model that maps images to high-dimensional embeddings. In practice, the fixed embedding model we use is *OpenCLIP-ViT-H-14* and the scoring model is trained by minimizing the classification cross-entropy loss. The performance of a generative model $p(I|U, R = 1)$ can be measured by

$$\text{Score}(p(I|U, R = 1)) = \mathbb{E}_U \mathbb{E}_{I\sim p(I|U,R=1)} \left[v(U, I)\right],$$

which can be estimated by: (i) fitting the verifier on training data, obtaining $\hat{v}$ and then (ii) empirically obtaining the estimator $\widehat{\text{Score}}(p(I|U, R = 1))$ by sampling some images per user. We use the verifier to compare different versions of REBECA and to test for its personalization capabilities.

Topic modelling

To further assess whether the generated images align with users' preferences, we employ contextual topic modeling (Angelov & Inkpen, 2024) to analyze users' preferences from multiple perspectives. Specifically, we generate image captions for the entire FLICKR dataset using GPT-4o-mini (OpenAI, 2024). For each image, we generate two distinct captions focusing on the content and style of that image. We then train a contextual topic model for each caption type. The topic model embeds each image into multiple contextual embeddings with *sentence-transformers* (Reimers & Gurevych, 2019). These embeddings are further projected into a lower-dimensional space using UMAP (McInnes et al., 2018) for clustering and topic identification. Two separate topic models are trained on the content and style captions from the FLICKR dataset. 345 topics are identified for image content and 255 topics are identified for image style. Each image used for training is assigned a topic distribution, representing its alignment with identified content and style topics. When presented with a new caption, the topic models map it to the most similar content and style topics from the learned topic distributions, without requiring additional training. See Appendix A.2 for more details on the prompt design and the learned topics.

Table 1. Comparison of FID and CMMD scores for several baselines and our method. Lower is better for both metrics.

Model	FID ↓	CMMD ↓
REBECA	**117.77**	**0.68**
T_0	121.50	1.13
T_1	131.75	0.88
T_2	145.45	1.56

4.5. Results

Global quality of generated images

Table 1 presents a comparison of different models based on FID (Fréchet Inception Distance) and CMMD (Conditional Maximum Mean Discrepancy), where lower values indicate better performance and a better alignment with the target distribution of images. The results demonstrate that REBECA achieves the best performance, with the lowest FID 117.77 and CMMD 0.68, indicating superior image quality and better alignment with user preferences. The T_2 baseline shows a clear degradation in performance, having the highest FID (145.45) and worst CMMD (1.56), indicating a significant drop in both image realism and conditional consistency despite a well-crafted prompt. These results reinforce that REBECA is the most effective approach, outperforming all baselines in global image quality.

Does REBECA generate personalized images?

Dependence between generated images and user IDs We start measuring and testing the degree of personalization of REBECA utilizing our trained verifier $\hat{v}$ and a statistical hypothesis testing. The degree of personalization can be measured by randomly assigning REBECA images to users and comparing the verifier scores with the case in which

Algorithm 1 Testing for REBECA personalization

Input: Users, REBECA model, verifier $\hat{v}$, significance level α, number of permutations $B = 1000$
for each user **do**
  Generate 30 images using REBECA
end for
Compute baseline performance $\widehat{\text{Score}}$ using verifier $\hat{v}$
for $b = 1$ **to** B **do**
  Randomly permute generated images across users
  Compute $\widehat{\text{Score}}_b$ using verifier $\hat{v}$
end for
Compute p-value:

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{1}\left[\widehat{\text{Score}} \leq \widehat{\text{Score}}_b\right]}{B+1}$$

if $p \leq \alpha$ **then**
  Reject null hypothesis $H_0 : U \perp\!\!\!\perp I$
  (that REBECA's images do not depend on users)
end if

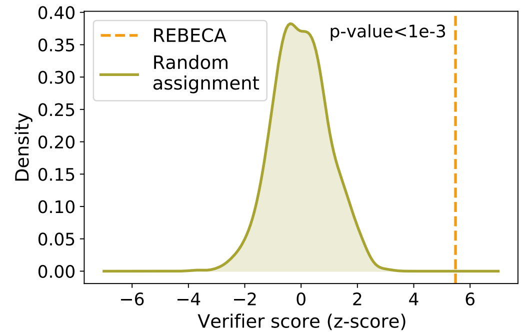

Figure 3. Measuring and testing for REBECA's personalization capabilities. Assigning the correct REBECA images vs randomly assigning them, increases the verifier score, on average, almost in 6 standard deviations and leads to a p-value $p < 10^{-3}$. This gives strong evidence that images and users are statistically dependent and that REBECA generates images tailored to users.

images are correctly assigned to the user they were created for; the difference of verifier scores (correct assigned minus randomly assigned) denotes how happier users get when the correct images are assigned to them versus when random images are assigned. To measure if the difference is statistically significant, we employ a hypothesis test measuring the degree of dependence between users U and generated images I; in statistical terms, we test the null hypothesis $H_0 : U \perp\!\!\!\perp I$ using a permutation test (Lehmann et al., 1986) and a significance level $\alpha \in (0, 1)$. The procedure is described in concrete terms in Algorithm 1. Figure 3 depicts our results and shows that, on average, assigning the correct REBECA images vs randomly assigning them, increases the verifier score, on average, almost 6 standard deviations and leads to a p-value $p < 10^{-3}$ (much smaller than the standard threshold $\alpha = .05$), giving strong evidence that images and users are statistically dependent and that REBECA generates images tailored to each user.

Topic Precision For each user, we define their liked topics as the set of topics that are assigned to some image that the user liked in the FLICKR data. We use the REBECA model to generate 50 images for each user conditioning on $(U, R = 1)$, implying that we expect the user to like the generated images. GPT-4o mini is then applied to caption those generated images from the content and style perspective with the same prompt as that used in FLICKR image caption generation. We assign the content and style topics for those captions and identify a generated image as a liked image in terms of content (resp. style) if its assigned content (resp. style) topics is a subset of the user's liked content (resp. style) topics. We define the content (resp. style) topic precision as the proportion of generated images that are also liked images for each user. Figure 5 show the distribution of content and style topic precision for each model. We can see that for content topics, REBECA, without any conditioning text prompt, outperforms baseline models T_0 and T_1, and has comparable performance to baseline model T_2, which benefits from a well-crafted text prompt as input to the text-to-image model. A possible explanation is that high-quality and visually appealing images often feature content that many users naturally like, such as cute animals or snowy mountains at sunrise. For style topics, we notice that REBECA outperforms all baseline models, implying that it does learn the styles the users prefer only from their like/dislike history. We also notice that model T_1 and T_2 are worse than model T_0, which does not use any text prompt, implying that artificially adding general prompts may generate nice images, but in a style that the user does not prefer.

ABLATIONS

We conduct additional ablation studies to gain deeper insights into REBECA's behavior under different conditions.

First, we examine the effect of coupling REBECA with prompts T_1 and T_2, previously used by our baselines. As shown in Table 4.5, REBECA achieves the best performance, yielding the lowest FID (117.77) and CMMD (0.68), indicating strong generative quality and alignment with user preferences. Interestingly, incorporating prompts T_1 and T_2 leads to performance degradation, despite their explicit emphasis on quality. This suggests that these prompts negatively impact user alignment, causing a shift away from the real data distribution. Additionally, we assess the impact of removing user embeddings while keeping the model architecture unchanged. The NoUE variant, which excludes

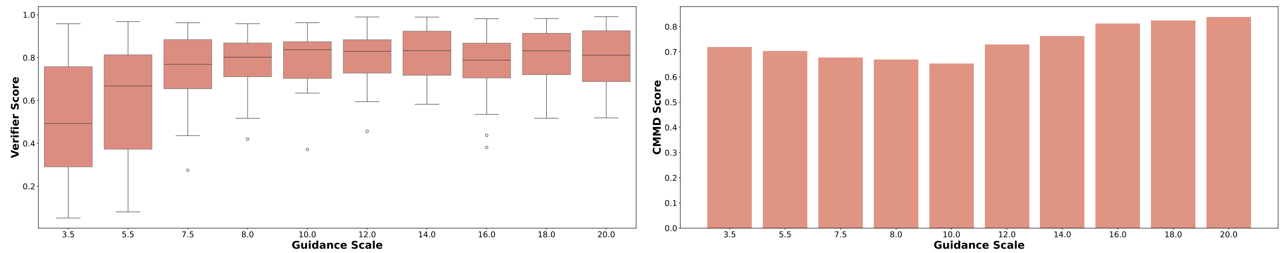

Figure 4. Impact of classifier-free guidance (CFG) values on: *(Left)* verifier score with increase in the (3.5, 10.0) range, followed by saturation, and *(right)* CMMD distance decreasing within the same range, reaching a minimum at $\omega = 10.0$, and increasing afterward.

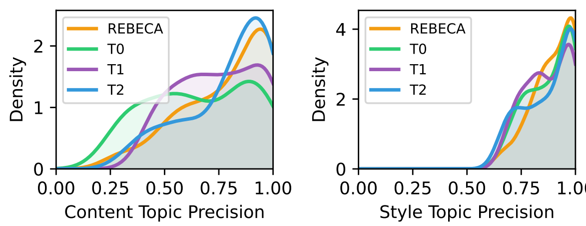

Figure 5. Precision of the generated images on content and style topics of the REBECA model and the baseline models. For content topics, REBECA models outperform the baselines T_0 and T_1 and has comparable performance to T_2, and using text prompts is better than no prompt. For style topics, REBECA models perform slightly better than the baselines, and using text prompts is worse than no use of prompts.

Method	FID ↓	CMMD ↓
REBECA	**117.77**	**0.68**
REBECA-T_1	124.22	0.78
REBECA-T_2	130.56	0.99
REBECA-NoUE	172.29	1.20

Table 2. Ablation studies of the REBECA model, evaluating varying prompt levels and the inclusion of user embeddings

user embeddings, performs significantly worse, with an FID of 172.29 and CMMD of 1.20. Since all REBECA variants outperform NoUE, this confirms that incorporating user information is crucial for enhancing both generation quality and alignment with user preferences.

We also analyze the effect of different classifier-free guidance (CFG) values for ω on the predicted user scores. As illustrated in Figure 4, while increasing ω within the range $[3.5, 10.0]$ leads to a substantial improvement in scores, meaning that generations with stronger user-conditioning signal give better results overall. The gains saturate beyond $\omega = 10$ and CMMD distance begins to rise, suggesting diminishing returns.

5. Discussion

In this paper, we introduce REBECA, a novel probabilistic framework for personalized recommendation beyond catalogs. Unlike traditional retrieval-based recommender systems, REBECA generates new items tailored to individual preferences using a diffusion-based generative model. By leveraging implicit user preference instead of explicit text prompts, our approach seamlessly integrates generative AI with recommendation pipelines.

Our experiments demonstrate that REBECA outperforms competing approaches that rely on textual prompts. Through topic modeling and a verifier-based evaluation, we established that our method generates images aligned with user interests, without requiring costly fine-tuning or extensive prompt engineering. Notably, we showed that introducing generic prompts can degrade personalization, highlighting the importance of direct preference modeling. Despite its success, developing personalized generative recommenders presents key challenges, particularly the lack of suitable image datasets that pair user preferences (e.g., ratings) with visual content. Future research could focus on collecting and curating benchmark datasets tailored to this task.

The effectiveness of REBECA in capturing user-specific content and style preferences suggests broader applicability to other domains, such as music, video, and textual content generation. Several directions remain open for future work:

- Extending REBECA to multimodal recommendations, incorporating text and audio to further enrich user experiences.
- Exploring dynamic personalization, where user preferences evolve over time, allowing the model to adapt accordingly.
- Developing fairness-aware generative recommenders, mitigating potential biases in the generated items.

Impact Statement

Generative modeling, including images and videos, has significant misuse potential. It can trigger negative consequences within the society in several ways. The primary concerns include various types of disinformation, but also the potential to amplify stereotypes and unwanted biases. While our advancements in personalized recommendation improve user experience, they may also inadvertently strengthen users' existing biases. Furthermore, fine-tuning such models carries the potential to align their outputs more closely with human values, but it also introduces ethical challenges that must be carefully managed. Recognizing these risks, we emphasize the need for responsible development, rigorous bias mitigation strategies, and ongoing evaluation to ensure that generative models serve society in a fair and beneficial manner.

References

Angelov, D. and Inkpen, D. Topic modeling: Contextual token embeddings are all you need. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13528–13539, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.790. URL https://aclanthology.org/2024.findings-emnlp.790/.

Cai, L., Choi, K., Hansen, M., and Harrell, L. Item response theory. *Annual Review of Statistics and Its Application*, 3 (1):297–321, 2016.

Castells, P. and Jannach, D. Recommender systems: A primer, 2023. URL https://arxiv.org/abs/2302.02579.

Chen, Y., Li, X., and Zhang, S. Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84:124–146, 2019.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Fletcher, R. and Nielsen, R. K. What does the public in six countries think of generative ai in news?, 2024. URL https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-05/Fletcher_and_Nielsen_Generative_AI_and_News_Audiences.pdf.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. Rethinking fid: Towards a better evaluation metric for image generation, 2024. URL https://arxiv.org/abs/2401.09603.

Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37, 2009.

Lehmann, E. L., Romano, J. P., and Casella, G. *Testing statistical hypotheses*, volume 3. Springer, 1986.

Li, Y., Liu, K., Satapathy, R., Wang, S., and Cambria, E. Recent developments in recommender systems: A survey, 2023. URL https://arxiv.org/abs/2306.12680.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Maia Polo, F., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024a.

Maia Polo, F., Xu, R., Weber, L., Silva, M., Bhardwaj, O., Choshen, L., de Oliveira, A. F. M., Sun, Y., and Yurochkin, M. Efficient multi-prompt evaluation of llms. *arXiv preprint arXiv:2405.17202*, 2024b.

McInnes, L., Healy, J., Saul, N., and Großberger, L. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., and Stoica, I. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024.

Paischer, F., Yang, L., Liu, L., Shao, S., Hassani, K., Li, J., Chen, R., Li, Z. G., Gao, X., Shao, W., Feng, X., Noorshams, N., Park, S., Long, B., and Eghbalzadeh, H. Preference discerning with llm-enhanced generative retrieval, 2024. URL https://arxiv.org/abs/2412.08604.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Pradeep, R., Hui, K., Gupta, J., Lelkes, A., Zhuang, H., Lin, J., Metzler, D., and Tran, V. How does generative retrieval scale to millions of passages? In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1305–1321, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.83. URL https://aclanthology.org/2023.emnlp-main.83/.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Ren, J., Shen, X., Lin, Z., Mech, R., and Foran, D. J. Personalized image aesthetics. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3732–3784. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0bc795afae289ed465a65a3b4b1f4eb7-Paper-Conference pdf.

Töscher, A., Jahrer, M., and Bell, R. M. The bigchaos solution to the netflix grand prize. *Netflix prize documentation*, pp. 1–52, 2009.

Wang, W., Xu, Y., Feng, F., Lin, X., He, X., and Chua, T.-S. Diffusion recommender model, 2023. URL https://arxiv.org/abs/2304.04971.

Wang, W., Lin, X., Feng, F., He, X., and Chua, T.-S. Generative recommendation: Towards next-generation recommender paradigm, 2024. URL https://arxiv.org/abs/2304.03516.

Wiles, O., Zhang, C., Albuquerque, I., Kajić, I., Wang, S., Bugliarello, E., Onoe, Y., Knutsen, C., Rashtchian, C., Pont-Tuset, J., and Nematzadeh, A. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings, 2024. URL https://arxiv.org/abs/2404.16820.

Xu, Y., Wang, W., Feng, F., Ma, Y., Zhang, J., and He, X. Diffusion models for generative outfit recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 1350–1359, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657719. URL https://doi.org/10.1145/3626772.3657719.

Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

Zhai, J., Liao, L., Liu, X., Wang, Y., Li, R., Cao, X., Gao, L., Gong, Z., Gu, F., He, J., Lu, Y., and Shi, Y. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xye7iNsgXn.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models, 2023.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A. Appendix

A.1. Training details

Our model is a diffusion-based generative model with a transformer backbone that operates in CLIP embedding space rather than pixel space. This design significantly reduces computational overhead while maintaining high expressiveness in learned representations. Notably, training each model requires less than 10 minutes to reach early stopping on a single NVIDIA RTX 4090 GPU.

Given the imbalance in the number of rated images per user in the FLICKR dataset, we employ strategic sampling to mitigate this issue. Specifically, in each batch, we sample an equal number of images per user to ensure balanced training.

For the transformer model within the diffusion framework, we conduct a grid search over key hyperparameters:

- **Number of layers:** {2, 4, 8, 16}
- **Number of heads:** {2, 4, 8, 16, 32}
- **Number of user tokens:** {1, 2, 4, 8}
- **Number of image tokens:** {1, 2, 4, 8}
- **Learning rate:** {1e-4, 1e-3}

Following this, we perform an additional grid search over:

- **Scheduler steps:** {1000, 2000, 4000, 6000}
- **Samples per user:** {30, 50, 80, 110, 140}

The selected model, achieving the lowest validation loss, is an 8-layer, 16-head decoder-only transformer with four user tokens, one image token, and one rating token. The optimal learning rate is 1e-4, and the best-performing configuration samples 80 images per user.

We use the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 1e-4 and a weight decay of 1e-5.

A.2. Topic modeling

We ask GPT-4o mini (OpenAI, 2024) to generate captions for each image using the following prompt:

Captioning prompt

Analyze the given image and generate two captions: 1. One describing the **content** of the image (what objects, scenes, or elements are present). 2. One describing the **style** of the image (artistic elements, mood, colors, or visual style).
Provide the captions in a dictionary format where the keys are ‘content‘ and ‘style‘.

Example Output 1:

```
{"content": "A bustling city street during sunset, featuring people walking on sidewalks, cars stopped at a traffic light, and tall buildings in the background. A dog is being walked by its owner in the foreground.", "style": "A warm and vibrant color palette with shades of orange and pink reflecting the sunset. The image has a modern, urban aesthetic with a lively and dynamic atmosphere."}
```

Example Output 2:

```
{"content": "A futuristic cityscape at night with towering skyscrapers illuminated by neon lights, flying vehicles zooming through the sky, and pedestrians walking on a glowing sidewalk.", "style": "A cyberpunk-inspired aesthetic with vivid neon colors like electric blue, hot pink, and lime green. The mood is gritty yet energetic, evoking a futuristic, high-tech vibe."}
```

Example Output 3:

```
{"content": "A dense jungle scene with tall trees, vines hanging from branches, a small stream cutting through the forest floor, and a tiger partially hidden among the bushes.", "style": "A richly detailed and vibrant depiction with deep greens and earthy tones. The style has a dramatic, cinematic feel with strong contrasts between light filtering through the trees and dark shadows."}
```

Now, provide the captions for the following image:

We trained two topic models separately for the content and style captions derived from the FLICKR dataset. Below are three content topics, each represented by their top-5 keywords:

- **Graduation-related**: *graduation gown, graduation gowns, graduation ceremony, celebrating graduation, graduation cap*.
- **Golf-related**: *golf course, golf cart, green turf, turf field, green grass*.
- **Cat-related**: *cat lounging, fur coat, domestic cat, fur trim, two cats*.

Similarly, three style topics and their top-5 keywords are shown below:

- **Aesthetic Design**: *decorative elements, intricate designs, visually appealing, Japanese aesthetics, visual appeal*.
- **Atmospheric Lighting**: *atmospheric depiction, twinkling lights, glowing lights, holiday cheer, brightly colored*.
- **Seasonal & Festive**: *holiday cheer, autumn leaves, Halloween themed, floral patterns, floral arrangement*.