DIFFUSION PRIORS FOR LIGHTWEIGHT PERSONALIZED IMAGE GENERATION

Gabriel A. Patron **Zhiwei Xu** **Ishan Kapnadak** **Felipe Maia Polo**
University of Michigan
{gapatron, xuzhiwei, kapnadak}@umich.edu, felipemaiapolo@gmail.com

ABSTRACT

Personalization is central to human-AI interaction, yet current diffusion-based image generation systems remain largely insensitive to user diversity. Existing attempts to address this often rely on costly paired preference data or introduce latency through Large Language Models. In this work, we introduce REBECA, a lightweight and scalable framework for personalized image generation that learns directly from implicit feedback signals such as likes, ratings, and clicks. Instead of fine-tuning the underlying diffusion model, REBECA employs a two-stage process: training a conditional diffusion model to sample user- and rating-specific image embeddings, which are subsequently decoded into images using a pretrained diffusion backbone. This approach enables efficient, fine-tuning-free personalization across large user bases. We rigorously evaluate REBECA on real-world datasets, proposing a novel statistical personalization verifier and a permutation-based hypothesis test to assess preference alignment. Our results demonstrate that REBECA consistently produces high-fidelity images tailored to individual tastes, outperforming baselines while maintaining computational efficiency[1][2].

1 INTRODUCTION

Personalization is central to human-AI interaction, as users exhibit diverse tastes, intents, and creative goals. However, current diffusion-based image generation systems remain largely insensitive to such user diversity, producing outputs that are visually impressive but not tailored. Addressing this limitation is crucial for advancing applications in creative tools, advertising, and content recommendation.

Recent work has begun to address personalization in diffusion models. *Personalized Preference Fine-tuning (PPD)* Dang et al. (2025) adapts diffusion models to user tastes through pairwise preference supervision, where users choose preferred images from pairs. While effective, this approach depends on costly paired data and model fine-tuning, limiting scalability. *Personalized Multimodal Generation (PMG)* Shen et al. (2024) instead leverages large language models (LLMs) to infer user preferences from behaviors such as clicks or chat histories, but introduces latency from LLM inference and relies heavily on textual cues rather than richer multimodal signals.

In this work, we introduce REBECA[3], a lightweight and scalable framework for personalized image generation with diffusion models. Unlike prior methods that rely on paired annotations or large language models, REBECA learns personalization directly from behavioral data such as likes, ratings, and clicks that are readily available in real-world social media platforms. The method operates in two stages:

1. Train a conditional diffusion to model the distribution $p_{\hat{\theta}}(I_e \mid U, R)$. Then samples image embeddings I_e conditioned on a user U and rating R.
2. Decode I_e into personalized images I using a pretrained diffusion model $p(I \mid I_e)$ that conditionally decodes the embedding.

[1]Code available in anonymized repository during review.

[2]This version of our work corresponds to a considerably expanded version of a previously introduced extended abstract. In Appendix C, we give a detailed list of the innovations in the current work.

[3]In Appendix B, we discuss in detail the motivations behind REBECA's name.

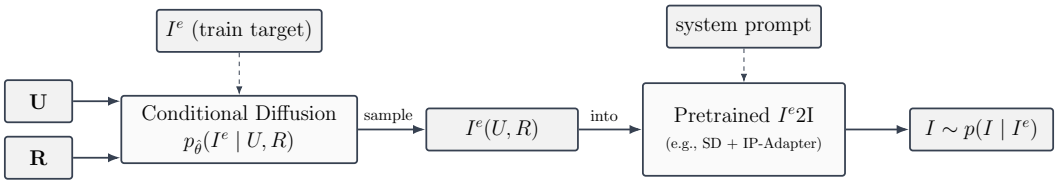


Figure 1: **REBECA overview.** *Training:* Conditional diffusion prior trained to generate personalized image embeddings from user IDs and ratings. *Inference:* Generated embeddings are decoded into images via a pretrained image decoder model.

Unlike prior work, REBECA does not rely on textual descriptions, paired preference labels, or per-user fine-tuning. Instead, we learn a single conditional diffusion prior whose geometry spans the preference manifolds of all users. Conditioning identifies a user-specific region of this shared space, enabling scalable, plug-and-play personalization.

We rigorously evaluate REBECA on real-world user datasets, measuring both personalization strength and visual fidelity. In addition to more traditional metrics such as recall, precision, and predicted image quality, we propose and employ a statistical personalization verifier and a permutation hypothesis test for personalized generations. Our results on synthetic and real data show that REBECA consistently produces images that align with individual user preferences. In summary, our contributions are threefold:

1. **Learning from implicit signals.** We introduce a single user-conditioned diffusion prior that models personalized CLIP-space embeddings directly from implicit feedback, eliminating the need for preference pairs, LLM mediation, or per-user fine-tuning.
2. **Lightweight, plug-and-play framework.** REBECA decouples personalization from the image generator: the learned prior produces embeddings that plug into any pretrained decoder with the same embedding class, enabling large-scale personalized generation with minimal compute.
3. **Rigorous evaluation.** We propose a statistical personalization verifier and a permutation-based hypothesis test to assess alignment with user preferences, complementing standard metrics, and show strong performance on both synthetic and real datasets. We believe our newly introduced evaluation procedure can be used by future work as a standard evaluation approach for personalized generations.

We include a detailed discussion of related work on Appendix A.

2 METHODOLOGY

We present the REBECA pipeline, a two-stage framework for personalized image generation from implicit feedback. In training, each interaction provides a tuple (I, U, R) where I is an image, $U \in \mathcal{U}$ is a *discrete user identity* (ID), and $R \in \mathcal{R}$ is a *discrete rating level* (e.g., like/dislike or binned scores).[4] Images are mapped to a lower-dimensional embedding space using CLIP Radford et al. (2021). A conditional diffusion model Ho et al. (2020) is then trained on these embeddings using classifier-free guidance Ho and Salimans (2022), conditioned on (U, R). At inference time, we condition on a chosen user $U{=}u$ and rating level $R{=}r$ to generate image embeddings, which are then decoded into images using a frozen generator. Unless stated otherwise, text prompts are held fixed (empty prompt); personalization is driven by conditioning on (U, R) in embedding space. REBECA consists of two key components:

1. **Personalized embedding generator.** A conditional diffusion prior samples CLIP-space embeddings conditioned on user IDs and ratings.
2. **Decoder model.** The generated embeddings are translated into images by a frozen text-to-image model (e.g., Stable Diffusion Podell et al. (2024); Rombach et al. (2022)) augmented with an IP-Adapter Ye et al. (2023) to enable image-conditioned generation.

[4] We focus on the closed-world setting where users at inference are drawn from the training user set; extending to unseen users via history encoders is left to future work.

See Figure 1 for an overview. These design choices avoid per-user fine-tuning of large diffusion backbones and enable flexible, language-free personalization in CLIP space.

Personalized image generation. We aim to sample personalized images I from $p(I \mid U, R)$, where $U \in \mathcal{U}$ denotes a user ID and $R \in \mathcal{R}$ a rating level. We introduce an intermediate CLIP embedding $I^e = \mathrm{CLIP}(I) \in \mathbb{R}^d$ and write

$$p(I \mid U, R) = \int p(I \mid I^e, U, R)\, p(I^e \mid U, R)\, dI^e. \tag{1}$$

We approximate this distribution with a two-stage model. First, a conditional diffusion prior $p_\theta(I^e \mid U, R)$ is trained to approximate $p(I^e \mid U, R)$ and generate personalized embeddings $I^e \sim p_{\hat{\theta}}(I^e \mid U, R)$. Second, a pretrained text-to-image model with an IP-Adapter defines a frozen decoder $p_\phi(I \mid I^e)$ that maps embeddings to images: $I \sim p_\phi(I \mid I^e)$. This corresponds to the modeling assumption that I^e is (approximately) sufficient for preference-relevant information, so that $p(I \mid I^e, U, R) \approx p_\phi(I \mid I^e)$, i.e., user preference enters through the conditional prior over embeddings, while the decoder remains user-agnostic and is not fine-tuned. Prior work shows that CLIP spaces exhibit smooth semantic directions and well-structured manifolds Radford et al. (2021); Ramesh et al. (2022); Levi and Gilboa (2025), motivating embedding-space modeling of $p(I^e \mid U, R)$. In section 4.5, we empirically demonstrate that this sufficiency approximation is adequate to induce personalization.

Conditional diffusion prior. Let $I^e \in \mathbb{R}^d$ denote a CLIP image embedding. We model the conditional prior $p_\theta(I^e \mid U, R)$ via a standard forward diffusion process:

$$q(I_t^e \mid I_0^e) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\, I_0^e,\, (1 - \bar{\alpha}_t)\mathbf{I})\,, \tag{2}$$

with $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The denoising network f_θ is trained to predict the clean embedding I_0^e directly from a noisy sample:

$$\mathcal{L}_{x_0}(\theta) = \mathbb{E}_{I_0^e,\, U, R,\, t,\, \epsilon}\left[\left\| f_\theta(I_t^e, t, U, R) - I_0^e \right\|_2^2\right], \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $t \sim \mathrm{Uniform}\{1, \ldots, T\}$. At inference, classifier-free guidance (CFG) is applied as

$$f_\theta^{(\omega)} = f_\theta(I_t^e, t, \varnothing, \varnothing) + \omega(f_\theta(I_t^e, t, U, R) - f_\theta(I_t^e, t, \varnothing, \varnothing))\,, \tag{4}$$

where $\varnothing$ denotes dropped conditioning (the unconditional branch) and ω controls conditioning strength. The final denoised embedding $\hat{I}_0^e$ is decoded using the frozen decoder $p_\phi(I \mid I^e)$, yielding personalized samples consistent with the target user and rating.

REBECA Prior Architecture. Our diffusion prior is intentionally lightweight: a 4.4M-parameter transformer with 6 PriorBlocks (Supplementary D), AdaLN-Zero conditioning, and a learned tokenizer that splits CLIP embeddings into tokens. Conditioning enters through user, rating, and timestep tokens combined via a small MLP. The small scale allows end-to-end training in < 10 minutes on a single RTX-4090, enabling rapid iteration and scalability to large user sets. A full description of our training protocol and final configuration may be found in Supplementary D.

3 REBECA UNDER CONTROL

3.1 DATASET AND SETUP

We construct a controlled setting using dSprites Matthey et al. (2017), restricting to red/blue hearts and squares (four shape–color pairs). Four synthetic users each prefer one pair (s_U, c_U), and we generate ratings $R \in \{0, 1\}$ via a simple probabilistic rule that assigns high probability (0.95) to the preferred pair, low probability (0.05) to the opposite, and intermediate probability (0.10) to mismatched shape or color. We sample 40,000 rated images and split them $90/10$.

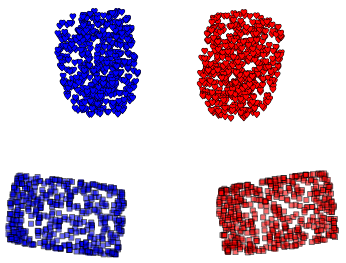


Figure 2: UMAP McInnes et al. (2018) projection of VAE embeddings. Color and shape clusters are cleanly separated.

3.2 IMAGE GENERATION

REBECA is implemented using a lightweight Variational Autoencoder (VAE) Kingma and Welling (2019; 2022) with a 32-dimensional latent space (Figure 2) trained on the dSprites dataset. We construct a small VAE solely to define a ground-truth latent space with a frozen decoder that reconstructs final images from embeddings. The encoder provides compact representations of dSprites, and the learned prior generates new embeddings aligned with user taste.

Table 1: **Precision@k and Recall@k** averaged over users in the controlled simulation. Recall is scaled by $\times 10^{-3}$. The mean baseline is deterministic and thus lacks diversity.

Method	k	Precision@k	Recall@k ($\times 10^{-3}$)
REBECA	1	0.782	0.167
	5	0.782	0.835
	10	0.789	1.685
	20	0.799	3.409
Mean (deterministic)	1	**1.000**	0.214
Random	1	0.290	0.062
	5	0.282	0.301
	10	0.287	0.613
	20	0.291	1.244

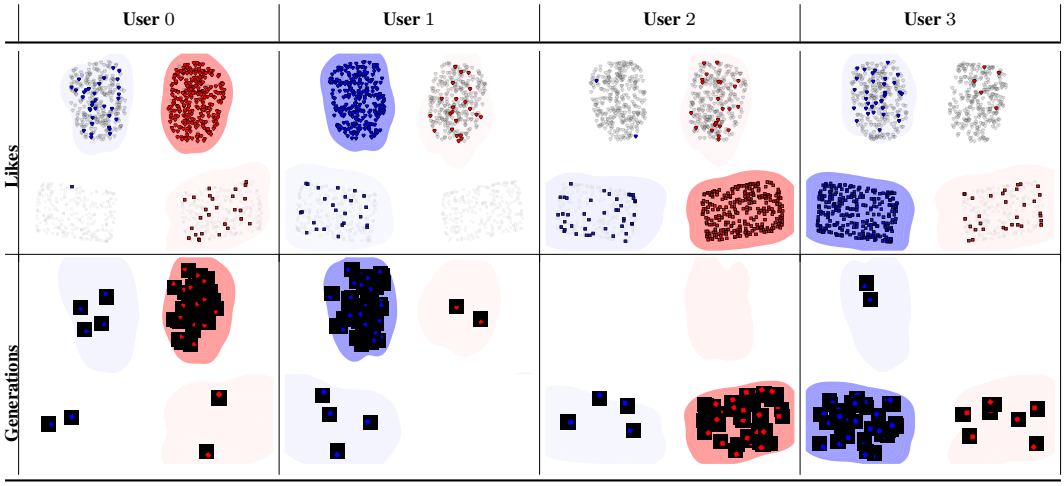


Figure 3: Per-user visualization in the controlled setting. Top: liked samples for each user. Bottom: images generated by REBECA using the frozen VAE decoder. REBECA captures each user's preference manifold while maintaining diversity.

3.3 Evaluation

Evaluation proceeds in two stages. (1) **Qualitative:** UMAP projections visualize whether user embeddings cluster coherently according to preferences. (2) **Quantitative:** REBECA is compared with (a) a mean-embedding baseline that generates a single prototype image per user and (b) a random baseline sampling uniformly from training data. For each user, REBECA and the random baseline generate $k \in \{1, 5, 10, 20, 25\}$ samples. We compare REBECA to (a) a mean-embedding baseline that generates a single prototype per user, and (b) a random baseline that samples uniformly from the training set. For $k \in \{1, 5, 10, 20\}$, we report Precision@k (fraction of generated images whose nearest neighbor in the test set belongs to the user's liked subset) and Recall@k (fraction of liked test images for which at least one generated sample is the nearest neighbor).

3.4 Results

Figure 3 shows that generated samples align with users' liked regions in latent space. Table 1 confirms that REBECA achieves the best precision-recall balance, capturing user tastes while preserving diversity.

This controlled setting validates that REBECA recovers user-specific preference manifolds when ground truth is known. In Section section 4 we show that the same structure appears in real-world behavioral data, using a learned verifier as a surrogate for unknown preferences.

4 REBECA IN THE WILD

4.1 DATASET AND SETUP

We evaluate on FLICKR-AES Ren et al. (2017), a benchmark introduced for *personalized image aesthetics* that captures systematic rater-specific differences in aesthetic judgment. The dataset contains 40,988 Creative Commons–licensed Flickr photos, each scored for aesthetic quality on a 1–5 scale by five distinct Amazon Mechanical Turk workers. In total, 210 workers contributed 193,208 ratings.

Implicit-feedback formulation. We treat each worker as a user and obtain implicit-feedback tuples (U, R, I), where $U \in \{1, \ldots, 210\}$ is a discrete user ID, I is an image, and R is that user's feedback on I. We compute deterministic CLIP image embeddings $I^e = \text{CLIP}(I) \in \mathbb{R}^d$ using *ViT-H-14* from OpenCLIP Cherti et al. (2023). Following prior work that binarizes ratings for preference modeling, we map the 1–5 scores to binary signals: $R = 1$ ("like") if the rating is ≥ 4, and $R = 0$ ("dislike") otherwise. We then train the embedding prior on triplets (U, R, I^e).

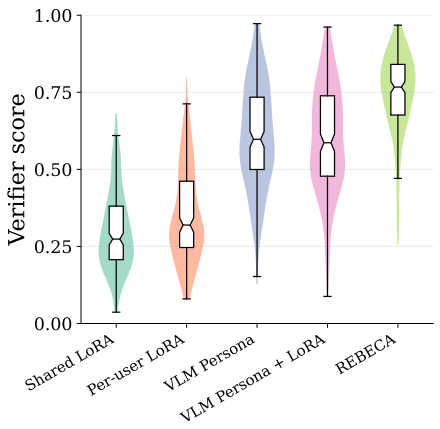


Figure 4: **Model comparison.** REBECA achieves the highest personalization scores, surpassing VLM-based baselines and LoRA variants.

Scope and limitations. FLICKR-AES provides an offline proxy for personalization: the preference signal corresponds to *aesthetic judgments* from annotators rather than long-horizon consumption behavior, contextual intent, or evolving tastes in a deployed recommender system. Accordingly, our claims on this dataset are scoped to learning *stable individual-level* preference structure from weak, scalar feedback. Importantly, captions are not available in FLICKR-AES; personalization is learned solely from users' historical ratings.

4.2 IMAGE GENERATION

To ensure that the sampled embeddings align with user preferences, we set $R = 1$ to indicate a strong preference signal, thereby prioritizing images that align with previously liked content. At first sight, it might appear that training on *liked* and *disliked* images to generate only positive samples is unnecessary. However, this decision is principled, and the reasoning is twofold. First, learning from both positive and negative feedback enables the model to build more informative user embeddings, capturing a fuller picture of individual preferences. Second, there is strong empirical evidence that conditional generative models consistently outperform their unconditional counterparts in terms of sample quality and alignment Dhariwal and Nichol (2021); Donahue and Simonyan (2019).

4.3 EVALUATION

All image generation uses a frozen *Stable Diffusion v1.5* decoder Podell et al. (2024); Rombach et al. (2022) at 512×512 resolution (fp16). Unless stated otherwise, we generate 25 images per user using 50 denoising steps and decoder classifier-free guidance (CFG) = 5. To isolate *user conditioning* from prompt engineering, we use an empty (fixed) system prompt for all methods and vary only the personalization mechanism (user conditioning, LoRA weights, or persona text). The diffusion prior outputs a personalized embedding I^e, which is provided to SD 1.5 equipped with an IP-Adapter Ye et al. (2023). Qualitative per-user grids (multiple samples per user) generated under identical decoding settings (fixed prompt, fixed seed schedule) and varying *prior* CFG are provided in Supplementary Sec. M. As prior CFG increases, generations for a given user become more concentrated and stylistically consistent, reflecting stronger conditioning on that user's preference signal; at lower CFG, samples are more diverse but less tightly aligned. This qualitative trend mirrors the quantitative precision–recall trade-off in Table 2: increasing CFG improves precision by focusing on dominant preference regions while slightly reducing recall due to reduced coverage of each user's broader preference manifold.

4.3.1 BASELINES AND COMPARABILITY

Our goal is to compare REBECA against strong alternatives that (i) operate with a frozen SD 1.5 decoder, and (ii) can be driven by the same supervision available in FLICKR-AES: user IDs and scalar ratings on images, *without* ground-truth captions or paired preference comparisons. We therefore evaluate two families of baselines: (a) diffusion-backbone adaptation (LoRA), and (b) prompt-mediated personalization derived from automatic image captioning (VLM personas). These baselines test whether REBECA's gains can be explained by per-user fine-tuning or by translating implicit feedback into text prompts.

Captions for prompt-based baselines. Several relevant personalization baselines require textual prompts or semantic descriptors. Because FLICKR-AES does not provide captions, we generate *pseudo-captions* using *LLaVA-1.5-7B* Liu et al. (2023) with a fixed JSON-style instruction that requests: a concise caption and short lists of objects, attributes, styles, and colors. We apply the *same* tagging pipeline to all prompt-based baselines to ensure consistency and avoid giving any method privileged metadata. Details of the prompt template and parsing rules are provided in Supplementary F.

Per-user LoRA. We train a separate LoRA adapter Hu et al. (2022) for each user using that user's rated images, keeping the SD 1.5 backbone and text encoder frozen. This baseline represents the strongest per-user fine-tuning strategy under the frozen-decoder constraint. Hyperparameters (rank, steps, warmup) are adapted to the number of images available per user to avoid overfitting and to keep compute comparable (Supplementary G.1).

Shared LoRA. We train a single LoRA adapter on the union of all users and load it into a fresh SD 1.5 pipeline at evaluation time (Supplementary G.2). This tests whether a *global* low-rank update can capture user variation without explicit user conditioning.

VLM Personas (prompt-mediated personalization). Following PMG-style prompt mediation Shen et al. (2024), we construct a per-user textual persona from LLaVA-generated captions/tags, including short descriptions plus positive/negative keyword sets. At generation time, we compose prompts by combining the persona text with sampled positive keywords, and use a fixed negative-prompt template containing standard degradation terms. This baseline tests whether translating implicit feedback into language is sufficient for personalization when the decoder is frozen. Full details are given in Supplementary G.3.

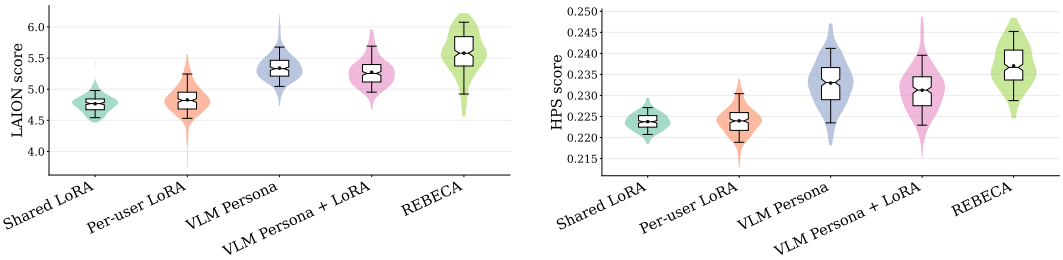
VLM Personas + LoRA. We combine prompt mediation with per-user LoRA by loading each user's LoRA adapter and prompting with that user's persona during generation. This hybrid baseline represents a strong "best of both worlds" alternative: backbone adaptation plus text-mediated preference cues.

Non-comparable or out-of-scope methods. We do not include methods whose training signal or interaction protocol is fundamentally different from our setting, e.g., approaches that require (i) user-authored prompts, (ii) paired preference comparisons, (iii) iterative online feedback loops, or (iv) specialized domains (e.g., constrained templates such as posters). Our focus is on a lightweight offline regime with implicit scalar feedback and a frozen decoder; we discuss these alternatives and their trade-offs in Appendix A.

4.3.2 EXTRA METRICS

Personalization Verifier. Since FLICKR-AES does not provide labels for user preferences on *generated* images, we cannot directly measure whether samples align with an individual user's taste. We therefore train a predictive *verifier* that estimates the probability that a user U will like an image I, a setup commonly used to evaluate generative models when direct human feedback is unavailable (Cobbe et al., 2021; Lightman et al., 2023).

Our verifier is based on *Neural Matrix Factorization* (He et al., 2017). It approximates $\mathbb{P}(R{=}1 \mid U, I)$ via $v(U, I) = g_\gamma\big(U, \texttt{CLIP}(I)\big)$, where g_γ is a neural network, each user ID U is mapped to



(a) LAION aesthetic predictor

(b) HPSv2 overall quality predictor

Figure 5: Quantitative comparison of aesthetic and quality predictors across personalization methods. (a) The LAION aesthetic predictor captures visual appeal independently of textual prompts. (b) The HPSv2 predictor measures prompt-conditioned human preference and image quality. Across both metrics, REBECA consistently achieves the best performance.

a trainable embedding, and $\texttt{CLIP}(\cdot)$ is a fixed image encoder.[5] The verifier is trained with binary cross-entropy on held-out user–image interactions and achieves ROC-AUC ≈ 0.87 on both train and test splits (class-1 prevalence $\approx 30\%$). Unless otherwise stated, we fit the verifier on the same training split used to train REBECA and baselines to maximize the number of labeled interactions; Appendix H shows no statistically significant train–test performance gap, suggesting limited overfitting.

We quantify personalization for a generative model $\hat{p}(I \mid U, R{=}1)$ by the expected verifier score

$$\text{Score}(\hat{p}(I \mid U, R{=}1)) = \mathbb{E}_{I\sim\hat{p}(I|U,R=1)}[\,\hat{v}(U, I)\,], \tag{5}$$

estimated by averaging $\hat{v}(U, I)$ over generated samples for each user.

Aesthetics and Overall Quality. Personalization methods should preserve key aspects of image quality, including aesthetics and overall visual appeal. To evaluate whether REBECA and the baselines maintain high-quality generation while achieving personalization, we employ the LAION aesthetic predictor (LAION-AI, 2022) and the Human Preference Score (HPSv2) (Wu et al., 2023) to assess image aesthetics and perceptual quality. The two methods produce scalar scores that can be used to compare the quality of generated images.

4.4 Results

Precision@k and Recall@k. For each user u, we construct a held-out evaluation pool $\mathcal{P}_u = \mathcal{L}_u \cup \mathcal{D}_u$, where $\mathcal{L}_u$ is the set of that user's *liked* test images (ground-truth positives / relevant set) and $\mathcal{D}_u$ is the set of that user's *disliked* test images (negatives). We generate m samples conditioned on $(u, R{=}1)$, embed each generated image and each image in $\mathcal{P}_u$ with the same CLIP encoder, and rank $\mathcal{P}_u$ by cosine similarity to each generated sample. Let $\text{Top-}k(g, u) \subseteq \mathcal{P}_u$ denote the k nearest neighbors of a generated sample g for user u. We define

$$\text{Prec@}k(u) = \tfrac{1}{m}\textstyle\sum_{g=1}^{m} \mathbf{1}[\text{Top-}k(g,u) \cap \mathcal{L}_u \neq \emptyset]\,,\ \text{Rec@}k(u) = \tfrac{1}{|\mathcal{L}_u|}\textstyle\sum_{\ell\in\mathcal{L}_u} \mathbf{1}[\exists g:\ \ell \in \text{Top-}k(g,u)]\,.$$

We report user macro-averages, i.e., $\text{Prec@}k = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \text{Prec@}k(u)$ and similarly for Rec@k, with F1@k computed from the macro-averaged precision and recall. Table 2 reports results over 210 users. Across both @1 and @5, REBECA matches or surpasses all baselines. At lower prior CFG values the model attains the highest recall, sampling broadly from each user's preference manifold but at lower precision. As CFG increases, generation concentrates around dominant preferences, raising precision while slightly reducing recall. Macro-F1 remains highest for REBECA, confirming the strongest overall precision-recall balance.

Personalization. We compare its performance against several baseline methods. Figure 4 presents box-plots showing the distribution of user scores for each generation approach. REBECA clearly

[5] We use `OpenCLIP-ViT-bigG-14` Cherti et al. (2023), a different backbone than in Section section 4.1, to reduce representation overlap.

Model	@1			@5		
	P	R	F1	P	R	F1
Shared LoRA	0.455	0.718	0.540	0.471	0.991	0.623
Per-user LoRA	0.485	0.713	0.557	0.470	0.986	0.621
VLM Persona	0.507	0.685	0.556	0.472	0.983	0.621
VLM Persona + LoRA	0.511	0.673	0.555	0.475	0.986	0.626
REBECA (CFG=3.0)	0.524	**0.767**	**0.605**	0.478	**0.997**	0.630
REBECA (CFG=5.0)	0.556	0.714	**0.605**	0.478	0.989	0.629
REBECA (CFG=7.0)	0.579	0.647	0.584	0.480	0.976	0.628
REBECA (CFG=9.0)	**0.605**	0.610	0.575	**0.484**	0.972	**0.631**

Table 2: Macro-averaged precision/recall/F1 over 210 users for retrieval from a per-user pool of held-out likes (relevant set) plus held-out dislikes using cosine similarity in CLIP space. Bold indicates best, underline indicates second-best per column.

outperforms all alternatives, with VLM-based methods serving as strong baselines. We also observe that training a separate LoRA per user improves personalization compared to using a shared LoRA across users, though the gain is modest. This suggests that while LoRA introduces efficient low-rank adaptation, it lacks the inductive biases embedded in REBECA that are crucial for personalized image generation. In Appendix J, we show disaggregated results.

Aesthetics and Overall Quality. Figure 5 reports results for both the LAION aesthetic predictor (left) and the HPSv2 human preference metric (right). The relative performance across methods closely mirrors what we observed in Figure 4, with REBECA consistently achieving the highest scores. Note that the HPSv2 metric is prompt-dependent: it evaluates the alignment between textual input and generated images. In our main evaluation (fig. 5), we use the generic prompt to capture overall quality "Realistic image, finely detailed, with balanced composition and harmonious elements." In section K, we further evaluate with an empty prompt. In that setting, all generative methods exhibit similar overall quality, indicating that even in the absence of explicit textual guidance, REBECA maintains high image quality while delivering personalization.

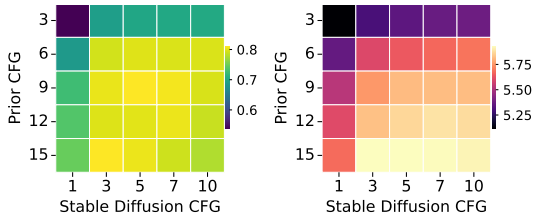


Figure 6: User-averaged scores by SD×Prior CFGs. Left: verifier score. Right: user-averaged LAION aesthetic score. Higher is better; color scales differ per metric.

Ablations. During generation, inference-time parameters such as classifier-free guidance (CFG) and system prompts may, in principle, influence the final output. We analyze REBECA along two axes.

(i) Dual guidance. A joint sweep over the diffusion-prior CFG and the Stable Diffusion CFG (Fig. 6) reveals a clear trade-off between personalization strength and image fidelity. Increasing the prior CFG amplifies the user-specific signal, while a higher SD CFG improves aesthetic quality but mildly attenuates personalization. This interaction motivates our choice of moderate CFG values in the main experiments.

(ii) Prompt control. We further test whether prompt engineering can enhance output quality by evaluating three increasingly structured system prompts (see Supplementary L.1). As shown in Fig. 7, all prompt levels yield nearly identical verifier and LAION scores across CFG settings, indicating that REBECA's personalization is effectively *prompt-independent*. Even strong descriptive and negative prompts do not modulate the personalization signal, which instead resides in the latent user-conditioned embedding distribution learned by the diffusion prior.

4.5 REBECA GENERATES PERSONALIZED IMAGES BEYOND GLOBAL AESTHETICS

A key concern is whether REBECA merely learns to generate *generally* aesthetic images (a global signal) rather than capturing *user-specific* preferences. We run two complementary diagnostics that isolate personalization from overall aesthetic quality (Fig. 8).

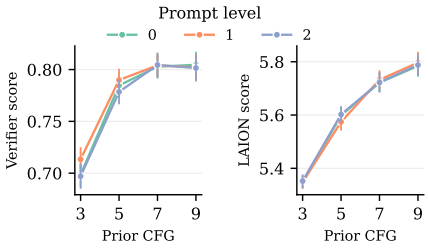


Figure 7: Prompt level for (a) verifier score and (b) LAION aesthetic quality. With REBECA, we do not observe a difference in scores with changing prompts.

User–image matching under randomization. If REBECA captures individual preferences, then a user's verifier score on their *own* held-out images should exceed the score obtained when the same images are assigned to *other* users. For each user u, we compute the median verifier score on correctly matched pairs (u, I) and compare it to the median score under a random reassignment of images to users. We summarize personalization as the per-user gap (correct minus random) and test $H_0 : U \perp\!\!\!\perp I$ via a permutation test (Lehmann et al., 1986) (Appendix I). Correct matches score nearly five standard deviations above random ($p < 10^{-3}$ at $\alpha = 0.05$), providing strong evidence that the verifier captures *user-specific* structure rather than a user-agnostic aesthetic signal. Notably, a purely aesthetic (user-agnostic) model would be invariant to user reassignment and would yield a near-zero gap in Fig. 8(a).

Aesthetics explains only part of the verifier signal. To quantify the extent to which the verifier could be explained by generic aesthetics, we correlate verifier scores with a standard LAION aesthetic predictor. The relationship is significant but modest ($R^2 = 0.32$), indicating that overall aesthetic quality contributes to satisfaction but does not fully account for the verifier's user-conditioned predictions. Taken together, the randomization test (which would fail under a purely global aesthetic model) and the limited correlation with generic aesthetics support that REBECA captures personalization beyond global quality improvements.

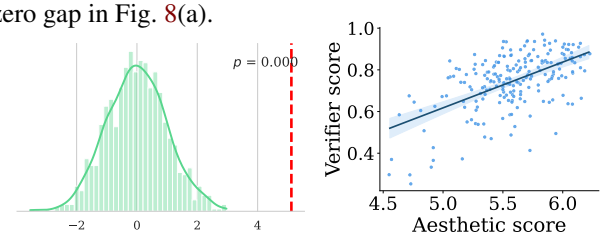


Figure 8: **Personalization is not reducible to global aesthetics.** (a) Permutation test for $H_0 : U \perp\!\!\!\perp I$: correctly matched user–image pairs score nearly five standard deviations above random. (b) Verifier vs. LAION aesthetic score: aesthetics explains part of the signal ($R^2 = 0.32$).

5 DISCUSSION

REBECA demonstrates that implicit feedback alone can drive scalable personalization for diffusion models. By learning a lightweight conditional prior over CLIP embeddings and decoding with a frozen backbone, it provides a simple and general framework for connecting recommender signals with generative models—without paired comparisons, per-user fine-tuning, or LLM mediation. Because the method operates in embedding space, it naturally extends to other modalities such as audio, video, or cross-modal generators equipped with pretrained encoders. We include a discussion on limitations and future work on Appendix B.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7727–7736, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8020–8030, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021.

Jeff Donahue and Karen Simonyan. *Large scale adversarial representation learning*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

LAION-AI. Aesthetic predictor. https://github.com/LAION-AI/aesthetic-predictor, 2022. Accessed: 2025-11-09.

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*. Springer, 1986.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *Forty-second International Conference on Machine Learning*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023.

Jiongnan Liu, Zhicheng Dou, Ning Hu, and Chenyan Xiong. Generate, not recommend: Personalized multimodal content generation, 2025.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 3833–3843, 2024.

Dimitri von Rütte, Elisabetta Fedele, Jonathan Thomm, and Lukas Wolf. Fabric: Personalizing diffusion models withnbsp;iterative feedback. In *Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part XX*, page 385–400, Berlin, Heidelberg, 2025. Springer-Verlag.

Xianquan Wang, Likang Wu, Shukang Yin, Zhi Li, Yanjiang Chen, Feng Hu, Yu Su, and Qi Liu. I-AM-G: Interest augmented multimodal generator for item personalization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21303–21317, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Yifei Xu, Xiaolong Xu, Honghao Gao, and Fu Xiao. Sgdm: An adaptive style-guided diffusion model for personalized text to image generation. *Trans. Multi.*, 26:9804–9813, 2024.

Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. Personalized image generation with large multimodal models. In *Proceedings of the ACM on Web Conference 2025*, page 264–274. ACM, 2025a.

Yiyan Xu, Wuqiang Zheng, Wenjie Wang, Fengbin Zhu, Xinting Hu, Yang Zhang, Fuli Feng, and Tat-Seng Chua. Drc: Enhancing personalized image generation via disentangled representation composition, 2025b.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models, 2024.

Yulai Zhao, Masatoshi Uehara, Gabriele Scalia, Sunyuan Kung, Tommaso Biancalani, Sergey Levine, and Ehsan Hajiramezanali. Adding conditional control to diffusion models with reinforcement learning, 2025.

A Related Work

Parameter-Efficient Fine-Tuning (PeFT). Adapters and low-rank modules provide an efficient means of specializing large text-to-image models without updating the full diffusion backbone Hu et al. (2022); Zhang et al. (2023); Ye et al. (2023). In diffusion models, IP-Adapter Ye et al. (2023) extends controllability to visual references by injecting image-conditioned cross-attention, while LoRA, Textual Inversion, and DreamBooth Hu et al. (2022); Gal et al. (2022); Ruiz et al. (2023); Mou et al. (2024) enable instance-level personalization from a small set of exemplars. However, these approaches typically require explicit reference content (prompts and/or example images) and thus do not directly address implicit-feedback settings where preferences must be inferred from behavior rather than provided as exemplars.

REBECA differs in where personalization lives: instead of adapting the generator to a user or providing exemplar references, it learns a compact user-conditioned *embedding prior* from behavioral data and reuses a frozen decoder. This yields a reusable personalization layer that scales across many users without per-user fine-tuning of the diffusion backbone.

Personalization via Iterative Feedback and Online Alignment. A complementary line of work personalizes generation through iterative feedback loops. FABRIC von Rütte et al. (2025) conditions diffusion sampling on reference images collected through multiple rounds of user feedback, improving results without updating model weights. Other approaches align diffusion models to external objectives using reinforcement learning or policy optimization, applying PPO-style updates to diffusion trajectories or denoising steps Black et al. (2024); Fan et al. (2023); Zhang et al. (2024), or introducing reward-conditioned objectives Zhao et al. (2025). These methods typically depend on prompt-conditioned rewards, dense evaluators, and/or repeated policy-improvement loops, which can be costly at scale.

In contrast, REBECA targets an *offline implicit-feedback* regime (likes/ratings/clicks) and learns a conditional distribution over user-preferred embeddings directly, avoiding iterative feedback loops and RL-style optimization while still enabling a controllable personalization strength via guidance.

Text-Mediated Personalization and Prompt Rewriting. Several recent works personalize text-to-image generation by mediating preferences through language. Tailored Visions Chen et al. (2024) rewrites user prompts using interaction histories, improving alignment by making prompts more expressive and user-specific. Style-guided personalization methods such as SGDM Xu et al. (2024) incorporate style signals from user-provided references to steer diffusion sampling toward consistent visual styles. While effective, these approaches either (i) assume prompt-based interaction and personalization expressed as text, or (ii) require explicit style exemplars at inference time.

REBECA instead learns personalization from implicit feedback without requiring users to author prompts; prompts (when used) can be held fixed, and personalization is driven by the learned conditional embedding prior.

Personalized Multimodal Generation with LMM/VLM Mediation. Another line of work uses large multimodal models (LMMs/VLMs) to infer or represent preferences from histories and then condition generation. PMG Shen et al. (2024) converts behavioral traces into natural language and extracts a mixture of keywords and embeddings to condition a generator. Pigeon Xu et al. (2025a) proposes an LMM-based pipeline with explicit preference alignment stages (including pairwise preference alignment) to guide personalized image generation from noisy histories. DRC Xu et al. (2025b) tackles entanglement in LMM guidance by explicitly disentangling style preferences and semantic intentions before composing latent instructions for generation. These methods are powerful but often introduce additional inference stages (LLM/LMM/VLM mediation) and/or rely on richer supervision such as pairwise comparisons.

REBECA differs fundamentally: it dispenses with LMM mediation and pairwise supervision, learning a lightweight user-conditioned diffusion prior directly in embedding space from scalar implicit feedback, then decoding with a frozen generator.

Generative Recommendation and Personalization Beyond Catalogs. Recent work argues for going beyond item retrieval toward generating new personalized multimodal content. Generate, Not Recommend Liu et al. (2025) frames a paradigm where large multimodal models generate per-

sonalized items (including images) from recommendation data. Related work in item personalization likewise augments generation with user interests inferred from interaction histories Wang et al. (2024). REBECA complements these directions by providing a minimal, plug-and-play mechanism for open-domain image generation from implicit feedback: a user-conditioned embedding prior that can be attached to off-the-shelf diffusion decoders without per-user backbone adaptation.

B Limitations and Future work

Limitations. REBECA is studied in an offline setting, where user preferences are learned from a fixed dataset, and the system is not designed to adapt as new users or interaction signals arrive rapidly. While this simplifies analysis, it limits applicability in online or streaming environments. We also do not address the cold start problem, where users provide very few interactions, making it difficult to infer meaningful preferences. In addition, our approach relies on CLIP embeddings to represent user taste, which may not capture all preference dimensions, and we do not model temporal preference drift or evolving user interests.

Future work. Our long-term vision is personalized generation for recommendation systems that go beyond selecting items from a fixed catalog, as reflected in the name REBECA, which stands for REcommendations BEyond CAtalogs. Future work will focus on working towards better recommendation systems; more concretely, next steps could focus on extending REBECA to online and continual learning settings, for example, via batching and selective retraining of lightweight components as new signals arrive, and on addressing cold start users through shared priors or transfer across users. Additional directions include modeling preference dynamics over time, integrating REBECA more tightly with recommender pipelines, and enabling generative systems that continuously create and adapt content in response to user feedback.

C Prior Disclosure and Delta vs. Preliminary Extended Abstract

Prior disclosure (anonymized). A publicly available extended abstract described a preliminary prototype of this direction. To preserve double-blind review, we do not cite or link it here. This submission is a substantially expanded and revised version, and we provide an explicit change log below to clarify the relationship and the substantive extensions.

High-level summary. Relative to the preliminary extended abstract, the present submission (i) finalizes the REBECA prior architecture and training recipe via systematic search and substantially improved stability/efficiency; (ii) introduces a new, stronger baseline suite to test prompt-mediated and fine-tuning alternatives under missing-caption conditions; and (iii) strengthens the evaluation protocol and diagnostics through redesigned model components, additional quantitative benchmarks, and expanded qualitative evidence.

Detailed delta (itemized).

1. **Finalized REBECA prior architecture via systematic search.** The preliminary extended abstract used an earlier prior design. This submission introduces the final lightweight attention-based diffusion prior with a learned tokenizer and reports an architecture/objective/noise-schedule search over depth/heads/width/token count and diffusion objectives/schedules, yielding a stable configuration that trains in < 10 minutes on a single RTX-4090.
2. **New baseline suite (not present previously).** This submission adds multiple strong baselines designed to stress-test REBECA under the same frozen-decoder setting: (i) shared LoRA, (ii) per-user LoRA, (iii) VLM-persona prompting, and (iv) VLM-persona + LoRA. Since the dataset lacks captions, we additionally introduce a consistent caption/tag extraction pipeline (fixed instruction, structured parsing) used *uniformly* across all prompt-based baselines.
3. **Redesigned controlled simulation to better match the frozen-decoder pipeline.** The controlled experiment is revised to more faithfully emulate "latent/embedding-space sampling $\rightarrow$ fixed decoder," replacing the earlier autoencoder-based bottleneck with a VAE-based setup and updated visualization.

4. **New quantitative retrieval benchmark (Precision@k / Recall@k / F1).** This submission introduces a full macro-averaged retrieval evaluation over users (e.g., @1 and @5), with explicit definitions of the per-user candidate pool and the relevant set (held-out likes) and a cosine-similarity ranking protocol.
5. **Verifier redesign and improved reporting.** The preference verifier architecture is updated from the earlier low-rank/matrix-factorization style to a neural matrix factorization model to better capture user–image interactions. We report verifier predictive performance (ROC-AUC) and add implementation details to reduce representation overlap (e.g., distinct CLIP backbone for the verifier).
6. **Strengthened diagnostics against the "global aesthetics" hypothesis.** The permutation test diagnostic was explored in the extended abstract; in this submission it is *re-run and strengthened* using the updated verifier and the revised evaluation protocol. We additionally include a correlation analysis between verifier scores and a standard aesthetic predictor, quantifying the extent to which global aesthetics explain (or fail to explain) the personalization signal.
7. **New inference-time ablations and prompt study.** This submission adds broader inference-time ablations (e.g., guidance settings) and a prompt-level study to disentangle prompt effects from user-conditioning effects under a frozen decoder.
8. **Substantially expanded qualitative evidence and revised figures.** We add large-scale qualitative grids showing multiple generations per user under identical decoding settings, and we revise core figures/diagrams for clarity. We additionally include a new embedding-navigation/editing experiment as an auxiliary capability analysis.

Bottom line. The preliminary extended abstract introduced the motivating direction and an early prototype. The present submission finalizes the architecture and training recipe, adds a new and stronger baseline suite, and substantially strengthens evaluation (quantitative benchmarks, improved verifier, re-run diagnostics, and expanded qualitative evidence).

D REBECA PRIOR ARCHITECTURE

Our conditional diffusion prior is a lightweight and designed for personalized CLIP embeddings. Given a noisy embedding I_t^e and conditioning, the model predicts the clean embedding I_0^e.

We first tokenize the CLIP embedding using a learned tokenizer that maps the 1D embedding into a small set of tokens. Each token is projected to a shared hidden dimension. Conditioning is injected through three tokens: user, rating, and timestep embeddings. These are mapped into a conditioning vector via a two-layer MLP.

The core of the model consists of L *PriorBlocks*, each combining:

- AdaLN-Zero layers for stable, scale-free conditioning,
- self-attention over the image embedding tokens,
- a gated cross-attention mechanism that selectively attends to the conditioning tokens, and
- a zero-initialized MLP residual block for controlled feature updates.

All residual pathways are initialized at zero, ensuring training stability and preventing early over-conditioning. After L blocks, the model projects tokens back to the original token dimension and merges them to reconstruct the predicted embedding. The entire architecture contains only 4.4M parameters, so each full training run takes under 10 minutes on a single RTX 4090.

E TRAINING PROTOCOL

E.1 HYPERPARAMETER SEARCH

Overview. REBECA's lightweight diffusion prior enables a fully exhaustive hyperparameter search, something typically infeasible for user–conditioned generative models. All models were trained on the same hardware (RTX 4090) using identical data splits and random seeds to ensure comparability.

Architectures. We evaluate a broad family of adapter architectures, including transformer–based variants, cross–attention mechanisms, and a direct residual diffusion prior (`rdp`), which emerged as the best-performing and most stable model. The grid includes variations in depth, attention width, and embedding dimensionality:

- **Depth:** $\{6, 8, 12\}$ layers
- **Heads:** $\{4, 8\}$ attention heads
- **Hidden size:** $\{128, 256\}$
- **Token count:** $\{16, 32\}$ (when applicable)

Diffusion and Objective. The prior is trained using several diffusion objectives to test robustness:

- ϵ–prediction
- sample–prediction
- v–prediction

We experiment with multiple noise schedules—`laplace`, `squaredcos_cap_v2`, and `epsilon`—and fix the number of timesteps to 1000 to avoid confounding training comparisons.

Optimization. All models use the `AdamW` optimizer and a `ReduceLROnPlateau` scheduler. Hyperparameters were intentionally kept narrow to isolate architectural effects:

- learning rate: 1×10^{-4}
- batch size: 64
- samples per user: 100
- no gradient clipping or additional normalization

User Conditioning. We sweep over configurations for user thresholding and normalization, though the final model uses no score normalization and no thresholding:

- normalization: `none`
- user threshold: 0

Compute and Stability. Every configuration trains in under 25 minutes on a single 4090 GPU, making the full grid search (dozens of runs) computationally tractable. This is a major advantage of `REBECA`'s formulation: the prior is small enough to train repeatedly, allowing principled exploration of design choices rather than relying on heuristics or one-off tuning.

Selection Criterion. For each run, we evaluate validation loss across objectives. We select the best performing models of each class and inspect a sample of 25 generated images, five for each of the first five users, for image quality.

E.2 Final Configuration

After completing the exhaustive grid search described above, a single model emerged as the most stable and best-performing across all evaluation metrics: the `rdp` (`REBECA` Diffusion Prior) with a lightweight 6-layer architecture.

Backbone. All REBECA results use `Stable-Diffusion v1.5` as the image generator. We load a standard IP-Adapter (`h94/IP-Adapter`, `ip-adapter_sd15.bin`) to provide the visual-conditioning interface, and disable the safety checker for reproducibility.

Winning REBECA Prior. The best configuration corresponds to the following hyperparameters:

- **Architecture:** `rdp` prior
- **Layers / Heads / Width:** 6 layers, 8 heads, hidden dimension 128
- **Tokens:** 32 learned latent tokens
- **Image embedding dim:** 1024 (CLIP-ViT-bigG)
- **Users:** 210
- **Score classes:** 2 (like/dislike)

Diffusion Objective. The winning configuration uses:

- **Prediction type:** `sample`
- **Timesteps:** 1000
- **Noise schedule:** `squaredcos_cap_v2`
- **Clip-sample:** disabled

The full diffusion scheduler is:

```
DDPMScheduler(
    num_train_timesteps=1000,
    beta_schedule="squaredcos_cap_v2",
    clip_sample=False,
    prediction_type="sample"
)
```

Weights. All final experiments load the trained prior from:

```
comprehensive_study_20250830_013540/
    modelrdp_num_layers6_num_heads8_hidden_dim128_tokens32_
    lr0.0001_optadamw_schreduce_on_plateau_bs64_
    nssquaredcos_cap_v2_ts1000_spu100_csFalse_objsample_normnone_uthr0
```

Compute. This model trains in under $\sim$10 minutes on a single RTX 4090, which enables the exhaustive search strategy and provides a key practical advantage over existing personalization pipelines.

F Image Tagging Pipeline

We employ the open-source checkpoint `llava-hf/llava-1.5-7b-hf` loaded with `transformers`. Each image is processed with the following fixed instructions:

```
You are a tagging engine. Return ONLY a JSON object with fields:
{"caption": str, "objects": [str], "attributes": [str],
 "styles": [str], "colors": [str]}.
Rules: lowercase; <=10 items total; no nulls.
```

G Baseline Specification

G.1 LoRA per User

We fit a single adapter per user and we employ each user's liked images as training data. Due to the variability in dataset sizes, we must adapt the training configuration for each user depending on the number of liked images. See table 3 for the various configurations. We adapt Hugging Face's Parameter Efficient Fine-Tuning (*PeFT*) script for Stable Diffusion to filter by user tags and loop over their IDs and configurations for training.

G.2 Shared LoRA

Collaborative filtering is a fundamental concept in Recommender Systems, and it posits a common latent representation for users. In this spirit, a single LoRA model with rank $r = 512$ is calibrated with all users simultaneously. For training hyperparameters, see Table 4.

G.3 VLM-Persona Generation

Prompt Construction. Each user's persona defines positive and negative keyword lists. When few positive terms exist, fallback categories ("nature, landscape, portrait, cityscape, animals") are used. The builder cycles deterministically through available terms and appends either the user's persona or a default stylistic tail.

Table 3: **Per-user LoRA training hyperparameters.** The LoRA rank (r), scaling factor (α), dropout, training steps, and warmup schedule are adapted based on the number of available user images (N_i).

N_i range	LoRA r	LoRA α	Dropout	Steps	Warmup
$N_i < 8$	8	8	0.10	1200	100
$8 \leq N_i \leq 24$	16	16	0.07	1500	120
$25 \leq N_i \leq 60$	32	32	0.05	1700	150
$61 \leq N_i \leq 120$	64	64	0.05	2000	200
$N_i > 120$	128	128	0.05	2500	200

Table 4: **Shared LoRA training hyperparameters.** The shared LoRA rank (r), scaling factor (α), dropout, training steps, and warmup schedule.

LoRA r	LoRA α	Dropout	Steps	Warmup
512	512	0.05	10000	1000

Implementation. We employ the `Diffusers` library for inference in half precision (fp16) on GPU, with the safety checker disabled for consistent reproducibility.

Table 5: **Fallback prompt components** used when user profiles lack sufficient detail.

Type	Default values or description
Positive keywords	{`nature, landscape, portrait, cityscape, animals`}
Style suffix	"high quality, detailed, natural lighting"
Negative prompt	{`low quality, blurry, deformed, overexposed, underexposed`}

Generation Summary.

- **Model:** Stable Diffusion v1.5 (`stable-diffusion-v1-5`)
- **Inference steps:** 50
- **Guidance scale:** 5.0
- **Images per user:** 25
- **Total users:** 210
- **Seeding:** $\text{seed}(u_i, j) = 42 + 10{,}000 \times i + j$
- **Output:** Serialized image bundles per user (`.imgs`)

This procedure, implemented in `vlm_personas.ipynb`, serves as the *text-conditioned personalization baseline* for comparison against LoRA-based and diffusion-prior models.

H VERIFIER DIAGNOSTICS

Figure 9 shows that the verifier achieves nearly identical performance on the training and test sets. In the second panel, we report the results of a bootstrap analysis of the test data, demonstrating that the training ROC-AUC lies within two standard errors of the test ROC-AUC. This indicates that the verifier does not overfit and generalizes well to unseen samples.

Algorithm 1 Prompt Builder for User u_i

Input: profile (persona, pos, neg), image index j
$k \leftarrow$ number of positive terms to include (1–2)
$P \leftarrow$ cycle through pos deterministically for k terms
if P is empty **then**
    $P \leftarrow$ random fallback from Table 5
end if
style $\leftarrow$ persona text if available else default suffix
prompt $\leftarrow P +$ style
negprompt $\leftarrow$ user negatives or default negatives
return (prompt, negprompt)

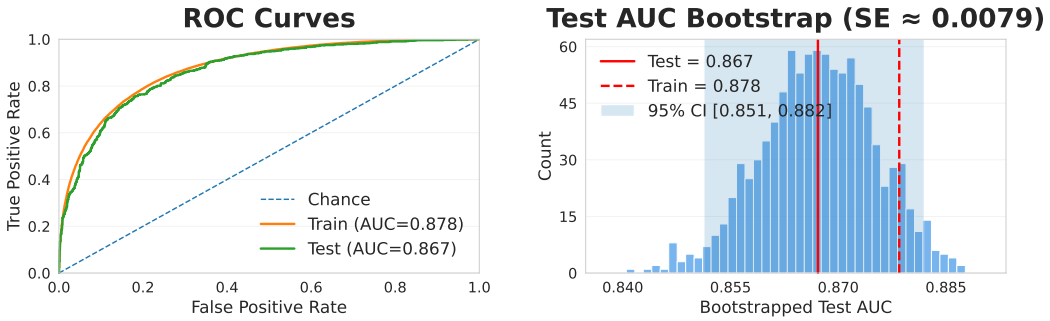


Figure 9: Verifier performance on the training and test sets. The two curves are nearly indistinguishable, and the ROC-AUC values are not statistically different.

I Formally testing for personalization

Algorithm 2 implements a permutation test used to formally verify the personalization aspect of REBECA. In Algorithm 2, $\widehat{\text{Score}}$ is the median verifier score across users and $\widehat{\text{Score}}_b$ is the median verifier score across users after permutation using the random seed equals b.

Algorithm 2 Testing for REBECA personalization

Require: 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

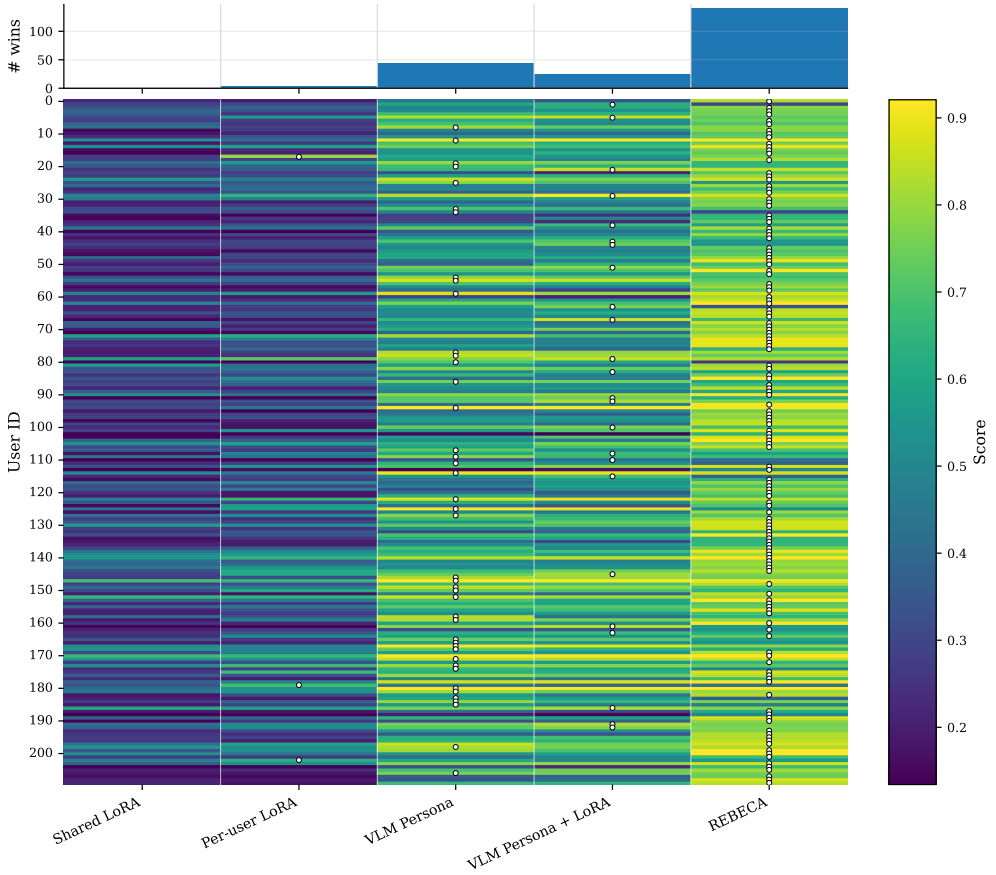


Figure 10: Verifier scores for each user across different generation methods. REBECA achieves the highest scores for most users, indicating stronger personalization performance.

J Extra plots for the personalization results

Figures 10 and 11 present the performance of each method across users in terms of verifier scores and relative rankings, respectively. In both cases, REBECA consistently outperforms all baseline methods for the vast majority of users.

K Extra plots for the aesthetics and overall quality results

Figure 12 reports the HPSv2 results obtained when using an empty prompt. In this zero-prompt setting, all generative methods display comparable overall quality. Notably, REBECA is capable of delivering personalized generations without hurting overall generation quality.

L Ablations

L.1 System Prompts

To support the prompt–control ablation reported in the main paper, we evaluate REBECA under three system–prompt levels c_t, ranging from no prompt to strongly opinionated aesthetic prompts with aggressive negative prompts. For all experiments, we freeze the best-performing diffusion-prior checkpoint and generate personalized embeddings for all 210 users, which are then decoded using a Stable Diffusion v1.5 pipeline with an IP-Adapter (identical sampling settings across conditions).

The three prompt levels used in the ablation are:

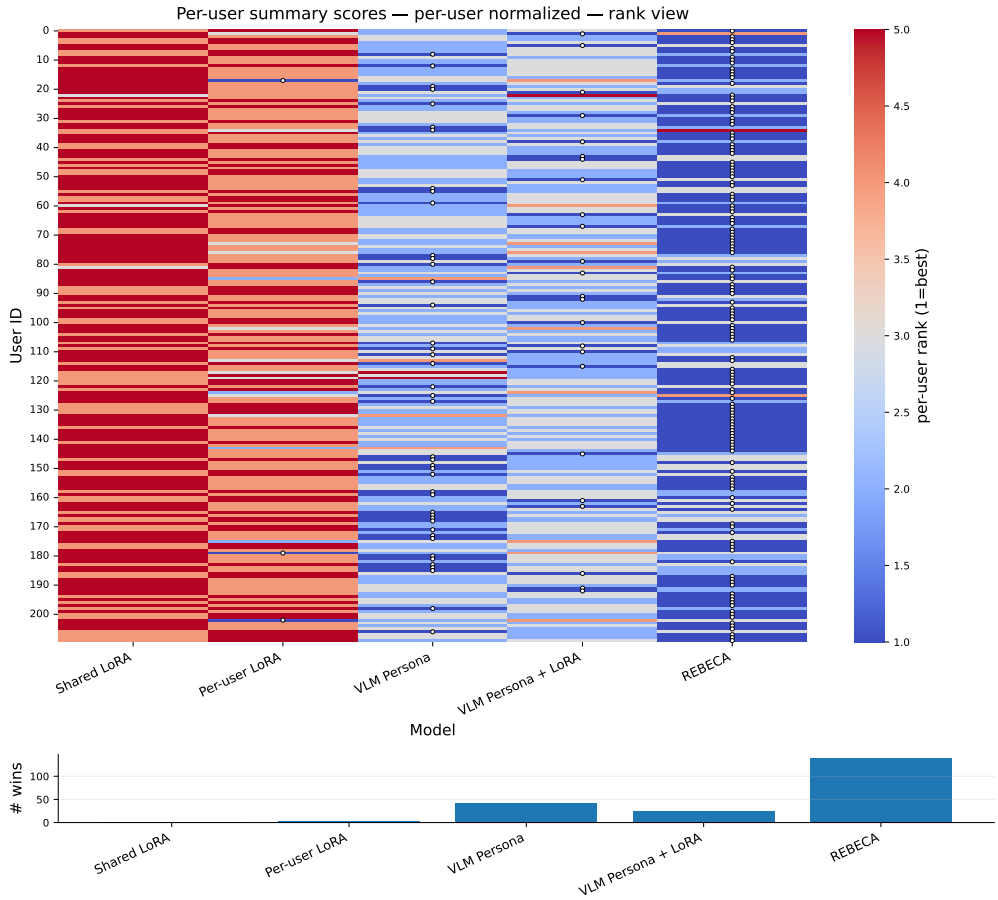


Figure 11: Relative ranking of generation methods per user. REBECA ranks highest for the majority of users, outperforming all baselines in personalized generation quality.

Level 0: No prompt.

- **Positive prompt:** ""
- **Negative prompt:** ""

Level 1: Mild quality prompt.

- **Positive prompt:** "high quality photo"
- **Negative prompt:** "bad quality photo, letters"

Level 2: Strong descriptive/negative prompts.

- **Positive prompt:** "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."
- **Negative prompt:** "deformed, ugly, wrong proportion, frame, watermark, low res, bad anatomy, worst quality, low quality"

For each prompt level, we generate 10 personalized images per user (2,100 images per condition). The outputs are subsequently evaluated by our verifier model. As discussed in the main text, we observe *no statistically significant effect* of prompt level c_t at any REBECA CFG value, indicating that personalization arises from the diffusion prior rather than from prompt engineering.

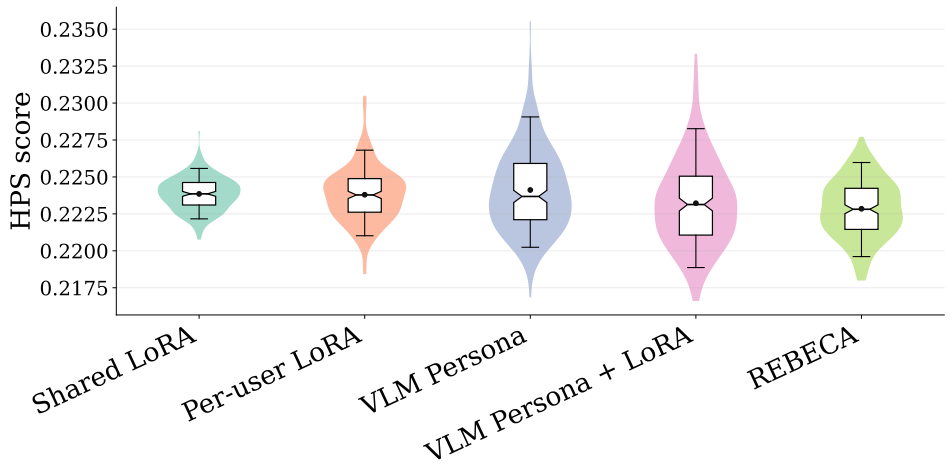


Figure 12: **HPSv2 results under the zero-prompt setting.** When no textual prompt is provided, all models achieve similar overall quality scores, indicating that prompt content mainly influences alignment rather than visual fidelity. REBECA maintains competitive quality while preserving personalization robustness.

M SAMPLES

In this Section, we add samples for the first twelve users by ID in our dataset. We generate the images below by fixing the seed and increasing the prior CFG.

With a lower CFG, the images are more diverse, capturing broader regions of the user preference manifold. Hence, the higher recall results in lower CFG values. As CFG increases, sampling concentrates around fewer high preference regions, and generations become more demarcated across users.

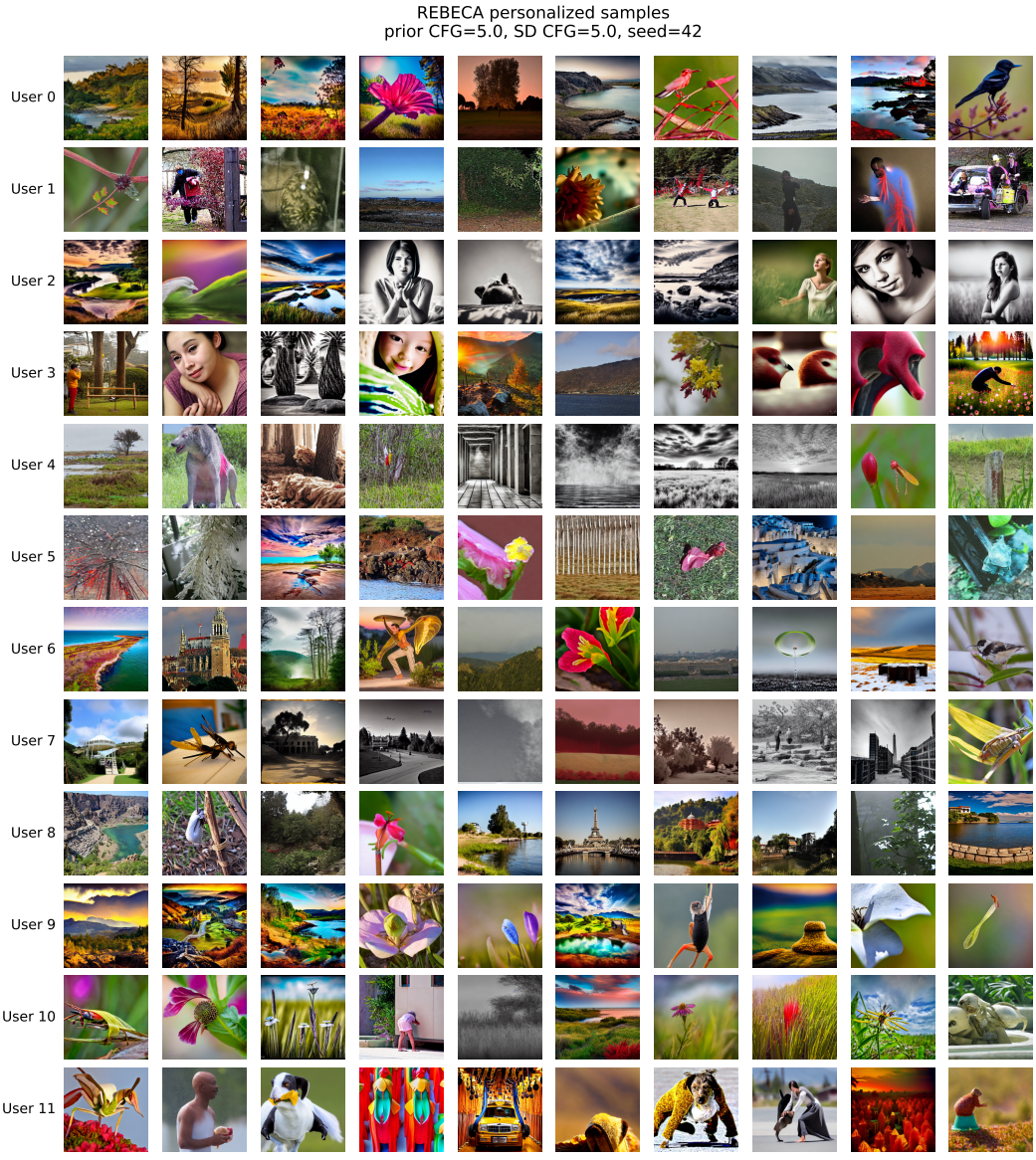


Figure 13: Sample images for users 1–12 with prior CFG 5.0.

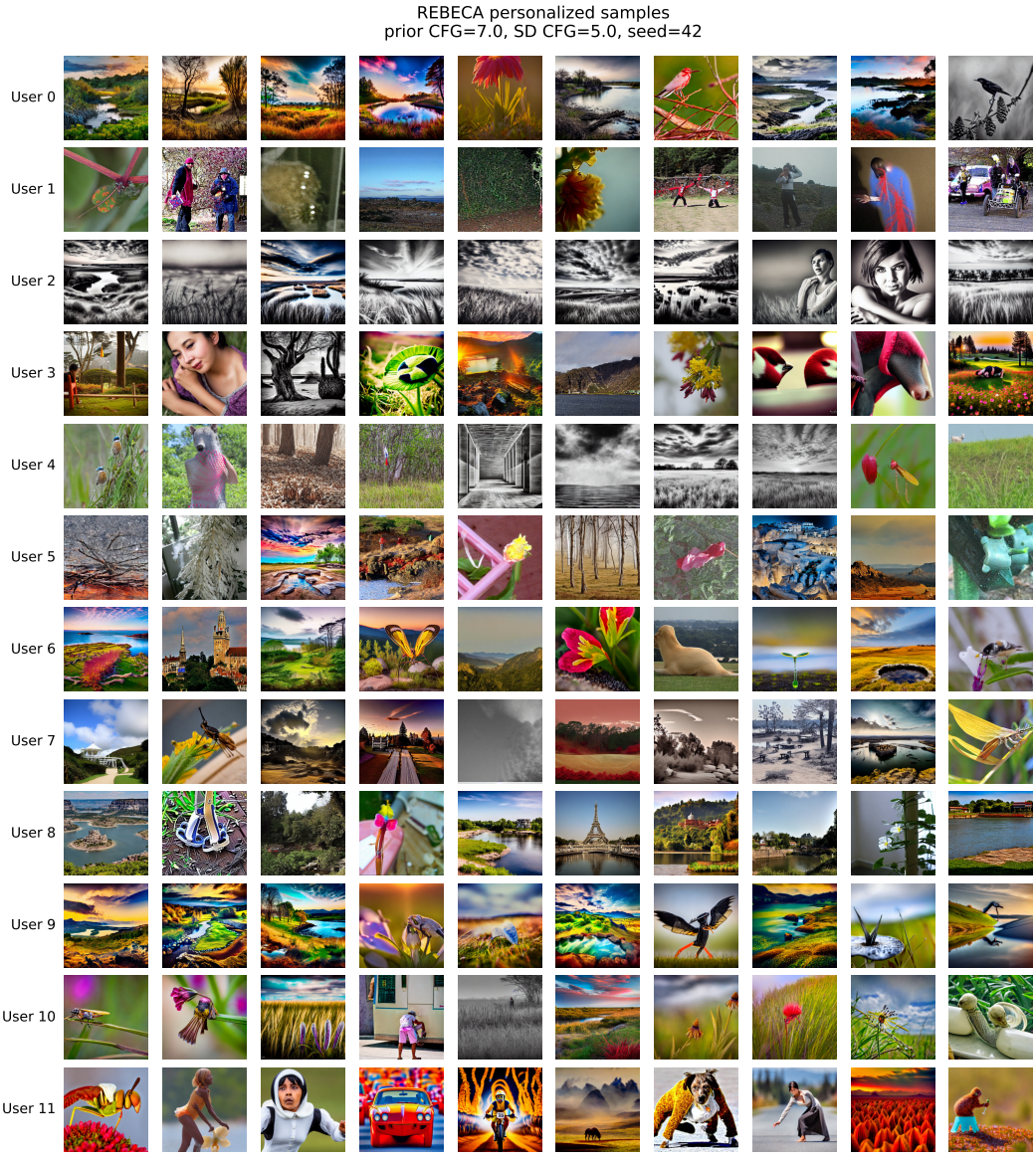


Figure 14: Sample images for users 1–12 with prior CFG 7.0.

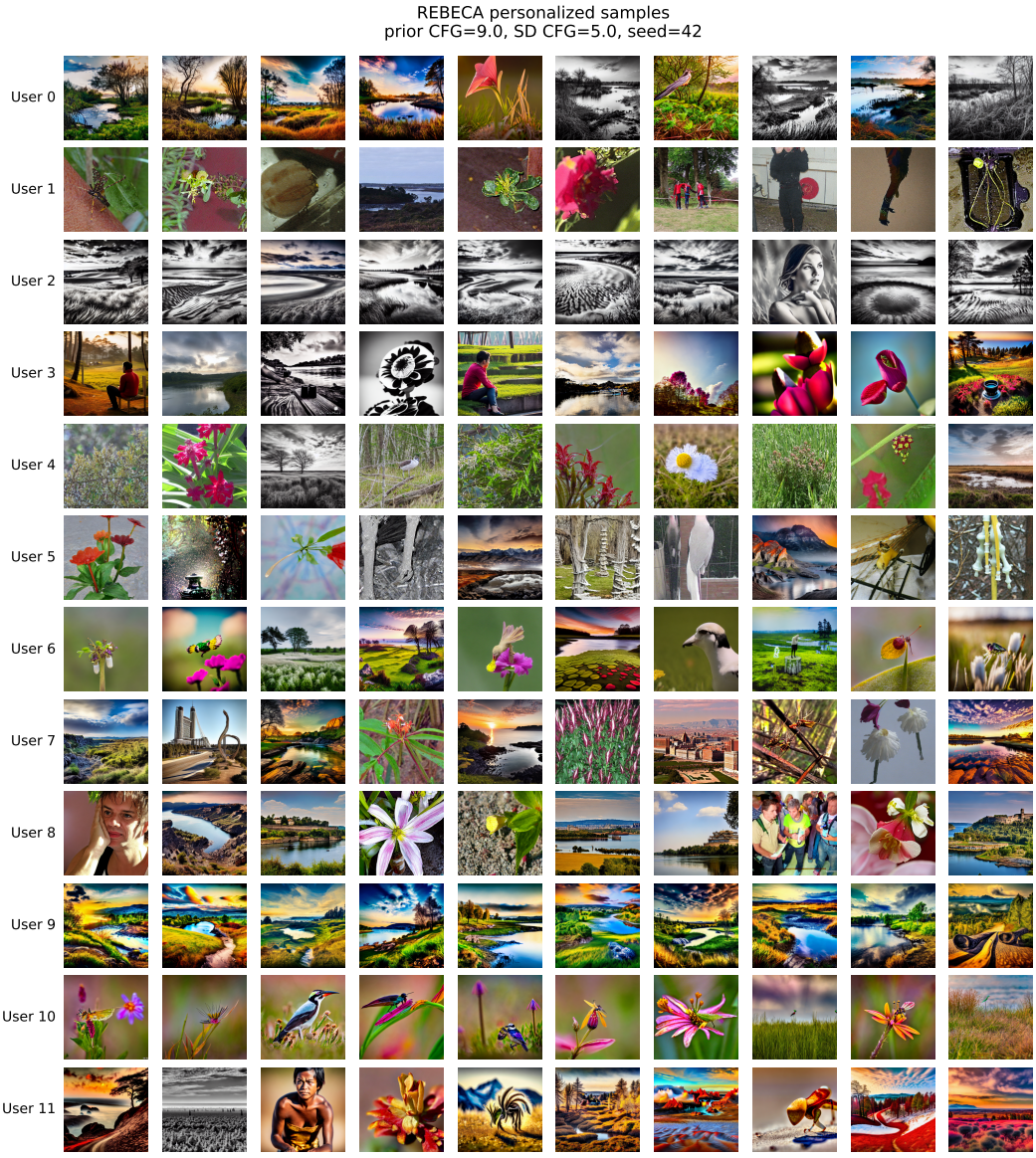


Figure 15: Sample images for users 1–12 with prior CFG 9.0.