PIPE: Personalized Image-generation via Preference Encoding

Moonkyung Ryu [1] **Chih-Wei Hsu** [1] **Avinab Saha** [1] **Ofir Nabati** [1] **Guy Tennenholtz** [1] **Junfeng He** [1] **Craig Boutilier** [1]

Abstract

While modern text-to-image (T2I) models excel at generating high-quality images, they are typically trained to optimize with respect to generalized, population-level preferences. This homogeneous approach ignores the diverse, individual tastes and aesthetic judgments of different users. In this work, we propose a novel framework that learns fine-grained user preferences without relying on computationally expensive visual language models (VLMs) or prompt-sensitive text profiles. Instead, we introduce a robust, continuous user representation that models a user's reward function as a linear combination of K base user types. We learn user-specific weights λ_u via logistic regression on pairwise preference data to construct a continuous user embedding. This embedding is integrated into the diffusion process via an IP-Adapter, and fine-tuned using Diffusion-DPO. Our approach consistently generates images aligned with individual reward functions, achieving a 66.2% win rate against a pretrained SDXL baseline and a 63.2% win rate against the state-of-the-art PPD framework.

1. Introduction

The prevailing paradigm in text-to-image (T2I) generation relies on diffusion models trained on massive datasets to satisfy homogeneous, population-level preferences. Techniques such as reinforcement learning from human feedback (RLHF) (Lee et al., 2023; Fan et al., 2023) and direct preference optimization (DPO) (Wallace et al., 2024) can align these models to produce images that the "average" user finds visually appealing and well-aligned with their text prompts. This approach, while effective for general image synthesis, overlooks the heterogeneity and individuality inherent in human aesthetic judgment and creative preferences.

[1]Google Research, Mountain View, CA, USA. Correspondence to: Moonkyung Ryu <mkryu@google.com>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

Our research goal is to move beyond this one-size-fits-all methodology. Specifically, we develop a novel framework, *PIPE (Personalized Image-generation via Preference Encoding)*, that models and learns diverse, fine-grained user preferences. This user-specific representation is integrated into a diffusion process to generate images that are not only plausible, but aligned with the unique tastes of that user. By capturing these personalized stylistic preferences, we elevate the utility and creative expressiveness of T2I models, allowing them to serve as more effective, personalized creative tools.

To achieve this, we adapt a pretrained T2I model to allow conditioning on our user representation, motivated by the recent state-of-the-art framework PPD (Dang et al., 2025), which uses an IP-Adapter (Ye et al., 2023) to inject user embedding into a diffusion model. However, PPD uses a highly constrained approach to learn user preference representations: for each user, a set of $N = 4$ few-shot examples—comprising a caption, a preferred image, and a dispreferred image—is processed by a pretrained visual language model (VLM) to generate a user profile in text. PPD uses the VLM hidden state of the final token in this text profile as the user preference representation. This VLM-based approach has several key drawbacks:

1. **Data Bottleneck:** Because the VLM's context window limits the number of examples that can be used, the amount of historical user preference data is very restricted, preventing robust, generalized learning.

2. **Inconsistent Profiling:** The generated text profiles can vary considerably for the same user depending on the subset of examples chosen.

3. **Semantic Gap:** The generated text profiles, even given a *full* history, will generally be lossy relative to the data used to generate them, and may fail to capture nuanced or latent user preferences.

4. **Computational Overhead:** VLM inference using multi-modal input data for each user is compute-intensive, slow, and will not scale to the demands of many real-time applications.

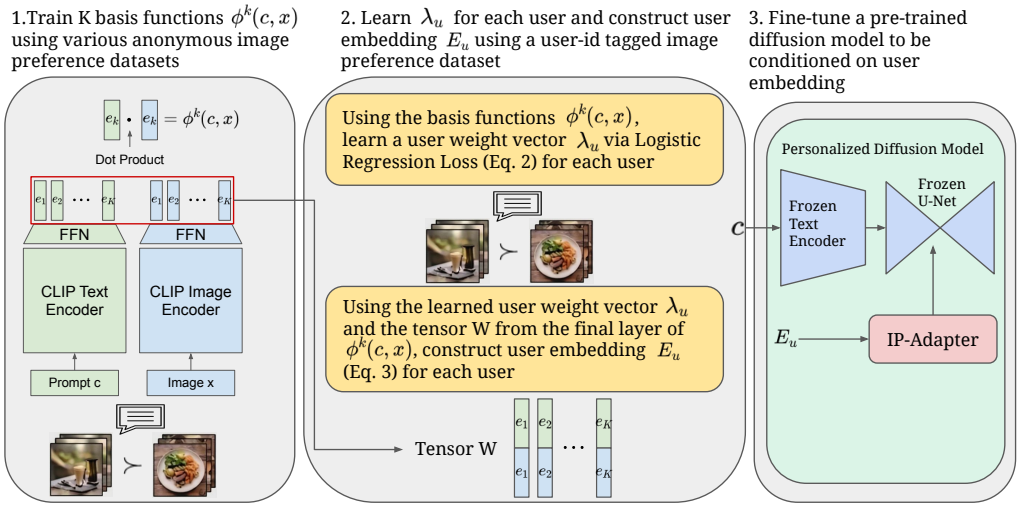


Figure 1. Our proposed framework. The PIPE user embedding E_u is derived from historical pairwise preferences and injected into the frozen U-Net via IP-Adapter cross-attention layers.

In this paper, we propose a novel method of learning continuous user preference representations that gracefully resolves these problems: PIPE enables highly scalable and effective personalization.

2. Method

A critical difference between our approach and prior work like PPD (Dang et al., 2025) is how we learn the representation for user preference on images. We abandon constrained text-profile generation in favor of a continuous, scalable embedding approach. To do so, our approach exploits prior user models trained on large T2I datasets over many users, such as the PASTA user model (Nabati et al., 2025).

2.1. PASTA: Personalized T2I

PASTA (Nabati et al., 2025) uses Reinforcement Learning (RL) to optimize agents that interact with users to *elicit* their image preferences for personalized T2I. One part of their methodology is the use of an expectation-maximization (EM) procedure to train a set of K discrete user types from a large pool of user image preference data, specifically, on these three large scale datasets: HPDv2 (Wu et al., 2023), Pick-a-Pic v2 (Kirstain et al., 2023), and Simulacra (Pressman et al., 2022). Each type $k \in K$ has a learned reward model $\phi^k(c,x)$ that associates a type-specific reward to any image x given prompt c. We extend this user-type model below.

2.2. PIPE User Model

The user types above are used for data generation/simulation in (Nabati et al., 2025), but they could also be used as reward models to fine-tune K distinct T2I models. However, the use of discrete user types limits the degree of personalization that can be injected into T2I, and the use of distinct models sacrifices generalization capability. Instead, we exploit the K types as the basis for a continuous user preference representation upon which T2I can be conditioned in a single model.

We use PASTA to establish a base user type model $\phi(c,x) \rightarrow \mathbb{R}^K$ that scores images x given prompts c for each of K foundational user types. This model uses frozen CLIP image and text encoders, training only the multi-layer perceptron (MLP) layers on top of each encoder.

We extend the PASTA user model by treating its discrete types as a *basis* for a continuous user type representation.[1] Specifically, we assume a user-specific weight vector $\lambda_u = (\lambda_u^1, \lambda_u^2, ..., \lambda_u^K)^\top$, where a user u's *personalized* reward function is:

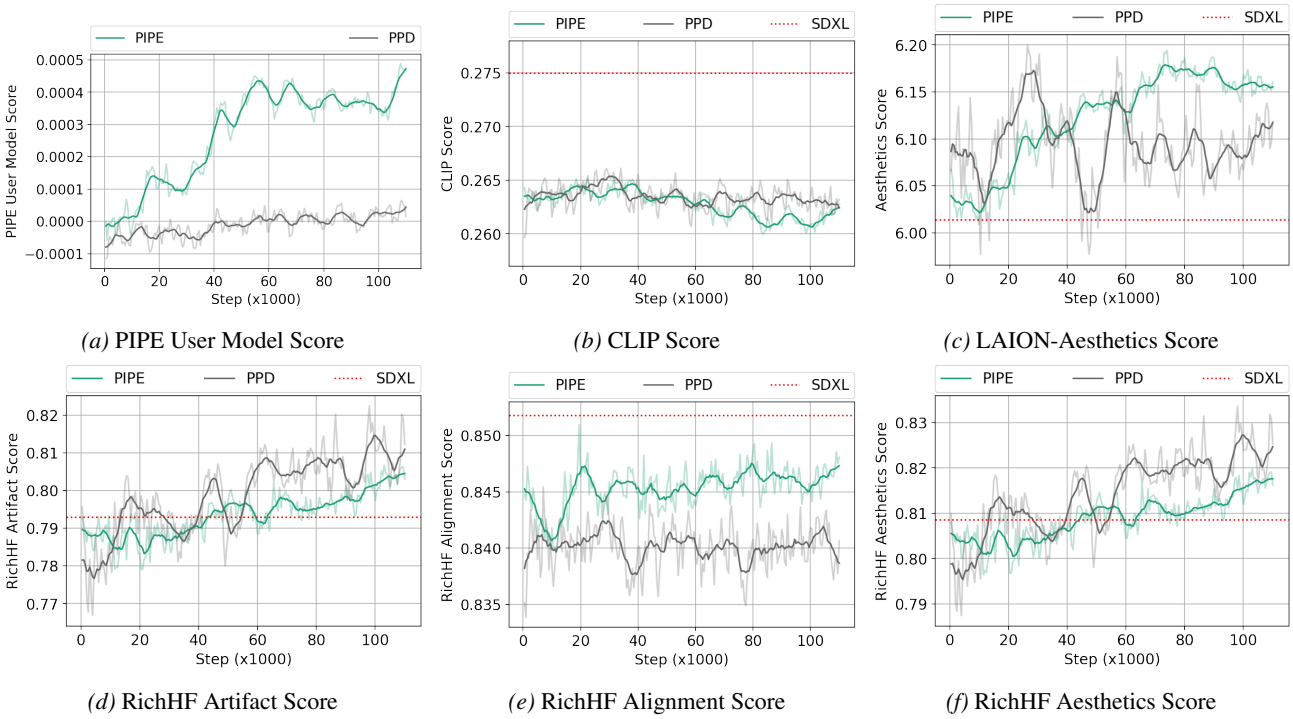
$$r_u(c,x) = \sum_{k=1}^{K} \lambda_u^k \phi^k(c,x) = \lambda_u^\top \phi(c,x) \tag{1}$$

Given a set of N_u pairwise image preferences from user u, where (x_1, x_2) means $x_1 \succ_u x_2$, we use logistic regression to estimate u's preference vector λ_u, minimizing:

$$\mathcal{L}(\lambda_u) = -\sum_{n=1}^{N_u} \log \sigma(\lambda_u^\top (\phi(c^n, x_1^n) - \phi(c^n, x_2^n))) \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function. This efficiently processes *all* historical user data, circumventing the context-

[1]Other bases could be used, and even optimized for our approach.



(a) PIPE User Model Score *(b)* CLIP Score *(c)* LAION-Aesthetics Score

(d) RichHF Artifact Score *(e)* RichHF Alignment Score *(f)* RichHF Aesthetics Score

Figure 2. Checkpoints during training are evaluated using 500 unseen prompts and 6 different models that measure alignment to individual user preference (a), text-image alignment (b, e), image aesthetics (c, f), and the presence of generative artifacts in images (d).

window limitations of VLMs. We refer to $r_u(c, x)$ as the *PIPE user model*.

2.3. PIPE User Representation and its Usage

Once λ_u is learned for a user u, we construct their user embedding E_u to represent their personal preferences. E_u is simply a linear combination of λ_u and the final-layer weights of the user type model:

$$E_u = \lambda_u^\top W \tag{3}$$

where W is a tensor representing the weights of the final feed-forward network layers.

Unlike the PPD representation, which may fluctuate wildly based on the specific few-shot examples drawn for u, E_u provides a highly stable, consistent, global representation of the user's aesthetic preferences. We refer to E_u as *PIPE user representation (or embedding)*. To personalize T2I, we inject cross-attention layers into a T2I foundation (diffusion) model using the IP-Adapter technique (Ye et al., 2023) to accept the user representation. The model is fine-tuned using diffusion-DPO (Wallace et al., 2024). We optimize only the parameters for the IP-Adapter layers while keeping the foundation model parameters frozen (see Fig. 1).

Method	Win Rate vs SDXL	Win Rate vs PPD
PPD (Dang et al., 2025)	56.8%	–
Ours (PIPE)	**66.2%**	**63.2%**

Table 1. Quantitative comparison of Win Rates against baselines evaluated on individual preference alignment.

3. Experiments

We demonstrate the performance of our model with experiments designed to assess its capabilities in personalized preference alignment compared to a pretrained model and PPD, a state-of-the-art personalization baseline.

3.1. Experiment Setting

Dataset. We use the Pick-a-Pic v2 (Kirstain et al., 2023), a pairwise preference dataset of the form $D = \{u^{(n)}, c^{(n)}, x_w^{(n)} \succ x_l^{(n)}\}_{n=1}^N$, where u is a user ID, c is a text caption, x_w is the winning (i.e., preferred) image and x_l is the losing image. The dataset contains 58K text prompts and 800K image pairs labeled by 5K unique users.

Foundation Model. We use the open-source text-to-image model SDXL (Podell et al., 2023) as our base foundation model, fine-tuned using IP-Adapters as above.

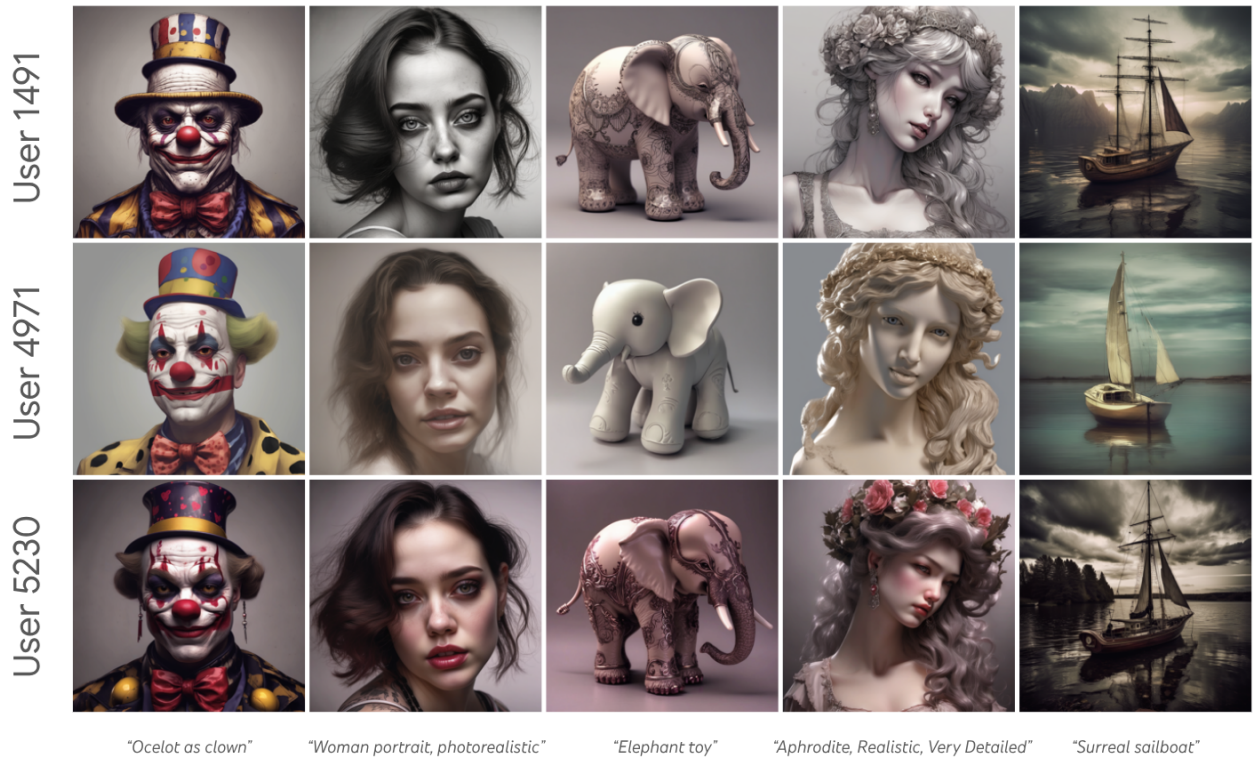


Figure 3. PIPE generates diverse images for the same prompts across users, surfacing their preferences.

Evaluation. The PIPE user model (Eq. 1) is used as the primary automated judge to measure alignment to individual preferences. We also use CLIP (Radford et al., 2021), LAION-Aesthetics (Schuhmann, 2022) and RichHF (Liang et al., 2024) to measure general image quality, such as text-image alignment, global image aesthetics, and the presence of generative artifacts.

3.2. PIPE User Model

We first assess the effectiveness of the learned PIPE user model (Eq. 1). The standard Pick-a-Pic test dataset only contains about 18% of the users found in the training data. To address this discrepancy, we grouped the training data by user ID and created a custom 9:1 train-to-test split, ensuring the same users are represented in both sets.

We optimize the user-specific weight vector λ_u for each u using Eq. 2 on this new training data split. During training, any paired images marked as a "tie" were randomly assigned a winning and losing label. We then evaluate the model's ability to accurately predict user image preferences on the test split, intentionally excluding any ties from the evaluation. The PIPE model achieves a prediction accuracy of **64.1%**, outperforming PickScore (Kirstain et al., 2023) which achieves a **56.8%** accuracy despite being trained on the same data.

3.3. User Preference Alignment

Next, we evaluate how well the fine-tuned text-to-image model aligns with personalized user preferences. Using 500 previously unseen prompts from the original Pick-a-Pic test dataset, we generate 500 corresponding images at various training checkpoints. Figure 2 tracks the performance of models fine-tuned with either the PIPE or PPD representations across multiple evaluation criteria during training.

For both the PIPE and PPD, fine-tuning the base model enhances personal preference alignment and overall image aesthetics. However, this improvement comes at the cost of a slight decrease in strict text-to-image alignment. To quantitatively compare performance, we select the optimal checkpoint for each method—this is the checkpoint with the highest score using the PIPE user model. As detailed in Table 1, fine-tuning a pre-trained model with PIPE user representations results in a substantial improvement in personal alignment. When relying on the PIPE user model as an automated judge, our fine-tuned SDXL model achieves a notable **66.2% win rate** against the baseline pre-trained SDXL model. Furthermore, our method secured a **63.2% win rate** when compared directly against the state-of-the-art PPD framework, which relies on the representation derived from VLMs.

We refer to Fig. 3 for examples of results generated by the PIPE user representation.

4. Conclusion

In this work, we proposed a scalable and robust framework for aligning text-to-image models with diverse, personalized aesthetic preferences. By extending the PASTA user model, we efficiently capture fine-grained individual tastes using a lightweight user representation vector learned directly from pairwise preference data. Our method circumvents the data bottleneck and prompt-inconsistency issues associated with VLM-based user profiling.

By projecting this representation into a dense user embedding and carefully injecting it into a diffusion model using an IP-Adapter and DPO-based training, we successfully condition the diffusion generation process on individual user preferences. Extensive evaluation demonstrates that our approach significantly outperforms both the foundation SDXL model and PPD, a state-of-the-art personalization baseline. This work marks a significant step toward unlocking text-to-image systems that are truly customized, high-fidelity creative tools capable of adapting to the nuance of individual user preferences. Our findings align with very recent advancements in the field, such as the Premier framework (Wang et al., 2026), which similarly captures individual preferences by injecting trainable user embeddings to a diffusion model. Future work will involve a direct quantitative comparison with the Premier framework and algorithms for user onboarding designed to elicit preference data more efficiently.

References

Dang, M., Singh, A., Zhou, L., Ermon, S., and Song, J. Personalized preference fine-tuning of diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 8020–8030, June 2025.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback. *arXiv preprint arxiv:2302.12192*, 2023.

Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., Ke, J., Dvijotham, K. D., Collins, K. M., Luo, Y., Li, Y., Kohlhoff, K. J., Ramachandran, D., and Navalpakkam, V. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19401–19411, June 2024.

Nabati, O., Tennenholtz, G., Hsu, C., Ryu, M., Ramachandran, D., Chow, Y., Li, X., and Boutilier, C. Preference adaptive and sequential text-to-image generation. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45362–45394. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/nabati25a.html.

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, 2023.

Pressman, J. D., Crowson, K., and Contributors, S. C. Simulacra aesthetic captions. Technical Report Version 1.0, Stability AI, 2022. url https://github.com/JD-P/simulacra-aesthetic-captions .

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Schuhmann, C. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed: 2024-05-24.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., and Naik, N. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238, June 2024.

Wang, Z., Wei, Y., Zhou, X., Zhang, T., Liang, T., Bai, Y., Zhang, H., and Zuo, W. Premier: Personalized preference modulation with learnable user embedding in text-to-image generation. *arXiv preprint arxiv:2603.20725*, 2026.

Wu, X., Sun, K., Zhu, F., Zhao, R., and Li, H. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023.

Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023.