RectifID: Personalizing Rectified Flow with Anchored Classifier Guidance

Zhicheng Sun[1] **, Zhenhao Yang**[3]**, Yang Jin**[1]**, Haozhe Chi**[1]**, Kun Xu**[2]**, Kun Xu**[2]**,**
Liwei Chen[2]**, Hao Jiang**[1]**, Yang Song, Kun Gai**[2]**, Yadong Mu**[1*]
[1]Peking University, [2]Kuaishou Technology,
[3]University of Electronic Science and Technology of China
{sunzc,myd}@pku.edu.cn

Abstract

Customizing diffusion models to generate identity-preserving images from user-provided reference images is an intriguing new problem. The prevalent approaches typically require training on extensive domain-specific images to achieve identity preservation, which lacks flexibility across different use cases. To address this issue, we exploit classifier guidance, a training-free technique that steers diffusion models using an existing classifier, for personalized image generation. Our study shows that based on a recent rectified flow framework, the major limitation of vanilla classifier guidance in requiring a special classifier can be resolved with a simple fixed-point solution, allowing flexible personalization with off-the-shelf image discriminators. Moreover, its solving procedure proves to be stable when anchored to a reference flow trajectory, with a convergence guarantee. The derived method is implemented on rectified flow with different off-the-shelf image discriminators, delivering advantageous personalization results for human faces, live subjects, and certain objects. Code is available at https://github.com/feifeiobama/RectifID.

1 Introduction

Recent advances in diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020) have ignited a surge of research into their customizability. A prominent example is personalized image generation, which aims to integrate user-defined subjects into the generated image. This plays a pivotal role in AI art creation, empowering users to produce identity-consistent images with greater customizability beyond text prompts. Nevertheless, there remain significant challenges in accurately preserving the subject's identity and being flexible to a variety of personalization needs.

Existing personalization methods are limited in these two aspects, as they require an extra finetuning or pre-training stage. For example, the pioneering works (Gal et al., 2023; Ruiz et al., 2023a) finetune conditional embeddings or model parameters per subject, resulting in suboptimal efficiency and identity consistency due to lack of domain knowledge. On the other hand, the recently prevailing tuning-free methods (Wei et al., 2023; Ye et al., 2023; Li et al., 2024; Wang et al., 2024b) pre-train a conditioning adapter to encode subject features into the generation process. However, their models must be pre-trained on extensive domain-specific data, *e.g.* LAION-Face 50M (Zheng et al., 2022), which is costly in the first place, and cannot be transferred flexibly across different data domains, *e.g.* from human faces to common live subjects and objects, and even to multiple subjects.

To address both challenges of identity consistency and flexibility, we advocate a training-free approach that utilizes the guidance of a pre-trained discriminator without extra training of the generative model. This methodology is well-known as classifier guidance (Dhariwal and Nichol, 2021), which modifies

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

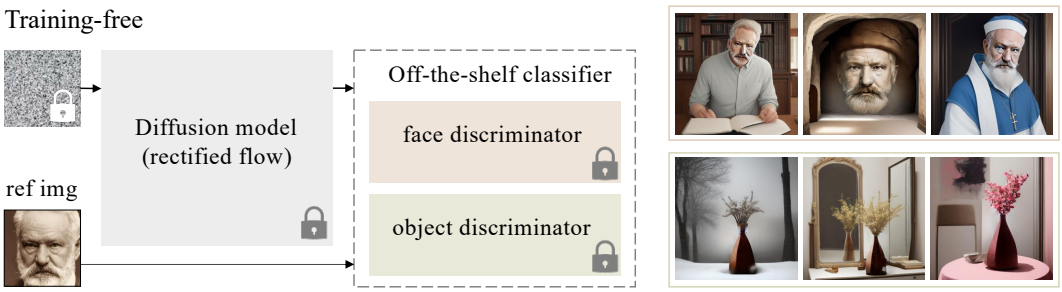


Figure 1: Illustration of training-free classifier guidance. Left: an off-the-shelf discriminator can be reused to steer the existing diffusion model, *e.g.* rectified flow, to generate identity-preserving images. Right: personalized image generation results for human faces and objects using our proposed method.

an existing denoising process using the gradient from a pre-trained classifier. The rationale behind our exploitation is twofold: first, it directly harnesses the discriminator's domain knowledge for identity preservation, which may be a cost-effective substitute for training on domain-specific datasets; secondly, keeping the diffusion model intact allows for plug-and-play combination with different discriminators, as shown in Fig. 1, which enhances its flexibility across various personalization tasks. However, the original classifier guidance is largely limited in the reliance on a special classifier trained on noised inputs. Despite recent efforts to approximate the guidance (Kim et al., 2022a; Liu et al., 2023b; Wallace et al., 2023; Ben-Hamu et al., 2024), they have mainly focused on computational efficiency, and have yet to achieve sophisticated performance on personalization tasks.

Technically, to extend classifier guidance for personalized image generation, our work builds on a recent framework named rectified flow (Liu et al., 2023a) featuring strong theoretical properties, *e.g.* the straightness of its sampling trajectory. By approximating the rectified flow to be ideally straight, the original classifier guidance is reformulated as a simple fixed-point problem concerning only the trajectory endpoints, thus naturally overcoming its reliance on a special noise-aware classifier. This allows flexible reuse of image discriminators for identity preservation in personalization tasks. Furthermore, we propose to anchor the classifier-guided flow trajectory to a reference trajectory to improve the stability of its solving process, which provides a convergence guarantee in theoretical scenarios and proves even more crucial in practice. Lastly, a clear connection is established between our derived anchored classifier guidance and the existing approximation practices.

The derived method is implemented for a practical class of rectified flow (Yan et al., 2024) assumed to be piecewise straight, in combination with face or object discriminators. This provides flexibility for a range of personalization tasks on human faces, live subjects, certain objects, and multiple subjects. Extensive experimental results on these tasks clearly validate the effectiveness of our approach. Our contributions are summarized as follows: (1) We propose a training-free approach to flexibly personalize rectified flow, based on a fixed-point formulation of classifier guidance. (2) To improve its stability, we anchor the flow trajectory to a reference trajectory, which yields a theoretical convergence guarantee when the flow is ideally straight. (3) The proposed method is implemented on a relaxed piecewise rectified flow and demonstrates advantageous results in various personalization tasks.

2 Background

Personalized image generation studies incorporating user-specified subjects into the text-to-image generation pipeline. To preserve the subject's identity, the seminal works Textual Inversion (Gal et al., 2023) and DreamBooth (Ruiz et al., 2023a) finetune conditional embeddings or model parameters for each subject, which imposes high computational costs. Subsequent literature resorts to more efficient parameters (Hu et al., 2022; Han et al., 2023; Yuan et al., 2023) or a pre-trained subject encoder (Wei et al., 2023; Ye et al., 2023) to allow personalization within a few minutes or even without tuning. At the other end, a recent trend is the reuse of existing discriminators to improve identity consistency, such as extracting discriminative face features as the condition (Ye et al., 2023; Wang et al., 2024b) or as a training objective for the encoder (Peng et al., 2024; Gal et al., 2024; Guo et al., 2024). However, these models require extensive pre-training on domain-specific data, *e.g.* LAION-Face 50M (Zheng et al., 2022). In contrast, our method is a training-free approach that exploits existing discriminators based on the recent rectified flow model, allowing flexible personalization for a variety of tasks.

Rectified flow is an instance of flow-based generative models (Song et al., 2021; Xu et al., 2022; Liu et al., 2023a; Albergo and Vanden-Eijnden, 2023; Lipman et al., 2023). They aim to learn a velocity field $\boldsymbol{v}$ that maps random noise $\boldsymbol{z}_0 \sim \pi_0$ to samples from a complex distribution $\boldsymbol{z}_1 \sim \pi_{\text{data}}$ via an ordinary differential equation (ODE):

$$d\boldsymbol{z}_t = \boldsymbol{v}(\boldsymbol{z}_t, t)dt. \tag{1}$$

Instead of directly solving the ODE (Chen et al., 2018), rectified flow (Liu et al., 2023a) simply learns a linear interpolation between the two distributions by minimizing the following objective:

$$\min_{\boldsymbol{v}} \int_0^1 \mathbb{E}\left[\|(\boldsymbol{z}_1 - \boldsymbol{z}_0) - \boldsymbol{v}(\boldsymbol{z}_t, t)\|^2\right] dt. \tag{2}$$

This procedure straightens the flow trajectory and thus allows faster sampling. Ideally, a well-trained rectified flow is a straight flow with uniform velocity $\boldsymbol{v}(\boldsymbol{z}_t, t) = \boldsymbol{v}(\boldsymbol{z}_0, 0)$ following:

$$\boldsymbol{z}_t = \boldsymbol{z}_0 + \boldsymbol{v}(\boldsymbol{z}_t, t)t. \tag{3}$$

Recently, rectified flow has shown promising efficiency (Liu et al., 2024b) and quality (Esser et al., 2024; Yan et al., 2024) in text-to-image generation. Our work extends its capabilities and theoretical properties to personalized image generation via classifier guidance.

Classifier guidance, initially proposed for class-conditioned diffusion models (Dhariwal and Nichol, 2021), introduces a test-time mechanism to adjust the predicted noise $\epsilon(\boldsymbol{z}_t, t)$ based on the guidance from a classifier. Given condition c and classifier output $p(c|\boldsymbol{z}_t)$, the adjustment is formulated as:

$$\hat{\epsilon}(\boldsymbol{z}_t, t) = \epsilon(\boldsymbol{z}_t, t) + s \cdot \sigma_t \nabla_{\boldsymbol{z}_t} \log p(c|\boldsymbol{z}_t), \tag{4}$$

where s denotes the guidance scale, and σ_t is determined by the noise schedule. Noteworthy, the condition c is not restricted to class labels, but can be extended to text (Nichol et al., 2022) and beyond. However, it is largely limited by the reliance on a noise-aware classifier for noised inputs $\boldsymbol{z}_t$, which restricts the use of most pre-trained discriminators that only predict the likelihood $p(c|\boldsymbol{z}_1)$ on clean images. Consequently, its usefulness is limited in practice. See Appendix B for more related work.

3 Method

This work aims at customizing rectified flow with classifier guidance. We show that the above limit of classifier guidance may be solved with a simple fixed-point solution for rectified flow (Section 3.1). To improve its stability, Section 3.2 proposes a new anchored classifier guidance with a convergence guarantee. Lastly, the implementation and applications are described in Sections 3.3 and 3.4.

3.1 Classifier Guidance for Rectified Flow

This section first derives the vanilla classifier guidance for rectified flow, and then present an initial attempt to remove the need for the noise-aware classifier $p(c|\boldsymbol{z}_t)$, which is based on a new fixed-point solution of classifier guidance assuming that the rectified flow is ideally straight.

The classifier guidance can be derived as modifying the potential associated with the rectified flow. According to the Helmholtz decomposition, a velocity field v may be decomposed into:

$$\boldsymbol{v}(\boldsymbol{z}_t, t) = \nabla_{\boldsymbol{z}_t} \phi(\boldsymbol{z}_t, t) + \boldsymbol{r}(\boldsymbol{z}_t, t), \tag{5}$$

where ϕ is a scaler potential and $\boldsymbol{r}$ is a divergence-free rotation field. They can be determined by solving the Poisson's equation $\nabla^2 \phi(\boldsymbol{z}_t, t) = \nabla \cdot \boldsymbol{v}(\boldsymbol{z}_t, t)$, but this is beyond our focus. We directly add a new potential proportional to the log-likelihood to simulate classifier guidance, as follows:

$$\hat{\boldsymbol{v}}(\boldsymbol{z}_t, t) = \nabla_{\boldsymbol{z}_t} \left[\phi(\boldsymbol{z}_t, t) + s \cdot \log p(c|\boldsymbol{z}_t)\right] + \boldsymbol{r}(\boldsymbol{z}_t, t), \tag{6}$$

where s denotes the guidance scale, and $\hat{}$ is used to distinguish the new flow from the original one. Subtracting the above two equations yields the vanilla classifier guidance, similar in form to Eq. (4):

$$\hat{\boldsymbol{v}}(\boldsymbol{z}_t, t) = \boldsymbol{v}(\boldsymbol{z}_t, t) + s \cdot \nabla_{\boldsymbol{z}_t} \log p(c|\boldsymbol{z}_t). \tag{7}$$

While this classifier guidance should allow for test-time conditioning of rectified flow, it cannot be applied in the absence of noise-aware classifier $p(c|\boldsymbol{z}_t)$. In the following, we show that this limitation may be overcome by exploiting the straightness property of rectified flow.

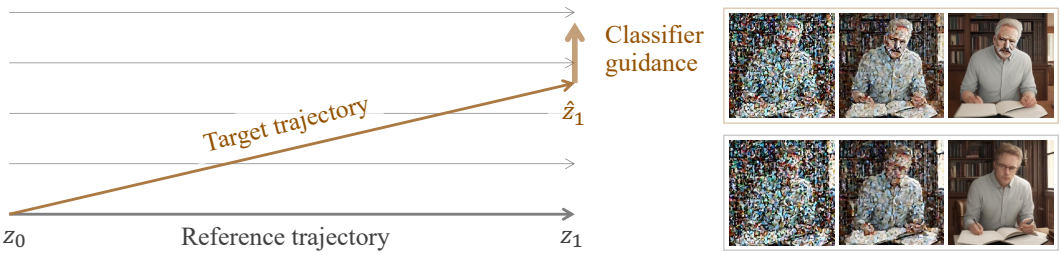


Figure 2: Illustration of anchored classifier guidance for rectified flow. Left: we propose to guide the flow trajectory while implicitly enforcing it to flow straight and stay close to a reference trajectory. Right: comparison of the new trajectory with the reference trajectory (in the last three sampling steps).

Attempt to bypass noise-aware classifier. We make a key observation that the intermediate classifier guidance $\nabla_{\boldsymbol{z}_t} \log p(c|\boldsymbol{z}_t)$ can be circumvented by approximating the new flow trajectory to be straight (an ideal guidance should preserve the properties of rectified flow) and focusing on the endpoint $\boldsymbol{z}_1$. Formally, substituting $t = 1$ in Eqs. (3) and (7) allows skipping any intermediate guidance terms:

$$\begin{aligned} \boldsymbol{z}_1 &= \boldsymbol{z}_0 + \hat{\boldsymbol{v}}(\boldsymbol{z}_1, 1) \\ &= \boldsymbol{z}_0 + \boldsymbol{v}(\boldsymbol{z}_1, 1) + s \cdot \nabla_{\boldsymbol{z}_1} \log p(c|\boldsymbol{z}_1). \end{aligned} \tag{8}$$

Interestingly, this turns out to be a fixed-point problem w.r.t. $\boldsymbol{z}_1$, suggesting that the classifier-guided flow trajectory could be solved iteratively by numerical methods such as the fixed-point iteration, without knowing the noise-aware classifier. This greatly enhances the flexibility of classifier guidance to a variety of off-the-shelf image discriminators. However, our further analysis reveals both empirical (Section 4.3) and theoretical evidence questioning the convergence of this iterative approach:

Proposition 1. *There exist Lipschitz continuous functions $\boldsymbol{v}(\boldsymbol{z}_1, 1)$ and $\nabla_{\boldsymbol{z}_1} \log p(c|\boldsymbol{z}_1)$, such that the fixed-point iteration for solving the target trajectory based on Eq.* (8) *is not guaranteed to converge by the Banach fixed-point theorem (Banach, 1922), irrespective of the choice of $s > 0$.*

Proof. Consider the following construction. Let $\boldsymbol{v}(\boldsymbol{z}_1, 1)$ and $\nabla \boldsymbol{z}_1 \log p(c|\boldsymbol{z}_1)$ be identical functions with a Lipschitz constant greater than 1. Then, the Lipschitz constant of the right-hand side of the fixed-point equation is greater than 1 for any $s > 0$. This violates the Banach fixed-point theorem's requirement for a Lipschitz constant strictly less than 1, thus convergence is not guaranteed. □

Proposition 1 shows that the derived fixed-point solution may not always be practical. Intuitively, even with a small perturbation at $\boldsymbol{z}_1$, the target flow trajectory estimated by Eq. (8) could diverge significantly after iterated updates, which hinders the controllability of rectified flow. This motivates us to anchor the target flow trajectory to a reference trajectory to stabilize its solving process.

3.2 Anchored Classifier Guidance

This section establishes a new type of classifier guidance based on a reference trajectory. The idea is to constrain the new trajectory to be straight and near the reference trajectory, as illustrated in Fig. 2. It provides a better convergence guarantee and a certain degree of interpretability.

Let $\hat{\boldsymbol{z}}_t$ and $\boldsymbol{z}_t$ represent two flow trajectories originating from the common starting point $\boldsymbol{z}_0$ with or without classifier guidance. The symbol $\hat{}$ denotes the new trajectory with classifier guidance. Assuming the two trajectories are close and straight (ideally preserving the characteristics of rectified flow), their difference can be estimated based on Eq. (7) and the first-order Taylor expansion:

$$\begin{aligned} \hat{\boldsymbol{v}}(\hat{\boldsymbol{z}}_t, t) - \boldsymbol{v}(\boldsymbol{z}_t, t) &= \boldsymbol{v}(\hat{\boldsymbol{z}}_t, t) + s \cdot \nabla_{\hat{\boldsymbol{z}}_t} \log p(c|\hat{\boldsymbol{z}}_t) - \boldsymbol{v}(\boldsymbol{z}_t, t) \\ &\approx [\nabla_{\boldsymbol{z}_t} \boldsymbol{v}(\boldsymbol{z}_t, t)] \, (\hat{\boldsymbol{z}}_t - \boldsymbol{z}_t) + s \cdot \nabla_{\hat{\boldsymbol{z}}_t} \log p(c|\hat{\boldsymbol{z}}_t) \\ &= [\nabla_{\boldsymbol{z}_t} \boldsymbol{v}(\boldsymbol{z}_t, t) t] \, (\hat{\boldsymbol{v}}(\hat{\boldsymbol{z}}_t, t) - \boldsymbol{v}(\boldsymbol{z}_t, t)) + s \cdot \nabla_{\hat{\boldsymbol{z}}_t} \log p(c|\hat{\boldsymbol{z}}_t) \\ &= [\boldsymbol{I} - \nabla_{\boldsymbol{z}_t} \boldsymbol{z}_0] \, (\hat{\boldsymbol{v}}(\hat{\boldsymbol{z}}_t, t) - \boldsymbol{v}(\boldsymbol{z}_t, t)) + s \cdot \nabla_{\hat{\boldsymbol{z}}_t} \log p(c|\hat{\boldsymbol{z}}_t), \end{aligned} \tag{9}$$

where the final step is derived from Eq. (3). From here, a new form of classifier guidance is obtained:

$$\hat{\boldsymbol{v}}(\hat{\boldsymbol{z}}_t, t) = \boldsymbol{v}(\boldsymbol{z}_t, t) + s \cdot [\nabla_{\boldsymbol{z}_0} \boldsymbol{z}_t] \, \nabla_{\hat{\boldsymbol{z}}_t} \log p(c|\hat{\boldsymbol{z}}_t). \tag{10}$$

This new classifier guidance anchors the target velocity to a predetermined reference velocity $\boldsymbol{v}(\boldsymbol{z}_t, t)$ that is dependent only on t and irrelevant to the current state $\hat{\boldsymbol{z}}_t$, thereby constraining the target flow trajectory near the reference trajectory and improving its controllability. Next, we extend its applicability to the more common scenarios where the noise-aware classifier $p(c|\hat{\boldsymbol{z}}_t)$ is absent.

Bypassing noise-aware classifier. To circumvent the intermediate classifier guidance, we follow the previous practice of substituting $t = 1$ into Eqs. (3) and (10), yielding a fixed-point problem w.r.t. $\hat{\boldsymbol{z}}_1$:

$$\hat{\boldsymbol{z}}_1 = \boldsymbol{z}_1 + s \cdot [\nabla_{\boldsymbol{z}_0} \boldsymbol{z}_1] \, \nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1). \tag{11}$$

As can be seen, the target endpoint $\hat{\boldsymbol{z}}_1$ is also anchored to a known reference point $\boldsymbol{z}_1$, which should enhance its stability in the solving process via fixed-point iteration or alternative numerical methods. Below, we exemplify its favorable theoretical property using the fixed-point iteration:

Proposition 2. *Suppose $\nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1)$ is Lipschitz continuous w.r.t. $\hat{\boldsymbol{z}}_1$, the fixed-point iteration to solve the target trajectory by Eq.* (11) *exhibits at least linear convergence with a properly chosen s.*

Proof. Denote the Frobenius norm of $\nabla_{\boldsymbol{z}_0} \boldsymbol{z}_1$ as L_1, and the Lipschitz constant of $\nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1)$ as L_2. The Lipschitz constant of the right side of the equation w.r.t. $\hat{\boldsymbol{z}}_1$ is upper bounded by $s \cdot L_1 \cdot L_2$. By choosing a sufficiently small $s < 1/(L_1 \cdot L_2)$, the Lipschitz constant of the right side is reduced to less than 1, thus ensuring linear convergence by the Banach fixed-point theorem. □

Interpretation of new classifier guidance. In addition to the above convergence guarantee, our new classifier guidance can be interpreted by connecting with gradient backpropagation. From Eq. (10) one could obtain an estimate of the intermediate classifier guidance (see Appendix A for derivation):

$$\nabla_{\hat{\boldsymbol{z}}_t} \log p(c|\hat{\boldsymbol{z}}_t) = [\nabla_{\boldsymbol{z}_t} \boldsymbol{z}_1] \, \nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1). \tag{12}$$

This suggests that our method is secretly estimating the intermediate classifier guidance with gradient backpropagation. While this is implicitly assumed or directly used in recent works that adapt classifier guidance to flow-based models (Wallace et al., 2023; Liu et al., 2023b; Ben-Hamu et al., 2024), it is explicitly derived here based on a very different assumption (the straightness of the flow trajectory). Such a connection helps to rationalize both our adopted assumption and the existing practice.

3.3 Practical Algorithm

Extension to piecewise rectified flow. The above analyses are performed based on the assumption that the rectified flow is well-trained and straight, which is often not the case in reality. In fact, existing rectified flow usually require multiple sampling steps due to the inherent curvature in the flow trajectory. Inspired by Yan et al. (2024), we adopt a relaxed assumption during implementation that the rectified flow is piecewise linear. Let there be K time windows $\{[t_k, t_{k-1})\}_{k=K}^{1}$ where $1 = t_K > \cdots > t_k > t_{k-1} > \cdots > t_0 = 0$, and the flow trajectory is assumed straight within each time window, then the inference procedure can be expressed as:

$$\boldsymbol{z}_t = \boldsymbol{z}_{t_{k-1}} + \boldsymbol{v}(\boldsymbol{z}_t, t)(t - t_{k-1}), \tag{13}$$

where k is the index of the time window $[t_k, t_{k-1})$ that t belongs to. Note that this framework is also compatible with the vanilla rectified flow by setting K to the number of sampling steps.

The previously derived fixed-point iteration in Eq. (11) cannot be applied directly, since its assumption that the target and reference trajectory segments share the same starting point (*e.g.* $\hat{\boldsymbol{z}}_{t_{k-1}} = \boldsymbol{z}_{t_{k-1}}$) may be violated after updates. A quick fix is to reinitialize the reference trajectory every round with predictions for updated target starting points. This allows to formulate the following problem:

$$\begin{aligned} \hat{\boldsymbol{z}}_{t_k} &= \boldsymbol{z}^e_{t_k} + s \cdot \left[\nabla_{\boldsymbol{z}_{t_{k-1}}} \boldsymbol{z}^e_{t_k}\right] \nabla_{\hat{\boldsymbol{z}}_{t_k}} \log p(c|\hat{\boldsymbol{z}}_{t_k}) \\ &= \boldsymbol{z}^e_{t_k} + s \cdot \left[\nabla_{\boldsymbol{z}_{t_{k-1}}} \boldsymbol{z}^e_1\right] \nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1), \end{aligned} \tag{14}$$

where the last step is obtained by recursively applying Eq. (12) to backpropagate the guidance signal, and a superscript e is introduced to denote the endpoint of the previous trajectory segment, as the above fix may disconnect different segments of the reference trajectory. Meanwhile, a straight-through estimator (Bengio et al., 2013) is applied to allow computing the Jacobian across different trajectory segments by estimating the Jacobian between the adjacent points $\boldsymbol{z}_{t_k}$ and $\boldsymbol{z}^e_{t_k}$ with $\boldsymbol{I}$.

Algorithm 1 Anchored Classifier Guidance

Input: rectified flow v, classifier $p(c|\cdot)$, sampling steps K, iterations N.
Initialize reference trajectory $\boldsymbol{z}_{t_k}^{[0]}$ from $\boldsymbol{v}$. ▷ Eq. (13)
Initialize target trajectory $\hat{\boldsymbol{z}}_{t_k}^{[0]} \leftarrow \boldsymbol{z}_{t_k}^{[0]}$.
for $i \leftarrow 0$ to $N-1$ **do**
  Update reference trajectory with predicted starting points $\boldsymbol{z}_{t_k}^{[i+1]}$. ▷ Eq. (15)
  Update target trajectory $\hat{\boldsymbol{z}}_{t_k}^{[i+1]}$ with classifier output $p(c|\hat{\boldsymbol{z}}_1^{[i]})$. ▷ Eq. (16)
Output: target trajectory $\hat{\boldsymbol{z}}_{t_k}^{[N]}$ subject to condition c.

Solving target flow trajectory. The target trajectory under classifier guidance, subject to Eq. (14), can be estimated iteratively by starting with $\hat{\boldsymbol{z}}_{t_k}^{[0]} = \boldsymbol{z}_{t_k}^{[0]}$ and performing the following iterations:

$$\boldsymbol{z}_{t_k}^{[i+1]} = \boldsymbol{z}_{t_k}^{[i]} + \underbrace{\boldsymbol{z}_{t_k}^{e[i+1]} - \boldsymbol{z}_{t_k}^{e[i]}}_{\text{current offset}} + \underbrace{\hat{\boldsymbol{z}}_{t_k}^{[i]} - \boldsymbol{z}_{t_k}^{e[i]}}_{\text{predicted update}}, \tag{15}$$

$$\hat{\boldsymbol{z}}_{t_k}^{[i+1]} = \boldsymbol{z}_{t_k}^{e[i+1]} + s \cdot \left[\nabla_{\boldsymbol{z}_{t_{k-1}}^{[i+1]}} \boldsymbol{z}_1^{e[i+1]}\right] \nabla_{\hat{\boldsymbol{z}}_1^{[i]}} \log p(c|\hat{\boldsymbol{z}}_1^{[i]}), \tag{16}$$

where the superscript $[i]$ is used to indicate the target and reference trajectories at the i-th iteration. Specifically, Eq. (15) implements the prediction of updated target starting points by extrapolating from history updates, and Eq. (16) tackles the derived problem. Note that there are more sophisticated methods for predicting target starting points and solving this problem, *e.g.* quasi-Newton methods, but we opt for simplicity here and leave their exploration to future work. The complete procedure for implementing the proposed classifier guidance is summarized by Algorithm 1.

3.4 Applications

The proposed algorithm is flexible for various personalized image generation tasks on human faces and common subjects. Given a reference image $\boldsymbol{z}_{\text{ref}}$ and our generated image $\hat{\boldsymbol{z}}_1$, we use their feature similarity on an off-the-shelf discriminator f, *e.g.* the face specialist ArcFace (Deng et al., 2019) or a self-supervised backbone DINOv2 (Oquab et al., 2023), as classifier guidance. In addition, to improve the guidance signal, a face detector or an open-vocabulary object detector g is employed to locate the identity-relevant region for feature extraction. Formally, the classifier output is as follows:

$$p(c|\hat{\boldsymbol{z}}_1^{[i]}) = \text{sim}\left(f \circ g(\hat{\boldsymbol{z}}_1^{[i]}), f \circ g(\boldsymbol{z}_{\text{ref}})\right). \tag{17}$$

More details are described in Appendix C. Notably, both configurations can be flexibly extended to a multi-subject scenario by incorporating a bipartite matching step between multiple detected subjects.

4 Experiments

4.1 Experimental Settings

Datasets. Our method does not involve training data, as it operates only at test time. For face-centric evaluation, we follow Pang et al. (2024) to evaluate on 20 prompts with the first 200 images from CelebA-HQ (Liu et al., 2015; Karras et al., 2018) as reference images. For subject-driven generation, we conduct qualitative studies on a subset of examples from the DreamBooth dataset (Ruiz et al., 2023b), spanning 10 subjects across two live subject categories and three object categories.

Metrics. Three metrics are considered: identity similarity, prompt consistency, and computation time. The first two are measured using an ArcFace model (Deng et al., 2019) and CLIP encoders (Radford et al., 2021), while the latter is tested on an NVIDIA A800 GPU. We reproduce the latest methods IP-Adapter (Ye et al., 2023), PhotoMaker (Li et al., 2024), and InstantID (Wang et al., 2024b) for a comprehensive comparison, and also include the existing baselines in Pang et al. (2024).

Implementation details. We experiment with a frozen piecewise rectified flow (Yan et al., 2024) finetuned from Stable Diffusion 1.5 (Rombach et al., 2022) with 4 equally divided time windows. The number of sampling steps is set to a minimum $K = 4$ given the memory overhead of backpropagation.

Table 1: Quantitative comparison for face-centric personalization. The inference time is measured in seconds on an NVIDIA A800. Unlike the previous state-of-the-art methods that require training on large face datasets (the number of images is listed for reference), our method achieves superior performance in a training-free manner, by exploiting the guidance from an off-the-shelf discriminator.

Method	Base model	Training	Identity ↑	Prompt ↑	Time ↓
Textual Inversion (Gal et al., 2023)	SD 2.1	-	0.2115	0.2498	6331
DreamBooth (Ruiz et al., 2023a)	SD 2.1	-	0.2053	0.3015	623
NeTI (Alaluf et al., 2023)	SD 1.4	-	0.3789	0.2325	1527
Celeb Basis (Yuan et al., 2023)	SD 1.4	-	0.2070	0.2683	140
Cross Initialtion (Pang et al., 2024)	SD 2.1	-	0.2517	0.2859	346
IP-Adapter (Ye et al., 2023)	SD 1.5	10M	0.4778	0.2627	**2**
PhotoMaker (Li et al., 2024)	SDXL	112K	0.2271	<u>0.3079</u>	<u>4</u>
InstantID (Wang et al., 2024b)	SDXL	60M	<u>0.5806</u>	0.3071	6
RectifID (20 iterations)	SD 1.5	-	0.4860	0.2995	9
RectifID (100 iterations)	SD 1.5	-	**0.5930**	0.2933	46
RectifID (20 iterations)	SD 2.1	-	0.5034	**0.3151**	20

Figure 3: Qualitative comparison for face-centric personalization. See Figs. 9 to 12 for more samples.

A naive implementation takes 14GB of GPU memory, which fits on a range of consumer-grade GPUs. More results on alternative rectified flows can be found in Appendix D.4. For hyperparameters, the guidance scale is fixed to $s = 1$ in quantitative evaluation. Meanwhile, for stability, the gradient is normalized following Karunratanakul et al. (2024). The number of iterations is set to $N = 100$.

4.2 Main Results

Face-centric personalization. Table 1 and Fig. 3 compare our method (denoted RectifID) with extensive baselines. Overall, our training-free approach achieves state-of-the-art performance in quantitative evaluations. Specifically, we observe that: (1) our SD 1.5-based implementation yields the highest identity similarity of all, and leads in prompt consistency among SD 1.x-based methods. It is also computationally efficient, *e.g.* taking less time than existing tuning-based methods, and outperforming the training-based IP-Adapter (Ye et al., 2023) in a near inference time of 9 seconds vs. 2 seconds. (2) By simply replacing the base diffusion model with SD 2.1 at its default image size, our prompt consistency further surpasses SDXL (Podell et al., 2024)-based models. Note, however,

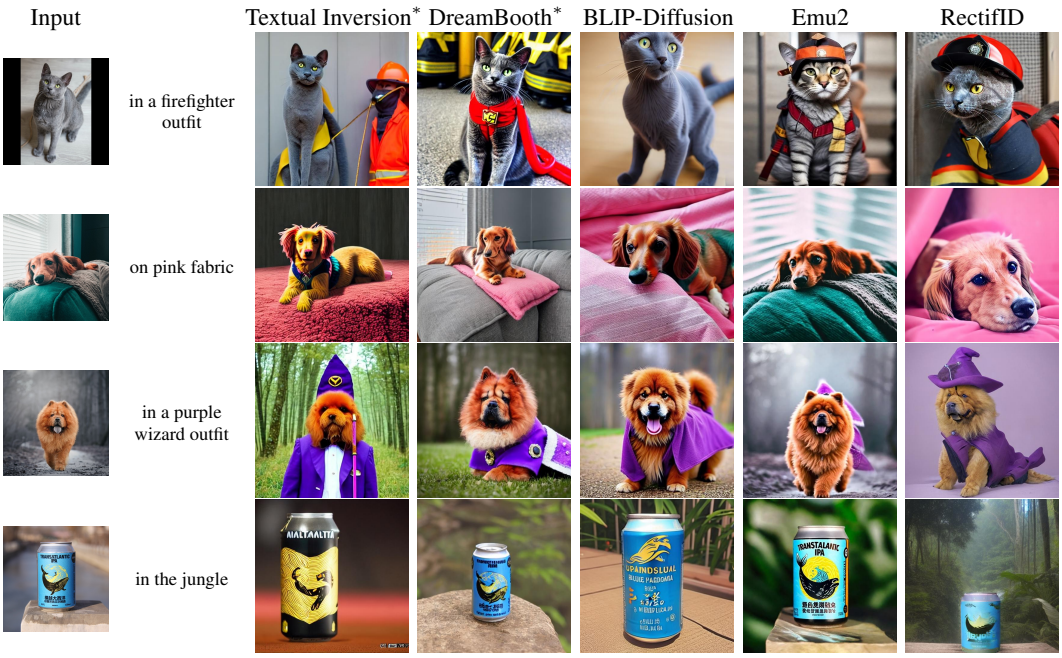


Figure 4: Qualitative comparison for subject-driven generation. * denotes finetuned with multiple images of the target subject to achieve sufficient identity consistency. See Fig. 13 for more samples.

Figure 5: Qualitative comparison for multi-subject personalization. See Fig. 14 for more samples.

that the rest of this paper still uses SD 1.5 for a fair comparison to SD 1.x-based baselines in various personalization tasks, excluding potential improvements from using better base models. (3) In general, our method takes a big step towards bridging the substantial performance gap with training-based personalization methods by exploring the effectiveness of training-free classifier guidance.

For face-centric qualitative comparison in Fig. 3, our method remains advantageous as its generated images by the guidance of the face discriminator exhibit high identity consistency. In comparison, InstantID (Wang et al., 2024b) delivers a near level of consistency by controlling face landmarks, but sometimes distorts the face shape (the first and third images) and contains much less natural variation. More generated samples are provided in Figs. 11 and 12 in the appendix.

Subject-driven generation. Our approach is flexibly extended beyond human faces towards more subjects, including certain common animals and regularly shaped objects. To validate our flexibility, Fig. 4 qualitatively compares it on three cats or dogs and a regularly shaped can, where the images generated by our method achieve highly competitive identity and prompt consistency. In comparison, the state-of-the-art method Emu2 (Sun et al., 2024), as a generalist multimodal large language model, yields high identity similarity largely by reconstructing the input image, which limits its usefulness. The tuning-based Textual Inversion (Gal et al., 2023) and DreamBooth (Ruiz et al., 2023a) only work well with multiple images and exhibit inferior prompt consistency due to finetuned model parameters or prompt embeddings. See Fig. 13 in the appendix for additional results from more subjects.

Figure 6: Comparison with alternative designs at varying guidance scale (or learning rate) and iterations. The prompts are "cave mural depicting a person" and "a person as a priest in blue robes". The base learning rate for gradient descent is 0.4, with momentum of 0.9 and an ℓ_2 regularizer of 1.0.

Multi-subject personalization. Our method can be further extended to multi-subject scenarios via a bipartite matching step. Figure 5 compares it to the domain experts FastComposer (Xiao et al., 2023) and Cones 2 (Liu et al., 2023c) on composing multiple faces, live subjects and objects. As can be seen, our method achieves overall advantageous identity consistency, in spite of differences in non-persistent attributes such as hairstyle. Image semantics and quality are also well preserved, as exemplified by the amusement park details in the first image, with some others even surpassing the SD 2.1-based specialized model Cones 2. More generated samples can be found in Fig. 14 in the appendix.

4.3 Ablation Study

To justify the effectiveness of our proposed classifier guidance, Fig. 6 and Table 2 compare it with two variants: the previously derived guidance without anchor, namely using Eq. (8), and a gradient descent method on the initial noise similar to DOODL (Wallace et al., 2023) and D-Flow (Ben-Hamu et al., 2024). The figure depicts that the gradient descent is unstable (left) and converges relatively slowly (right) despite using momentum and ℓ_2 regularization. And its identity preservation is sensitive to the learning rate. Though our new fixed-point formulation allows for a more stable layout, the initial version fails to converge as the face feature keeps drifting. In contrast, our full method exhibits better stability (left) and faster convergence (right) by implicitly regularizing the flow trajectory to be close and straight. This is further supported by the quantitative comparison, where our method delivers better identity and prompt consistency than the alternatives. Further analysis for hyperparameter sensitivity is provided in Appendix D.3.

Table 2: Quantitative comparison with alternative designs. The number of iterations is 100, and the remaining settings for gradient descent follow Fig. 6.

Method	Identity ↑	Prompt ↑
Gradient descent of noise	0.5249	0.2842
RectifID w/o anchor	0.1158	0.2916
RectifID	**0.5930**	**0.2933**

4.4 Generalization

To validate the generalizability of our approach to broader application scenarios, we have extended it to more controllable generation tasks by directly using the guidance functions from Universal Guidance (Bansal et al., 2024). The experimental results under the guidance of segmentation map or style image are illustrated in Fig. 7. As shown, our classifier guidance can perform both tasks without any additional tuning, faithfully following the various forms of control signals provided by the user. This confirms the adaptability of our approach for various controllable generation tasks. Additional generalization analysis of our method for broader diffusion models is presented in Appendix D.1.

(a) Segmentation map

(b) Style transfer

Figure 7: Experimental results for more controllable generation tasks. The first column shows the guidance, and the rest are the generated results. Our method is extended to various controllable generation tasks by incorporating the guidance functions from Universal Guidance (Bansal et al., 2024).

5 Conclusion

This work presents a training-free personalized image generation method using anchored classifier guidance. It extends the applicability of the original classifier guidance based on two key findings: first, by developing on a rectified flow framework assuming ideal straightness, the classifier guidance can be transformed into a new fixed-point formulation involving only clean image-based discriminators; secondly, anchoring the flow trajectory to a reference trajectory greatly improves its solving stability. The derived anchored classifier guidance allows flexible reuse of existing image discriminators to improve identity consistency, as validated by extensive experiments on various personalized image generation tasks for human faces, live subjects, certain objects, and multiple subjects.

Acknowledgement: This research work is supported by National Key R&D Program of China (No. 2022ZD0160305), a research grant from China Tower Corporation Limited, an internal grant (No. 2024JK28) and a grant from Beijing Aerospace Automatic Control Institute.

References

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics*, 42(6):1–10, 2023.

Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial FC: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1445–1449, 2021.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, 2024.

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-Flow: Differentiating through flows for controlled generation. In *International Conference on Machine Learning*, 2024.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. PhotoVerse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6572–6583, 2018.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, pages 25683–25696, 2022.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. LCM-lookahead for encoder-based text-to-image personalization. In *Proceedings of the European Conference on Computer Vision*, 2024.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations*, 2022.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. PuLID: Pure and lightning ID customization via contrastive alignment. In *Advances in Neural Information Processing Systems*, 2024.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, et al. Manifold preserving guided diffusion. In *International Conference on Learning Representations*, 2024.

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

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

Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022a.

Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. DiffFace: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*, 2022b.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, pages 30146–30166, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Hanwen Liu, Zhicheng Sun, and Yadong Mu. Countering personalized text-to-image generation with influence watermarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12257–12267, 2024a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023a.

Xingchao Liu, Lemeng Wu, Shujian Zhang, Chengyue Gong, Wei Ping, and Qiang Liu. FlowGrad: Controlling the output of generative ODEs with gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24335–24344, 2023b.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024b.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *Advances in Neural Information Processing Systems*, pages 57500–57519, 2023c.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3730–3738, 2015.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *Proceedings of the European Conference on Computer Vision*, pages 728–755, 2022.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DiffEditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8488–8497, 2024.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804, 2022.

OpenAI. GPT-4V(ision) system card. `https://openai.com/research/gpt-4v-system-card`, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

Lianyu Pang, Jian Yin, Haoran Xie, Qiping Wang, Qing Li, and Xudong Mao. Cross initialization for personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. PortraitBooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *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, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

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*, 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, 2023a.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. HyperDreamBooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Proceedings of the European Conference on Computer Vision*, 2024.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pages 25278–25294, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pages 32483–32498, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11918–11930, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhentao Tan and Yadong Mu. Learning solution-aware transformers for efficiently solving quadratic assignment problem. In *International Conference on Machine Learning*, 2024.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7280–7290, 2023.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency model. In *Advances in Neural Information Processing Systems*, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024b.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. In *Advances in Neural Information Processing Systems*, pages 16782–16795, 2022.

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. PeRFlow: Piecewise rectified flow as universal plug-and-play accelerator. In *Advances in Neural Information Processing Systems*, 2024.

Lingxiao Yang, Shutong Ding, Yifan Cai, Jingyi Yu, Jingya Wang, and Ye Shi. Guidance with spherical gaussian constraint for conditional diffusion. In *International Conference on Machine Learning*, 2024.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023.

Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. In *Advances in Neural Information Processing Systems*, pages 72958–72982, 2023.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023.

A Detailed Derivations

Most of the equations in our paper are accompanied by their derivations, except for Eqs. (11) and (12). For the sake of completeness, their derivations are supplemented here.

Equation (11): We substitute $t = 1$ into Eqs. (3) and (10) to obtain:

$$\begin{aligned}\hat{\boldsymbol{z}}_1 &= \boldsymbol{z}_0 + \hat{\boldsymbol{v}}(\hat{\boldsymbol{z}}_1, 1) \\ &= \boldsymbol{z}_0 + \boldsymbol{v}(\boldsymbol{z}_1, 1) + s \cdot [\nabla_{\boldsymbol{z}_0} \boldsymbol{z}_1] \nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1) \\ &= \boldsymbol{z}_1 + s \cdot [\nabla_{\boldsymbol{z}_0} \boldsymbol{z}_1] \nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1).\end{aligned} \tag{18}$$

Equation (12): By applying Eq. (10) twice and utilizing the straightness property of rectified flow, we have:

$$\begin{aligned}\nabla_{\hat{\boldsymbol{z}}_t} \log p(c|\hat{\boldsymbol{z}}_t) &= 1/s \cdot [\nabla_{\boldsymbol{z}_t} \boldsymbol{z}_0] \left(\hat{\boldsymbol{v}}(\hat{\boldsymbol{z}}_t, t) - \boldsymbol{v}(\boldsymbol{z}_t, t)\right) \\ &= 1/s \cdot [\nabla_{\boldsymbol{z}_t} \boldsymbol{z}_0] \left(\hat{\boldsymbol{v}}(\hat{\boldsymbol{z}}_1, 1) - \boldsymbol{v}(\boldsymbol{z}_1, 1)\right) \\ &= [\nabla_{\boldsymbol{z}_t} \boldsymbol{z}_0] [\nabla_{\boldsymbol{z}_0} \boldsymbol{z}_1] \nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1) \\ &= [\nabla_{\boldsymbol{z}_t} \boldsymbol{z}_1] \nabla_{\hat{\boldsymbol{z}}_1} \log p(c|\hat{\boldsymbol{z}}_1).\end{aligned} \tag{19}$$

B Related Work on Classifier Guidance

Since the proposal of classifier guidance (Dhariwal and Nichol, 2021) which uses a special noise-aware classifier as training-free guidance for diffusion models, many efforts have been made in order to extend its applicability to off-the-shelf loss guidance. They can be grouped into three categories: (1) Early literature focuses on simpler objectives for linear inverse problems such as image super-resolution, deblurring, and inpainting (Chung et al., 2022, 2023; Wang et al., 2023; Zhu et al., 2023). (2) These methods can be extended to more complex discriminators through various approximations. Yu et al. (2023); Song et al. (2023) use Tweedie's formula and Monte Carlo method, respectively, to estimate the integrated classifier guidance. Bansal et al. (2024); He et al. (2024) perform updates directly in the clean data space, with the latter imposing an additional manifold constraint. Similarly, Gaussian spherical constraint is explored in Yang et al. (2024). Mou et al. (2024) advance these techniques to more versatile editing tasks. The above methods are further unified in Ye et al. (2024). (3) A recent line of work directly uses gradient descent with specific diffusion models. To enable their gradient computation, DiffusionCLIP (Kim et al., 2022a) relies on shortened ODE trajectories, while FlowGrad (Liu et al., 2023b) adopts a non-uniform ODE discretization and decomposed gradient computation. DOODL (Wallace et al., 2023), DNO (Karunratanakul et al., 2024) and D-Flow (Ben-Hamu et al., 2024) use invertible models or flow models to backpropagate gradient to the initial noise. Our proposed method is related to the third category, focusing on rectified flow whose approximation remains understudied. Moreover, it features a fixed-point formulation with a convergence guarantee for ideal rectified flow, which allows potentially better stability over the existing approaches, *e.g.* gradient descent on initial noise, as empirically validated in Section 4.3.

Recent studies also explore the use of pre-trained classifiers to guide personalized image generation, but mostly during model training. DiffFace (Kim et al., 2022b) and PhotoVerse (Chen et al., 2023) directly apply a face discriminator to noised images to compute an identity loss. PortraitBooth (Peng et al., 2024) improves loss quality by computing it at less noisy stages. More relevant to our work, LCM-Lookahead (Gal et al., 2024) and PuLID (Guo et al., 2024) utilize distilled diffusion models to generate images in few steps, allowing for direct gradient backpropagation of image-space losses. However, these personalization methods must first be trained on extensive face recognition data, *e.g.* LAION-Face 50M (Zheng et al., 2022). In comparison, we harness off-the-shelf face discriminators at test time based on the methodology of classifier guidance, enabling more flexible customization without training. And it can be generalized to other subjects by simply replacing the classifier.

C Experimental Settings

Method. We mainly experiment on the recently proposed piecewise rectified flow (Yan et al., 2024). The model is finetuned from Stable Diffusion 1.5 (Rombach et al., 2022) on the LAION-Aesthetic-5+ dataset (Schuhmann et al., 2022) without special pre-training on human faces or subjects. We adopt a fixed image size of 512×512, number of sampling step $K = 4$, and a classifier-free guidance scale

of 3.0 during quantitative evaluation. The newly incorporated anchored classifier guidance uses a default guidance scale $s = 1.0$ and number of iterations $N = 100$. For qualitative studies, a few results are generated with a slightly different guidance scale $s = 0.5$ or 2.0 for better visual quality or identity consistency. But in general, our method is not very sensitive to these hyperparameters. In terms of computational and memory overhead, our method takes less than 0.5s per iteration on an NVIDIA A800 GPU and fits on consumer-grade GPUs such as NVIDIA RTX 4080, the latter of which may be further improved with gradient checkpointing or the techniques in Liu et al. (2023b).

For face-centric personalization, our method is implemented with the antelopev2 model pack from the InsightFace library.[2] Specifically, it detects and crops the face regions with an SCRFD model (Guo et al., 2022), and then extracts face features using an ArcFace model (Deng et al., 2019) trained on Glint360K (An et al., 2021), consistent with most personalization methods that use a face model (Ye et al., 2023; Wang et al., 2024b; Gal et al., 2024). The resulting face feature is compared with the reference image to compute a cosine loss, which serves as the classifier guidance signal.

For subject-driven generation, we use an open-vocabulary object detector OWL-ViT (Minderer et al., 2022) to detect the region of interest, and extract visual features with DINOv2 (Oquab et al., 2023). The extracted feature is compared with the reference to calculate a cosine loss as the guidance signal. While this guidance already works well for various live subjects including multiple dogs and cats, we add an optional ℓ_1 loss with a coefficient of 10.0 to help preserve the identity of certain objects, such as cans, vases and duck toys. The current implementation is still limited in not capturing the details of some irregularly shaped objects, *e.g.* plush toys, and we expect to resolve this issue with an improved discriminator, or by combining with existing image prompt techniques (Ye et al., 2023).

For multi-subject personalization, we consider a simplified case of exactly two subjects and perform the following: first detect the two subjects, then enumerate all possible bipartite matches with the reference subjects to minimize the matching cost. For more complex scenarios, a possible workaround is to formulate it as an quadratic assignment problem and apply efficient solvers (Tan and Mu, 2024).

Evaluation. For face-centric personalization, we follow Pang et al. (2024) in evaluating on the first 200 images in the CelebA-HQ dataset (Liu et al., 2015; Karras et al., 2018) with 20 text prompts including 15 realistic prompts and 5 stylistic prompts. The evaluation process reuses the code from Celeb Basis (Yuan et al., 2023),[3] which first detects the face region using a PIPNet (Jin et al., 2021) with a threshold of 0.5 and then computes the cosine similarity on face features extracted by an ArcFace model (Deng et al., 2019). It should be noted that our method adopts a different face detector and different alignment and cropping methods, so it does not overfit the evaluation protocol. For the baselines, in addition to those compared in Pang et al. (2024), we also evaluate the recently proposed IP-Adapter (Ye et al., 2023), PhotoMaker (Li et al., 2024), and InstantID (Wang et al., 2024b) on their recommended settings. The number of their sampling steps is set to 30 for a fair comparison. The checkpoint version of IP-Adapter is ip-adapter-full-face_sd15. The image size is set to 512×512 for IP-Adapter and 1024×1024 for PhotoMaker and InstantID based on SDXL (Podell et al., 2024). In the qualitative analysis, we also include Celeb Basis (Yuan et al., 2023) as a baseline method.

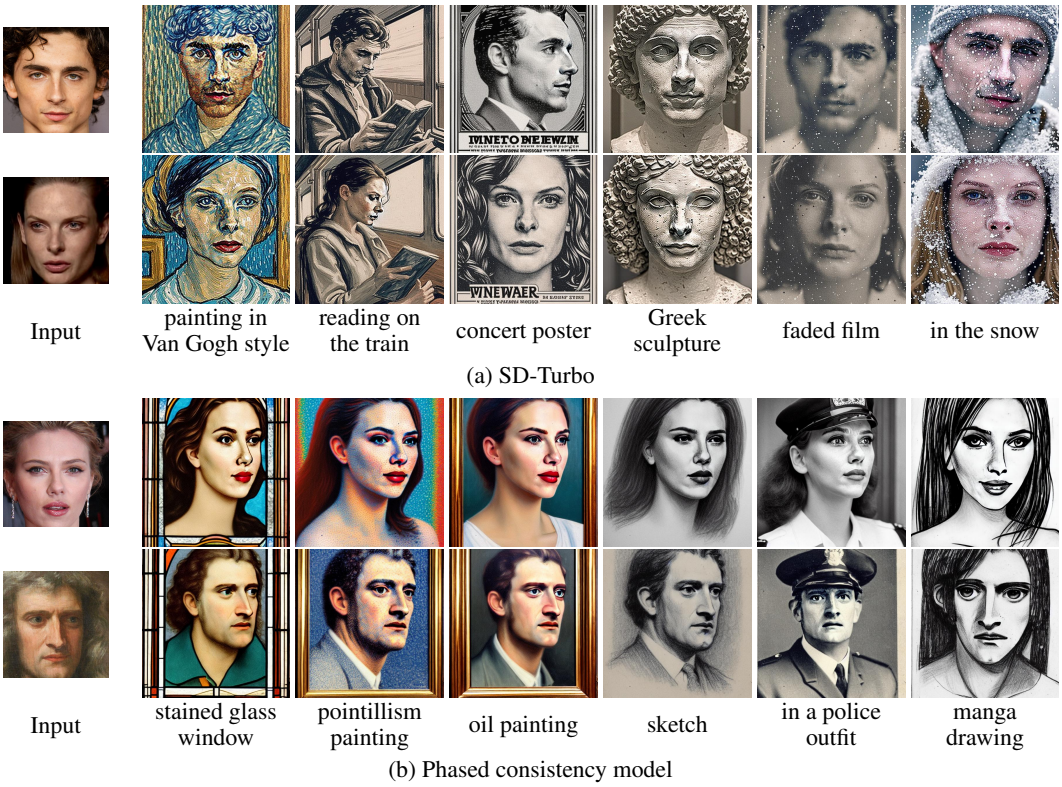
For subject-driven generation, we perform a qualitative rather than a quantitative comparison on a subset of the DreamBooth dataset (Ruiz et al., 2023a) due to the previously mentioned limitations. Nevertheless, many subjects are considered during the qualitative study, including 7 live subjects from two categories (cats and dogs) and 3 regularly shaped objects from different categories (cans, vases, and teapots). For the baselines, we incorporate Textual Inversion (Gal et al., 2023), DreamBooth (Ruiz et al., 2023a), BLIP-Diffusion (Li et al., 2023), and Emu2 (Sun et al., 2024) for extensive comparison. Their diffusion models and hyperparameter settings follow the official or Diffusers implementation.[4]

Licenses. The piecewise rectified flow (Yan et al., 2024) used in the main experiments is released under the BSD-3-Clause License and the 2-rectified flow (Liu et al., 2023a, 2024b) used in Appendix D.4 is released under the MIT License. The InsightFace library for face detection and recognition is released under the MIT License, while its pre-trained models are available for non-commercial research purposes only. The OWL-ViT (Minderer et al., 2022) and DINOv2 (Oquab et al., 2023) models for object detection and feature extraction are released under the Apache-2.0 License. For evaluation, the code of Celeb Basis (Yuan et al., 2023) is licensed under the MIT License. The

[2] `https://github.com/deepinsight/insightface`

[3] This is mentioned at `https://github.com/lyuPang/CrossInitialization#metrics`.

[4] For example, we use `https://huggingface.co/docs/diffusers/training/text_inversion`.



(a) SD-Turbo

(b) Phased consistency model

Figure 8: Generalization to few-step diffusion models, including SD-Turbo (Sauer et al., 2024) and phased consistency model (Wang et al., 2024a), both distilled from SD and using 4 sampling steps in inference. The results show that our method is effective for personalizing broader diffusion models.

CelebA-HQ (Liu et al., 2015; Karras et al., 2018) and DreamBooth (Ruiz et al., 2023a) datasets for quantitative and qualitative evaluation are released under the CC BY-NC 4.0 and CC-BY-4.0 licenses.

D Additional Results

D.1 Generalization

While our classifier guidance is derived based on rectified flow, the same idea can be generalized to some few-step diffusion models by assuming straightness of their trajectories within each time step. We empirically demonstrate this in Fig. 8 with two popular few-step diffusion models, SD-Turbo (Sauer et al., 2024) and phased consistency model (Wang et al., 2024a). As the results indicate, our method effectively personalizes these diffusion models to generate identity-preserving images. We will continue to explore this approach for other generative models in future research.

D.2 More Visualizations

More examples of our generated image are provided in Figs. 9 to 14. For face-centric personalized image generation, it is shown that our method can follow a variety of text prompts to generate both realistic or stylistic images while preserving the user-specified identity from a diverse group of people. Although there exist minor differences in the person's age and hairstyles, the face looks very similar to the reference image. In particular, the method demonstrates good generalizability among different piecewise rectified flows based on SDXL (Podell et al., 2024) and SD 1.5 (Rombach et al., 2022). For subject-driven generation, new results are presented from additional subject types, including different breeds of cats and dogs, and some regularly shaped objects such as vases, demonstrating the flexibility of our approach across different use cases. For multi-subject generation, it naturally blends multiple human faces or live subjects into a single image while maintaining the visual quality

and semantics, revealing a wider range of potentially interesting applications. Overall, our method demonstrates to be effective and flexible for various personalization tasks.

D.3 Ablation Study

In addition to the ablation experiments in the main paper, we perform a sensitivity analysis of the two hyperparameters in our method, namely the guidance scale s and the number of iterations N. Specifically, we study the effect of different s under a fixed $N = 20$ and then the effect of different N under a fixed $s = 1.0$. The results are summarized in Fig. 15. As can be observed, increasing both hyperparameters from 0 leads to a significant improvement in identity consistency, confirming the effectiveness of our proposed classifier guidance. Also, the performance is stable over a fairly wide range of hyperparameters, indicating that our approach is not very sensitive to hyperparameters. Furthermore, we find that the use of a small classifier guidance scale is actually beneficial for prompt consistency, possibly because it enhances the visual features, as demonstrated in Fig. 2.

D.4 Experiments with 2-Rectified Flow

Figures 16 and 17 present additional qualitative results on a vanilla 2-rectified flow (Liu et al., 2023a, 2024b) finetuned on Stable Diffusion 1.4 (Rombach et al., 2022). As can be seen, our method continues to deliver satisfactory identity preservation when moving to a different model, despite a noticeable drop in the generation quality. Concretely, it integrates target subjects with some quite challenging prompts, such as a person swimming or getting a haircut, while showing very little interference with the original background, *e.g.* jungles and cityscapes. These results clearly validate the effectiveness of our proposed method in alternative rectified flow models.

Note that we also experimented with $K = 1$ on a single-step InstaFlow (Liu et al., 2024b) distilled from this 2-rectified flow, but found that it tended to converge to slightly distorted images. This may be attributed to the larger modeling error inherent in InstaFlow's distillation process, which reduces the effectiveness of our approach assuming each flow trajectory segment is well-trained and straight.

E Broader Impacts and Limitations

Broader impacts. The proposed method can be integrated with the emerging rectified flow-based models to enhance identity preservation and versatile control over existing AI art production pipelines. However, as a training-free personalization technique, it may increase the risk of image faking and have negative societal effects. Some immediate remedies include text-based prompt filters and AI-generated image detection, but it remains an open problem for a more principled solution, for which we advocate further research on data watermarking and model unlearning as potential mitigations. To further clarify it, we provide a more detailed explanation of these mitigations below:

- Prompt filtering, model unlearning: Since our method keeps the diffusion model intact, existing techniques for regulating diffusion models can be applied seamlessly, including prompt filters or unlearning methods. The former can be applied explicitly like the text filters in SD models, or implicitly via CFG as in Schramowski et al. (2023). The latter approaches involve finetuning the diffusion model to remove the ability to generate harmful content (Gandikota et al., 2023; Kumari et al., 2023).
- Data watermarking: To prevent misuse of personal images, one could add a protective watermark to their images (Van Le et al., 2023; Liu et al., 2024a). With this watermark or perturbation, the image can no longer be learned by common personalization methods. However, it is unclear how robust these watermarks are to training-free methods such as ours. An alternative watermarking scheme is to embed special watermark to the images generated by our proposed model, which would be invisible to the users yet identifiable by us (i.e., the tech provider). Images with such watermarks will be marked as being artificial.
- AI-generated image detection: As a post-hoc safety measure, it helps to distinguish fake images generated by the attackers from real images. Beyond above watermark-based scheme, more sophisticated data-driven methods have attracted increasing interest from the AI community. Despite that current methods still lack accuracy, we believe that developing reliable and widely available AI-generated image detectors is an important research direction.

Limitations. Our theoretical guarantee is limited to ideal rectified flow and cannot be generalized to more complex flow-based models. Empirically, anchoring the new flow trajectory to a reference trajectory only proves effective for faces, live subjects and certain regularly shaped objects, and remains insufficient for many objects with large structural variations, *e.g*. plush toys. Furthermore, while our method is training-free, its inference time has yet to match several training-based baselines, which may be addressed by applying more advanced numerical solvers to the derived problem.

Another important issue is the lack of pre-trained discriminators. To address this in the short term, we suggest first training a specialized discriminator and then applying our classifier guidance. There are two reasons for doing this instead of finetuning the generator directly: (1) training/finetuning a discriminator is usually more efficient and stable than training/finetuning a generator; (2) it can take full advantage of domain images that have no captions or even labels by using standard contrastive learning loss. In the future, scaling up vision-language models may be a general solution for these domains. The current models such as GPT-4V (OpenAI, 2023) have demonstrated certain generalizability across visual understanding tasks. As they continue to improve in generalizability and robustness, they will become a viable source for guiding diffusion models in new domains.

Figure 9: Additional face-centric personalization results with piecewise rectified flow (Yan et al., 2024) based on SDXL (Podell et al., 2024). Our method is compatible with more advanced base models and provides sophisticated personalization results.

Figure 10: Additional face-centric personalization results with piecewise rectified flow (Yan et al., 2024) based on SDXL (Podell et al., 2024). Our method is compatible with more advanced base models and provides sophisticated personalization results.

Figure 11: Additional face-centric personalization results with piecewise rectified flow (Yan et al., 2024), which is based on Stable Diffusion 1.5 (Rombach et al., 2022). Our method achieves high identity consistency. See Fig. 16 for results on the vanilla 2-rectified flow (Liu et al., 2023a, 2024b).

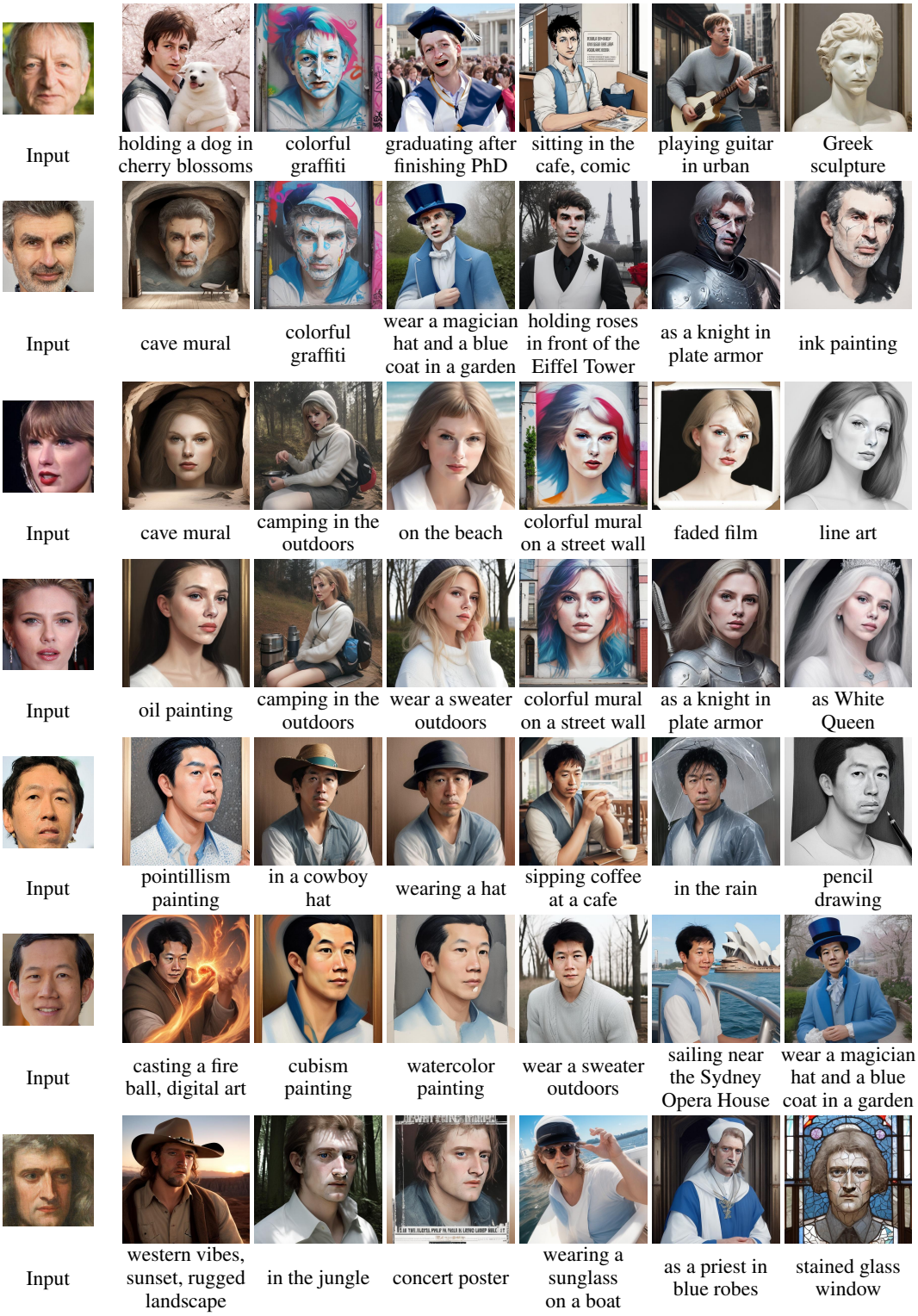


Figure 12: Additional face-centric personalization results with piecewise rectified flow (Yan et al., 2024), which is based on Stable Diffusion 1.5 (Rombach et al., 2022). Our method achieves high identity consistency. See Fig. 16 for results on the vanilla 2-rectified flow (Liu et al., 2023a, 2024b).

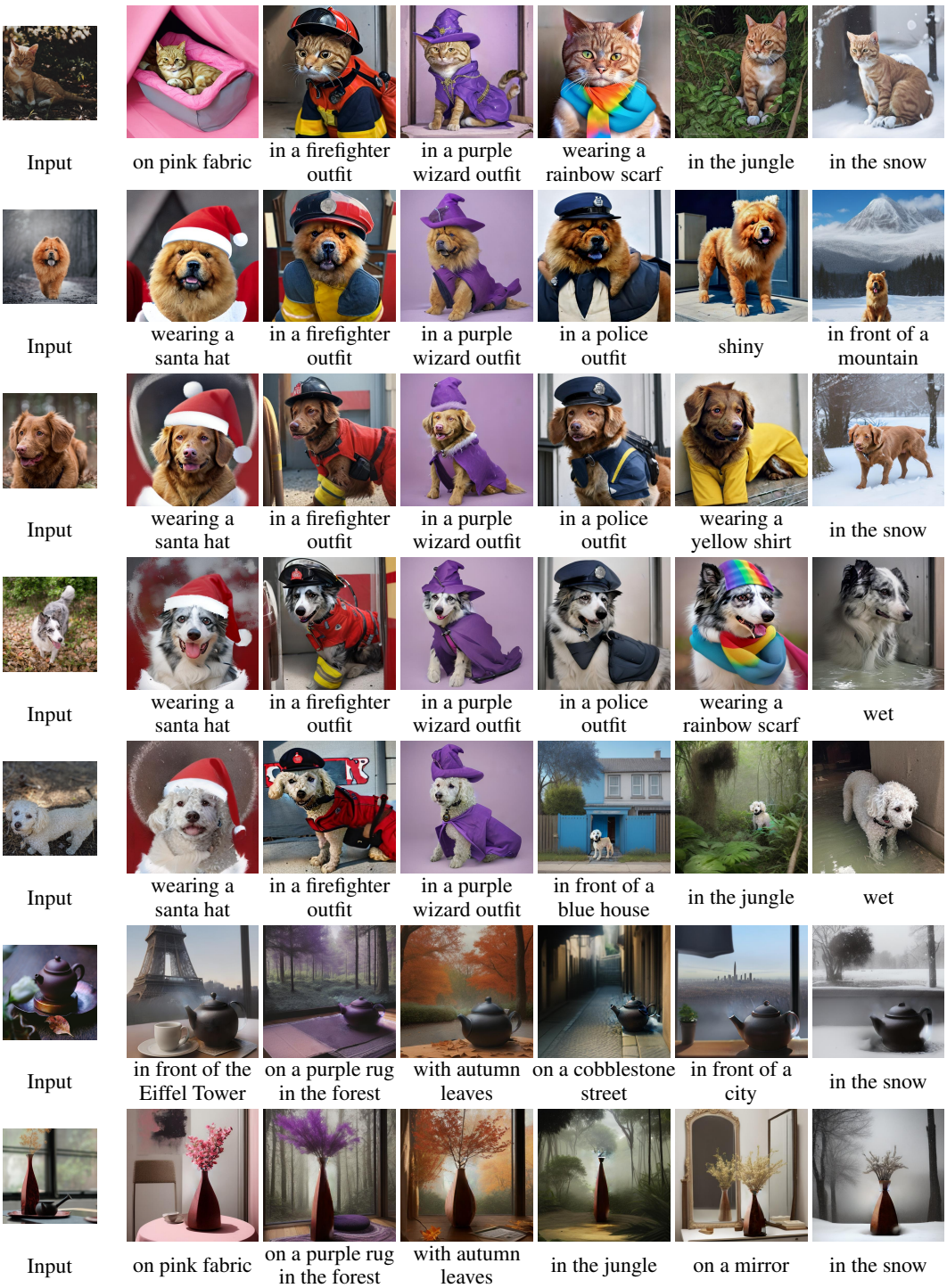


Figure 13: Additional subject-driven generation results with piecewise rectified flow (Yan et al., 2024), which is based on Stable Diffusion 1.5 (Rombach et al., 2022). Our approach preserves the identity of both live subjects and some regularly shaped objects. Please see Fig. 17 for more examples using the vanilla 2-rectified flow (Liu et al., 2023a, 2024b).

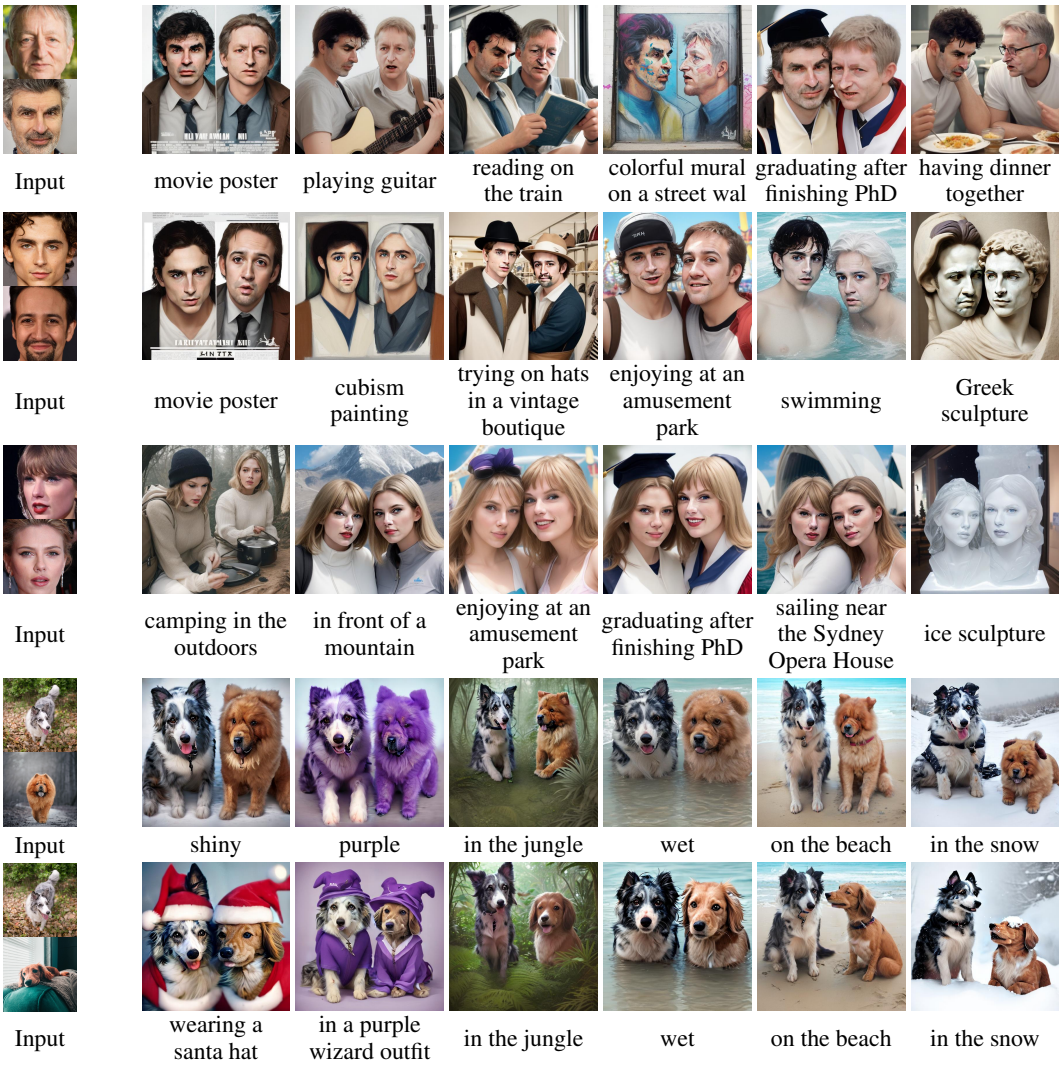


Figure 14: Additional multi-subject personalization results with piecewise rectified flow (Yan et al., 2024), which is based on Stable Diffusion 1.5 (Rombach et al., 2022). Our approach can naturally compose multiple subjects into the generated image while preserving their identities.

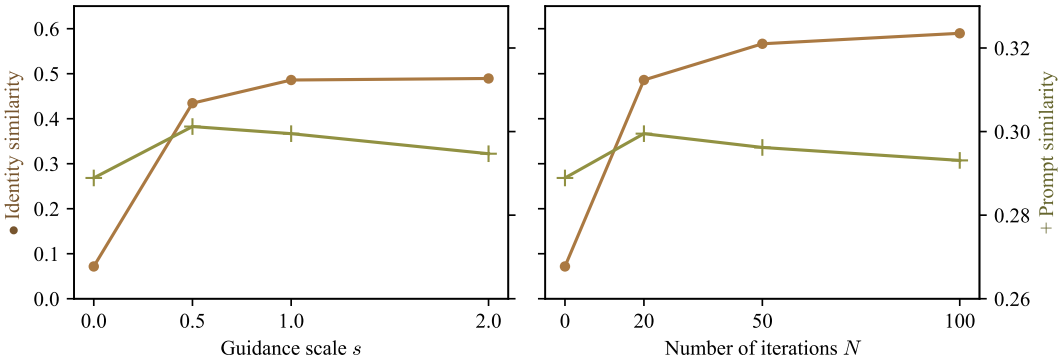


Figure 15: Ablation study of hyperparameters. Left: ablation study of guidance scale s under $N = 20$. Right: ablation study of the number of iterations N under $s = 1.0$. Our method remains effective over a reasonably wide range of hyperparameters.

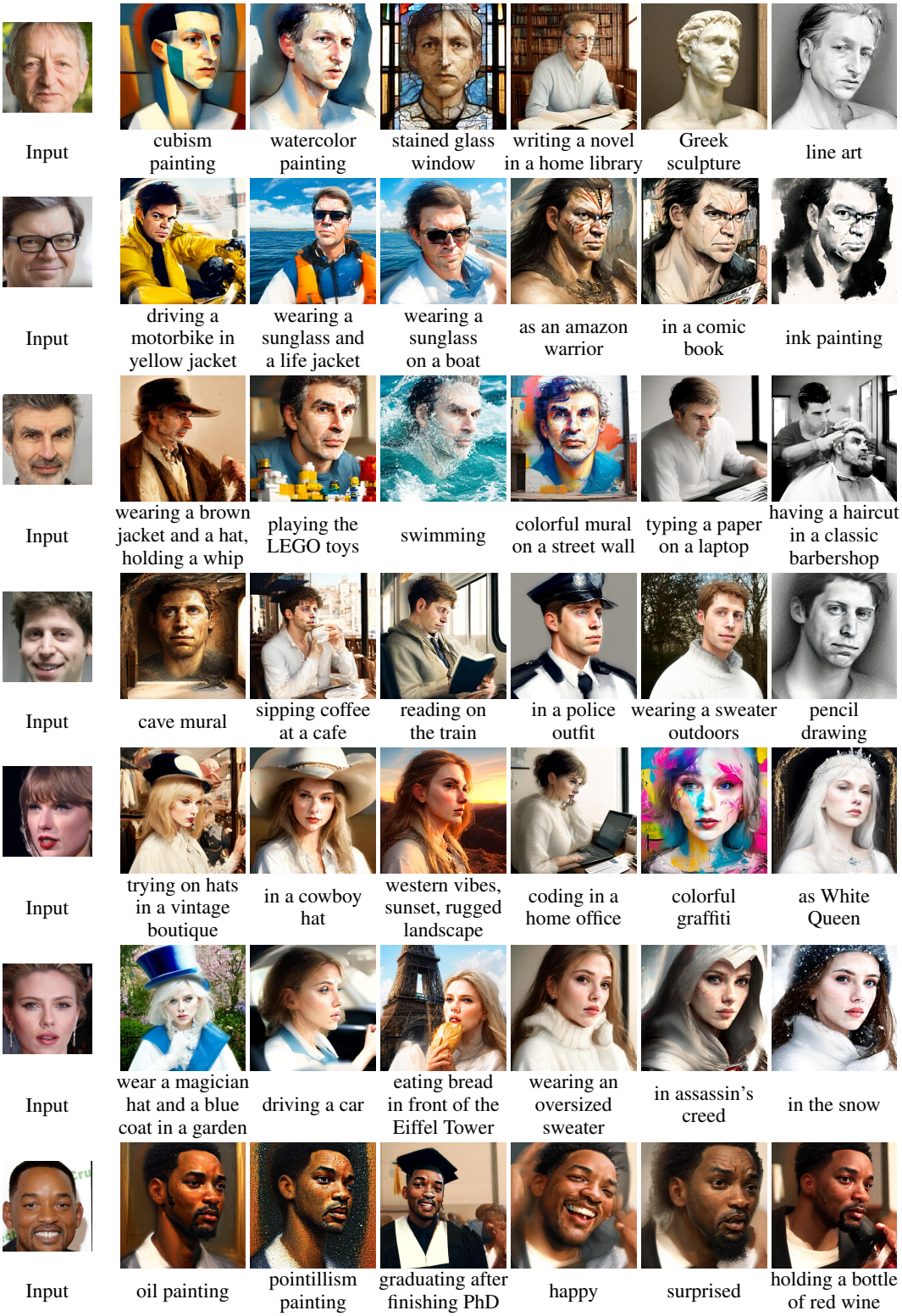


Figure 16: Face-centric personalization results with vanilla 2-rectified flow (Liu et al., 2023a, 2024b), which is based on Stable Diffusion 1.4 (Rombach et al., 2022). Our method preserves their identities well while remaining faithful to the text prompt during generation.

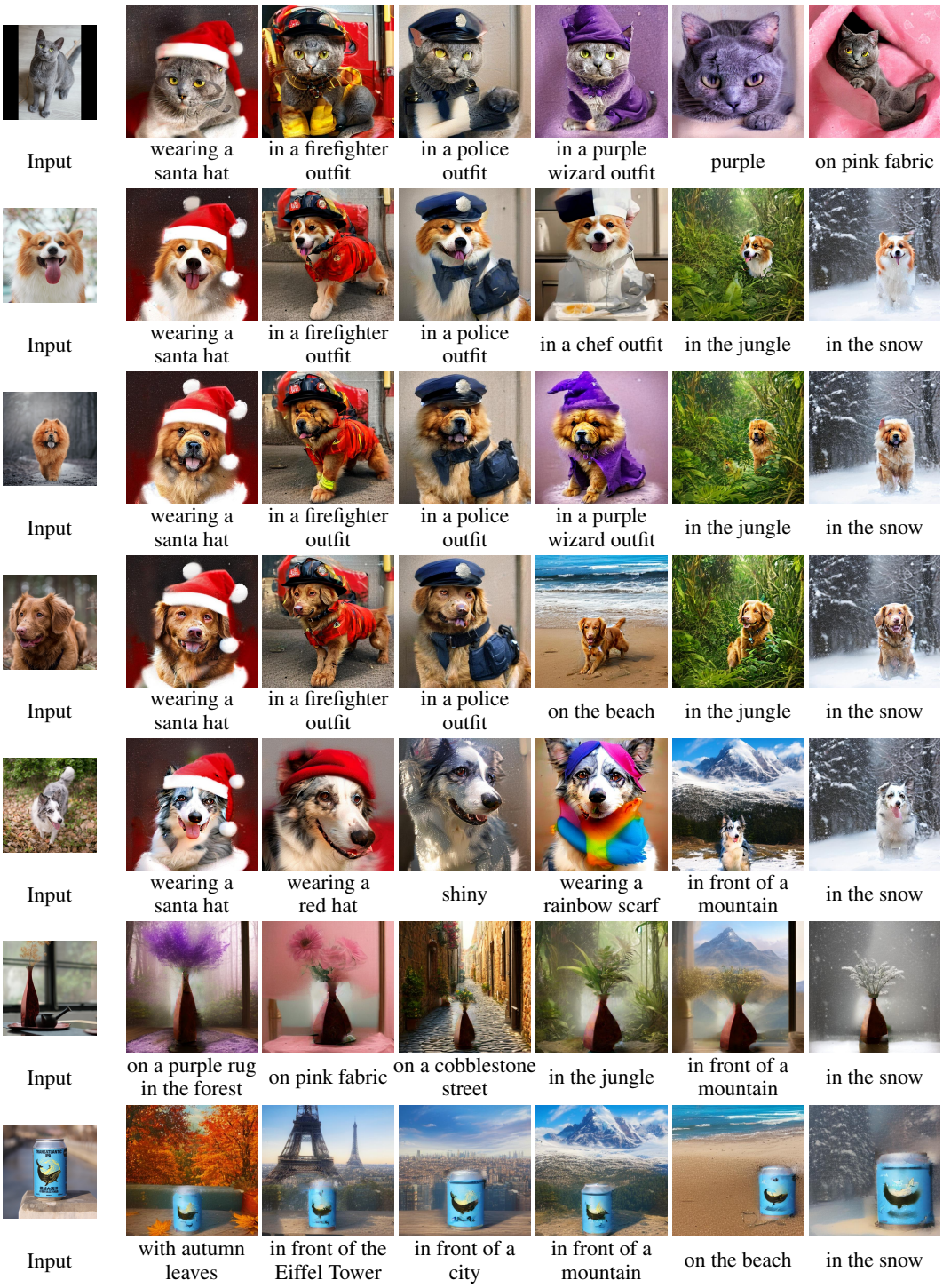


Figure 17: Subject-driven generation results with vanilla 2-rectified flow (Liu et al., 2023a, 2024b), which is based on Stable Diffusion 1.4 (Rombach et al., 2022). Examples for additional categories of cats, dogs, and objects are included to demonstrate the effectiveness of our approach.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims and contributions of this paper are summarized at the end of the introduction (Section 1).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in **??**.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions and proofs for Propositions 1 and 2 are provided in the main paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have detailed these information in Section 4.1 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code to reproduce all results is available at this anonymized link `https://github.com/feifeiobama/RectifID`, with sufficient instructions in the README file.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings are specified in Section 4.1 and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report the results from a single experimental run, following many baselines in the field. The experiments use a fixed seed of 42.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are detailed in Section 4.1 and Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research complies with the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential societal impacts are discussed in **??**.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not release any datasets or pre-trained models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original paper that produced the assets used in this paper and summarized their licenses in Appendix C.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is available at this anonymized repository `https://github.com/feifeiobama/RectifID`, along with a README documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.