Curriculum Multi-Reward Reinforcement Learning for Customized Text-to-Image Generation

Anonymous submission

Abstract

Customized text-to-image generation aims to generate images that contain customized subjects according to given prompts, which has garnered considerable interest. Nevertheless, existing methods struggle to produce images that simultaneously fulfill requirements across multiple evaluation dimensions, such as prompt fidelity, subject fidelity, and human preference. To tackle this problem, we propose CurCustomizer, a curriculum multi-reward reinforcement learning framework for customized text-to-image generation. In this two-stage method, we first fine-tune diffusion models for customization. Then we adopt diverse evaluation metrics as reward signals and further fine-tune the model through reinforcement learning for reward maximization. To tackle the challenge of reward conflict during multi-reward optimization, we design a curriculum reweighting mechanism to adaptively balance the rewards rather than simply adding them up, which helps each reward be sufficiently optimized. Empirical experiments show that CurCustomizer consistently outperforms existing customized generation methods and generates images of satisfactory quality across various evaluation dimensions. The code is in the Supplementary Material.

1 Introduction

Customized text-to-image generation offers a valuable path for individuals outside the professional art domain to express their personalized interests and creativity. Due to the widespread demand for customized creation, this research field has attracted considerable attention. The typical procedure for implementing such generation involves providing a limited number (e.g., 3-6) of images depicting the same subject and assigning a unique token (e.g., [V]) to represent the subject so that the subject and the token are bound together and the fine-tuned generative model is capable of rendering images specific to the desired subject rather than an arbitrary one. Existing methods can be categorized into three branches (Cao et al. 2024) based on how to control the customized condition: tuning-based, model-based, and training-free. The most prominent one is the tuning-based, which maps the target subject into a unique token by fine-tuning a subset of parameters within the diffusion pipeline. This type of methods is represented by Textual Inversion (Gal et al. 2022), which fine-tunes the text encoder, and DreamBooth (Ruiz et al. 2023), which fine-tunes the U-Net model.

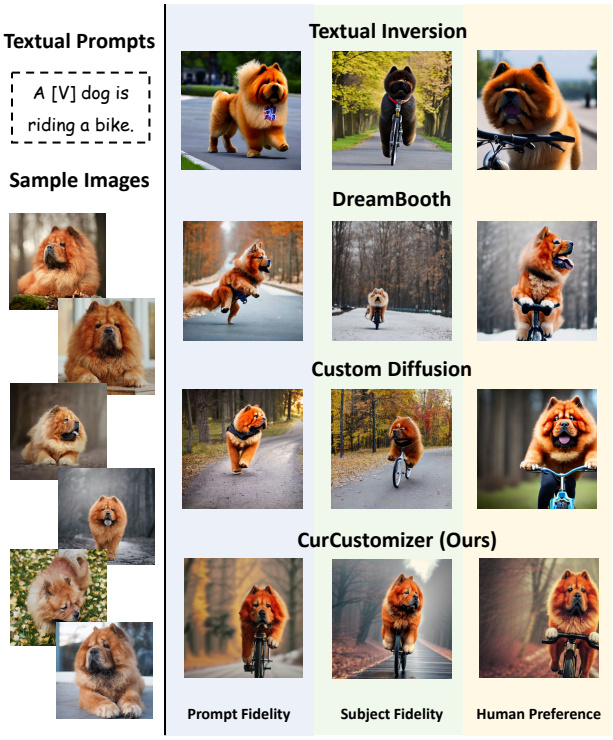


Figure 1: Comparison between existing customized generation methods and our CurCustomizer across three evaluation dimensions: prompt fidelity, subject fidelity, and human preference. Existing methods demonstrate limitations in achieving high performance across all three dimensions.

Despite the advancements in customized text-to-image generation, the images produced by these methods often fall short in simultaneously satisfying the criteria across multiple evaluation dimensions, including prompt fidelity, subject fidelity, and human preference, as illustrated in Figure 1. It is observed that existing representative methods may encounter the problem where the generated images are not consistent with the provided prompt (e.g., not riding a bike), the target subject (e.g., not the target dog), or human cognitive expectations (e.g., abnormal body parts).

Several prior works have explored this problem with the

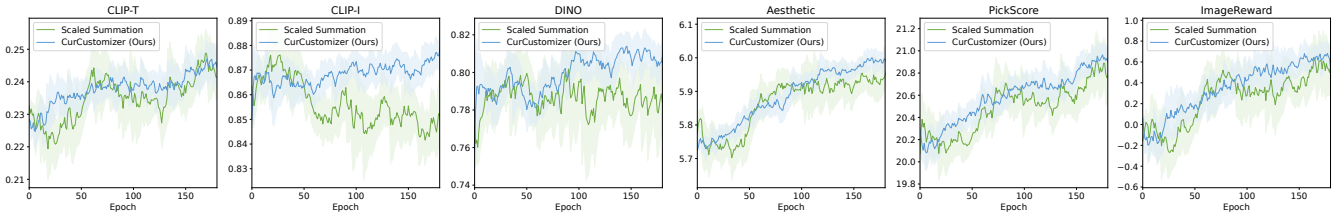


Figure 2: Changes in metrics for the baseline (scaled summation as reward) and our CurCustomizer during reinforcement learning fine-tuning. The lines are smoothed for clarity.

prevailing view suggesting that pre-trained generative models tend to overfit on the limited sample images (Kumari et al. 2023; Han et al. 2023; Tewel et al. 2023; Chae et al. 2023) or on the subject-irrelevant information within those images (Chen et al. 2023; Avrahami et al. 2023; Li et al. 2023; Motamed, Paudel, and Van Gool 2023; Cai et al. 2024). Although these advanced works have made efforts to address the overfitting problem through techniques such as preservation mechanisms (Ruiz et al. 2023; Kumari et al. 2023; Han et al. 2023; Tewel et al. 2023; Qiu et al. 2024; Wang et al. 2023), disentanglement learning (Chen et al. 2023; Cai et al. 2024), and reinforcement learning (Chae et al. 2023), they are not able to improve the quality of generated images across various evaluation dimensions simultaneously. Actually, it is non-trivial to achieve a comprehensive optimization of multiple metrics due to possible conflicts among them. To explain this issue, we illustrate the reinforcement learning process with the reward as the summation of multiple metrics (all scaled to $[0, 1]$) in Figure 2. These metrics include CLIP-T (Radford et al. 2021) for prompt fidelity; CLIP-I (Radford et al. 2021) and DINO (Caron et al. 2021) for subject fidelity; and Aesthetic (Schuhmann 2022), PickScore (Kirstain et al. 2024), and ImageReward (Xu et al. 2024) for human preference. While most metrics show increment, CLIP-I, and DINO consistently decrease, indicating that the model is optimized towards prompt fidelity and human preference at the expense of subject fidelity degradation.

To address the aforementioned problem and challenge, we propose CurCustomizer, a curriculum multi-reward reinforcement learning framework for customized text-to-image generation, which can improve images across various quality aspects simultaneously. Overall, we adopt a two-stage process: first, fine-tuning the diffusion model with sample images and a textual prompt containing the unique token to enable customized generation, and second, further fine-tuning through reinforcement learning using multiple evaluation metrics as rewards to enhance the customized generation. During reinforcement learning, we design a curriculum reweighting mechanism to address the conflicts among various metrics. The key insight lies in increasing the weight of the metric on which the model's performance declines or increases slowly during training. This enables the model to achieve comprehensive improvement across all metrics, analogous to students devoting more efforts to their weaker courses to improve their overall grades.

To validate the efficacy of our method, we conduct experiments on DreamBench (Ruiz et al. 2023), a dataset comprising tens of subjects and prompts for robust evaluation. The comparative empirical results show that our method brings notable enhancements to existing customized text-to-image generation, and the analytical experiments demonstrate how our method balances multiple rewards within the reinforcement learning framework. In summary, our contributions are outlined as follows.

- To the best of our knowledge, this work is the first one to investigate multi-reward reinforcement learning for customized text-to-image generation.
- We propose a curriculum reweighting method to adaptively balance different rewards and improve them all during reinforcement learning fine-tuning.
- Empirical experiments demonstrate the benefit and improvement our method brings across various quality criteria compared to existing methods.

2 Related Work

2.1 Customized Text-to-Image Generation

The development of customized generation techniques owes much to the advancements of text-to-image generation, where diffusion models pre-trained on large-scale datasets have succeeded in generating photo-like images (Nichol et al. 2021; Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022). Considering the requirements for customized creation and the inherent limitations of text in expressing complex concepts, researchers explore the use of sample images to identify user-defined subjects for tailored generation, known as customized or subject-driven text-to-image generation. Seminal works such as Textual Inversion (Gal et al. 2022) and Dreambooth (Ruiz et al. 2023) manage to embed the concept of the target subject into the generation process by fine-tuning text encoders and U-Net models respectively. Following these two works, diverse methods have been continuously emerging, which can be divided into three categories, distinguished by their conditional score prediction (Cao et al. 2024). The first is tuning-based, which selectively tunes a subset of parameters within the diffusion pipeline. The second one is model-based, exemplified by InstantBooth (Shi et al. 2023) and ELITE (Wei et al. 2023), which incorporates additional encoders for subject embedding. The third one is training-free, represented

by Re-Imagen (Chen et al. 2022), which extracts subject information directly from reference images.

2.2 Reinforcement Learning for Fine-Tuning

Reinforcement learning (RL) is a learning paradigm that trains an agent to learn a policy for maximum reward, mimicking the way humans learn from trial-and-error processes. Inspired by reinforcement learning from human feedback (RLHF) applied in language modeling, recent works have begun investigating RL fine-tuning for text-to-image generation. Fan et al. (2023) first regard diffusion denoising as a sequential decision-making process and apply RL to diffusion models by integrating policy gradient and GAN training. Subsequent approaches, such as DPOK (Fan et al. 2024), DDPO (Black et al. 2023), and DRaFT (Clark et al. 2023), incorporate additional rewards, particularly human preference metrics, and propose innovative policy gradient algorithms for diffusion denoising. Based on the works above, Parrot (Lee et al. 2024) introduces a multi-reward RL framework employing batch-wise Pareto optimal selection for text-to-image generation, and InstructBooth (Chae et al. 2023) applies RL fine-tuning to personalized text-to-image generation. Both works are closely aligned with ours. Nevertheless, compared to Parrot, we focus on customized text-to-image generation, and our curriculum reweighting method markedly diverges from the Pareto selection. Compared to InstructBooth, our emphasis is on multi-reward RL rather than single-reward RL, along with the identification and resolution of the conflict problem among evaluation metrics. To clarify our contribution, we provide a direct comparison with them in the Supplementary Material.

2.3 Curriculum Learning

Curriculum Learning (CL) is a training strategy that instructs machine learning models to learn in a meaningful order, similar to the way humans learn from curricula. Fundamentally, the key elements of CL encompass a measurer to distinguish the difficulty and a scheduler to organize the learning sequence, thereby facilitating the suitable curricula design (Wang, Chen, and Zhu 2021). Bengio et al. (2009) first propose a formal definition of CL and design a simple baseline, whose measurer and scheduler for curriculum are completely pre-defined, named BabyStep. Subsequently, this field has witnessed the emergence of various CL methods at the data, model, and task levels (Soviany et al. 2022), alongside their utilization across various domains such as computer vision (Sangineto et al. 2018; Guo et al. 2018), natural language processing (Platanios et al. 2019; Liu et al. 2018), audio processing, graph learning (Li, Wang, and Zhu 2023) and reinforcement learning (Narvekar et al. 2020; Portelas et al. 2020). To the best of our knowledge, there is currently no related work on CL at the level of evaluation metrics nor exploring its application in the field of customized text-to-image generation.

3 Preliminary

For simplicity of description, in this section, we review the general formulation of Latent Diffusion Model (LDM) (Rombach et al. 2022). Given a large-scale dataset with image-condition pairs $\{(x, y)\}$, the conditioning input y of the image x can be text, semantic maps, image-to-image translation tasks, and so on. LDM consists of an encoder $\mathcal{E}$ to encode x into a latent representation $z = \mathcal{E}(x)$, a decoder $\mathcal{D}$ to reconstruct the image from the latent $\mathcal{D}(z) \approx x$, a conditioning model c_θ to project y to a conditioning representation $c_\theta(y)$, and a U-Net based (Ronneberger, Fischer, and Brox 2015) diffusion model ϵ_θ to conduct diffusion denoising process in the latent space. The optimization objective is to predict the noise $\epsilon \sim \mathcal{N}(0, I)$, which is added to z at the time step $t \in [0, T]$ as z_t, and it can be formulated to:

$$\min \mathbb{E}_{z,y,\epsilon,t}\Big[||\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))||_2^2\Big]. \tag{1}$$

At training phase, both ϵ_θ and c_θ are jointly optimized via Equation (1). At the inference phase, randomly sampled noise is iteratively denoised to produce a latent z_0, which is then decoded to a newly generated image $\hat{x} = \mathcal{D}(z_0)$.

4 Method: CurCustomizer

4.1 Customized Text-to-Image Generation

The first stage of CurCustomizer is to embed the target subject into the generative model. Specifically, in this study, we employ the tuning-based method, which is widely utilized for customized text-to-image generation. Based on the notation in Section 3, we formulate this process as follows.

Given a limited set of images $\{x_k\}_{k=1}^K$ depicting the target subject, where K is the number of provided images, typically ranging from 3 to 6, and a text prompt P containing the unique token [V] as the conditioning input $y \triangleq P$, a pre-trained LDM ϵ_θ comprising a text encoder Γ as the conditioning model $c_\theta \triangleq \Gamma$ is fine-tuned using a similar scheme to that of the pre-training stage by minimizing the reconstruction loss:

$$\min \mathbb{E}_{z,P,\epsilon,t}\Big[||\epsilon - \epsilon_\theta(z_t, t, \Gamma(P))||_2^2\Big]. \tag{2}$$

Various methods differ in the trainable parameters. For example, Textual Inversion (Gal et al. 2022) fine-tunes the text encoder, Dreambooth (Ruiz et al. 2023) fine-tunes the entire U-Net, and Custom Diffusion (Kumari et al. 2023) fine-tunes key and value matrices of cross-attention layers.

Furthermore, preservation (Ruiz et al. 2023) or regularization (Kumari et al. 2023) losses are integrated into this stage to address the language drift issue (Lee, Cho, and Kiela 2019; Lu et al. 2020):

$$\begin{aligned}\min \mathbb{E}_{z,z',P,P',\epsilon,\epsilon',t,t'}\Big[&||\epsilon - \epsilon_\theta(z_t, t, \Gamma(P))||_2^2 + \\ &\lambda||\epsilon' - \epsilon_\theta(z'_{t'}, t', \Gamma(P'))||_2^2\Big],\end{aligned} \tag{3}$$

where $z' = \mathcal{E}(x')$ is the latent embedding of the external images, whether generated or collected, belonging to the same class as the target subject, P' is the text prompt without the unique token, and λ is the regularization coefficient.

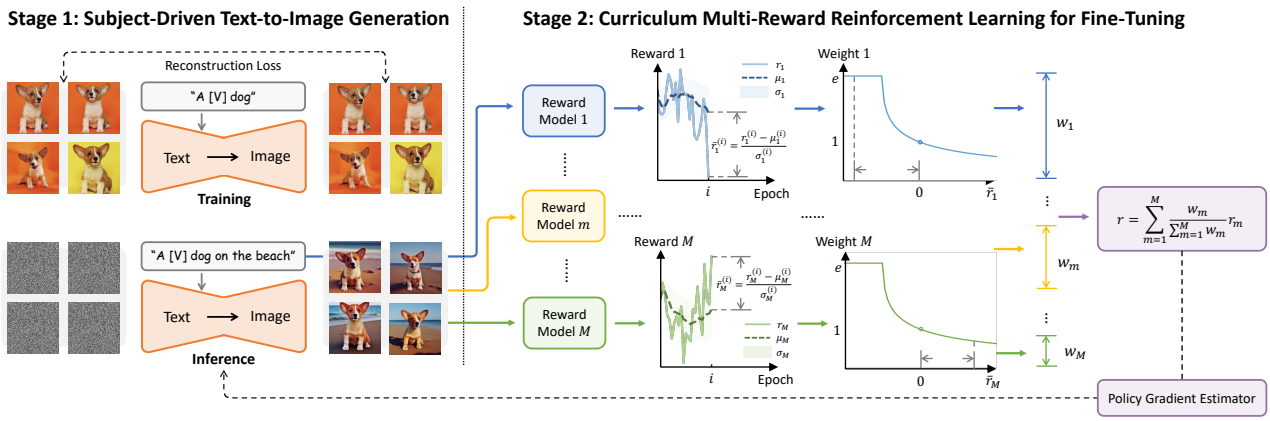


Figure 3: The framework of CurCustomizer.

4.2 Reinforcement Learning for Fine-Tuning

This stage is to further enhance the model for more refined images under diverse quality criteria. Given that image quality metrics, particularly human preference, are not differentiable, it is intuitive to employ reinforcement learning to directly optimize the model with these metrics.

Considering the diffusion denoising process as a sequential decision-making problem, where random noise z_T iteratively follows the parameterized transition distribution $p_\theta(z_{t-1}|z_t, y)$ and generates a trajectory $\tau = \{z_T, ..., z_0\}$ to yield the final latent z_0 for decoding into a new image $\hat{x} = \mathcal{D}(z_0)$, it can be formulated as a Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, \rho_0, \mathcal{P}, \mathcal{R}) = (\{s_t\}, \{a_t\}, \rho_0, \mathcal{P}(s_{t+1}|s_t, a_t), \mathcal{R}(s_t, a_t))$, respectively representing the state space, the action space, the distribution of initial states, the transition kernel, and the reward function:

$$\begin{aligned} &\rho_0 \triangleq \big(p(y), \delta_T, \mathcal{N}(0, I)\big), \quad \pi(a_t|s_t) \triangleq p_\theta(z_{t-1}|z_t, y), \\ &s_t \triangleq (z_t, t, y), \quad \mathcal{P}(s_{t+1}|s_t, a_t) \triangleq (\delta_{z_{t-1}}, \delta_{t-1}, \delta_y), \\ &a_t \triangleq z_{t-1}, \quad \mathcal{R}(s_t, a_t) \triangleq \begin{cases} r(z_0, y) & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{4}$$

where $\pi(a_t|s_t)$ refers to the policy at time step t, δ_z is the Dirac delta distribution at z, y is the conditioning input.

Since the cumulative reward $\mathcal{R}(s_t, a_t)$ equals $r(z_0, y)$, the optimization objective of reinforcement learning for diffusion denoising can be aligned with that of the MDP (Fan et al. 2024; Black et al. 2023):

$$\max \mathbb{E}_{p(\tau|\pi)} \left[\sum_{t=0}^{T} \mathcal{R}(s_t, a_t) \right] = \max \mathbb{E}_{p(y), p_\theta(z_0|y)} \big[r(z_0, y) \big], \tag{5}$$

which can be optimized by the policy gradient estimator (Williams 1992; Mohamed et al. 2020):

$$\mathbb{E} \left[\sum_{t=0}^{T} \nabla_\theta \log p_\theta(z_{t-1}|z_t, y) r(z_0, y) \right]. \tag{6}$$

4.3 Curriculum Multi-Reward Reweighting

Multiple Rewards. In reinforcement learning, a crucial aspect lies in designing the reward $r(z_0, y)$. Previous studies typically focus on a single reward at a time (Fan et al. 2024; Black et al. 2023; Xu et al. 2024; Chae et al. 2023), which may not adequately fulfill the demands of customized text-to-image generation, as numerous factors can influence the quality of a customized image. In this paper, we select diverse metrics including prompt fidelity (CLIP-T), subject fidelity (CLIP-I and DINO), and human preference (Aesthetic, PickScore, and ImageReward) as multiple rewards to facilitate more effective reinforcement learning fine-tuning. Detailed descriptions of these metrics are provided in the Supplementary Material.

Rewards as Curricula. During each epoch of fine-tuning, we regard rewards as curricula for the generative model and assign adaptive weight w_m to the m^{th} reward r_m as its importance. This allows the model to learn in an organized manner, maximizing the reweighted rewards:

$$\max_r r = \max_{r_1, ..., r_M} \sum_{m=1}^{M} w_m r_m, \tag{7}$$

where M is the total number of rewards. The next two paragraphs introduce how we measure the model's performances on all curricula and how to schedule among them.

Curriculum Measurer. Before assigning the weights, it is essential to assess how well the model performs on the metrics associated with the rewards. Directly comparing their absolute values is not feasible due to the considerable variation in the value range of each metric. Therefore, we propose to evaluate the model's performance changes on each metric by comparing its current value $r_m^{(i)}$ to its exponential moving average (EMA) $\mu_m^{(i)}$ under the constrain of its exponential moving standard deviation (EMSD) $\sigma_m^{(i)}$, i.e., calculating its

Z-Score $\overline{r}_m^{(i)}$ at each epoch (Finch 2009):

$$\begin{aligned}\mu_m^{(i)} &= (1-\alpha_m)\mu_m^{(i-1)} + \alpha_m r_m^{(i)},\\ \sigma_m^{(i)} &= (1-\alpha_m)(\sigma_m^{(i-1)} + \alpha_m (r_m^{(i)} - \mu_m^{(i)})^2),\\ \overline{r}_m^{(i)} &= \left(r_m^{(i)} - \mu_m^{(i)}\right)/\sigma_m^{(i)},\end{aligned} \tag{8}$$

where $\alpha_m \in [0,1]$ is a hyper-parameter for the decay factor of the m^{th} reward r_m, and i denotes the fine-tuning epoch. A relatively small or even negative value of $\overline{r}_m^{(i)}$ indicates that the model fails to make sufficient progress on the m^{th} reward in the current epoch, so we need to encourage the model to learn more from this reward by increasing the weight w_m. Conversely, a large positive value of $\overline{r}_m^{(i)}$ indicates significant improvement, suggesting a decrease in weight. Using this curriculum measurer, we then implement curriculum schedulers for all rewards to balance them and prevent the possible reward conflict, as described in the following paragraph.

Curriculum Scheduler. To implement the idea of increasing weights for poor metrics and decreasing them for good metrics, we formulate the curriculum scheduler as:

$$\forall\, \overline{r}_m \le \overline{r}_{m'},\ \text{s.t. } w_m \ge w_{m'} \ge 0. \tag{9}$$

The simplest possible resolution of Equation (9) is:

$$\begin{aligned}&\min_{w_1,\dots w_M} \sum_{m=1}^{M} w_m \overline{r}_m,\ \text{s.t.} \sum_{m=1}^{M} w_m = 1\\ \Longrightarrow\quad & w_m = \begin{cases}1 & \text{if } m = \arg\min_m \overline{r}_m\\ 0 & \text{otherwise}\end{cases}.\end{aligned} \tag{10}$$

This solution essentially selects the worst metric as the reward while ignoring other metrics, which is detrimental to the optimization of all metrics. Therefore, we propose a soft version by adding a regularizer $\beta_m(\log w_m)^2$, which encourages w_m to be close to 1 controlled by β_m:

$$\begin{aligned}&\min_{w_1,\dots w_M} \sum_{m=1}^{M} w_m \overline{r}_m + \beta_m(\log w_m)^2\\ \Longrightarrow\quad & w_m = \begin{cases}e & \overline{r}_m/\beta_m \le -2/e\\ e^{-\mathcal{W}(\overline{r}_m/\beta_m)} & \overline{r}_m/\beta_m > -2/e\end{cases},\end{aligned} \tag{11}$$

where $\beta_m \ge 0$ is a hyper-parameter, and $\mathcal{W}$ refers to Lambert W function. The detailed calculation process for Equation (10) and (11) are provided in the Supplementary Material, along with the visualization of the function between w_m and $\overline{r}_m$, illustrating that w_m satisfies Equation (9). Furthermore, to account for the relative performance changes between metrics, we introduce L1-normalization:

$$w_m \leftarrow \frac{w_m}{\sum_{m=1}^{M} w_m}. \tag{12}$$

Curriculum in Practice. Combining the multiple rewards and their corresponding weights, we can derive the aggregated reward r with Equation (7). However, the weights cannot eliminate the magnitude difference among the rewards. Following the standard practice in reinforcement learning[1], we normalize r_m to have zero mean and unit variance within one epoch as the practical reward. Furthermore, this normalization is on the same per-prompt basis as used in DDPO (Black et al. 2023). A detailed description of this process is provided in the Supplementary Material.

Algorithm 1: CurCustomizer

Require: a pre-trained diffusion model ϵ_θ, sample images $\{x_k\}_{k=1}^K$, text prompt P with a unique token [V], preservation loss coefficient λ, moving average factor α_m and curriculum reweighting coefficient β_m.

1: **Fine-tune** the diffusion model for customized text-to-image generation with Eq. (3).
2: **while** not convergent **do**
3: Derive multiple target rewards r_m;
4: Calculate the Z-Score of each reward with Eq. (8);
5: Update reward weight w_m with Eq. (11) and (12);
6: Aggregate all the reweighted rewards with Eq. (7);
7: Fine-tune the diffusion model via RL with Eq. (6).
8: **end while**
9: **Return** the fine-tuned customized generation model.

5 Experiments

5.1 Experimental Setup

Datasets. We utilize the DreamBench dataset introduced by Dreambooth (Ruiz et al. 2023). It is a collection of 30 subjects downloaded from Unsplash[2], including various objects such as backpacks, toys, and bowls, as well as live subjects such as dogs and cats. Additionally, it provides 25 text prompts related to recontextualization, accessorization, and property modification. Following previous studies, we generate 4 images per text prompt for each target subject, resulting in a total of 3000 images for comprehensive evaluation.

Comparable Methods. We compare our method with the following state-of-the-art methods based on the latent diffusion model proposed for customized text-to-image generation. Textual Inversion (Gal et al. 2022) inverts subjects to new pseudo-words in the embedding space of a pre-trained text-to-image model by fine-tuning the embedding lookup table of the text encoder. Dreambooth (Ruiz et al. 2023) binds subjects with unique identifiers by fine-tuning the text-to-image diffusion with reconstruction loss and class-specific prior preservation loss. Custom Diffusion (Kumari et al. 2023) introduces new modifier tokens for subjects by optimizing key and value projection matrices in the diffusion model cross-attention layers.

Evaluation Metrics. We evaluate the quality of customized generation in terms of prompt fidelity (CLIP-T), subject fidelity (CLIP-I and DINO), and human preference

[1] https://github.com/pytorch/examples/tree/main/reinforcement_learning

[2] https://unsplash.com

Table 1: Quantitative comparison between different methods on DreamBench. For Avg-Rank, a lower value indicates better performance, while for others, a higher value represents better results. We highlight the best one for each metric in bold.

	CLIP-T	CLIP-I	DINO	Aesthetic	PickScore	ImageReward	Avg-Rank (↓)
Textual Inversion	0.2730	0.7710	0.5665	5.2046	20.692	-0.8794	3.6462
Dreambooth	0.2883	0.7903	0.6322	5.3304	20.277	-0.1666	2.1308
Custom Diffusion	0.3018	0.7555	0.5624	5.2426	21.711	0.7353	2.5846
CurCustomizer	**0.3157**	**0.8026**	**0.6605**	**5.6174**	**21.944**	**0.8994**	**1.6385**

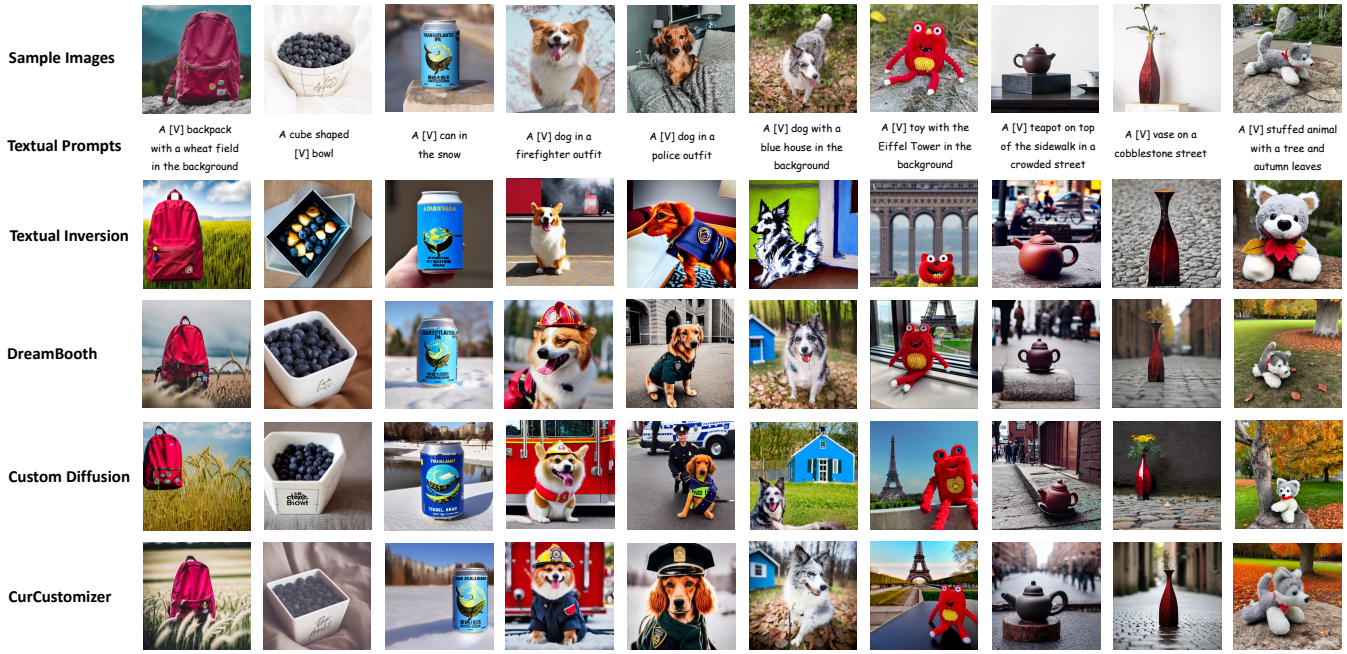


Figure 4: Customized text-to-image generation examples of different subjects given sample images and textual prompts. More results are provided in the Supplementary Material.

(Aesthetic, PickScore, and ImageReward). Apart from these optimized metrics, we introduce another metric named Average Human Rank (Avg-Rank) by inviting 30 users to rank the images generated by different methods in order to evaluate the comprehensive satisfaction of humans. The details of this user study are provided in the Supplementary Material.

5.2 Main Results

Quantitative Results. Table 1 reports the values of metrics across diverse quality dimensions. According to the results, Dreambooth has its edge on subject fidelity and aesthetics, while Custom Diffusion excels in prompt fidelity and human preference. Nevertheless, CurCustomizer outperforms them and obtains the highest values across all metrics, presenting notable and competitive performances. It is worth noting that CurCustomizer ranks top in the user study although Avg-Rank is not directly optimized, which further demonstrates the improvements CurCustomizer brings.

Qualitative Results. Figure 4 illustrates the quality of customized images generated by different methods. It is observed that Textual Inversion struggles to align with the given images and prompts. Custom Diffusion is able to generate images consistent with textual prompts but presents deficiencies in maintaining subject fidelity. Dreambooth performs well in both prompt fidelity and subject fidelity, but there is still room for improvement in human preference. Compared with these baselines, our CurCustomizer not only exhibits good alignment with subjects and prompts but also achieves excellent performance in terms of aesthetics, atmosphere, and light and shadow.

Rewards and Weights. To further analyze how CurCustomizer works, we take *dog2* in DreamBench as the subject and illustrate the changes in the values and weights of different rewards in Figure 5. It can be observed that the metrics with relatively slow growth tend to have higher weights, and vice versa. For example, the weights of CLIP-I mainly range from 0.75 to 1.45, while the weights of Aesthetic primarily range from 0.65 to 0.95. By increasing the weight of CLIP-I, CurCustomizer makes the optimization process more focused on improving the CLIP-I values, which are harder to optimize, thereby balancing rewards and resolving the conflicts between the rewards.

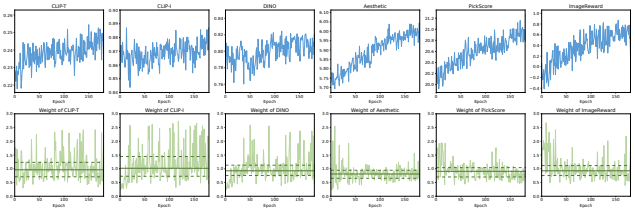


Figure 5: Changes in the values and weights of different rewards. In the bottom figures, the horizontal solid lines represent the median, and the horizontal dashed lines represent the quartiles.

5.3 Ablation Studies

As stated in Section 4.3, CurCustomizer employs curriculum reweighting to address reward conflict and follows the same per-prompt normalization as DDPO to eliminate the magnitude difference among the rewards. To demonstrate the effectiveness of these two parts, we take *dog2* in DreamBench as the subject and illustrate the ablation study in Figure 6. It is clear that without curriculum reweighting, CLIP-I and DINO show a declining trend due to the reward conflict. Without per-prompt normalization, each metric cannot be optimized effectively because customized images with different prompts may exhibit large variations in the metrics. Per-prompt normalization can alleviate these differences, leading to more stable and robust optimization.

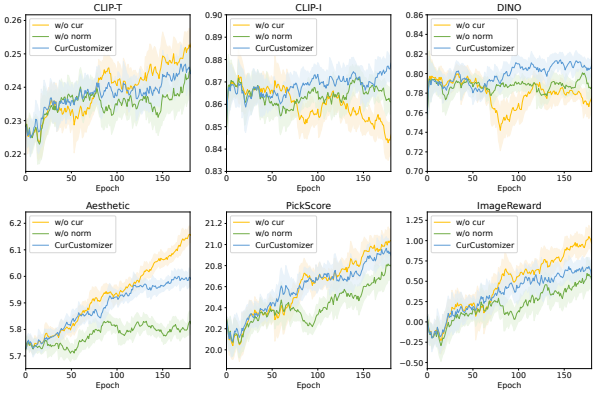
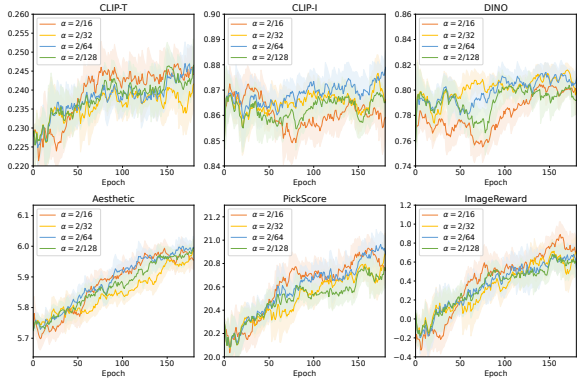


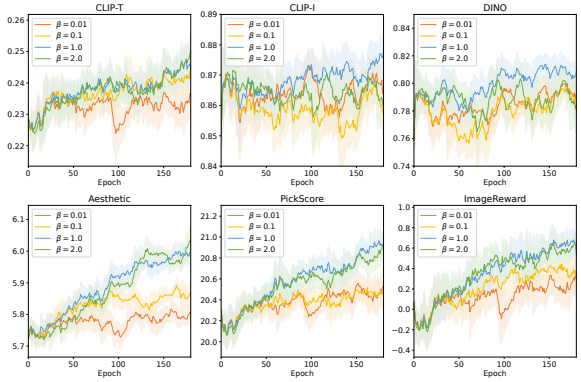
Figure 6: Ablation Study. CurCustomizer is compared with no curriculum reweighting (w/o cur) and no per-prompt normalization (w/o norm). The lines are smoothed for clarity.

5.4 Hyper-Parameters Sensitivity

We conduct experiments to study the sensitivity of results to α and β in curriculum reweighting. We first fix $\beta = 1.0$ to change α, and then fix $\alpha = 2/64$ to change β. During this process, all other training hyper-parameters are kept the same. The results are illustrated in Figure 7. It is found that α has a greater impact on the metrics that are hard to optimize, and β has the opposite effect. It is because as α increases, the window of the moving average decreases, making the method's perception of the overall trend weaker, thus failing to improve metrics that are difficult to optimize. On the other hand, as β decreases, the variance of weights increases, causing the weights of all metrics to fluctuate widely. Consequently, the weights of metrics that are easy to optimize become too small, resulting in poor overall performance.



(a) Keep $\beta = 1.0$ and compare $\alpha \in \{2/16, 2/32, 2/64, 2/128\}$.



(b) Keep $\alpha = 2/64$ and compare $\beta \in \{0.01, 0.1, 1.0, 2.0\}$.

Figure 7: Hyper-parameter sensitivity results. We fix one hyperparameter and adjust the other to demonstrate their impact on the results. The lines are smoothed for clarity.

6 Conclusion

In this paper, we study the multi-reward reinforcement learning fine-tuning for customized text-to-image generation. We select diverse evaluation metrics tailored to customized generation tasks serving as rewards in our framework. Additionally, we introduce a curriculum reweighting method to address the conflicts among different rewards. By adaptively increasing the weights of the metrics where the model performance declines throughout training, our approach facilitates comprehensive reward enhancement. Empirical comparisons and analytical experiments are conducted to validate the efficacy of our proposed method. A possible direction for future investigation is to extend curriculum learning in reinforcement learning fine-tuning by progressively introducing more refined images and complex prompts for further improvement on customized generation.

References

Avrahami, O.; Aberman, K.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, 1–12.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Cai, Y.; Wei, Y.; Ji, Z.; Bai, J.; Han, H.; and Zuo, W. 2024. Decoupled textual embeddings for customized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 909–917.

Cao, P.; Zhou, F.; Song, Q.; and Yang, L. 2024. Controllable Generation with Text-to-Image Diffusion Models: A Survey. *arXiv preprint arXiv:2403.04279*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chae, D.; Park, N.; Kim, J.; and Lee, K. 2023. InstructBooth: Instruction-following Personalized Text-to-Image Generation. *arXiv preprint arXiv:2312.03011*.

Chen, H.; Zhang, Y.; Wu, S.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2023. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *The Twelfth International Conference on Learning Representations*.

Chen, W.; Hu, H.; Saharia, C.; and Cohen, W. W. 2022. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

Clark, K.; Vicol, P.; Swersky, K.; and Fleet, D. J. 2023. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*.

Fan, Y.; and Lee, K. 2023. Optimizing ddpm sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*.

Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; and Lee, K. 2024. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36.

Finch, T. 2009. Incremental calculation of weighted mean and variance. *University of Cambridge*, 4(11-5): 41–42.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Guo, S.; Huang, W.; Zhang, H.; Zhuang, C.; Dong, D.; Scott, M. R.; and Huang, D. 2018. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European conference on computer vision (ECCV)*, 135–150.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7323–7334.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2024. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Lee, J.; Cho, K.; and Kiela, D. 2019. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*.

Lee, S. H.; Li, Y.; Ke, J.; Yoo, I.; Zhang, H.; Yu, J.; Wang, Q.; Deng, F.; Entis, G.; He, J.; et al. 2024. Parrot: Pareto-optimal Multi-Reward Reinforcement Learning Framework for Text-to-Image Generation. *arXiv preprint arXiv:2401.05675*.

Li, H.; Wang, X.; and Zhu, W. 2023. Curriculum Graph Machine Learning: A Survey. *arXiv preprint arXiv:2302.02926*.

Li, Y.; Liu, H.; Wen, Y.; and Lee, Y. J. 2023. Generate anything anywhere in any scene. *arXiv preprint arXiv:2306.17154*.

Liu, C.; He, S.; Liu, K.; Zhao, J.; et al. 2018. Curriculum Learning for Natural Answer Generation. In *IJCAI*, 4223–4229.

Lu, Y.; Singhal, S.; Strub, F.; Courville, A.; and Pietquin, O. 2020. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, 6437–6447. PMLR.

Mohamed, S.; Rosca, M.; Figurnov, M.; and Mnih, A. 2020. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132): 1–62.

Motamed, S.; Paudel, D. P.; and Van Gool, L. 2023. Lego: Learning to Disentangle and Invert Concepts Beyond Object Appearance in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2311.13833*.

Narvekar, S.; Peng, B.; Leonetti, M.; Sinapov, J.; Taylor, M. E.; and Stone, P. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *The Journal of Machine Learning Research*, 21(1): 7382–7431.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Platanios, E. A.; Stretcu, O.; Neubig, G.; Poczos, B.; and Mitchell, T. M. 2019. Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*.

Portelas, R.; Colas, C.; Weng, L.; Hofmann, K.; and Oudeyer, P.-Y. 2020. Automatic curriculum learning for deep rl: A short survey. *arXiv preprint arXiv:2003.04664*.

Qiu, Z.; Liu, W.; Feng, H.; Xue, Y.; Feng, Y.; Liu, Z.; Zhang, D.; Weller, A.; and Schölkopf, B. 2024. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sangineto, E.; Nabi, M.; Culibrk, D.; and Sebe, N. 2018. Self paced deep learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 41(3): 712–725.

Schuhmann, C. 2022. Laion aesthetics. https://laion.ai/blog/laion-aesthetics/. Aug 2022.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*.

Soviany, P.; Ionescu, R. T.; Rota, P.; and Sebe, N. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6): 1526–1565.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Wang, X.; Chen, Y.; and Zhu, W. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 4555–4576.

Wang, Z.; Wei, W.; Zhao, Y.; Xiao, Z.; Hasegawa-Johnson, M.; Shi, H.; and Hou, T. 2023. HiFi Tuner: High-Fidelity Subject-Driven Fine-Tuning for Diffusion Models. *arXiv preprint arXiv:2312.00079*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8: 229–256.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

A Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)

Does this paper make theoretical contributions? (yes, if a new method can be regarded as a new theory.)
If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes)
- All novel claims are stated formally (e.g., in theorem statements). (yes)
- Proofs of all novel claims are included. (yes)
- Proof sketches or intuitions are given for complex and/or novel results. (yes)
- Appropriate citations to theoretical tools used are given. (yes)
- All theoretical claims are demonstrated empirically to hold. (yes)
- All experimental code used to eliminate or disprove claims is included. (yes)

Does this paper rely on one or more datasets? (yes)
If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (yes)
- All novel datasets introduced in this paper are included in a data appendix. (yes)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (yes)

Does this paper include computational experiments? (yes)
If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (yes).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (no)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes)