PrefPaint: Aligning Image Inpainting Diffusion Model with Human Preference

Kendong Liu[1][*], **Zhiyu Zhu**[1][*][†], **Chuanhao Li**[2][*], **Hui Liu**[3],
Huanqiang Zeng[4], **Junhui Hou**[1]

[1]City University of Hong Kong [2]Yale University [3]Saint Francis University [4]Huaqiao University
`{kdliu2-c, zhiyuzhu2-c}@my.cityu.edu.hk`
`chuanhao.li.cl2637@yale.edu  h2liu@sfu.edu.hk  jh.hou@cityu.edu.hk`

Abstract

In this paper, we make the first attempt to align diffusion models for image inpainting with human aesthetic standards via a reinforcement learning framework, significantly improving the quality and visual appeal of inpainted images. Specifically, instead of directly measuring the divergence with paired images, we train a reward model with the dataset we construct, consisting of nearly 51,000 images annotated with human preferences. Then, we adopt a reinforcement learning process to fine-tune the distribution of a pre-trained diffusion model for image inpainting in the direction of higher reward. Moreover, we theoretically deduce the upper bound on the error of the reward model, which illustrates the potential confidence of reward estimation throughout the reinforcement alignment process, thereby facilitating accurate regularization. Extensive experiments on inpainting comparison and downstream tasks, such as image extension and 3D reconstruction, demonstrate the effectiveness of our approach, showing significant improvements in the alignment of inpainted images with human preference compared with state-of-the-art methods. This research not only advances the field of image inpainting but also provides a framework for incorporating human preference into the iterative refinement of generative models based on modeling reward accuracy, with broad implications for the design of visually driven AI applications. Our code and dataset are publicly available at `https://prefpaint.github.io`.

1 Introduction

Image inpainting, the process of filling in missing or damaged parts of images, is a critical task in computer vision with applications ranging from photo restoration [1, 2] to content creation [3, 4]. Traditional approaches have leveraged various techniques, from simple interpolation [5–7] to complex texture synthesis [8, 9], to achieve visually plausible results. The recent advent of deep learning, particularly diffusion models, has revolutionized the field by enabling more coherent and contextually appropriate inpaintings [3, 10–12]. Despite these advancements, a significant gap remains between the technical success of these models and their alignment with human aesthetic preferences, which are inherently subjective and intricate. As shown in Fig. 1, the existing stable diffusion-based inpainting model tends to generate weird and discord reconstruction.

Human preference for visual content is influenced by complex and mutual factors, including but not limited to personal background, experiences, and the context in which the content is viewed [13, 14]. This makes the task of aligning inpainting models with human preference particularly challenging, as

---

[*]Equal Contribution
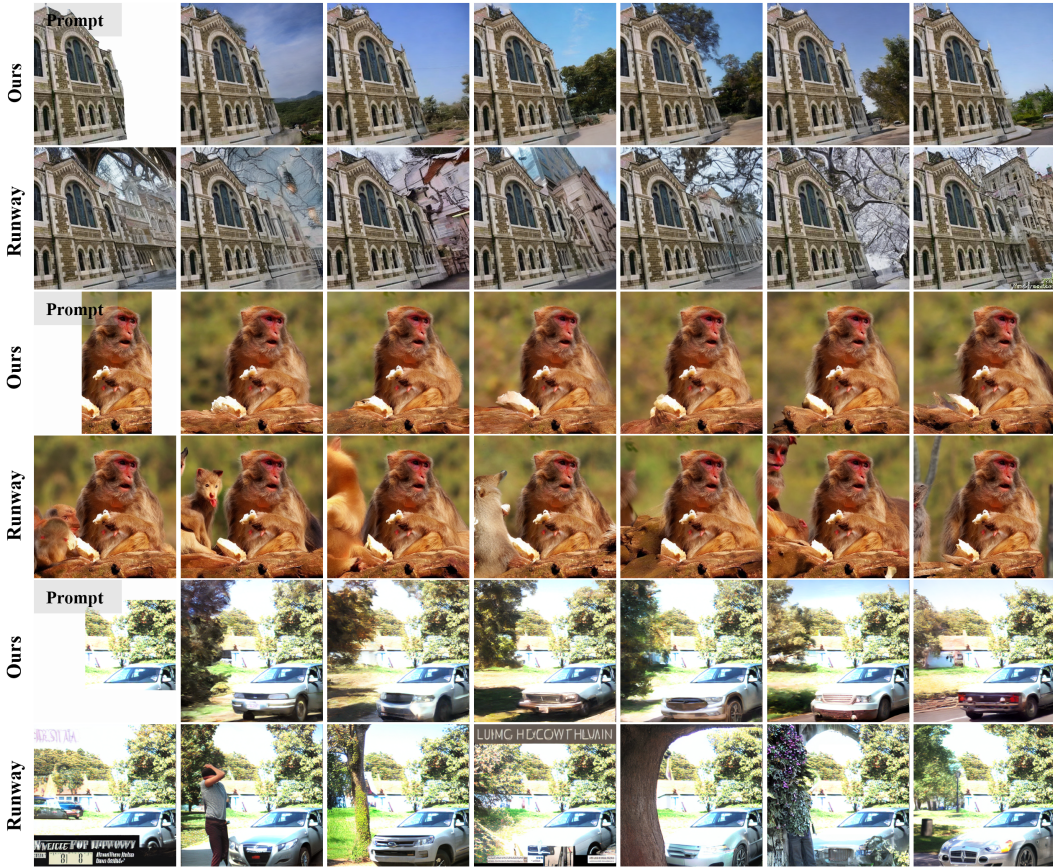[†]Corresponding Author.

Figure 1: Visual comparisons of the results by the diffusion-based image inpainting model named "*Runway*", and the aligned model through the proposed method.

it requires the model not only to understand the content of the missing parts but also to predict and adapt to the diverse tastes of its users.

Inspired by recent advancements in the reinforced alignment of large pre-trained models [15–19], we propose to align diffusion models for image inpainting with human preference through reinforcement learning. Specifically, our approach is grounded in the hypothesis that by incorporating human feedback into the training loop, we can guide the model toward generating inpainted images that are not only technically proficient but also visually appealing to users. Technically, we formulate the boundary of each reward prediction in terms of the model's alignment with human aesthetic preferences, thereby leveraging the accuracy of the reward model to amplify the regularization strength on more reliable samples. Extensive experiments validate that the proposed method can consistently reconstruct visually pleasing inpainting results and greatly surpass state-of-the-art methods.

In summary, the main contributions of this paper lie in:

- we make the *first attempt* to align diffusion models for image inpainting with human preferences by integrating human feedback through reinforcement learning;
- we theoretically deduce the accuracy bound of the reward model, modulating the refinement process of the diffusion model for robustly improving both efficacy and efficiency; and
- we construct a dataset containing 51,000 inpainted images annotated with human preferences.

2   Related Work

Reinforcement Learning & Model Alignment. Reinforcement learning [20–22] is a paradigm of machine learning where an agent learns to make decisions by taking actions in an environment to maximize some notion of cumulative reward. The foundational theory of reinforcement learning is rooted in the concepts of Markov decision processes [23], which provide a mathematical structure

for modeling decision-making in environments with stochastic dynamics and rewards [24, 25]. With the advent of deep learning, deep reinforcement learning [21, 26] has significantly expanded the capabilities and applications of traditional reinforcement learning. The integration of neural networks with reinforcement learning, exemplified by the Deep Q-Network [27, 28] algorithm, has enabled the handling of high-dimensional state spaces, which were previously intractable. Subsequent innovations, including policy gradient methods [29, 30] like Proximal Policy Optimization [31] and actor-critic frameworks [32, 33] like Soft Actor-Critic [34], have further enhanced the efficiency and stability of learning in complex environments.

The recent surge in popularity of large-scale models has significantly underscored the importance of reinforcement learning in contemporary AI research and application [15, 16]. As these models, including large language models [35, 36] and deep generative networks [37–39], become more prevalent, reinforcement learning is increasingly employed to fine-tune, control, and optimize their behaviors in complex, dynamic environments. This integration is particularly visible in areas such as natural language processing, where reinforcement learning techniques are used to improve the conversational abilities of chat-bots and virtual assistants, making them more adaptive and responsive to user needs. [17–19] Moreover, in the realm of content recommendation and personalization, reinforcement learning algorithms are instrumental in managing the balance between exploration of new content and exploitation of known user preferences, significantly enhancing the user experience. The growing intersection between large models and reinforcement learning not only pushes the boundaries of what's achievable in AI but also amplifies the need for sophisticated reinforcement learning techniques that can operate at scale, adapt in real-time, and make decisions under uncertainty, thereby marking a pivotal evolution in how intelligent systems are developed and deployed.

Diffusion Model. Diffusion models have recently emerged as a powerful class of generative models [38–42], demonstrating remarkable success in generating high-quality, coherent images [43, 44]. These models work by gradually transforming a distribution of random noise into a distribution of images, effectively 'diffusing' the noise into structured patterns [45, 46]. In the context of image inpainting, diffusion models offer a significant advantage by leveraging their generative capabilities to predict and fill in missing parts of images in a way that is contextually and visually coherent with the surrounding image content [47, 48].

Recent studies have showcased the potential of diffusion models in achieving state-of-the-art results in image inpainting tasks, outperforming previous generative models like Generative Adversarial Networks in terms of image quality and coherence [49, 50]. However, while these models excel in technical performance, there remains a gap in their ability to cater to diverse human aesthetic preferences. Most existing works focus on the objective quality of inpainting results, such as fidelity to the original image and coherence of the generated content, with less attention given to subjective satisfaction or preference alignment.

3 Proposed Method

Due to the random masking within the task of image inpainting, there may not be a definitive causal relationship between the known and inpainted contents, which manifests as a one-to-many issue. Consequently, stringent per-pixel regularization inherently results in unnatural reconstruction, which significantly diverges from samples conforming to human preference. To this end, we propose a reinforcement learning-based alignment process involving human preferences to fine-tune pre-trained diffusion models for image inpainting, aiming to improve the visual quality of inpainted images (Sec. 3.1). More importantly, we theoretically deduce the upper bound on the error of the reward model (Sec. 3.2). Based on this deduction, we formulate a reward trustiness-aware alignment process that is more efficient and effective (Sec. 3.3).

3.1 Reinforced Training of Diffusion Models for Image Inpainting

Diffusion models [38, 40] iteratively refine a randomly sampled standard Gaussian distributed noise, resulting in a generated image. To adjust the distribution of diffusion models, we introduce human feedback rewards to measure and regularize the distribution of model sampling outputs. To be specific, instead of applying standard policy-gradient descent [51] that is hard to converge to high-quality models, inspired by classical methods, e.g., TRPO/PPO [31, 52], which introduces a model trust region to avoid potential model collapse during the training process, we achieve the reinforced

training of a diffusion model as

$$\nabla_{\boldsymbol{\theta}}\mathcal{J}(\boldsymbol{x}) = -\int_{P_{\boldsymbol{\theta}'}} \frac{\nabla_{\boldsymbol{\theta}}P_{\boldsymbol{\theta}}(\boldsymbol{x})}{P_{\boldsymbol{\theta}'}(\boldsymbol{x})}\mathcal{R}(\boldsymbol{x}) + \kappa\nabla_{\boldsymbol{\theta}}\mathcal{D}(P_{\boldsymbol{\theta}'}|P_{\boldsymbol{\theta}}), \tag{1}$$

where $\mathcal{R}(\cdot)$ represents the reward model, which quantifies the quality of diffusion samples; $\boldsymbol{x}$ is the reconstructed sample; $P_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{\theta}'}$ represent the probabilistic functions of training and reference models parameterized with $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, respectively; $\nabla_{\boldsymbol{\theta}}$ calculates the derivative on $\boldsymbol{\theta}$; $\mathcal{D}(\cdot|\cdot)$ is the divergence measurement of the given two distributions for regularizing the distribution shifting; the hyperparameter κ balances the two terms. Inspired by recent work [53, 54], we take the same step to measure the probability of each diffusion step via calculating the probability of perturbation noise, e.g., the DDIM sampling algorithm [40],

$$\boldsymbol{x}_{t-1} = \underbrace{\sqrt{\alpha_{t-1}}\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{x}_t)}_{\bar{\boldsymbol{x}}_{t-1}} + \underbrace{\sigma_t\epsilon_t}_{\widetilde{\boldsymbol{x}}_{t-1}}, \tag{2}$$

where $\epsilon_{\boldsymbol{\theta}}^{(t)}(\cdot)$ is the noise estimate diffusion network; t is the diffusion step; the scalar α_t controls the noise to signal ratio; and $\sigma_t\epsilon_t$ represents a Gaussian noise to increase the sampling diversity. Thus, the result of each reverse step consists of a deterministic component $\bar{x}_{t-1}$ and a probabilistic component $\widetilde{x}_{t-1}$. Since $\boldsymbol{x}_{t-1} \sim \mathcal{N}(\bar{\boldsymbol{x}}_{t-1}, \sigma_t\mathbf{I})$, we calculate the density of each step $P = \Phi\left(\frac{\boldsymbol{x}_{t-1}-\bar{\boldsymbol{x}}_{t-1}}{\sigma_t}\right)$ to approximate the probability, where $\Phi(\cdot)$ denotes the density function of standard Gaussian distribution.

3.2 Bounding Reward Model Error

The precision of the reward model $\mathcal{R}(\cdot)$ assumes a pivotal function within this learning framework as it directs the optimization trajectory. In this section, we theoretically derive its error upper bound, which can facilitate the reinforced training process in terms of both efficacy and efficiency.

Denote by $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{N \times D}$ and $\boldsymbol{y} = [y_1, y_2, \cdots, y_N] \in \mathbb{R}^{N \times 1}$ a set of N inpainted images of dimension D by diffusion models and corresponding ground-truth reward values from human experts, respectively. By dividing the reward model into two parts, i.e., feature extractor $\mathcal{F}(\cdot)$ and linear regression weights $\boldsymbol{\psi} \in \mathbb{R}^{D' \times 1}$, we thus can represent it as $\langle \mathcal{F}(\boldsymbol{x}_t), \boldsymbol{\psi} \rangle$, where $\langle \cdot, \cdot \rangle$ calculates the inner product of two vectors. We can formulate the learning process of the last weights as the result of the following optimization process:
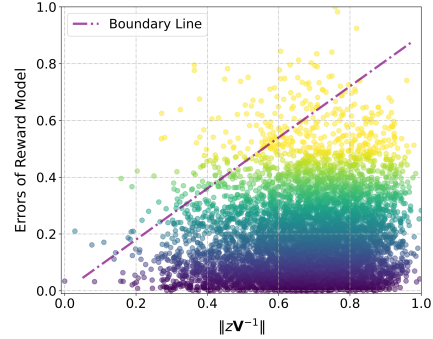
Figure 2: Experimental plot of reward prediction error *vs.* $\|z\|_{\mathbf{V}^{-1}}$ on the validation set, where a dashed line is an upper boundary of error, positively relative to $\|z\|_{\mathbf{V}^{-1}}$.

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{\arg\min}\left\|\langle\mathcal{F}(\mathbf{X}), \boldsymbol{\psi}\rangle - \boldsymbol{y}\right\|_2^2 + \lambda\|\boldsymbol{\psi}\|_2^2, \tag{3}$$

where $\|\cdot\|_2$ is the ℓ_2 norm of a vector, $\lambda > 0$, and the second term is used for alleviating the over-fitting phenomenon. Thus, we have

$$\hat{\boldsymbol{\psi}} = (\mathbf{Z}^\mathsf{T}\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^\mathsf{T}\boldsymbol{y}, \tag{4}$$

$$\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_* = (\mathbf{Z}^\mathsf{T}\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^\mathsf{T}\boldsymbol{\zeta} - \lambda(\mathbf{Z}^\mathsf{T}\mathbf{Z} + \lambda\mathbf{I})^{-1}\boldsymbol{\psi}_*, \tag{5}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times D'} = \mathcal{F}(\mathbf{X})$ is the set of embeddings of $\mathbf{X}$. $\boldsymbol{\psi}_*$ is the ideal weight with $\boldsymbol{y} = \mathbf{Z}\boldsymbol{\psi}_* + \boldsymbol{\zeta}$ and $\boldsymbol{\zeta}$ a noise term between $\mathbf{Z}\boldsymbol{\psi}_*$ and $\boldsymbol{y}$. Then, the following term also holds:

$$\boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{\psi}} - \boldsymbol{z}^\mathsf{T}\boldsymbol{\psi}_* = \boldsymbol{z}^\mathsf{T}(\mathbf{Z}^\mathsf{T}\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^\mathsf{T}\boldsymbol{\zeta} - \lambda\boldsymbol{z}^\mathsf{T}(\mathbf{Z}^\mathsf{T}\mathbf{Z} + \lambda\mathbf{I})^{-1}\boldsymbol{\psi}_*, \tag{6}$$

where $\boldsymbol{z} \in \mathbb{R}^{D' \times 1} = \mathcal{F}(\boldsymbol{x})$ is the reward embedding of a typical input sample. Based on Cauchy-Schwarz inequality, there is

$$|\boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{\psi}} - \boldsymbol{z}^\mathsf{T}\boldsymbol{\psi}_*| \le \|\boldsymbol{z}\|_{\mathbf{V}^{-1}}(\|\mathbf{Z}^\mathsf{T}\boldsymbol{\zeta}\|_{\mathbf{V}^{-1}} + \lambda^{1/2}\|\boldsymbol{\psi}_*\|_2), \tag{7}$$
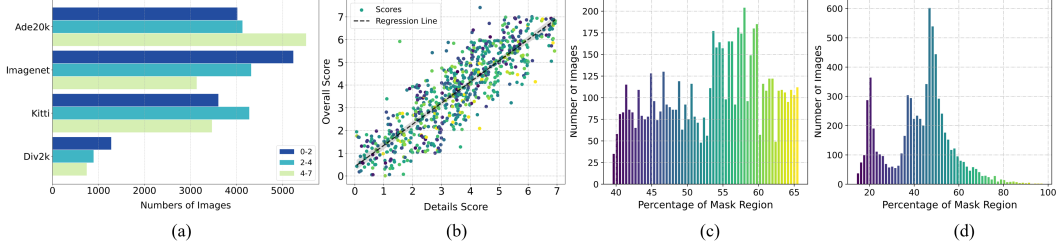
4

Figure 3: Statistical characteristics of the dataset we constructed. **(a)** the score distribution of the images across different selected datasets; **(b)** the comparison between the distribution of the average score and score for details; **(c)** and **(d)** show the numbers of images with different mask ratios on the outpainting and warping splits, respectively.

where $\mathbf{V} = \mathbf{Z}^\mathsf{T}\mathbf{Z} + \lambda\mathbf{I}$ and $\|\boldsymbol{z}\|_{\mathbf{V}^{-1}} := \|\boldsymbol{z}^\mathsf{T}\mathbf{V}^{-1}\|_2$. Moreover, based on *Theorem 2* of [55], for $\delta > 0$, we have $1 - \delta$ probability to make following inequality stands

$$\|\mathbf{Z}^\mathsf{T}\epsilon\|_{\mathbf{V}^{-1}} \le B\sqrt{2\log(\frac{\det(\mathbf{V})^{1/2}\det(\lambda\mathbf{I})^{-1/2}}{\delta})}, \tag{8}$$

where $\det(\cdot)$ computes the determinant of the matrix, and B is a scalar. We have

$$|\boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{\psi}} - \boldsymbol{z}^\mathsf{T}\boldsymbol{\psi}_*| \le \|\boldsymbol{z}\|_{\mathbf{V}^{-1}} \underbrace{\left(B\sqrt{2\log(\frac{\det(\mathbf{V})^{1/2}\det(\lambda\mathbf{I})^{-1/2}}{\delta})} + \lambda^{1/2}\|\boldsymbol{\psi}_*\|_2 \right)}_{C_{bound}}. \tag{9}$$

Due to C_{bound} being constant after the training of the reward model, we can conclude that

$$\sup_{\boldsymbol{z}\sim p(\boldsymbol{z})} |\boldsymbol{z}^\mathsf{T}\hat{\boldsymbol{\psi}} - \boldsymbol{z}^\mathsf{T}\boldsymbol{\psi}_*| \propto \|\boldsymbol{z}\|_{\mathbf{V}^{-1}}, \tag{10}$$

where $\sup$ represents the upper bound. Such a theoretical bound is also experimentally verified in Fig. 2.

3.3 Reward Trustiness-Aware Alignment Process

According to Eq. (10), it can be known that the error of the reward model is bounded by $\|\boldsymbol{z}\|_{\mathbf{V}^{-1}}$, there might be a relatively large error for those $\|\boldsymbol{z}\|_{\mathbf{V}^{-1}}$ quite large. Thus, we further propose a weighted regularization strategy to amplify the penalty strength of those samples in the high-trust region. Specifically, we calculate the amplification factor as

$$\gamma = e^{-k\|\boldsymbol{z}\|_{\mathbf{V}^{-1}}+b}, \tag{11}$$

where the hyperparameters $k = 0.05$ and $b = 0.7$ are used for scaling the regularization strength. We then define the overall gradient of the reward trustiness-aware alignment process as

$$\nabla_{\boldsymbol{\theta}}\mathcal{J}'(\boldsymbol{x}) = \gamma\nabla_{\boldsymbol{\theta}}\mathcal{J}(\boldsymbol{x}), \tag{12}$$

where $\mathcal{J}'(\boldsymbol{x})$ is the weighted reward loss for sample $\mathbf{x}$. For the updating of the reference model $\boldsymbol{\theta}'$, we adopt [53] to update the model in each optimization step. Correctly amplifying the scaling factor can both speed up the convergence speed and reconstruction effectiveness, as experimentally validated in Sec. 5.2.

4 Human Preference-Centric Dataset for Reward

We first randomly selected 6,000, 4,000, 6,000, and 1,000 images with diverse content from ADE20K [56, 57], ImageNet [58], KITTI [59], and Div2K [60, 61] datasets, respectively. We then applied the operations described below to generate prompt images (i.e., incomplete images), which were further fed into the diffusion model for image inpainting named *Runway* [62], producing inpainted images. To mitigate the potential bias of the reward model on different rewards, we

Table 1: Quantitative comparisons of different methods. $\star$ indicates the small model (non SD-based); "S" is the number of sampling times. For the calculation of WinRate, we first derive the best sample of the compared method among S sampling times. Then, we calculate it as $\frac{T_w}{T}$, where T_w indicates the number of compared samples that surpass the results of *Runway* ($S = 1$) and T is the total number of prompts. "↑ (resp. ↓)" means the larger (resp. smaller), the better. We normalized the predicted reward values with the dataset distribution. "Var" calculates the variance of different sampling times, showing the consistency of generation quality. (See the *Supplementary Material* for more details.)

| Prompt Methods | Outpainting Prompts | | | | | Warping Prompts | | | | |
| | WinRate (%) ↑ | | | Reward | | WinRate (%) ↑ | | | Reward | |
Metrics	$S = 1$	$S = 3$	$S = 10$	Mean↑	Var↓	$S = 1$	$S = 3$	$S = 10$	Mean↑	Var↓
Runway [62]	$--$	73.40	89.32	$--$	0.07	$--$	75.74	91.42	$--$	0.06
SD v1.5 [63]	11.95	20.67	30.24	-0.43	0.05	11.38	21.22	32.85	-0.38	0.06
SD v2.1 [64]	10.73	18.51	26.82	-0.44	0.04	11.68	22.11	34.22	-0.36	0.06
SD xl [65]	14.56	22.58	31.09	-0.31	0.04	15.43	25.43	36.77	-0.26	0.05
SD xl ++ [66]	21.15	33.25	45.51	-0.13	0.05	18.66	30.53	43.07	-0.18	0.04
Compvis [67]	50.51	66.39	78.21	+0.03	0.03	47.35	65.08	78.01	-0.01	0.04
Kandinsky [68]	14.06	22.73	32.16	-0.37	0.04	11.38	19.46	29.20	-0.42	0.05
MAT $\star$ [69]	15.06	17.97	20.51	-0.40	0.01	7.17	9.97	12.96	-0.56	0.01
Palette $\star$ [41]	10.96	16.92	21.37	-0.38	0.02	13.41	20.18	27.37	-0.34	0.03
Ours	**70.16**	**84.65**	**93.14**	**+0.38**	**0.01**	**72.38**	**87.10**	**93.85**	**+0.36**	**0.01**

Metric / Model	T2I [54]	CLIP [70]	BLIP [71]	Aes. [72]	CA [73]	IS [74]	Rank
SDv1.5	-1.67	0.19	0.44	4.52	0.38	17.07	5.17
SDv2.1	-1.37	0.20	0.45	4.62	0.39	17.07	4.33
Kand.	-3.49	0.18	0.39	**5.19**	0.39	17.06	5.33
SD xl ++	0.63	0.21	0.46	4.77	0.40	18.95	3.17
Runway	3.16	0.22	0.48	4.61	0.43	20.30	2.33
Platte	-1.76	0.22	0.46	4.08	0.37	16.24	5.33
Ours	**4.49**	**0.23**	**0.49**	4.55	**0.45**	**23.71**	**1.67**

Table 2: Comparison across metrics: higher values are better for all metrics except "Rank".

Figure 4: WinRate comparison heat-map between different methods.

repetitively generated from a given prompt image three distinct inpainted images. Consequently, we obtained 51,000 inpainted images in total, which were scored by human experts following the criteria described below.

Generation of Incomplete Images. We considered two distinctive image completion patterns: inpainting and outpainting. For inpainting, we simulated warping holes on images by changing the viewpoints, where we derived the depth of the scene using [75]. Following past practice [76], we defined a camera sequence that forms a sampling grid with three columns for yaw and three rows for pitch. The resulting nine views feature yaw angles within a ±0.3 range (i.e., a total range of 35°) and pitch angles within a ±0.15 range (i.e., a total range of 17°). As the range of views increases, the task of inpainting becomes progressively more challenging. For outpainting, we randomly masked the boundary of the image through two types of random cropping methods: (1) square cropping, where the size of the prompt ranges from 15% to 25% of the image size (512×512) randomly; (2) rectangular cropping, where the height of the prompt matches the image size, while the width is randomly sampled between 35% and 40% of the image size. Each cropping method accounts for half of the outpainting prompts.

Scoring Criteria. In the scoring process, we employed three criteria, i.e., (**1**) *structural rationality* representing the rationality of the overall structure, whether illogical objects and structures are generated; (**2**) *feeling of local texture* showing whether strange textures are generated that do not conform to the characteristics of the object; and (**3**) *overall feeling* indicating the impression capturing the overall feeling upon first glances at the image. The score value is in the range of 1 to 7, indicating
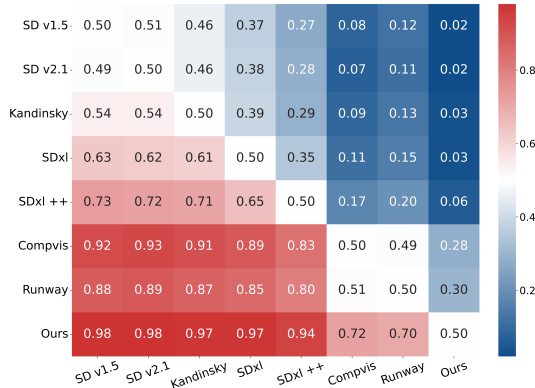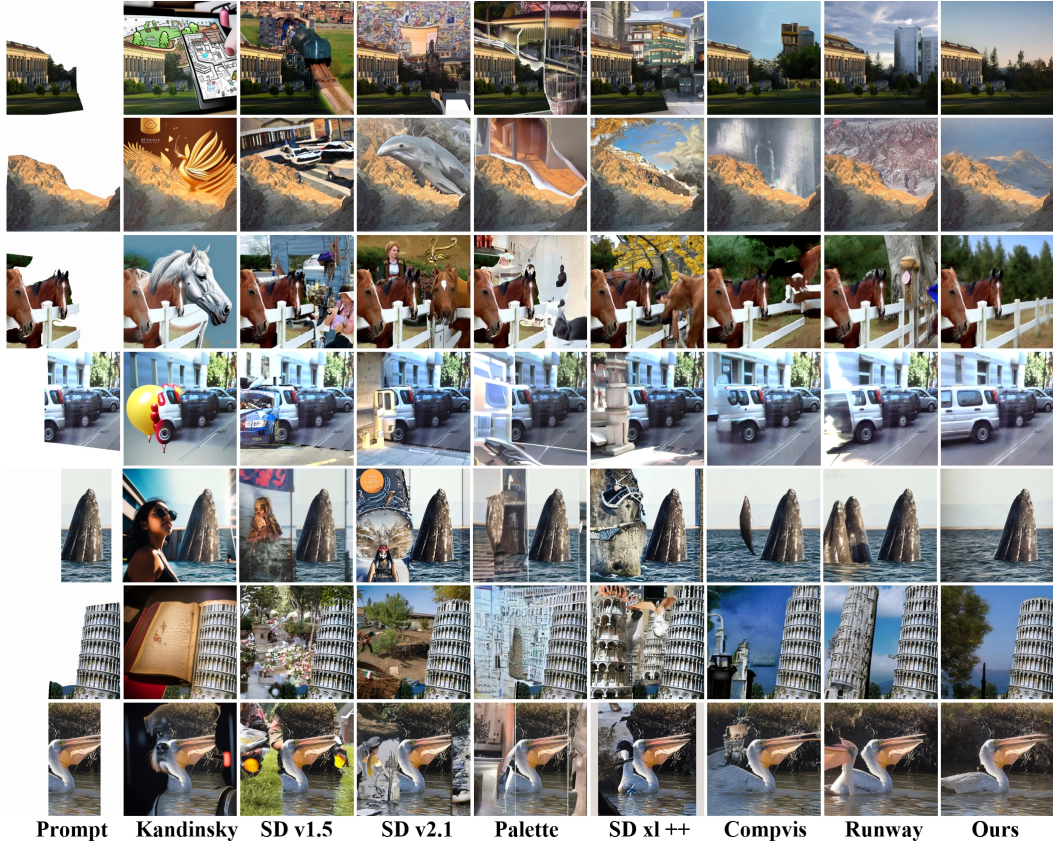
Figure 5: Visual comparisons of our approach and SOTA methods. The prompted images of 5th and 7th rows are generated by boundary cropping, while the remaining rows by warping. All images were generated with the same random seeds.

from the least to most favorable. The final score for reward model training is derived by averaging those three scores using the weights of $[0.15, 0.15, 0.7]$. We refer readers to the *Supplementary Material* for more details about dataset process and labeling scheme.

Statistical Characteristics of our Dataset. As illustrated in Fig. 3 (a), the score distribution of images from different datasets is generally uniform. Meanwhile, from Fig. 3 (b), we can see that the details and overall score are independent of each other, showing the correctness of the scoring scheme and the necessity of each score. Finally, Figs. 3 (c) and (d) show the ratio of images with different mask sizes, where it can be seen that the outpainting has a more uniform distribution than warping based hole, which rely on depth map and may not have uniform distribution.

5 Experiments

Evaluation Metrics. We adopted seven metrics to measure the quality of inpainted images, i.e., the predicted reward value by our trained reward model , T2I reward [54], CLIP [70], BLIP [71], Aesthetic (Aes.) [72], Classification Accuracy (CA) [73], and Inception Score (IS) [74]. Specifically, T2I reward, CLIP, BLIP, and Aes. directly measure the consistency between the semantics of inpainted images and the language summary of corresponding prompt images. While, CA and IS indicate the quality of the generative model. Due to the fact that our inpainting reward directly measures the reconstruction quality, we adopted it as the principle measurement of our experiment.

Implementation Details. We partitioned the dataset in Sec. 4 into training, validation and testing sets, containing 12,000, 3,000 and 2,000 prompts (with 36,000, 9,000 and 6,000 images), respectively. The reinforcement fine-tuning dataset contains the prompts from the original reward training dataset. We employed the pre-trained CLIP (ViT-B) checkpoint as the backbone of our reward model $\mathcal{R}(\cdot)$ with the final MLP channel equal to 256. We utilized a cosine schedule to adjust the learning rate. Notably, we achieved optimal preference accuracy by fixing 70% of the layers with a learning rate of $1e - 5$ and a batch size of 5. We trained the reward model with four

Table 3: **Left**: Ablation studies on amplification factors, where "static" refers to employing a constant factor to replace γ in Eq. (12). The column "Factor" indicates an average magnitude of amplification strength, i.e., $\mathbb{E}_{\mathbf{z}\sim\mathbf{P}(\mathbf{z})}(\gamma)$. For our method, we coordinate the value of k in Eq. (11) to change the $\mathbb{E}(\gamma)$ shown in the Table. "Acl." signifies acceleration, calculated by $\frac{T_b}{T_m} - 1$ with T_b and T_m being the convergence iterations of baseline and compared methods, respectively. For all metrics, the larger, the better. **Right**: Performance of the reward model trained with two manners based on pre-trained CLIP [70] with various fix rates (FRs). "Acc" and "Var" stand for the accuracy and variance of the reward estimation, respectively. "Bd." is the ratio of data below the same upper boundary. (The underlined settings are selected.)

	RL	Amp.	Factor	WinRate	Reward	**Acl.**
a)	×	- -	- -	50.00%	0.00	- -
b)	✓	×	1.0	69.57%	+0.36	00.00%
c)	✓	Static	1.4	65.95%	+0.34	+84.28%
d)	✓	Static	1.6	65.11%	+0.35	**+203.53 %**
e)	✓	Ours	1.4	**71.27%**	**+0.37**	+106.40%
f)	✓	Ours	1.6	70.47%	+0.36	+102.83%

Training	Classification-driven		Regression-driven	
Width	MLP-128	MLP-256	MLP-128	MLP-256
Metrics ↑	Acc. Var.	Acc. Var.	Acc. Var. Bd.	Acc. Var. Bd.
FR. = 0.3	73.49 0.24	74.29 0.18	74.68 0.41 92.1	74.88 0.38 93.2
FR = 0.5	74.73 0.25	**75.32 0.26**	74.78 0.44 **92.3**	75.32 0.44 94.4
FR = 0.7	**74.89 0.28**	74.53 0.25	**75.91 0.46** 85.0	<u>75.94</u> <u>0.45</u> <u>97.3</u>
FR = 0.9	74.09 0.16	72.76 0.14	74.73 0.41 78.8	74.83 0.40 **98.0**

NVIDIA GeForce RTX 3090 GPUs, each equipped with 20GB of memory. With the trained reward model, we subsequently fine-tuned the latest diffusion-based image inpainting model, namely *Runway* [62], on four 40GB NVIDIA GeForce RTX A6000 GPUs as our PrefPaint. During fine-tuning, we employed half-precision computations with a learning rate of $2e - 6$, and a batch size of 16.

5.1 Results of Image Inpainting

We compared our PrefPaint with SOTA methods both quantitatively and qualitatively to demonstrate the advantage of our method. We compared the

Table 4: Performance under various parameterizations of amplification factor γ, where $f = \|\boldsymbol{z}\|_{\mathbf{V}^{-1}}$. (The underlined settings are selected.)

Amplification factor (γ)	k	b	WinRate↑	Reward↑
e^{-kf+b}	0.050	0.70	**71.27%**	**0.37**
e^{-kf+b}	0.065	0.90	70.47%	0.36
$e^{-kf} + b_1/f + b_2$	0.100	{0.10, 0.85}	70.07%	0.37
$e^{-kf} + b_1/f + b_2$	0.120	{0.80, 0.85}	69.95%	0.36
$-kf + b$	1.900	0.06	60.28%	0.28
b	–	1.43	65.95%	0.34

WinRate and Reward under various sampling steps in Table 1 and Fig. 4, where it can be seen that our method actually greatly improves the probability of high-quality human-preferred samples, since with a single inference step, our method achieves more than a 70% WinRate, which is similar to the baseline model (*Runway*) with $S = 3$. The higher reward value and lower variance also indicate that our method can consistently generate high-quality samples. Comparisons with other metrics in Table 2 further support the superiority of the proposed method. In addition, we visually compared the inpainted images by different methods in Fig. 5, where it can be seen that our method can generate more meaningful and reasonable content, which is consistent with the style of the prompt region. We refer readers to the *Supplementary Material* for more details about more comparison results.

Finally, we have also carried out a user study to evaluate our superiority. We have randomly selected about 130 groups of results and conducted a user study involving 10 users. The WinRate map, as depicted in Fig. 11, demonstrates that our reward scoring is highly accurate and capable of assessing the results of inpainting under criteria based on human preference. We also present some examples scored by our reward model in Fig. 8 to illustrate the scoring criteria system.

5.2 Ablation Study

Reward Trustiness-Aware Scheme. We validated the advantages of our reward trustiness-aware training scheme. As shown on the left side of Table 3 **b)** and **e)**, we can see that this scheme improves the reward by 1.7% and accelerates the algorithm efficiency of 106.40%. Although a constant amplification, such as **c)** and **d)** in Table 3, can expedite the training process even faster than the proposed method, it compromises model accuracy, as evidenced by the reduced WinRate and Reward. This demonstrates the necessity and efficacy of adaptively managing the sampled trajectory and underscores the superiority of the proposed method.

Training Manner of Reward Model. We investigated two types of strategies to train the reward model $\mathcal{R}(\cdot)$ i.e., classification-driven and regression-driven. Specifically, the former classifies the

Figure 6: Results of image FOV enlargement by our method on two scenes (**a**) and (**b**), where the prompt region (the given image) is delineated by the central area between white dashed lines.
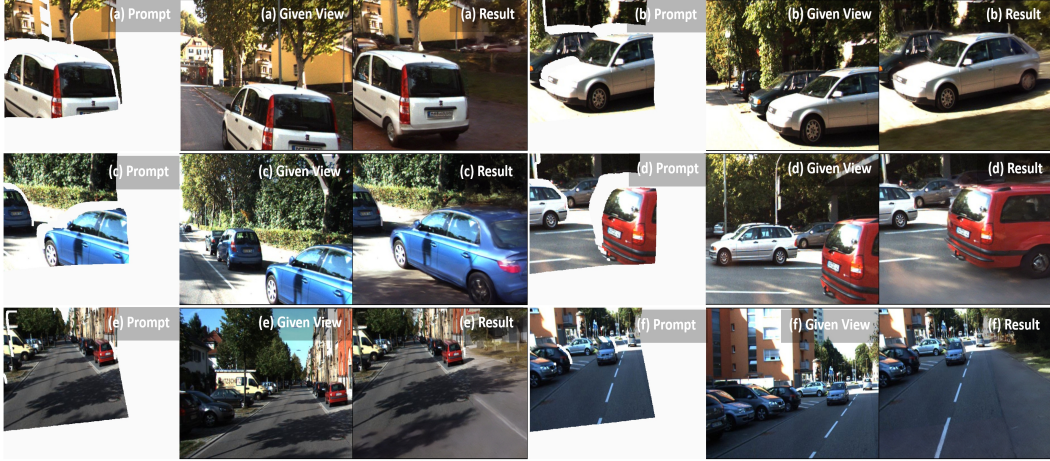

Figure 7: Novel view synthesis on KITTI dataset of 6 scenes from (**a**) to (**f**). For each scene, we give the "Prompt", which is warped from the "Given View", with the white regions referring to holes/missing regions. "Result" is our in-painting result from "Prompt". Note that for the synthesized novel view, there is no ground-truth available.

discrete scores of annotated reward samples, while the latter directly makes a regression on the scores. As listed on the right side of Table 3, it can be seen that the regression-driven training generally outperforms the classification-driven one on reward accuracy. Moreover, the larger variances from the regression-driven training show the strong discriminative ability. Based on accuracy and boundary performance, we finally selected a fixed rate of 0.7 and MLP-256 with the regression-driven training manner as the configuration of our reward model.

Parameterize amplification factor γ. To parameterize the amplification factor in Eq. (10), we investigated various functions to parameterize as shown in Table 4. The experimental results indicate that the exponential function provides the best regularization effect. In contrast, the linear function and static constant do not fully exploit the regularization effect of the reward upper boundary.

5.3 Further analysis

Reward errors distribution. We make statistics of reward estimation errors, and the results are shown in Fig. 9. Although the proportion of very large error samples is not large, the incremental performance of our method lies in a more suitable choice of amplification function, as evidenced by Table 4.

Necessity of Our dataset. Although existing metrics such as BERT Score [77] provide a general measure of quality, we emphasize that our dataset is specifically tailored for the task of image inpainting, where human-labeled scores are both essential and more precise. To substantiate this claim, we conduct a comparison between the BERT score and our dataset's score, as illustrated in Fig. 10. The results reveal a significant divergence between human judgments and BERT's preferences, underscoring the necessity and superior accuracy of the proposed dataset for this specific task.
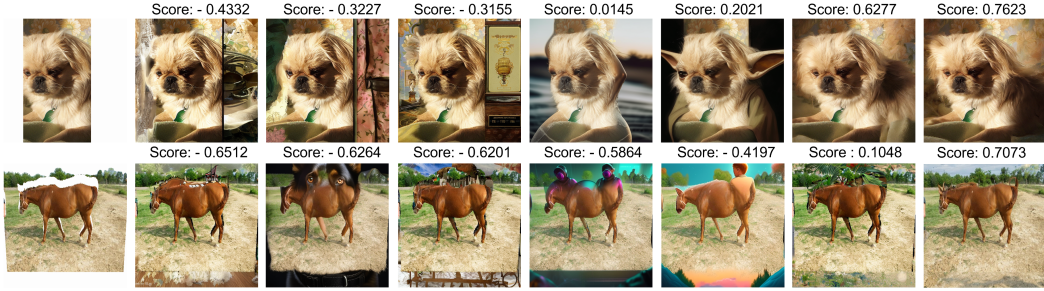
Figure 8: Visualization of various image in-painting results and associated rewards from our model. Our model effectively evaluates in-painting reconstructions based on human preference.
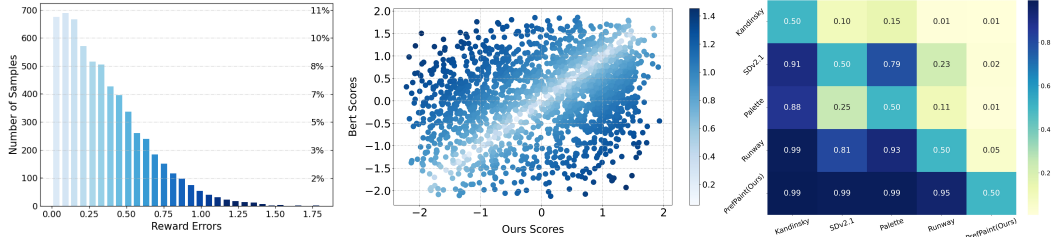

Figure 9: Reward error distributions of the proposed reward model. The distribution of reward error percentages is depicted on the y-axis to the right.

Figure 10: Plot of GT reward by our dataset (x-axis) and the Bert Score of (y-axis) on the validation set, where each point indicates a sample. The darker indacated the larger error.

Figure 11: WinRate heat-map of user study. The winrate at specific locations shows the ratio of superior top methods to those at the bottom.

5.4 Applications of Image FOV Enlargement and Novel View Synthesis

We also applied our approach to two additional tasks: (1) image field of view (FOV) enlargement, where we iteratively extended the boundaries of a typical image strip in the horizontal direction to create a wider FOV image; and (2) novel view synthesis, where we warped a given image using the predicted depth image through the method in [75] to generate a novel viewpoint and subsequently applied diffusion models to fill the holes/missing regions of the Warped view. As depicted in Fig. 6 (a)-a oil painting and (b)-a realistic photography, our method yields natural and visually pleasing results that can be seamlessly integrated with the prompt regions. As illustrated in Figs. 7, our method can generate more reasonable novel views and a visually appealing reconstruction. We refer readers to the *Supplementary Material* for more details about more application visual demonstrations.

6 Conclusion and Discussion

We have presented PrefPaint, an innovative scheme that leverages the principles of the linear bandit theorem to align diffusion models for image inpainting with human preferences iteratively. By integrating human feedback directly into the training loop, we have established a dynamic framework that not only respects the subjective nature of visual aesthetics but also continuously adapts to the evolving preferences of users. We conducted extensive experiments to demonstrate the necessity for alignment in the task of image inpainting and the significant superiority of PrefPaint both quantitatively and qualitatively. We believe that our method, along with the newly constructed dataset, has the potential to bring significant benefits to the development of visually driven AI applications.

PrefPaint aligns with a distribution that corresponds to the preferences of a specific group. However, it is important to recognize that individual preferences for image styles vary. Therefore, after achieving alignment with the general preferences of a group, it is advisable to develop personalized rewards and a corresponding reinforced alignment model to ensure complete alignment with the preferences of each user. Recently, the implementation of a reward-free alignment process, as discussed in [16], has gained popularity. Consequently, exploring reward model-free training methods for alignment in diffusion-based models represents a promising avenue for future research. Furthermore, there is an opportunity to explore the potential applications of our in-painting algorithm for additional 3D reconstruction tasks.

Acknowledgement

References

[1] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7508–7517, 2020.

[2] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7760–7768, 2020.

[3] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12608–12618, 2023.

[4] David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lempitsky. Dinar: Diffusion inpainting of neural textures for one-shot human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7062–7072, 2023.

[5] Efsun Karaca and M Alper Tunga. Interpolation-based image inpainting in color images using high dimensional model representation. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 2425–2429. IEEE, 2016.

[6] Pablo Arias, Vicent Caselles, Gabriele Facciolo, Vanel Lazcano, and Rida Sadek. Nonlocal variational models for inpainting and interpolation. *Mathematical Models and Methods in Applied Sciences*, 22(supp02):1230003, 2012.

[7] Mashail Alsalamah and Saad Amin. Medical image inpainting with rbf interpolation technique. *International Journal of Advanced Computer Science and Applications*, 7(8), 2016.

[8] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14134–14143, 2021.

[9] Jitesh Jain, Yuqian Zhou, Ning Yu, and Humphrey Shi. Keys to better image inpainting: Structure and texture go hand in hand. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 208–217, 2023.

[10] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior guided gan based semantic inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13696–13705, 2020.

[11] Xian Zhang, Xin Wang, Canghong Shi, Zhe Yan, Xiaojie Li, Bin Kong, Siwei Lyu, Bin Zhu, Jiancheng Lv, Youbing Yin, et al. De-gan: Domain embedded gan for high quality face image inpainting. *Pattern Recognition*, 124:108415, 2022.

[12] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3190–3199, 2023.

[13] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[14] Prem Nandhini Satgunam, Russell L Woods, P Matthew Bronstad, and Eli Peli. Factors affecting enhanced video quality preferences. *IEEE transactions on image processing*, 22(12):5146–5157, 2013.

[15] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[18] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[19] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

[20] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[21] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[22] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, et al. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354, 2018.

[23] Eugene A Feinberg and Adam Shwartz. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012.

[24] Martijn Van Otterlo and Marco Wiering. Reinforcement learning and markov decision processes. In *Reinforcement learning: State-of-the-art*, pages 3–42. Springer, 2012.

[25] Ather Gattami, Qinbo Bai, and Vaneet Aggarwal. Reinforcement learning for constrained markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 2656–2664. PMLR, 2021.

[26] Hao-nan Wang, Ning Liu, Yi-yun Zhang, Da-wei Feng, Feng Huang, Dong-sheng Li, and Yi-ming Zhang. Deep reinforcement learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 21(12):1726–1744, 2020.

[27] Hao Yi Ong, Kevin Chavez, and Augustus Hong. Distributed deep q-learning. *arXiv preprint arXiv:1508.04186*, 2015.

[28] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[29] Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Başar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8767–8775, 2021.

[30] Shauharda Khadka and Kagan Tumer. Evolution-guided policy gradient in reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[32] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*, 2016.

[33] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[34] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[35] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[36] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023.

[37] Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008, 2021.

[38] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[39] 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. PMLR, 2015.

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[41] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[42] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[43] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024.

[45] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

[46] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[47] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[48] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4334–4343, 2024.

[49] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024.

[50] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Sven Nebelung, Jakob Nikolas Kather, et al. Diffusion probabilistic models beat gans on medical images. *arXiv preprint arXiv:2212.07501*, 2022.

[51] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[52] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[53] Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. *arXiv preprint arXiv:2401.12244*, 2024.

[54] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[55] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[57] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[59] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[60] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[61] Radu Timofte, Shuhang Gu, Jiqing Wu, Luc Van Gool, Lei Zhang, Ming-Hsuan Yang, Muhammad Haris, et al. Ntire 2018 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[62] Runway. Runway model on huggingface, 2023. URL https://huggingface.co/runwayml/stable-diffusion-inpainting.

[63] Runway. Runway model on huggingface, 2023. URL https://huggingface.co/runwayml/stable-diffusion-v1-5.

[64] StabilityAI. Stabilityai model on huggingface, 2023. URL https://huggingface.co/stabilityai/stable-diffusion-2-1.

[65] StabilityAI. Stabilityai model on huggingface, 2023. URL https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0.

[66] Diffuser. Stabilityai model on huggingface, 2023. URL `https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1`.

[67] Compvis. Compvis github website, 2023. URL `https://github.com/CompVis/latent-diffusion`.

[68] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

[69] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[70] 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. PMLR, 2021.

[71] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[72] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[73] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[74] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[75] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.

[76] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2383–2393, 2023.

[77] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[78] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*, 2023.

[79] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024.

[80] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: high-resolution and prompt-faithful text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2312.14091*, 2023.

[81] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[82] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024.

Appendix

Table of Contents

A Dataset Details

In this section, we visualize our labeling platform and the corresponding labeling scheme.

A.1 Labeling Scheme

The reward score contains the following 3 terms:

Structural Rationality. It measures the correctness of the model understanding of image content. **Score (0-2):** Failure generation, e.g. strange and weird objects; random patterns completed; inappropriate strings; abrupt bar completions. Objects are generated but obviously do not belong to the scene, e.g. There are shelves unique to homes on the side of the highway; there is a marble floor blocking the front of the washing machine. **Score (3-5):** The understanding of the scene is roughly correct, e.g. the inpainted sky next to the sky pattern; the sea behind the beach. **Score (6-7):** The overall structure of the generated scene is reasonable and do not seem to be any major problems. To illustrate the scoring scheme, we show some examples and corresponding reasons in Table A-1.

Table A-1: Illustration of different scoring examples on structural rationality.

Prompt with Reconstruction	Prompt with Reconstruction
Score: 0. **Reason:** there should be no marble in front of the washing machine. Wrong semantics and wrong understanding of the scene.	**Score:** 1. **Reason:** the appearance of racks on the highway is unreasonable, semantically incorrect, and misunderstanding of the scene.
Score: 3. **Reason:** the model understands that this is a volleyball match and try to complete relevant people.	**Score:** 4. **Reason:** Although the quality of the details generated here is not rated high, the model understands that the scene is on the beach, and the overall structure is rated.
Score: 6. **Reason:** although there are some flaws, I understand the content of the highway scene and complete the relevant content.	**Score:** 7. **Reason:** Understand the scenario and correctly add relevant content.

Feeling of Local Texture. It measures whether the texture of the object is correct and consistent with objective facts. **Score (0-2):** Failure generation (e.g. strange and weird objects; random patterns

completed; inappropriate strings; abrupt bar completions). The generated objects do not conform to objective facts (e.g. a dog has two heads; a person without an upper body). The texture of the generated objects is very poor (e.g. two car bodies stuck together; poor quality sky). **Score (3-5):** The generated objects are generally reasonable (e.g. a complete leopard, but with very long legs; a complete corn, but the corn kernels are strange; there are flaws in the complete object;). The generated texture is not obtrusive but will have flaws if you zoom in. **Score (6-7):** Details are complete and reasonable. We also show some examples and corresponding reason in Table A-2.

Table A-2: Illustration of different scoring examples on the feeling of local texture.

Prompt with Reconstruction	Prompt with Reconstruction
Score: 0. **Reason:** it can be seen that the model understands the scene correctly and wants to add a dog, but the quality is not high and does not conform to the objective facts.	**Score:** 1. **Reason:** it wants to generate people but the generation quality is not high.
Score: 4. **Reason:** The completion is generally reasonable.	**Score:** 4. **Reason:** The completion is generally reasonable.
Score: 6. **Reason:** it wants to patch up the leopard but the legs is not consistent with common sense, but it can also be seen that it is a complete leopard.	**Score:** 7. **Reason:** The completion quality is good. good.

Overall Feeling. The overall feeling given by the picture, whether it is reasonable and consistent with the objective facts. We directly show some examples and corresponding reasons in Table A-3.

A.2   Labeling Platform

The visual demonstration of labeling platform is shown in Fig. A-1. In the right-side, we aggregate three reconstruction results with the prompt images. The labeling is annotated in the left-side, where for each image, people need to label 3 scores by aforementioned 3 different criteria, ranging from 0 to 7.

Table A-3: Illustration of different scoring examples on the overall feeling.

Prompt with Reconstruction (from left to right are Recon-1, 2, 3 and the prompt)

Score: 3, 2, 6. **Reason:** The quality of Recon-3 is significantly better than Recon-1, and there are no obvious flaws, so it has the highest score of 6.

Score: 5, 3, 1. **Reason:** Recon-1 is the most reasonable, but the steel basin structure is a bit strange. Recon-2 may want to make up for the melon seeds. The details are quite good but the scenes are relatively rare. Recon-3 is very strange.

Score: 1, 1, 4. **Reason:** The overall structure of Recon-1 is wrong, and the quality of the corn heads in Recon-2 is relatively poor, and it looks like there is glue stuck on it. Recon-3 has a reasonable structure, but the proportion of the head is relatively large, and the grains of the corn head are strange, so the score of only 4 is not very high.

Score: 1, 7, 2. **Reason:** The shape of the car on the right side of Recon-1 is too weird and affects the look and feel. Recon-2 is of very good quality and very natural. The house on the right side of Recon-3 has a poorer perspective.

Score: 0, 4, 1. **Reason:** Recon-1 was randomly generated, Recon-3 tried to generate a traffic light, but failed. Similar branches in Recon-2 don't look so inconsistent.
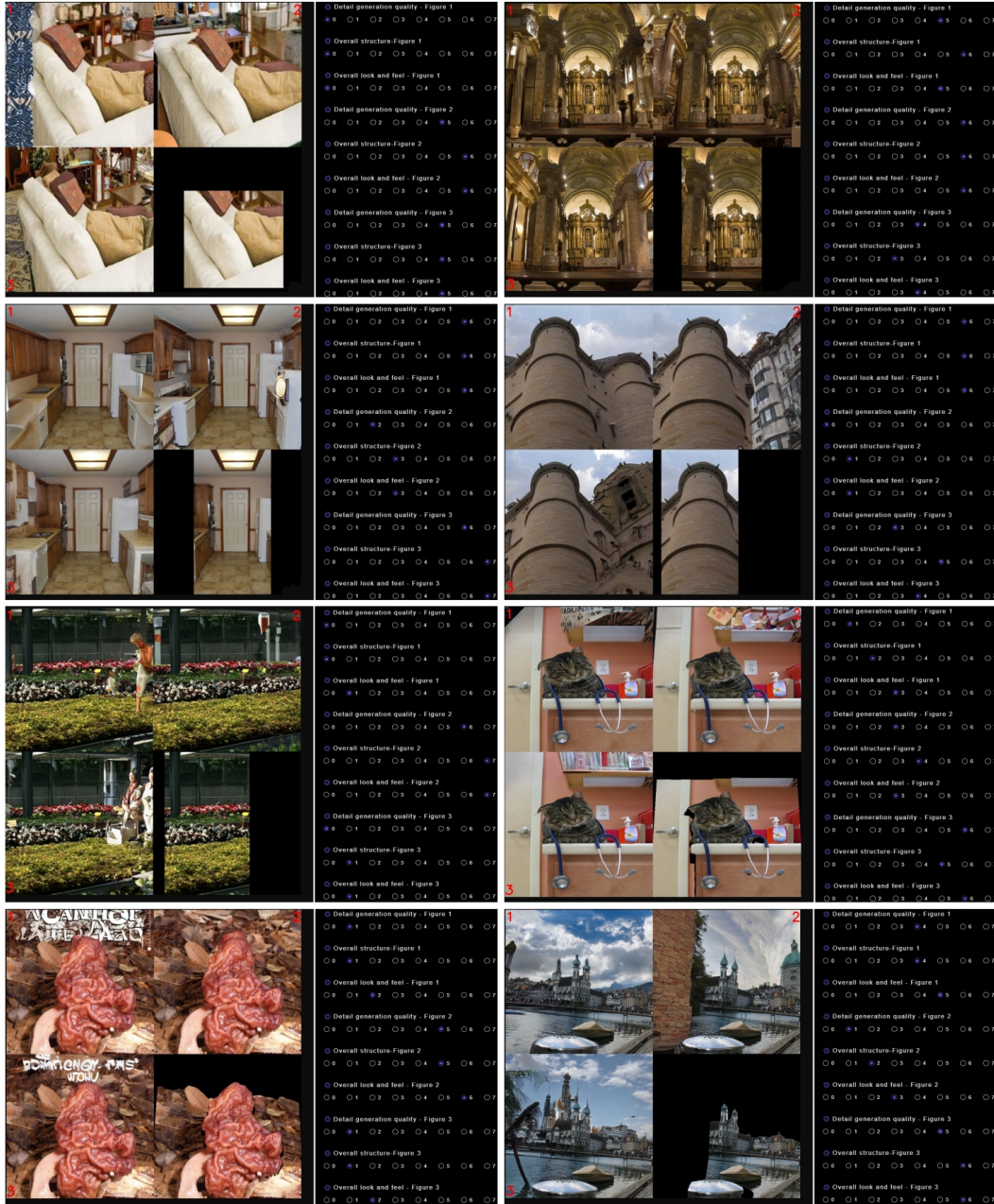
Figure A-1: Labeling platform demonstration. In each group, from top left to bottom right, are reconstruction 1, 2, 3 and the prompt image.

A.3 License

The license of the proposed dataset is **CC BY-NC 4.0**, which allows creators to share their work with others while retaining certain rights but gives a restriction on commercial use.

B   More Application Visual Demonstrations

To further verify the effectiveness of the proposed method, this section further visualizes more scenes for the application of the proposed method (as demonstrated in Sec.5.4).

B.1   Application of Image FOV Enlargement

For the task of Image FOV Enlargement, the results are shown in Fig. B-2. We randomly select some picture from Internet and apply center crop on those pictures. We can see that no matter the style of prompt image, e.g., nature photography, oil painting or Chinese painting, the proposed method can consistently generate meaningful and visually pleasing results, which demonstrates the superority of the proposed method.
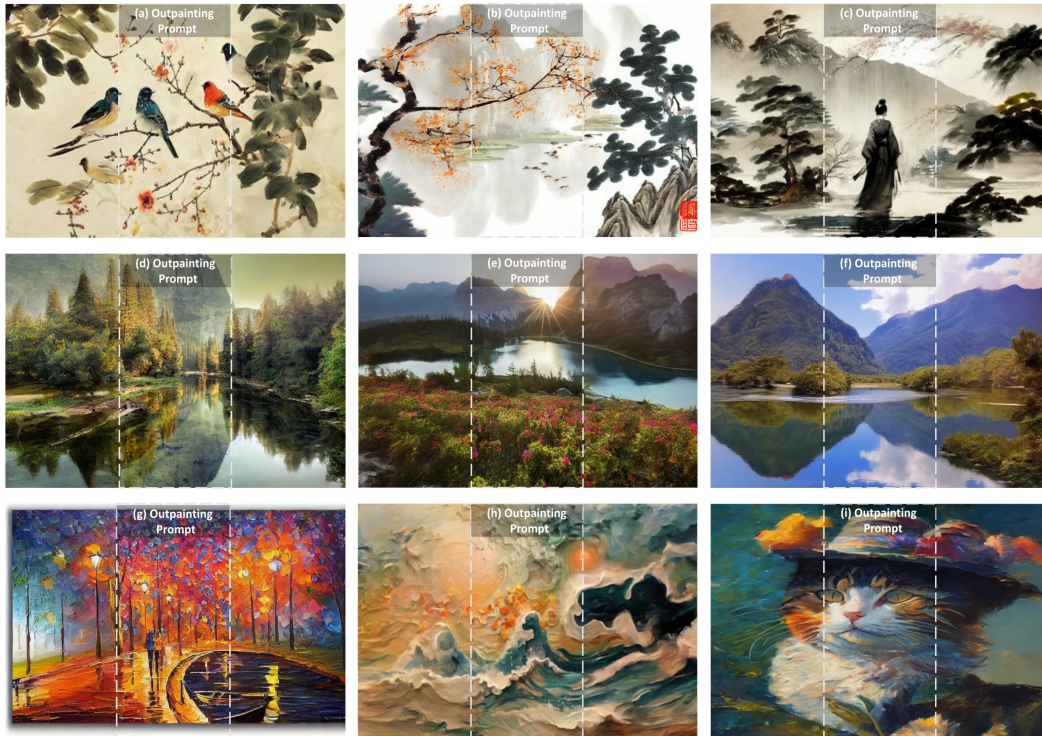

Figure B-2: Application of FOV enlargement, where we visualize 9 scenes from (a) to (i) with corresponding prompt cropped image and enlarged result.

B.2   Application of Novel View Synthesis

We also visualize more novel view synthesis examples on the KITTI and DIV2K datasets in Figs. B-3 and B-4. The warping-induced inpainting hole is much more irregular than the FOV Enlargement, which makes the task more challenging. However, the proposed method successfully filled up the missing region. The resulting high-quality reconstructions verify the superiority of the proposed method.

Figure B-3: Application of novel view synthesis on KITTI dataset, where we visualize 8 scenes from (a) to (h) with corresponding prompt warped image, given view and reconstruction result.
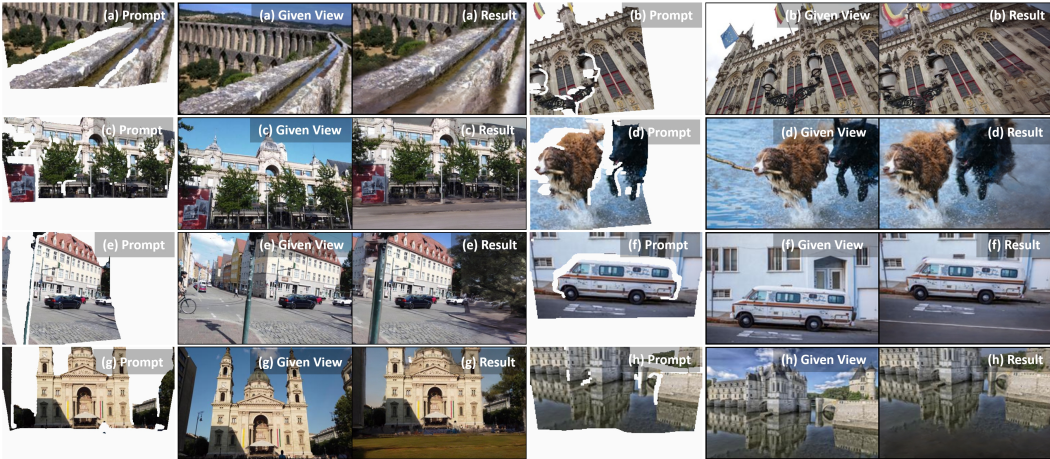

Figure B-4: Application of novel view synthesis on DIV2K dataset, where we visualize 8 scenes from (a) to (h) with corresponding prompt warped image, given view and reconstruction result.

C   Reward Scoring Statistics & Reward Normalization Configuration

C.1   Score Distribution across Different Categories

We count the ratio of prompt with different semantic class in Fig. C-5. Most methods, including SOTA diffusion-based methods, e.g., stable diffusion or Runway, have a large ratio of samples with negative rewards. However, the proposed method achieves the most positive rewards, as shown in the last figure, which validates the necessity and effectiveness of applying such an alignment task to the diffusion in-painting model.
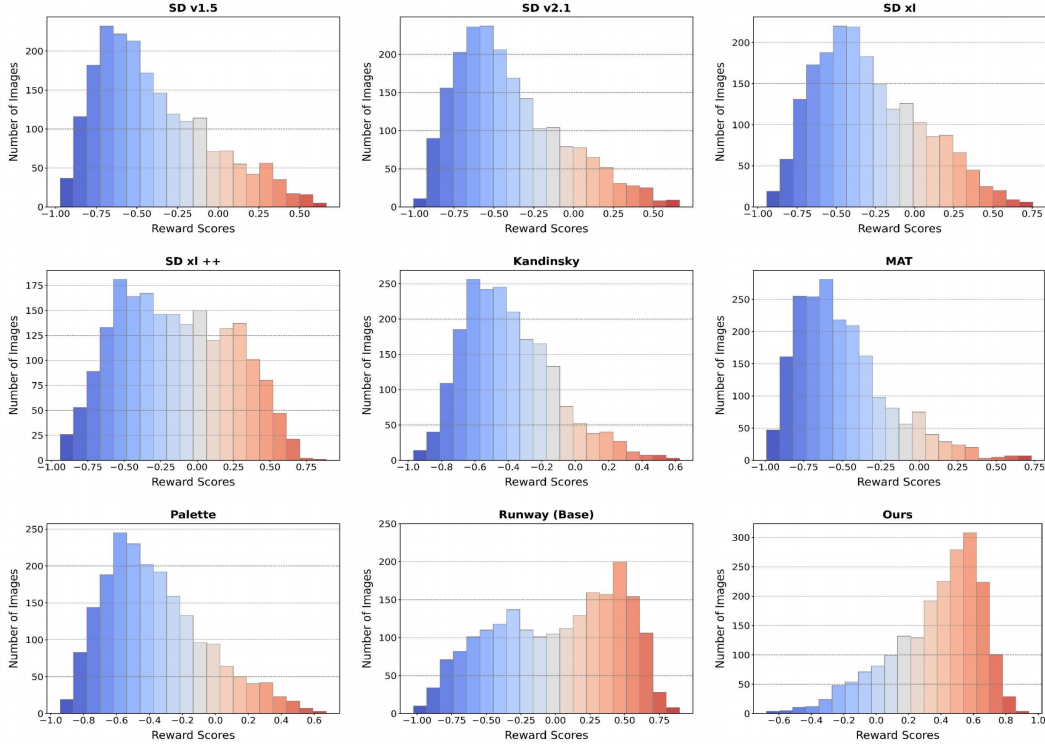
Figure C-5: Detailed score statistics of the proposed dataset.

C.2 Normalization Factors

To facilitate the training of reward model, we normalize the score with $\frac{s-mean}{var}$, where s is raw score data. To remove the data bias for training the reward model, the normalization factors are calculated within different datasets and inpainting patterns. We show the resulting factors in Table C-4. The relatively balanced factors indicate that our labels do not have large biases on different datasets or inpainting pattern.

Dataset	ADE20K		KITTI		ImageNet		Div2K	
	Warping	Outpainting	Warping	Outpainting	Warping	Outpainting	Warping	Outpainting
Mean	3.46	3.12	3.02	2.87	2.85	2.50	2.99	2.34
Variance	2.77	4.42	3.04	2.69	3.03	3.08	3.26	3.60

Table C-4: The normalization factor of score to facilitate for reward model training.

D   More Visualizations

D.1   More Visualizations of Comparisons

In this section, we visually compare the reconstruction results of different methods. As shown in Figs. D-6 and D-7. The comparison of different methods in Figs. D-6 and D-7 indicate that the proposed method could generate high-quality reconstruction compared with other SOTA methods. Notably, the results from our approach show improved clarity and detail retention, which are essential for practical applications.

| Prompts | Compvis | Kandinsky | Palette | Runway | SD xl++ | SD v2.1 | Ours |
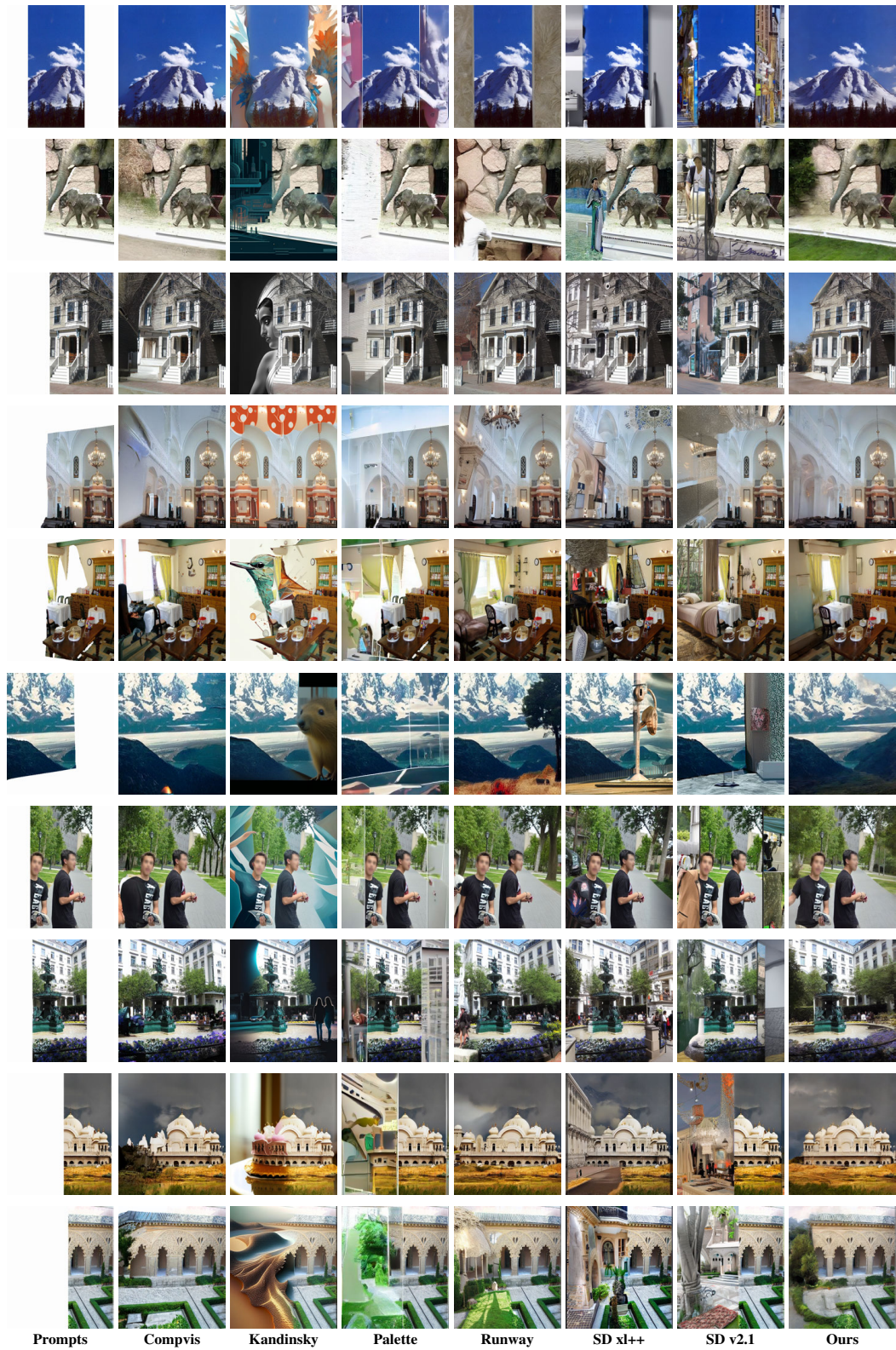
Figure D-6: Qualitative comparison between the proposed method and other SOTA methods.

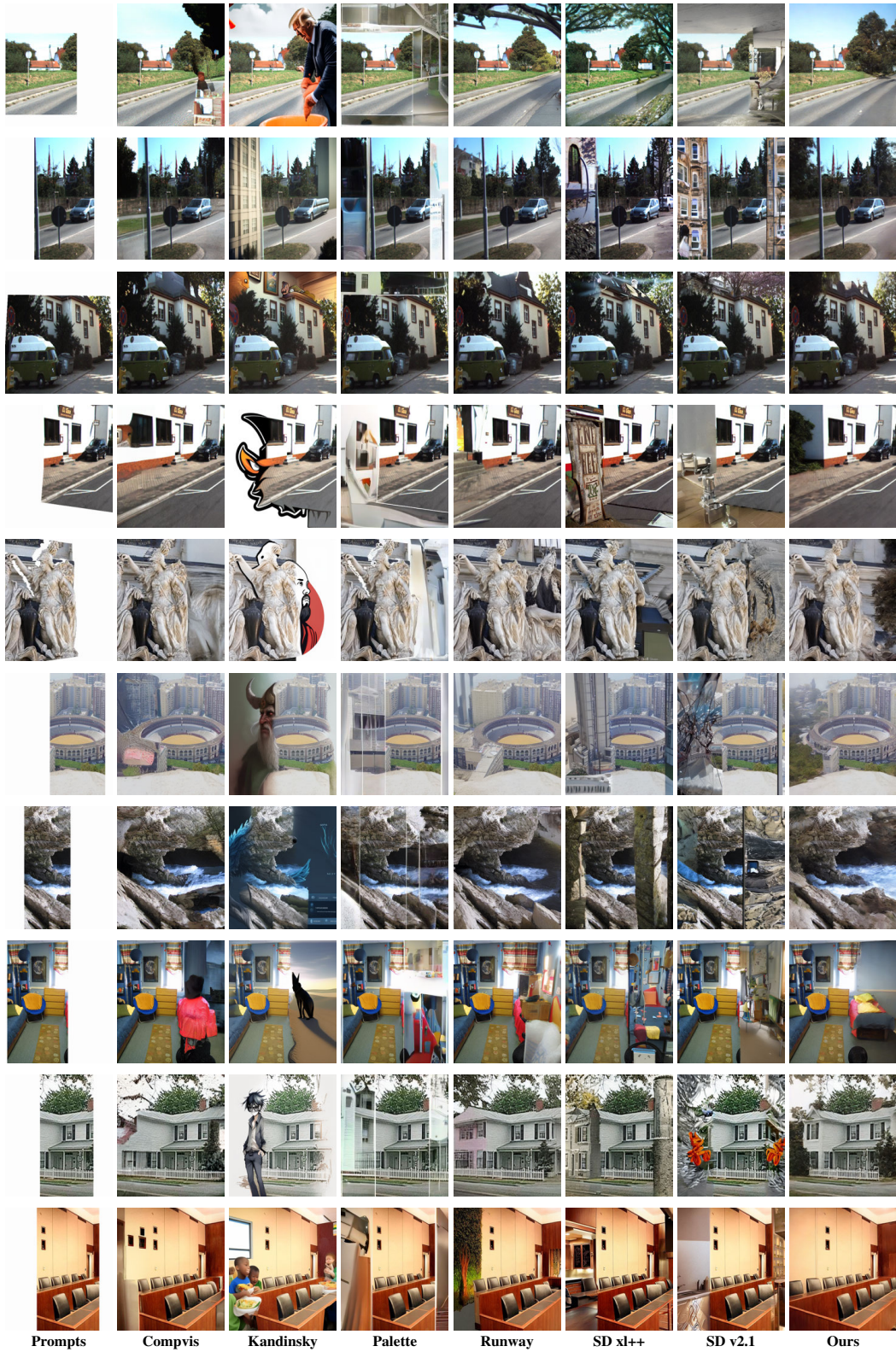| Prompts | Compvis | Kandinsky | Palette | Runway | SD xl++ | SD v2.1 | Ours |

Figure D-7: Qualitative comparison between the proposed method and other SOTA methods.

D.2   More Visualizations of Multiple Sampling

To verify the robustness and consistency of the inpainting model after alignment with our method. We run 5 times with different random seeds on the same prompt to derive different reconstruction results. The results shown in Fig. D-8 demonstrate the stability of the proposed method. We can see the proposed method generate high-quality and visually pleasing results under different conditions.
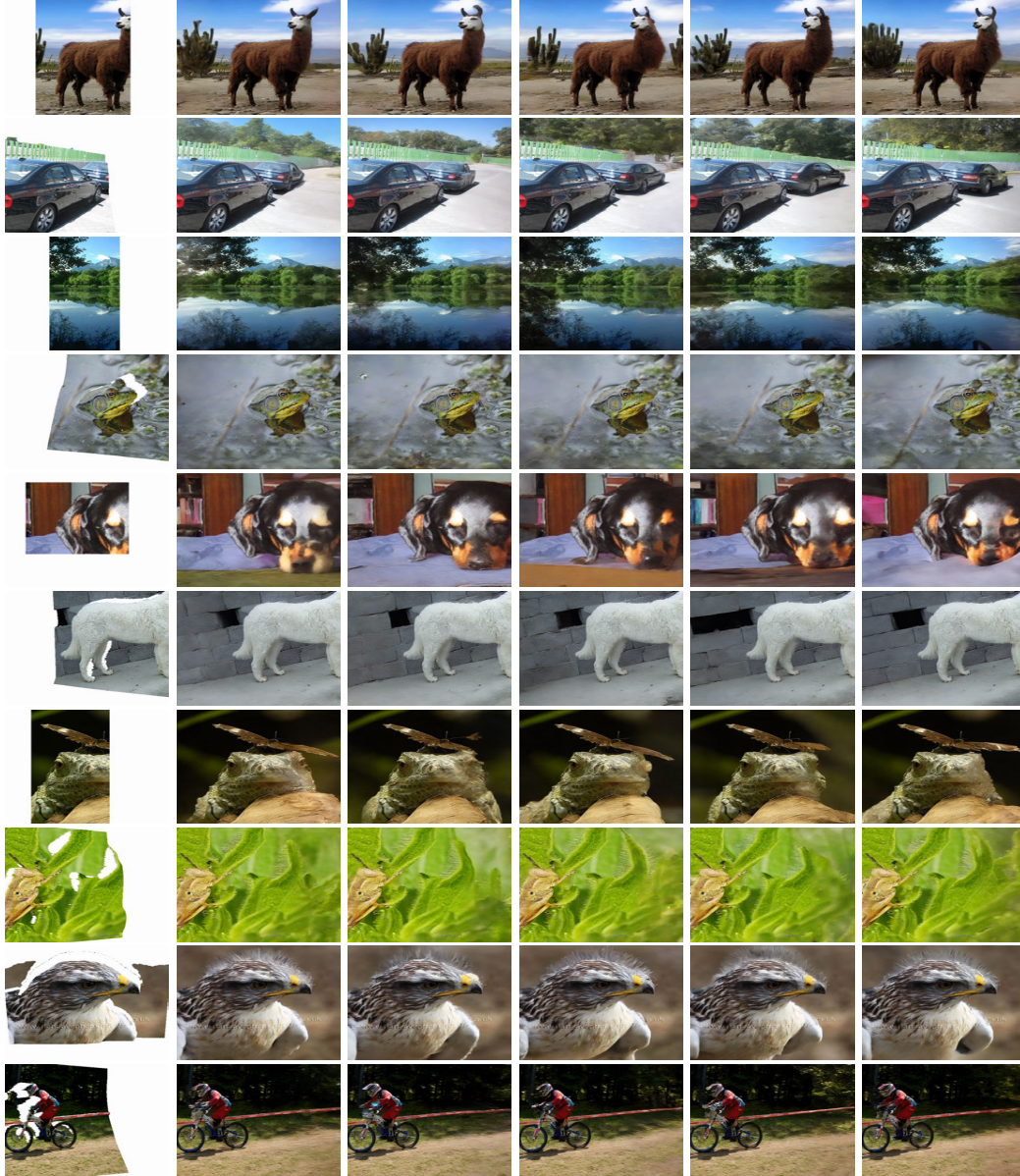
Figure D-8: Qualitative comparison of different sampling times of the proposed method.

E   More experimental results

E.1   Comparisons with Additional inpainting Methods

We have experimentally compared with some inpainting Methods, i.e., PowerPaint [78], Brush-Net [79] and Hdpaint [80] . For all these methods, we assessed performance using their publicly released models. As shown in the Table E-1 and Table E-2, our method significantly outperforms all

the compared methods, particularly in perceptual metrics such as T2I and our Reward score. Besides, BrushNet employs roughly twice the number of parameters as our model, and HdPaint requires 15 times more computational time (experimented on NVIDIA GeForce RTX A6000 GPUs), making it impractical for real-world applications.

Table E-1: Comparison across metrics: higher values are better for all metrics except "# Param".

Metrics	T2I	BLIP	CLIP	CA(Incep.)	Reward	# Param(M)	Infer. Time(s)
PowerPaint(v-1)	-4.44	0.46	0.21	0.42	-0.057	819.72	**5**
PowerPaint(v-BrushNet)	-3.84	0.46	0.20	0.42	-0.036	1409.88	16
BrushNet	1.26	0.46	0.22	0.43	0.137	1409.86	16
HdPaint	-4.57	0.47	0.21	0.44	-0.059	**451.47**	60
PrefPaint(Ours)	**11.60**	**0.49**	**0.23**	**0.45**	**0.374**	819.72	**5**

Table E-2: Quantitative comparisons of different methods. "S" is the number of sampling times. For the calculation of WinRate, we first derive the best sample of the compared method among S sampling times.

WinRate↑ (v.s. Runway)	S=1	S=2	S=3
PowerPaint(v-1)	27.06%	39.92%	47.38%
PowerPaint(v-BrushNet)	29.86%	43.12%	52.01%
BrushNet	49.49%	62.83%	69.22%
HdPaint(ds8-inp)	33.37%	43.41%	49.03%
PrefPaint(Ours)	**71.27%**	**85.88%**	**93.50%**

E.2 Comparisons with other reinforcement learning methods.

we have experimentally compared our method with some other reinforcement learning methods in Table E-3. Specifically, we simply summarize the implementation of each method. Human Preference Score [13] learns a negative prompt to map the diffusion process to low-quality samples. Then, in the inference process, the negative sample is utilized in the classifier-free guidance (CFG) to push the generation trajectory away from low-quality samples. ImageReward [54] trains a reward model and then applies the reward model as a loss metric to end-to-end optimize the diffusion model accompanied by a reconstruction loss. We also conduct the ablation study on reward training strategy in Table 3 in our paper. Our method employs a regression-driven training strategy, while ImageReward a classification-drive strategy. DPOK [81] simultaneously optimizes the whole trajectory of a reverse diffusion process and utilizes the KL divergence to panel the regularization, avoiding a large distribution shift. D3PO [82] adopts the RL strategy from direct performance optimization (DPOK), directly optimizes the model on the reward labeling data to minimize the probability of low-quality samples and increase the probability of high-quality samples.

E.3 Compared with SD xl ++

Currently, their is also a kind of diffusion-inpainting network like SD xl ++ [66], which requires a complete image to redraw the mask region and can generate a visually pleasing result. However, their performance would be dramatically reduced without the complete image. To validate the characteristics of SD xl ++, we carry out experiments to examine its performance.

Table E-3: Quantity evaluations of different reinforcement learning methods.

Methods	WinRate	T2I	Reward	CLIP	BLIP	CA
Human Preference Score	58.03%	-16.67	0.26	0.20	0.47	0.40
ImageReward	65.10%	13.12	0.29	0.22	0.48	0.44
DPOK (KL weight=0.1)	64.59%	11.43	0.32	0.21	0.48	0.43
DPOK (KL weight=1.0)	62.67%	9.36	0.30	0.21	0.48	0.43
D3PO	59.74%	-19.20	0.26	0.21	0.46	0.41
PrefPaint	**71.27%**	**21.53**	**0.37**	**0.23**	**0.49**	**0.45**

Table E-4: Alignment evaluation across various generative methods, where ⋆ represents the small model (non SD-based); "*S*" refers to the number of sampling times. For comparing the WinRate, the best inpainting results are selected through multiple sampling and compared with the *Runway* Inpainting model with only 1 sampling. "↑ (resp. ↓)" means the larger (resp. smaller), the better. The WinRate is calculated against Runway [62].

Prompt Methods	Outpainting Prompts		Warping Prompts	
Metrics	WinRate (%) ↑	Rewards Mean↑	WinRate (%) ↑	Rewards Mean↑
SD v1.5 w/ xl ++	21.15	-0.13	18.66	-0.18
SD v2.1 w/ xl ++	26.07	-0.07	22.26	-0.12
Kandinsky w/ xl ++	17.09	-0.29	12.04	-0.37
MAT ⋆ w/ xl ++	19.12	-0.34	8.58	-0.56
Palette ⋆ w/ xl ++	27.99	-0.07	25.82	-0.09
Runway w/ xl ++	56.94	+0.03	45.71	+0.02
Ours	70.16	+0.38	72.38	+0.36

F   Details & Reward

F.1   Training Curve

We present training curves of various methods for comparison. The curves for '1.4BaseLine' and '1.4Boundary' (Ours) demonstrate our modification's acceleration, reaching 0.35 rewards first. Our approach converges with the fewest training iterations.
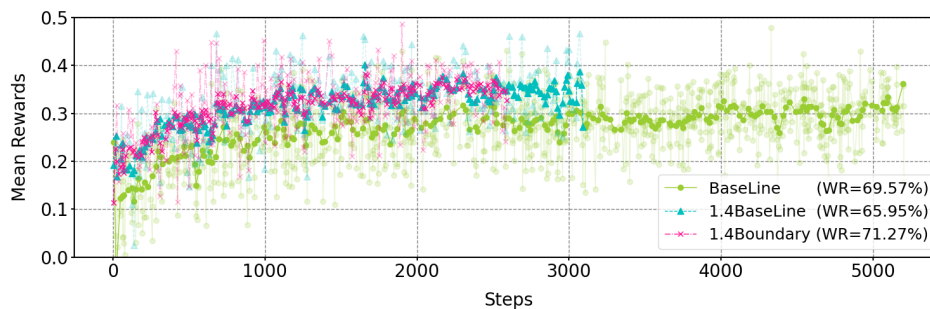

Figure F-1: Training curves of different experimental setups.

F.2 Details between Eq. 4-5

According to the [55], we derive a similar process for the linear reward regression on feature space, where $\mathbf{Z}$ denotes the latent of an image.

$$\hat{\psi} = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Y} \tag{13}$$

$$= (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^T(\mathbf{Z}\psi_* + \zeta) \tag{14}$$

$$= (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}[(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I} - \lambda\mathbf{I})\psi_* + \mathbf{Z}^T\zeta] \tag{15}$$

$$= \psi_* - \lambda(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\psi_* + (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\zeta \tag{16}$$

$$\hat{\psi} - \psi_* = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^T\zeta - \lambda(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\psi_*, \tag{17}$$

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 authors correctly summarized the contribution in the abstract and introduction sections.

   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 authors have discussed the limitations in the section of the conclusion and discussion.

   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: [NA]

Justification: No theoretical result.

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 detailedly illustrated our implementation details.

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: We have given the example data and code for supplementary material.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have given the implementation details.

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: [NA]

Justification: N.A.

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: We have provided the computational resources.

   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 follows the instructions for preparing the paper.

   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: [NA]

    Justification: N.A.

    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: N.A.

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 utilized papers.

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: We have discussed the new assets.

    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: [Yes]

    Justification: We have given the instructions in the appendix.

    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: N.A.

    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.