DreamAlign: Dynamic Text-to-3D Optimization with Human Preference Alignment

Gaofeng Liu[1], Zhiyuan Ma[2*] , Tao Fang[1*]

[1]Department of Automation, Shanghai Jiao Tong University, Shanghai 201100, China.
[2]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
{diehualong, tfang}@sjtu.edu.cn, mzyth@tsinghua.edu.cn

Abstract

Recent years have witnessed the remarkable success of Text-to-3D generation, particularly with the rise of mainstream conditional diffusion models (DMs). Though achieving substantial progress, existing methods still face a knotty ***"human preference"*** dilemma, that is the 3D contents generated by the models often deviate greatly from the desired effects (e.g., perspective, aesthetics, shading, appearance, etc.) due to the lack of attention to human preferences. To mitigate the limitation of data deficiency and enable human preference learning, we first elaborately curate the HP3D, a text-to-3D dataset with expert preference annotations which is initally captioned by the multimodal large model LLava and then refined by human expert. Based on such a brand-new HP3D, we further propose DreamAlign, a reward-free method that does not require designing any complex reward models whereas only by introducing a light-weight lora adapter and then designing a novel direct 3D preference optimization (D-3DPO) algorithm for training. Moreover, in the stage of text-to-3D we design an additional Preference Contrastive Feedback training for score distillation sampling, which enables the generated 3D objects to align the human preferences (e.g., aesthetics, material, etc.). Extensive experiments demonstrate that DreamAlign consistently achieves state-of-the-art performance on generative effects and human preference alignment across various benchmark evaluations.

Introduction

3D asset creation is widely applied in numerous fields such as gaming, animation, simulated environments, and art production. However, high-quality 3D creation requires excellent artistic creativity and specialized modeling skills, making it difficult to meet multitude demands. Some recent works (Gupta et al. 2023; Chen et al. 2023; Jun and Nichol 2023; Liu et al. 2023; Nichol et al. 2022b; Poole et al. 2022; Qian et al. 2023; Shen et al. 2021; Wang et al. 2023, 2021) have leveraged the success of generative modeling (Ho, Jain, and Abbeel 2020) and massive 3D datasets (Luo et al. 2024; Deitke et al. 2023; Collins et al. 2022) to achieve efficient 3D content generation. Different from others, optimization-based 2D lifting method (Jun and Nichol 2023; Li et al. 2023; Nichol et al. 2022a; Poole et al. 2022; Shi et al. 2023; Wang et al. 2024) have received much attention due to their ability to create 3D assets from textual prompts adpoting text-2D prior to optimize differentiable 3D representations. While this framework is capable of generating high-fidelity multi-view consistent 3D content, the significant limitation is that it ignores the consistency of 3D content properties with human aesthetics. As shown in Figure 2, asymmetric lighting causes shadows behind the teddy bear, perspective offsets make the flattened diary difficult to view, and blurring between appearance and caption makes the black cats incompatible with human aesthetics. Therefore, we argue that the consistency evaluations encompasses not only multi-view and 3D-text alignment, but also the alignment efficacy

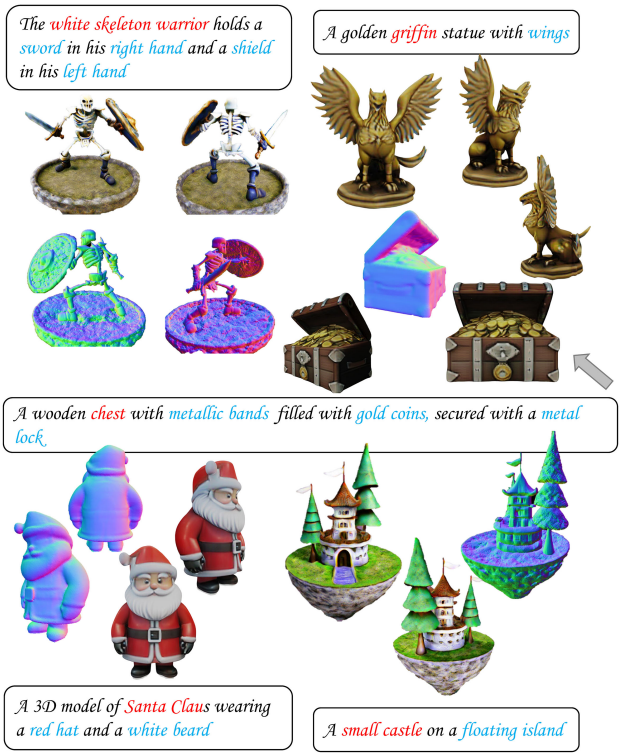

Figure 1: Text-to-3D generation of our DreamAlign. (Text marked in red represents 3D object, while text marked in blue represents adjectives.)

*Corresponding authors.

between the attributes of the 3D content and human preferences, such as style, shading, geometry, and appearance.

Recently, Human Feedback Reinforcement Learning (RLHF) has been successfully applied to various Large Language Models (LLMs) (Ouyang et al. 2022; Achiam et al. 2023) and text-to-image models (Black et al. 2023; Wada et al. 2024; Zhu et al. 2023; Ma et al. 2024c). However, reward-based RLHF approaches require massive cost to create score-annotated datasets for the training of the reward model, but also suffer from the risks of model collapse due to only supporting training on relatively narrow types of feedback datasets. Inspired by the success of DPO in diffusion models (Wallace et al. 2024), we introduce **DreamAlign**, a fresh text-to-3D preference tuning technique enables the model efficiency and low-cost training without the need for additional training of complex reward models. Firstly, we created a 3D dataset HP3D with human preference annotation. Following Cap3D (Luo et al. 2024), we selected 367 high-quality samples with prompts consistent with 3D content. In order to supplement comparative data, we employed advanced multi-modal model LLava (Liu et al. 2024a,b) to obtain captions with different attributes of the same semantics, and then utilize a multi-view diffusion model to generate 3D contents with easy to distinguish preferences. After expert sorting and filtering, a refined dataset of 1300 3D entries with human preferences annotation was obtained. Based on the 3DHF dataset, we first fine-tuned the multi-view diffusion model using the D-3DPO algorithm. More importantly, by employing a 2D lifting approach, the multi-view diffusion model can effortlessly integrate the human preferences to the generation of 3D content. To further exploit the capabilities of DreamAlign, we incorporated contrastive learning into 3D generation process, termed as Preference for Contrastive Feedback Learning. We elaborately designed a Preference Loss and integrated it into the SDS pipeline for 3D generation.

In Figure 1, we showcase several impressive cases from DreamAlign, which exhibits our method not only generates 3D content that aligns with the textual descriptions, but also consistently matches human preferences across various attributes. Additionally, we conducted both quantitative and qualitative comparisons of our text-to-image and text-to-3D instance generation methods against other techniques. Furthermore, a user study was conducted to analyze preferences, which indicates that DreamAlign is capable of producing high-quality 3D content that closely aligns with human preferences in various aspects.

Related Work

Text-to-image generation. The successful application of diffusion models (Ho, Jain, and Abbeel 2020; Meng et al. 2023; Ma et al. 2022, 2024b; Rombach et al. 2022; Sohl-Dickstein et al. 2015; Ma et al. 2024a) in text-to-2D image generation has been widely recognized and popular. Typically, these models consist of forward process, which incrementally introduces noise into the original data, and reverse process that samples and denoises from random Gaussian noise. Recent advances in research have demonstrated the formidable ability of diffusion models in generative tasks,

Figure 2: Text-to-3D generation results for MVDream (Shi et al. 2023). It remains a challenge to align 3D attributes (e.g., perspective, aesthetics, shading, appearance, etc.) with human preferences. Asymmetric lighting causes shadows behind the teddy bear, perspective offsets make the flattened diary difficult to view, and blurring between appearance and caption makes the black cats incompatible with human aesthetics.

allowing inpainting (Lugmayr et al. 2023; Xie et al. 2023) and editing (Brooks, Holynski, and Efros 2023; Kawar et al. 2023; Ma, Jia, and Zhou 2024) of diverse and high-fidelity images that faithfully adhere to textual prompts. The impressive performance in 2D image generation also suggests a substantial potential for diffusion models in 3D content creation (Poole et al. 2022; Wang et al. 2023). Compared to previous models, our DreamAlign has been trained on the 3DHF dataset with preference ranking, making it better suited for generating 3D content that aligns with human preferences.

2D-lifting Text-to-3D generation. Recently, the powerful text-to-image generation models (Rombach et al. 2022) have been extensively utilized for converting text into 3D content (Lin et al. 2023; Poole et al. 2022; Shi et al. 2023; Wang et al. 2024, 2023), rapidly becoming a hotbed of research. Compared to 2D datasets, the diversity of 3D datasets is relatively limited (Chang et al. 2015). To address this, DreamFusion (Poole et al. 2022) has introduced a new methodology, called SDS (Score Distillation Sampling), which extracts scores from priors of 2D diffusion models to guide the optimization of 3D representations such as NeRF (Mildenhall et al. 2021), showing promising results. Subsequently, researchers have made numerous efforts to improve the quality and 3D consistency of the generated content (Huang et al. 2023; Wang et al. 2024) using distillation-based approaches (Lin et al. 2023; Poole et al. 2022; Zhu, Zhuang, and Koyejo 2023; Zhuang et al. 2023). For instance, some

studies have fine-tuned 2D diffusion models with extensive 3D data to produce images consistent across multiple viewpoints (Shi et al. 2023; Long et al. 2024). Although these methods have produced impressive 3D content, they still fall short in terms of thematic diversity, adherence to the provided texts, and consistency with human preferences or intentions (Ma et al. 2021, 2023). To overcome these challenges, we have employed the PreferenceLoss framework to bridge the gap between 3D and 2D generation, ensuring that the 3D content generated from text aligns with human preferences.

Method

In this study, we explore how to align the attributes of text-to-3D with human preferences. Traditional Reinforcement Learning from Human Feedback (RLHF) methods require the construction of a 3D dataset with expert ratings during the fine-tuning stage and then training of a reward model to fit human preferences. This approach is not only costly but also prone to reward hacking (Skalse et al. 2022). We propose DreamAlign, a novel framework that directly optimizes diffusion models on 3D data ranked by human preferences. This method only requires the introduction of a lightweight LoRA adapter, followed by the application of a 3D preference optimization algorithm to train the existing diffusion model without the need for additional reward model training. The fine-tuned text-to-image model can generate text-to-3D content through optimization-based 2D lifting method. To differentiate between the text-to-image and text-to-3D content of DreamAlign, we refer to them as DreamAlign-2D and DreamAlign-3D, respectively.

In Section , we introduce the process of creating HP3D dataset with human preference labels and the DreamAlign-2D framework. After training DreamAlign-2D, we present Preference for Contrastive Feedback Learning in Section to further optimize text-to-3D utilizing the 2D lifting method.

DreamAlign for Text-to-image Generation

Data preparation. We carefully construct the HP3D dataset with comparative expert labels for training. The HP3D dataset is based on the Cap3D dataset, which only includes 3D objects and their text captions, lacking comparative data and human preference labels. Therefore, the Cap3D is not suitable for DreamAlign training to learn human preference. To supplement the comparative data, we utilize advanced multi-modal models to generate additional text captions and multi-view images of 3D objects as shown on the left of Figure 3.

Specifically, the construction process can be divided into four steps: filtering, rendering, generating descriptions and comparative data, and expert ranking. (1) To ensure diversity in the selected 3D objects, we employed the K-center algorithm, combined with the CLIP features of the text, and selected 10,000 text-3D object pairs from Cap3D. Due to the significant noise in the Cap3D dataset, we filtered these 10,000 pairs and chose 376 high-quality objects that matched the captions. (2) In the rendering step, the 3D objects obtained from the screening phase are densely rendered into 32 multi-view images with a resolution of 512x512. During the 360-degree orbit of the camera, the azimuth starts from the front view (90°) and the elevation is randomly selected from $[0, 30]$, while saving the external parameters of the camera corresponding to each viewpoint. (3) To accurately label the preference ordering of different 3D content corresponding to the same caption, we employed a multi-modal large model LLava to generate $2 \sim 3$ captions with different attributes of the same semantics for each original caption. Subsequently, these captions are used to generate preference-distinguishable multi-view images through a multi-view diffusion model. (4) During the final ranking stage, we hired ten researchers in the field of computer vision to score 3D content. They were required to consider three aspects such as text-3D object consistency, multi-view consistency, and aesthetics, awarding scores from 0 to 7. Subsequently, the multi-view images were ranked on the basis of these scores. Ultimately, we obtained the HP3D dataset labeled with preferences. For more detailed information, please refer to the supplementary materials.

Diffusion Models. The diffusion model consists of two main components: a forward noising process and a reverse denoising process. Given a noise scheduling function α_t and σ_t, during the forward process, samples are taken from the data distribution $x_0 \sim q(x)$ and random Gaussian noise is incrementally added over T steps. Each noise addition is considered a Markov decision process:

$$\begin{aligned} q(x_t|x_{t-1}) &= \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 \boldsymbol{I}) \\ q(x_{1:T}|x_0) &= \prod_{t=1}^{T} q(x_t|x_{t-1}). \end{aligned} \tag{1}$$

The denoising process of the diffusion model with parameter ϕ can be described through a discrete-time reverse process with a Markov structure $p_\phi(\boldsymbol{x}_{0:T}) = \prod_{t=1}^{T} p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ where

$$p_\phi(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\phi(x_t), \sigma^2_{t|t-1}\frac{\sigma^2_{t-1}}{\sigma^2_t}\boldsymbol{I}). \tag{2}$$

Optimization based D-3DPO. As previously mentioned, we have meticulously created a dataset denoted as $\mathcal{D} = \{(x_w, x_l, D, C)\}_1^N$, where N represents the number of 3D assets in the dataset, D is the description of the 3D asset x, and C is the camera extrinsic matrix. We randomly select four orthogonal camera views from the training set to render images as a single training sample, which can be formalized as $\left\{I_w^j, I_l^j, D, c^j\right\}_1^4$. Here, I_w and I_l represent the "winning" and "losing" images as judged by experts, respectively, and c is the camera position corresponding to the image. With these image-text-camera-preference label pairs, we can leverage the preference labels to train the model using the D-3DPO algorithm. This enables the model to implicitly learn and generate images with attributes that align with user aesthetics.

As shown on the right side of Figure 3, we did not train the U-Net model directly within the stable diffusion framework. Instead, we integrated LoRA (Low-Rank Adaptation)

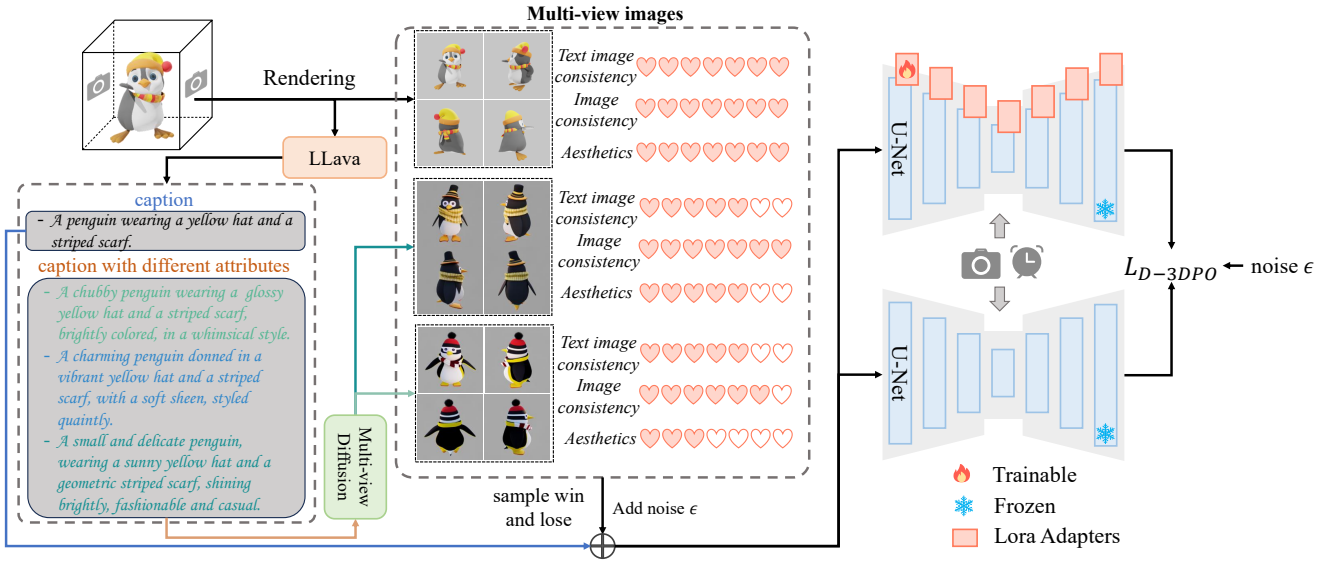

Figure 3: Overall framework of DreamAlign-2D. Left: The data preparation process and data flow of DreamView-2D. Original 3D objects from the Cap3D dataset are first rendered as multi-view images, and then annotated with appropriate captions using the LLava technique. These captions are semantically consistent, yet differ in detailed descriptions, allowing the multiview images generated by MVDream to exhibit distinct, recognizable human preference attributes. These data are subsequently ranked by experts. Right: Based on the HP3D dataset, the D-3DPO algorithm is applied to fine-tune U-Net modules to learn capabilities that mimic human preferences. LoRA adapters (Hu et al. 2021) are added to the attention layers of U-Net, reducing training requirements while preserving the model's original functionalities.

adapters into the attention layers of the U-Net. This approach preserves the original diffusion model's capability in generating 2D images. This not only maintains the high performance of the original diffusion model, but also significantly improves training efficiency, allowing for faster convergence with fewer resources.

During training, the diffusion model receives input as $\{I_w, I_l, c, D, t\}$, where t denotes the sampling step. Unlike RLHF-based methods, which use reinforcement learning from human feedback to inject content generation capabilities, our approach employs the D-3DPO algorithm. The D-3DPO algorithm enables direct training by learning the latent reward model from the existing dataset, avoiding the need for additional reward model training and simplifying the process. The optimization objective of the D-3DPO applied to a diffusion model with parameter ϕ is:

$$
\begin{aligned}
&L_{D-3DPO}(\phi) = \\
&-\mathbb{E}_{(\boldsymbol{I}_0^w, \boldsymbol{I}_0^l)\sim\mathcal{D}, t\sim\mathcal{U}(0,T), \boldsymbol{I}_t^w\sim q(\boldsymbol{I}_t^w|\boldsymbol{I}_0^w), \boldsymbol{I}_t^l\sim q(\boldsymbol{I}_t^l|\boldsymbol{I}_0^l)} \\
&\quad \log\sigma\left(-\beta T\omega(\lambda_t)\right) \\
&\quad \left(\|\epsilon^w - \epsilon_\phi(I_t^w, t)\|_2^2 - \|\epsilon^w - \epsilon_{\text{ref}}(I_t^w, t)\|_2^2\right. \\
&\quad \left. - \left(\|\epsilon^l - \epsilon_\phi(I_t^l, t)\|_2^2 - \|\epsilon^l - \epsilon_{\text{ref}}(I_t^l, t)\|_2^2\right)\right)
\end{aligned} \tag{3}
$$

We omitted D and c for compactness. The ϵ_θ and ϵ represent the noise predicted by the DreamAlign and the randomly generated noise, respectively. The ϵ_{ref} is used as a reference model, which has the structure of the original diffusion model. $\boldsymbol{x}_t^* = \alpha_t \boldsymbol{x}_0^* + \sigma_t \boldsymbol{\epsilon}^*$, $\boldsymbol{\epsilon}^* \sim \mathcal{N}(0, I)$ is a draw from $q(x_t^*|x_0^*)$. $\omega(\lambda_t)$ is a pre-specified weighting function, $\lambda_t = \alpha_t^2/\sigma_t^2$ is the signal-to-nosie ratio. The hyperparameter β controls regularization. This loss function incentivizes ϵ_θ to focus more on enhancing the denoising process for I_t^w compared to I_t^l. Through the above optimization objectives, the model is expected to be biased towards generating images of the "wining" data distribution.

DreamAlign for Text-to-3D Generation

Based on the Score Distillation Sampling technique, we develop the DreamAlign-3D, which utilizes the DreamAlign-2D model to extract and apply prior knowledge to guide the 3D generation process. Additionally, we introduce the Preference Contrastive Learning Feedback, which enhances the aesthetic appeal of the generated 3D content by encouraging it to closely align with the conditional distribution of DreamAlign-2D.

Score Distillation Sampling. Score Distillation Sampling (SDS) is a method that extracts insights from a pre-trained 2D diffusion model to enhance 3D representations, greatly advancing the development of 3D generation technology (Poole et al. 2022; Wang et al. 2024; Zhu, Zhuang, and Koyejo 2023; Lin et al. 2023). Utilizing a pretrained diffusion model ϕ as a score function $\hat{\epsilon}_\phi(x_t; y, t)$ can effectively predict sampling noise ϵ related to a given noisy image x_t, conditional embeddings y, and specific noise levels t. For DreamAlign, conditional embeddings y include text D and camera position c. The 3D asset is represented as a differentiable image parameterization, where a differentiable generator renders 2D images $x = \text{g}(\theta, \text{c})$ based on parameters θ and camera position c. The gradient obtained through Score Distillation Sampling provides direction guidance for

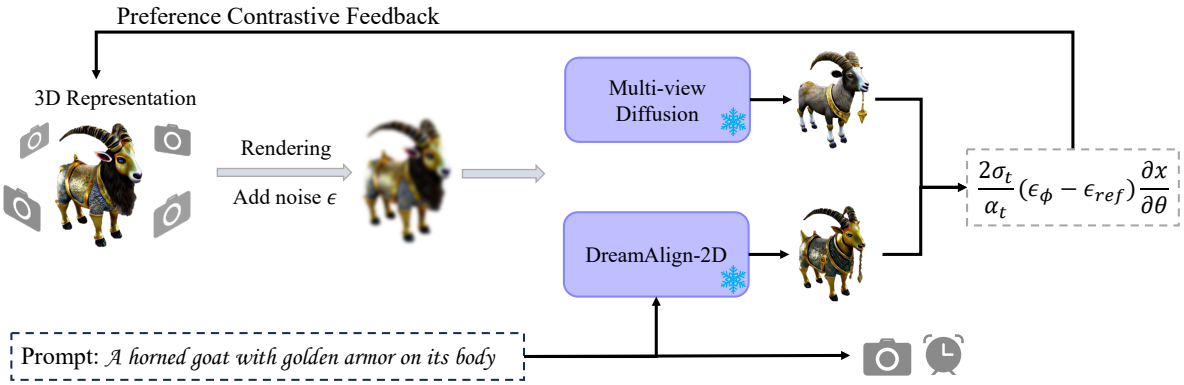

Figure 4: The overall framework of DreamAlign-3D is based on DreamAlign-2D, adopting an enhancement approach from 2D to 3D for generating three-dimensional content, thereby inheriting prior user preferences. To further minimize the discrepancy between generated 3D content, we have designed a Preference Contrastive Feedback Learning.

the given variable rendering mapping function.

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, x) = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(x_t; y, t) - \epsilon)\frac{\partial \text{g}(\theta, \text{c})}{\partial \theta}\right] \tag{4}$$

where $\omega(t)$ is a weighting function.

Preference Contrastive Learning Feedback The primary challenge in aligning 3D generative models with user preferences is the difficulty in accurately estimating the human preference distribution conditioned on user prompts. Although DreamAlign-2D has alleviated this issue to some extent after D-3DPO fine-tuning, the error between its estimated human preference distribution and the true value objectively exists. To further bridge the gap between the conditional distribution derived from DreamAlign-2D and the actual human preference distribution, we have introduced a preference contrastive learning feedback mechanism, as shown in Figure 4. Our proposed preference contrastive loss replaces the original SDS loss. Assuming that DreamAlign-2D ϕ is capable of generating 3D content that aligns with user aesthetics, and given that the multi-view diffusion model, utilized as a reference, lacks this capability, we aim for the distribution of images rendered by the 3D representation to be closer to DreamAlign-2D and further from multi-view diffusion model. To this end, we use the KL divergence to measure the distance between these distributions. Following this approach, the preference contrastive loss is formalized as follows:

$$\begin{aligned} L_{PCL} =& \mathbb{E}_{t,\mathbf{c},\epsilon}(\\ & \text{D}_{\text{KL}}(q(x_{t-1}|x_{0,t})\|p_\phi(x_{t-1}|x_t)) - \\ & \text{D}_{\text{KL}}(q(x_{t-1}|x_{0,t})\|p_{ref}(x_{t-1}|x_t))) \end{aligned} \tag{5}$$

Based on equation 1 and 2 and algebra, the gradient of the loss function for the 3D representation rendering function can be derived as follows:

$$\nabla_\theta \mathcal{L}_{PCL}(\mathbf{x} = g(\theta)) = \mathbb{E}_{t,\mathbf{c},\epsilon}\left[\frac{2\sigma_t}{\alpha_t}(\epsilon_\phi - \epsilon_{ref})\frac{\partial \mathbf{x}}{\partial \theta}\right] \tag{6}$$

For a detailed derivation process, please refer to the appendix. From the perspective of the error between predicted images, we have also derived equation 6, which validates the correctness of this gradient.

Experiments

Implementation Details

DreamAlign-2D We trained the DreamAlign-2D model using the HP3D dataset and Pick-a-Pic dataset (Kirstain et al. 2023). The model employs MVDream as backbone, which is capable of generating multiview consistent 3D content. DreamAlign-2D was trained on two V100 GPU, with a total batch size of 256 (pairs), a local batch size of 1 pair, and gradient accumulation over 128 steps. The learning rate was set at $1e-5$. In the configuration of the LoRA Adapter structure , the rank was set as 128 and the learning rate scaling factor was established at 256. For the D-3DPO algorithm, the β (divergence penalty parameter) was set at 5000. Each 3D object was rendered as the camera orbited 360 degrees, starting from a frontal view with the elevation angle randomly selected from the range $[0, 30]$ degrees. During training, we randomly selected images from four orthogonal views and resized the images to 256×256 pixels, with the corresponding camera extrinsics normalized onto a unit sphere. During inference, we utilized a DDIM sampler with a classifier-free guidance (CFG) scale of 7.5, sampling steps were set as 50, and random sampling of the camera's elevation angle from $[0, 30]$ degrees, while the azimuth angles were fixed at four orthogonal views.

DreamAlign-3D We implemented DreamAlign-3D based on threestudio and replaced Stable Diffusion in MVDream with our DreamAlign-3D for text-to-3D generation. We employed an implicit volumetric representation for the three-dimensional content. During the optimization process, we used the AdamW optimizer for 10,000 steps. Additionally, we adopted two techniques: firstly, we applied linear annealing to the maximum and minimum time steps of SDS; secondly, we set the rendering resolution to 64×64 for the first 5,000 steps and gradually increased it to 256×256. In most cases, the camera was set as a frontal view when the phase angle was 90 degrees.

Figure 5: Text-to-3d generation for Dreamalign-3D. RGB images from different viewpoints and the corresponding normal pictures are shown. We have highlighted the subjects and the adjectives in captions.

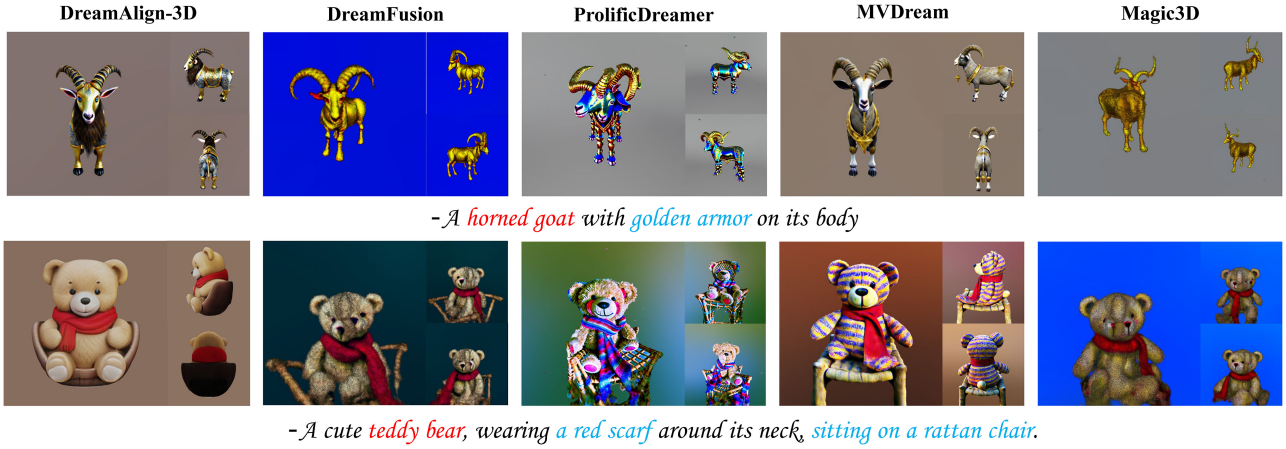

Figure 6: Compare with four baselines in text to 3D generation.

Text-to-Image Generation

In this section, we utilize CLIP and ImageReward to evaluate the image generation capabilities of DreamAlign-2D and explore whether DreamAlign-2D masters the generation of multi-view images that are consistent with human aesthetics.

Quantitative comparison. We evaluated our model using our validation set, quantitatively assessing the results from the perspectives of consistency and aesthetic appeal. Evaluation metrics included the CLIP text score, CLIP image score, and ImageReward (Xu et al. 2024), which are used to measure the alignment between the generated images, the corresponding text, and the ground truth images. For each text prompt, we generated four images with orthogonally positioned camera views. Since SD-v2.1 can only produce one image at a time, we augmented the text with descriptions of camera positions to ensure the generated images were as orthogonal as possible. To ensure fairness in evaluation, five results will be generated for each text caption. The quantitative results are displayed in Table 1.

Table 1 shows that our DreamAlign-2D achieves the best results on all metrics compared to MVDream and SD-v2.1. In terms of text and image scores, DreamAlign-2D improves on generative power of MVDream and even outperforms Ground Truth. The performance of SD-v2.1 is the worst for multi-view consistency. It is trained solely on 2D data, so it lacks generalization capabilities in 3D multi-view generation tasks. Additionally, with respect to the ImageReward evaluation DreamAlign-2D significantly outperforms other methods, demonstrating that it produces multi-view images that are more aligned with human aesthetic preferences.

Method	CLIP ↑			ImageReward ↑
	Text	Views Image	GT Image	
Ground Truth	33.61	81.29	1.00	-0.05
SD-v2.1	33.75(±0.58)	70.28(±2.87)	62.47(±3.21)	0.11(±0.05)
MVDream	31.90(±2.64)	73.12(±1.41)	69.64(±1.27)	-0.25(±0.18)
DreamAlign-2D	**35.85(±1.02)**	**78.29(±1.13)**	**72.28(±1.15)**	**0.29(±0.09)**

Table 1: Comparisons on Image Synthesis Quality. We compare DreamAlign-2D with MVDream and SD-v2.1 on our validation set. Text refers to the CLIP score between the textual description and the generated image, Views Image refers to the CLIP score for consistency across neighboring orthogonal views of the generated image, and **GT Image** indicates the CLIP score between the generated image and the ground truth.

Text-to-3D Generation

In this section, DreamAlign-3D demonstrates its results in generating 3D objects from text and is compared with three baseline methods: ProlificDreamer (Wang et al. 2024), DreamFusion (Poole et al. 2022) and MVdream (Shi et al. 2023). All methods were implemented using the open-source threesutido (Guo et al. 2023) library.

Qualitative results. Figure 5 displays several qualitative results obtained from DreamAlign-3D. The content specified in the text accurately appears in the generated 3D assets, The content specified in the text appears exactly as it should in the generated 3D assets, such as the warrior's sword and shield, and the triangular scarf of black cat. In addition to maintaining a high degree of textual consistency, our method also ensures instance-level consistency across multiple viewpoints, avoiding problems with multiple faces or feet. Furthermore, DreamAlign not only exhibits impressive consistency capabilities but also creatively generates elements not explicitly described in the text, without deviating from the original semantic content, thereby making the generated 3D assets more appealing to human aesthetics. In particular, our approach does not require explicit treatment of lighting, but effectively solves problems such as asymmetric shadows that occur in 3d content. For example, the teddy bears etc. generated by the MVDream method in Figure 2 have obvious shadows when viewed from different angles. But the 3d content generated by DreamAlign has no shadows present at any angle.

Moreover, we compared DreamAlign-3D with other text-to-3D methods. As demonstrated in Figure 6, the intuitive visual evaluation can be conducted through orthogonal views of the front, back, and side. The results indicate that all methods are capable of generating subjects. For simpler texts, all methods successfully include the main subject described in the prompts. But DreamFusion, Magic3D and DrolificDreamer are prone to multi-faceted problems, although the 3D content generated by ProlificDreamer is a bit more vibrant in color. The 3D content generated by MVDream also basically meets the description in the prompt, but is a bit worse than DreamAlign in performance. For example, the goat doesn't wear the full golden armor as required in the prompt, and the goat's beard turns into a golden

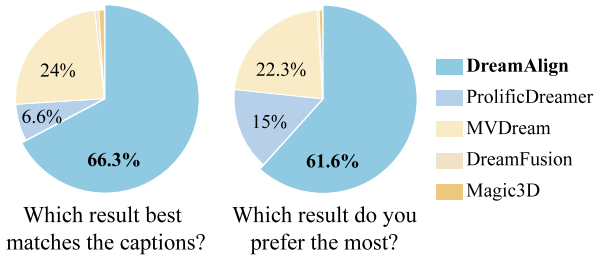

Figure 7: The result of user study.

pendant. The teddy bear created by MVDream has circles of blue stripes on its surface, probably due to the influence of the **cute** in the prompt, which obviously doesn't fit the human aesthetic. More importantly, the teddy bear's legs were ignored. However, when dealing with more complex texts, other methods often overlook some details, except for our approach. In contrast, DreamAlign-3D not only accurately generates the 3D content described by the text but also aligns more closely with human aesthetic preferences.

User study. To more comprehensively assess human preferences for the outputs generated by various methods, we conducted a user study to gather preference data. We selected 30 text captions from GPTEval3D and our validation set, consisting of 10 each of short, medium, and long descriptions. Each participant was asked two questions: (1) Which result best matches the description? (2) Which result do you prefer the most? We invited ten researchers from the field of computer vision, collecting a total of 300 responses. As shown in the Figure 7, 60.3% of users believed that our DreamAlign-3D method generated content that was more consistent with the text descriptions. Additionally, 68.6% of users preferred our method. This indicates that our approach not only maintains consistency with the text but also generates 3D content that better aligns with user aesthetics.

Conclusion

In this work, we propose a novel framework named DreamAlign, which can directly optimize text-to-image models on data with contrastive labels to better meet human preferences. We introduce LoRA adapters in the attention module of stable diffusion and create a HP3D dataset with expert contrastive labels. Using the D-3DPO algorithm for training, the generated images align more closely with the user preferences in various attributes. DreamAlign applies the 2D lifting method directly to the 3D object generation process. We introduce preference contrastive feedback learning to further reduce the gap between 2D and 3D generation. The potential of DreamAlign to judge user preferences is well preserved in 3D generation. Extensive quantitative and qualitative results demonstrate the superiority of our approach in text-to-3D generation tasks. Our DreamAlign provides a general pathway for producing 3D assets that meet user aesthetic preferences.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 41571402), the Science Fund for Creative Research Groups of the National Natural Science Foundation of China (No. 61221003), the National Natural Science Foundation of China (No. 62406161), China Postdoctoral Science Foundation (No. 2023M741950), and the Postdoctoral Fellowship Program of CPSF (No. GZB20230347).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22246–22256.

Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Vicente, T. F. Y.; Dideriksen, T.; Arora, H.; et al. 2022. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21126–21136.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Guo, Y.-C.; Liu, Y.-T.; Shao, R.; Laforte, C.; Voleti, V.; Luo, G.; Chen, C.-H.; Zou, Z.-X.; Wang, C.; Cao, Y.-P.; and Zhang, S.-H. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio.

Gupta, A.; Xiong, W.; Nie, Y.; Jones, I.; and Oğuz, B. 2023. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Wang, J.; Shi, Y.; Qi, X.; Zha, Z.-J.; and Zhang, L. 2023. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*.

Jun, H.; and Nichol, A. 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

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

Li, W.; Chen, R.; Chen, X.; and Tan, P. 2023. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9970–9980.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. R. 2023. Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Luo, T.; Rockwell, C.; Lee, H.; and Johnson, J. 2024. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36.

Ma, Z.; Jia, G.; Qi, B.; and Zhou, B. 2024a. Safe-SD: Safe and Traceable Stable Diffusion with Text Prompt Trigger for Invisible Generative Watermarking. *arXiv preprint arXiv:2407.13188*.

Ma, Z.; Jia, G.; and Zhou, B. 2024. AdapEdit: Spatio-Temporal Guided Adaptive Editing Algorithm for Text-Based Continuity-Sensitive Image Editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4154–4161.

Ma, Z.; Li, J.; Li, G.; and Cheng, Y. 2022. UniTranSeR: A unified transformer semantic representation framework for multimodal task-oriented dialog system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 103–114.

Ma, Z.; Li, J.; Zhang, Z.; Li, G.; and Cheng, Y. 2021. Intention reasoning network for multi-domain end-to-end task-oriented dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2273–2285.

Ma, Z.; Yu, Z.; Li, J.; and Li, G. 2023. HybridPrompt: bridging language models and human priors in prompt tuning for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13371–13379.

Ma, Z.; Yu, Z.; Li, J.; and Zhou, B. 2024b. LMD: faster image reconstruction with latent masking diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4145–4153.

Ma, Z.; Zhao, L.; Qi, B.; and Zhou, B. 2024c. Neural Residual Diffusion Models for Deep Scalable Vision Generation. *arXiv preprint arXiv:2406.13215*.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14297–14306.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022a. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. P.-E. 2022b. A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qian, G.; Mai, J.; Hamdi, A.; Ren, J.; Siarohin, A.; Li, B.; Lee, H.-Y.; Skorokhodov, I.; Wonka, P.; Tulyakov, S.; et al. 2023. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*.

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

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34: 6087–6101.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Skalse, J.; Howe, N.; Krasheninnikov, D.; and Krueger, D. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35: 9460–9471.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Wada, Y.; Kaneda, K.; Saito, D.; and Sugiura, K. 2024. Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13559–13568.

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

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2023. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12619–12629.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

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

Zhu, J.; Zhuang, P.; and Koyejo, S. 2023. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*.

Zhu, Z.; Zhao, H.; He, H.; Zhong, Y.; Zhang, S.; Yu, Y.; and Zhang, W. 2023. Diffusion models for reinforcement learning: A survey. *arXiv preprint arXiv:2311.01223*.

Zhuang, J.; Wang, C.; Lin, L.; Liu, L.; and Li, G. 2023. Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers*, 1–10.