RealMat: Realistic Materials with Diffusion and Reinforcement Learning

XILONG ZHOU, Max Planck Institute for Informatics, Germany
PEDRO FIGUEIREDO, Texas A&M University, USA
MILOŠ HAŠAN, Adobe Research, USA
VALENTIN DESCHAINTRE, Adobe Research, UK
PAUL GUERRERO, Adobe Research, UK
YIWEI HU, Adobe Research, USA
NIMA KHADEMI KALANTARI, Texas A&M University, USA

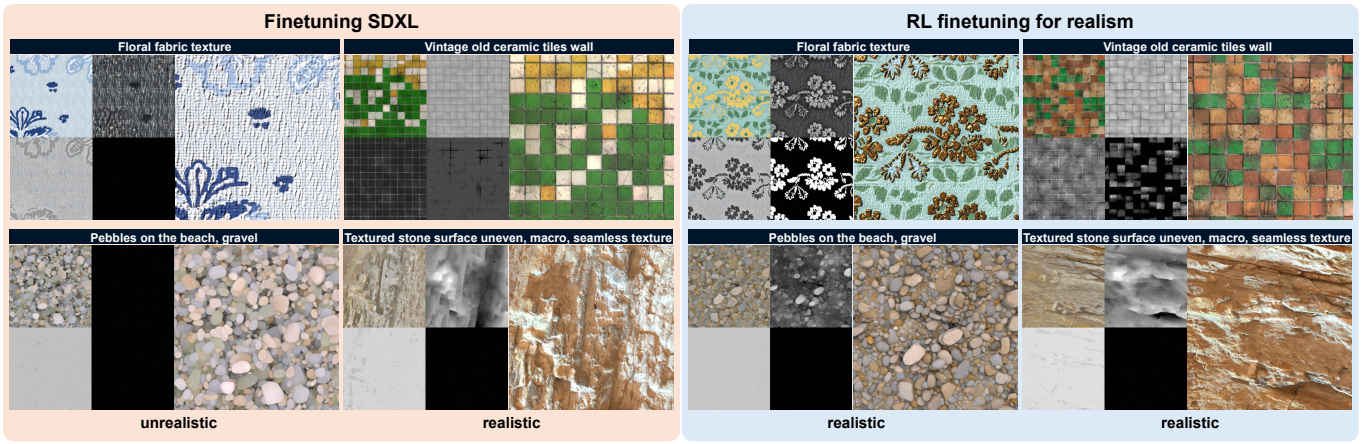

Fig. 1. We propose RealMat, a diffusion-based realistic text-to-material generator. We first finetune a Stable Diffusion XL (SDXL) model pretrained on images to generate detailed material maps using synthetic training data (left). Although high-quality materials are generated in many cases, this finetuning shifts the distribution of SDXL toward a more synthetic appearance. To address this, we further finetune our model using reinforcement learning (RL) with a realism reward function (right). We show samples and maps (in clockwise order from top left: albedo, height, metallicity, and roughness) with fixed text prompts and seeds. With RL, the generated material distribution shifts towards more realistic overall.

Generative models for high-quality materials are particularly desirable to make 3D content authoring more accessible. However, the majority of material generation methods are trained on synthetic data. Synthetic data provides precise supervision for material maps, which is convenient but also tends to create a significant visual gap with real-world materials. Alternatively, recent work used a small dataset of real flash photographs to guarantee realism, however such data is limited in scale and diversity. To address these limitations, we propose RealMat, a diffusion-based material generator that leverages realistic priors, including a text-to-image model and a dataset of realistic material photos under natural lighting. In RealMat, we first finetune a pretrained Stable Diffusion XL (SDXL) with synthetic material maps arranged in 2×2 grids. This way, our model inherits some realism of SDXL while learning the data distribution of the synthetic material grids. Still, this creates a realism gap, with some generated materials appearing synthetic. We propose to further finetune our model through reinforcement learning (RL), encouraging the generation of realistic materials. We develop a realism reward function for any material image under natural lighting, by collecting a large-scale dataset of realistic material images. We show that this approach increases generated materials' realism compared to our base model and related work.

CCS Concepts: • **Computing methodologies** → **Reflectance modeling**.

Additional Key Words and Phrases: materials, SVBRDF, generative models, diffusion models, reinforcement learning

ACM Reference Format:
Xilong Zhou, Pedro Figueiredo, Miloš Hašan, Valentin Deschaintre, Paul Guerrero, Yiwei Hu, and Nima Khademi Kalantari. 2025. RealMat: Realistic Materials with Diffusion and Reinforcement Learning. *ACM Trans. Graph.* 1, 1 (July 2025), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Xilong Zhou, Max Planck Institute for Informatics, Saarbrücken, Germany; Pedro Figueiredo, Texas A&M University, College Station, USA; Miloš Hašan, Adobe Research, San Jose, USA; Valentin Deschaintre, Adobe Research, London, UK; Paul Guerrero, Adobe Research, London, UK; Yiwei Hu, Adobe Research, San Jose, USA; Nima Khademi Kalantari, Texas A&M University, College Station, USA.

ACM 0730-0301/2025/7-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Material reflectance properties play a key role in the appearance of a 3D scene. Material design by artists traditionally required manual

work by experts and specialized software tools. More recently, generative techniques have been proposed [Guerrero et al. 2022; Guo et al. 2020; Vecchio et al. 2024b; Xue et al. 2024; Zhou et al. 2022]. However, these still fall short of achieving sufficient realism, since they are trained on synthetic data. Material generation exclusively from real photos has also been explored [Zhou et al. 2023a], but this approach still suffers from limited real training data. We propose RealMat, which leverages the priors learned by a text-to-image diffusion model (we use Stable Diffusion XL [Podell et al. 2023; Rombach et al. 2022]), combined with a reinforcement learning (RL) approach [Black et al. 2023] to further improve the realism of the generated materials.

Finetuning diffusion models has been proposed for various tasks, showing that the priors from the original models can usually be preserved while adapting the model to new domains. However, these tasks typically still target RGB images, while materials (SVBRDFs) require multiple channels (typically diffuse albedo, heights/normals, roughness, metallicity, etc.). To adapt SDXL to generate materials rather than RGB images, we propose to organize the albedo, height map, roughness and metallicity into a 2×2 grid within a single RGB image. Using this approach, we can finetune SDXL using synthetic material data, while retaining the diffusion model prior. This allows our model to generate a much wider range of appearances with higher realism than a model trained from scratch for SVBRDFs [Vecchio et al. 2024b]. The finetuning process on synthetic data nonetheless pushes the model towards the generation of somewhat more synthetic appearances. To counteract this, we propose to use a reinforcement learning approach inspired by DDPO [Black et al. 2023] to further finetune our model towards generating materials with higher realism. Using RL allows us to apply a realism reward directly to the output of the multi-step inference process of a diffusion model. This approach better reflects the quality observed at inference time than traditional training methods, where a loss is applied only to the coarse output of a single denoising step. We design a reward function that assigns higher rewards to materials that appear more realistic. Specifically, we train a linear layer that takes CLIP image features [Radford et al. 2021] of the material image under natural lighting, and outputs a realism score. The reward model is trained on a newly filtered and labeled dataset of material images.

Combining the finetuning stage using our material grid approach and the RL stage with a realism reward, RealMat learns to generate photorealistic and diverse materials. We evaluate our approach with qualitative comparisons and a user study. In summary, we make the following contributions:

- We propose a simple yet effective strategy to finetune a pretrained text-to-image diffusion model to the SVBRDF map generation task, using a 2×2 grid approach.
- We develop a reward function that can accurately evaluate the realism of rendered materials under natural lighting, and use it in a reinforcement learning (RL) stage to enhance the realism of the final model.

2 RELATED WORK

Image Generation. Image generation has been a long-standing task in computer vision and computer graphics. Variational Autoencoders (VAEs) [Kingma 2013] and Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] have been widely used for image generation applications. However, VAEs often generate blurry images due to the difficulty of modeling complex joint distribution in image space. GANs can produce sharp images, but they face many challenges like training instability and mode collapse. In recent years, Diffusion Models (DM) [Ho et al. 2020] have stood out in image generation tasks due to stable training and high-quality results. With a well-trained noise estimation model, DMs progressively transform input noises into high-quality images through a multi-step denoising process. Built upon DMs, Latent Diffusion Models (LDMs) [Rombach et al. 2022] were proposed to perform denoising steps in latent space, significantly reducing computational and memory cost for high-resolution (512×512 or 768×768) image generation. To further lift LDMs to higher resolution, Stable Diffusion XL (SDXL) [Podell et al. 2023] utilizes a larger UNet and a second text encoder. Leveraging the priors learned from a very large image dataset [Schuhmann et al. 2022] SDXL can generate $1k \times 1k$ high-fidelity and photo-realistic RGB images. Therefore, we build RealMat on top of SDXL to leverage its strengths, including strong generalization, realism, and high-resolution generation.

Material Acquisition. Material reflectance property is a key factor determining the appearance of virtual scenes. Obtaining high-quality photo-realistic materials have been a challenging task in computer graphics industry. Traditional acquisition approaches rely on bulky and expensive hardware such as gonioreflectometer [Guarnera et al. 2016; Matusik et al. 2003] to extensively sample the light and view directions to extract Spatially Varying Bidirectional Reflectance Distribution Functions (SVBRDFs) from real material samples. Recently, with the development of deep learning techniques, many learning-based methods have been proposed for material acquisition. These methods train material priors using synthetic or real dataset to extract SVBRDFs through either feed-forward process [Deschaintre et al. 2018, 2019; Guo et al. 2021; Li et al. 2017, 2018; Martin et al. 2022; Nie et al. 2024; Wang et al. 2023; Zhang et al. 2023; Zhou and Kalantari 2021, 2022], optimization [Gao et al. 2019; Guo et al. 2020; Henzler et al. 2021; Luo et al. 2024; Zhou et al. 2023b], or denoising process [Sartor and Peers 2023; Vecchio et al. 2024a]. As opposed to our goal, these methods target the acquisition of SVBRDFs from input photographs, while our approach focuses on generating new materials from text prompts.

Material Generation. An alternative method to obtaining materials is through material generation. In the industry, the material creation process heavily relies on complex procedural node graphs, which require significant expertise. To simplify the material creation process, various learning-based material generators have been proposed, either using GANs [Guo et al. 2020; Zhou et al. 2022] trained on synthetic data or tailoring the training procedure to work exclusively with real materials captured using flash photographs [Zhou et al. 2023a] for improved realism. An alternative research direction targets procedural material generation, using generic material

graphs and user-driven segmentation [Hu et al. 2022] or direct generation using transformer architectures [Guerrero et al. 2022; Hu et al. 2023]. With the exception of PhotoMat [Zhou et al. 2023a], which is trained on a limited dataset, these material generators are trained on synthetic dataset, creating a visual gap to real-world material.

Recently, DMs have emerged as state-of-the-art generative models, leading to their use in multiple material generators. MatFuse and MatGen [Vecchio et al. 2024a,b] train an LDMs-based material generator from scratch with multiple encoders conditioned on text, image, or sketch. Closer to our approach, Text2Mat [He et al. 2023] and ReflectanceFusion [Xue et al. 2024] propose to leverage pretrained diffusion models on large-scale image datasets. These approaches first generate a latent representation of a natural image or center-flashed material image through pretrained DMs or finetuning DMs. At the second stage, they train a separate decoder to extract SVBRDF parameters, as done in the material acquisition task. With this process, these methods benefit from the realistic prior of DMs; however, the reliance on an "acquisition" step leads to typical "baked lighting" limitations. In contrast, we propose to directly finetune a DM to output material parameters, benefiting from the network's prior without the typical acquisition limitations, and introduce a Reinforcement Learning (RL) approach to improve realism.

Reinforcement Learning of DMs. RL has been widely used in large language models and computer vision tasks. With the success of DMs in text-to-image generation applications, several RL-based algorithms have been proposed to improve the DMs with a goal of finetuning a model to achieve higher reward instead of fitting a target distribution. More specifically, Lee at al. [2023] incorporate human feedback to improve text-to-image alignment of DMs. They finetune DMs on fixed dataset sampled from pretrained DMs using reward-weighted regression (RWR) method, where denoising process is reframed as one-step Markov Decision Process (MDP). Built upon RWR, Dpok [Fan et al. 2024] utilizes new samples from pretrained DMs and include KL-regularization during finetuning process. Both RWR and Dpok are designed from a specific training prompt each training loop and rely on approximate log-likelihood due to single-step MDP. In comparison, denoising diffusion policy optimization (DDPO) [Black et al. 2023] treats the denoising process as a multi-step MDP, and thus is more accurate than RWR and can incorporate many training prompts at once. Xie et al. [2024] have demonstrated that DDPO can be applied to improve 3D consistency of generated multi-view images. To improve realism of a material generator, we build RealMat upon DDPO, with new base model and self-defined realism reward trained using large-scale real dataset.

3 PRELIMINARIES

In this section, we briefly introduce two techniques that are central to our approach: image diffusion models [Ho et al. 2020] and Denoising Diffusion Policy Optimization (DDPO) [Black et al. 2023].

3.1 Diffusion Models

Diffusion models [Ho et al. 2020] represent a distribution $p(x_0|c)$ of images x_0 conditioned on c (e.g. a text prompt embedding) by learning to approximately invert a *forward process* $q(x_t|x_0)$ that maps images x_0 to noisy versions of the images x_t, with a noise strength given by a *time step* $t \in [0, T]$. To invert this mapping, a network $f_\theta(x_t, t, c)$ is trained with the following loss to approximate x_0 given a noisy image x_t, the time step t, and the condition c:

$$\mathbb{E}_{(x_0,c)\sim p(x_0,c), x_t\sim q(x_t|x_0), t\sim\mathcal{U}(0,T)}\left[\|f_\theta(x_t,t,c)-x_0\|^2\right]. \quad (1)$$

Images x_0 can then be obtained by iteratively sampling small denoising steps. We use a Markovian (i.e. probabilistic) DDPM sampler [Ho et al. 2020]:

$$p_\theta(x_{t-1}|x_t,t,c) = \mathcal{N}(a_t x_t + b_t f_\theta(x_t,t,c),\ \sigma_t^2\mathbf{I}), \quad (2)$$

where the factors a_t and b_t, as well as the variance σ_t^2 are chosen according to the denoising schedule of the sampler, and determine a step size that trades off between quality and total number of steps needed. The sequence of images $\tau = (x_T, \ldots, x_0)$ forms a denoising trajectory in image space that starts with pure noise x_T and ends with the fully denoised image x_0.

Our diffusion model is based on Stable Diffusion XL (SDXL) [Podell et al. 2023], a latent diffusion model where images x are represented in a latent space. This model can handle high-resolution $1\text{k}\times 1\text{k}$ generation. We additionally use Zero-SNR [Lin et al. 2024a], a technique to further improve inference quality.

3.2 Denoising Diffusion Policy Optimization

Recently, it has been shown that Proximal Policy Optimization (PPO) [Schulman et al. 2017], a form of reinforcement learning, can be applied to refine diffusion models in an approach called Denoising Diffusion Policy Optimization (DDPO) [Black et al. 2023]. DDPO allows us to optimize diffusion models to maximize a potentially non-differentiable reward that can be computed on the image we obtain after multiple denoising steps (Eq. 2). This contrasts with the traditional loss used to train diffusion models, which optimizes only a single denoising step (Eq.1).

In PPO, a *policy* is trained to perform *actions* that maximize the *accumulated reward* over a sequence of actions. In DDPO, the policy corresponds to the denoising steps $p_\theta(x_{t-1}|x_t,t,c)$ and actions correspond to images x_{t-1} from the denoising trajectory, so that the denoising trajectory $\tau = (x_{T-1}, \ldots, x_0)$ is the sequence of actions that results in the denoised image x_0. DDPO then aims at maximizing the expected accumulated reward $R(\tau)$ over denoising trajectories τ:

$$J(p_\theta) \coloneqq \mathbb{E}_{\tau\sim p_\theta}\left[R(\tau)\right], \quad (3)$$

In DDPO, $R(\tau)$ is computed based on only the final denoised image x_0 (i.e., no reward is given for intermediate denoising steps). We will define the exact reward we use in Section 4.3.

To optimize this objective, we follow the formulation of Schulman et al. [2017], which shows that the gradient of $J(p_\theta)$ w.r.t. the parameters θ is:

$$\nabla_\theta J(p_\theta) = \mathbb{E}_{\tau\sim p_\theta}\left[\sum_{t=T}^{1} R(\tau)\nabla_\theta \log p_\theta(x_{t-1}|x_t,t,c)\right]. \quad (4)$$

The expectation in this equation is approximated using multiple trajectory samples, and the resulting gradient is used to finetune the denoiser parameters θ. Optimization with this gradient can further be stabilized using trust region clipping and importance sampling

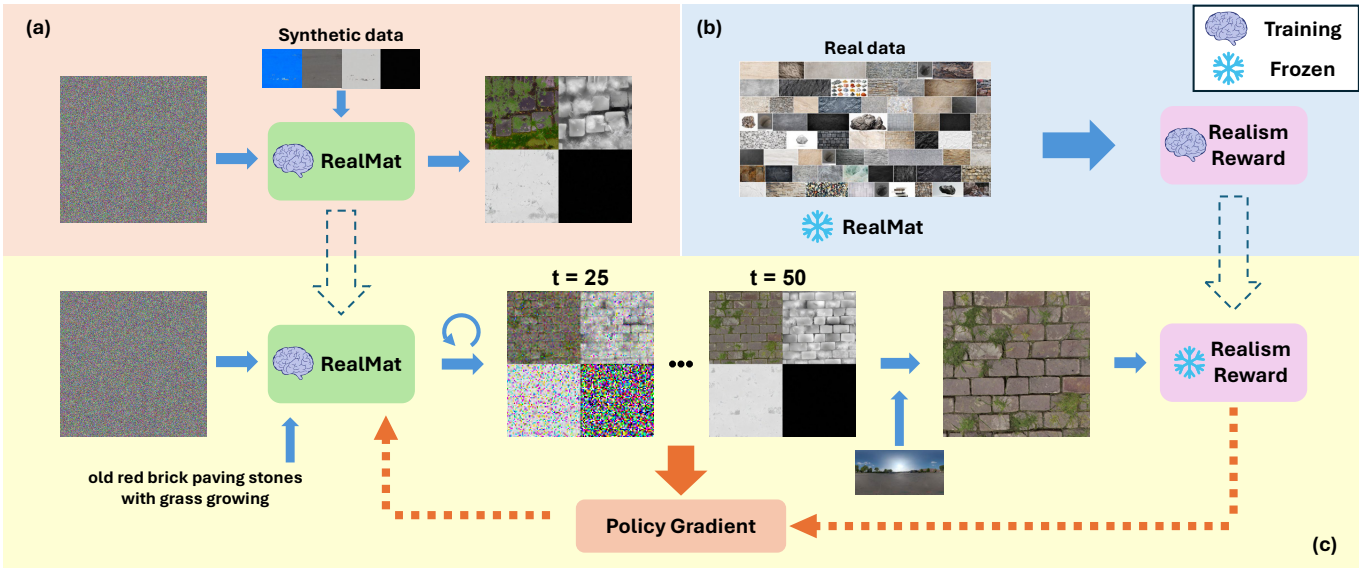

Fig. 2. A visual illustration of RealMat. (a) In the first stage, we finetune SDXL for text-to-material generation using synthetic SVBRDF maps arranged in 2×2 grids; (b) next, we train a realism reward (score) using a mixture of real photographs and synthetic data; (c) in a second finetuning stage, we use the reward in a reinforcement learning (RL) strategy to further push the generated distribution towards more realistic materials.

strategies, as described in DDPO [Black et al. 2023]. Note that the gradient of R is not needed in the computation.

4 REALMAT

Our goal is to train a diffusion-based SVBRDF generator that produces high-fidelity, diverse, and photo-realistic materials from text prompts. The main challenge in this task is to preserve priors from existing data sources. *Pre-trained diffusion models* like Stable Diffusion XL [Podell et al. 2023] offer great text alignment, diversity, and realism, but are meant for general RGB images and cannot generate material maps. *Synthetic material datasets* [Adobe 2023; Vecchio and Deschaintre 2024] contain ground truth SVBRDF maps, but are limited in size and do not fully bridge the realism gap. *Photographed material datasets* [Zhou et al. 2023a] offer realism, but do not have SVBRDF ground truth and are limited in size.

Our key insight is that the large diffusion models and the combination of real and synthetic data can be used to generate realistic SVBRDFs. We use a two-stage approach. In the first stage, we finetune SDXL [Podell et al. 2023] to generate SVBRDFs using a synthetic material dataset (Section 4.2). This model inherits the text alignment, diversity, and realism from SDXL, but the tuning process moves the distribution towards a more synthetic appearance, creating a realism gap. In the second stage, we additionally leverage a dataset of material images (under natural lighting, not flash) to design a realism reward function that we can use to improve realism using reinforcement learning with DDPO (Section 4.3). Note that we propose a two-stage approach because the realism reward function cannot be directly used as a loss in a single finetuning process, as it needs to be evaluated on the material we obtain after multiple denoising steps, and it may not be (easily) differentiable. We show an overview of our approach in Fig. 2.

4.1 Datasets

Synthetic Data. We follow previous work [Vecchio et al. 2024a] and use 8,615 material graphs to generate $\sim$ 126,000 material variations which we render under randomly selected natural environment maps to create $\sim$ 800,000 pairs of renderings and material properties. We augment this data through crops and rotations. This data is used in our first stage of finetuning SDXL.

Generated Materials. Following our first stage of finetuning, we can easily sample thousands of materials from text. These materials, however, have varying degrees of realism. To train a realism reward function, we randomly sample 2,000 materials, each of which is rendered under 10 random environment lighting, for a total of 20,000 renderings in our Generated Materials dataset.

Real Photographs. We filter 800,000 material photographs from an internal image database through a nearest neighbor CLIP-based search using FAISS [Johnson et al. 2019]. Note that although the database contains high-quality images in general, not all material samples are guaranteed to be realistic. Therefore, a classifier is still necessary to filter out realistic data from the Real Photographs dataset.

Data labeling. We further filter the Generated Materials and Real Photographs datasets for realism. Because manually annotating both datasets entirely is impractical, we propose to train classifiers separately on each dataset to estimate the realism labels. Initially, each classifier is trained on a small subset of 2000 images manually annotated as realistic or unrealistic. To prevent overfitting to these small training sets, we employ simple linear classifiers that take CLIP embeddings of the images as input and produce probabilities indicating realism. After training, each classifier predicts realism scores for the remaining unlabeled images in its respective dataset.

We then apply empirically determined thresholds (0.2/0.4 for the Real Photographs/Generated Materials) to generate binary realism labels. The labels are used to train our realism reward function in Stage 2 (Section 4.3).

4.2 Stage 1: Finetuning SDXL

We choose SDXL, trained on the large natural image dataset Laion-5B [Schuhmann et al. 2022], as it offers good realism, diversity, and text alignment, and is able to generate 1k × 1k images; it also empirically responds well to the RL stage below.

Inspired by the 3D generation method [Li et al. 2023], in which four multi-view images of an object are generated as a grid, we represent SVBRDFs as a 2×2 grid: albedo, height, roughness and metallic (as shown in Fig. 4). We then finetune SDXL using these SVBRDF grids. In practice, this lets us trade spatial resolution for the capacity to generate more channels.

Although our base model inherits much of the diversity and realism of SDXL, the finetuning process still moves the distribution towards a more synthetic appearance. We demonstrate this behavior in Fig. 4, with synthetic (top row) and realistic examples (bottom row) generated after this finetuning stage.

4.3 Stage 2: Finetuning for Realism

In Stage 2, we refine our material generator to close the realism gap created in Stage 1. The unknown setting the photos were taken in (lighting conditions, camera parameters, etc.) makes any attempt to use a differentiable renderer to supervise renders of the generated SVBRDFs difficult. In the absence of direct supervision and differentiability, we use our realism reward function. Note that even a differentiable reward could not be directly used as a loss in the regular single-step finetuning setup of a diffusion model, as it needs to be evaluated on the material we obtain after multiple denoising steps. We instead propose to use an RL strategy to finetune our material generator with the goal of maximizing the realism reward.

Realism reward function. The reward function has the same architecture as our realism classifiers for datasets filtering, without the sigmoid activation after the last layer since we target a reward score rather than a classification. Here again, we use the CLIP [Radford et al. 2021] features as input as they have been shown to perform well on materials [Yan et al. 2023] and have been used to define a general image aesthetic score [Black et al. 2023]. In summary, our score function $r_\phi(I)$ is a linear layer, mapping CLIP features to a realism score. We use an MSE loss to train the reward function, with an additional Total Variation (TV) [Johnson et al. 2016] regularization, encouraging similar scores for similar inputs. Our complete loss $\mathcal{L}_r$ for the reward function training is:

$$\mathcal{L}_r = \lambda_1 \| r_\phi(I) - l_{gt} \|_2 + \lambda_2 \mathcal{L}_{tv}\big(\mathrm{CLIP}(I)\big). \tag{5}$$

where l_{gt} represents the thresholded binary realism labels, $\mathcal{L}_{tv}$ is the TV loss term for CLIP features, and λ_1 and λ_2 weight the MSE and TV loss terms. The accumulated reward R in Eq. 4 is then defined as:

$$R(\tau) \coloneqq r_\phi\Big(g\big(d(x_0), L\big)\Big) \text{ with } L \sim \mathcal{U}(\mathbf{L}), \tag{6}$$

where x_0 is the final denoised latent image, d is the VAE decoder, and $g(X, L)$ renders material maps grid X. The rendering uses a lighting environment L, sampled uniformly from a set $\mathbf{L}$ of 200 natural lighting environments. Note that rewards for intermediate diffusion steps are not needed.

Training prompts. To ensure that RL training improves realism where it is most needed, we design a set of "training prompts" which are sampled during the training. We first take 1,000 text prompts from the image descriptions, covering 16 material categories (brick, ceramic, fabric, flower, granite, grass, ground, leather, leaves, marble, metal, paper, pebble, rock, stone and wood). For each text prompt, we sample ten materials using our finetuned material SDXL and evaluate their realism score with our realism reward function. We then select 6 or 7 prompts with the lowest realism within each category, for a total of 100 prompts. This ensures improved realism across a sufficiently diverse range of appearances, especially for concepts that typically lacked realism before RL. We include the list of prompts in the supplemental materials and evaluate the effect of changing the number of training prompts in Section 6.3.

5 IMPLEMENTATION

SDXL finetuning. We finetune with the 1024×1024 material grids from the Synthetic Data. We create corresponding text prompts by combining material category names (e.g."wood" or "metal") with the material names. We finetune SDXL using AdamW [Loshchilov 2017] with a learning rate of 2e−6 and batch size of 120. We set terminal signal-noise ratio (SNR) to zero [Lin et al. 2024b], which helps to generate a high-contrast range (e.g. colorful albedo and solid black metallicity in the same grid). We finetune SDXL for 400k iterations (∼ 7 days) on 24 80GB A100 GPUs.

Training realism reward function. We train our realism classifiers and realism reward function with a learning rate of 1e−3 and batch size of 64. The weights of MSE term λ_1 and TV term λ_2 are set as 1 and 100 respectively. The training takes around 6 hours on a single 40GB A100 GPU.

RL training. We use 32 80GB A100 GPUs. We use 50 denoising steps. We use a batch size of 128 (4 per GPU) and gradient accumulation over two minibatches. Similar to DDPO, instead of finetuning the entire base model, we freeze it and only finetune a Low Rank Adaptation (LoRA) [Hu et al. 2021] with rank 4 to reduce memory and computation cost. We finetune the model with RL for 110 epochs (∼ 18 hours) with a fixed learning rate of 3e−4.

Rendering. We use Mitsuba 3 [Jakob et al. 2022] to render materials on fully displaced meshes under 200 randomly selected realistic HDR environment maps from Poly Haven [env 2024]. The environments are normalized to preserve the same irradiance on a flat plane per RGB channel, to prevent overall intensity and color differences.

6 RESULTS AND DISCUSSION

In this section, we first evaluate our realism reward function numerically and visually. Next, we show sampled results of RealMat from both stages, and perform a RL finetuning user study to evaluate the effect of the second finetuning stage. In addition, we ablate the effect

of training prompt number on the realism RL finetuning process and the effect of TV regularization in the reward function. Finally, we compare RealMat against previous methods qualitatively and through a second user study, demonstrating that our method can generate realistic and diverse materials.

6.1 Realism Reward Function

We first evaluate the realism reward function on both real and synthetic test sets. We find that on average, the reward function estimates a realism score of 0.73 for a real test set (photographs taken by us) and 0.516 for a synthetic test set from Vecchio and Deschaintre [2024]. After normalization to the range [0,1] based on max/min values (0.878/0.343), the average scores for the real and synthetic sets are 0.723 and 0.324, respectively. This shows that our reward function successfully acts as a material realism estimator. In addition, in Fig. 5 we show visual examples from real and synthetic test sets with the estimated scores.

6.2 Results of RealMat

Finetuning with Synthetic Data. We show materials and corresponding renderings sampled after the first stage finetuning in Fig. 4. Our model inherits the diversity and details of SDXL, and can often generate high-quality 512×512 materials with strong text alignment and coherent material channels. Still, this stage creates some realism gap, moving the distribution toward a more synthetic appearance, which motivates the second finetuning stage.

RL Finetuning for Realism. We show the progressive improvement of RealMat during the RL finetuning with our realism reward in Fig. 10. The leftmost image in each row is a rendering of a material sampled from the model after the first stage of RealMat, before RL finetuning, and the rightmost image after the RL finetuning is complete. We keep the prompt and seed fixed in each row. We can see that the realism and visual appearance of synthetic materials (first three rows) gradually improve from left to right as the RL finetuning progresses. The last row, which is already realistic after the first finetuning stage, remains realistic throughout the RL finetuning. We also show a side-by-side comparison of renderings in the initial and final RL finetuning step, along with the corresponding realism score in Fig. 3. The renderings of materials sampled after RL finetuning display more realistic features such as natural-looking leaves, wavy sand, and cracks in tiles.

RL finetuning user study. To further evaluate the impact of our realism RL finetuning, we conduct a user study asking participants to assess the realism of paired materials. In our user study interface (see left in Fig. 6), given a material pair and its text prompt, the user is asked to select which material looks more realistic: "left", "right" or "similar". These paired materials are rendered under the same lighting, with one material sampled from the model at the initial step and the other at the final RL fine tuning step, with the same random seed. The materials of each pair are randomly ordered. We have in total 21 sub-surveys, each of which contains approximately 20 material pairs. Participants can take more than one sub-surveys if they desire.

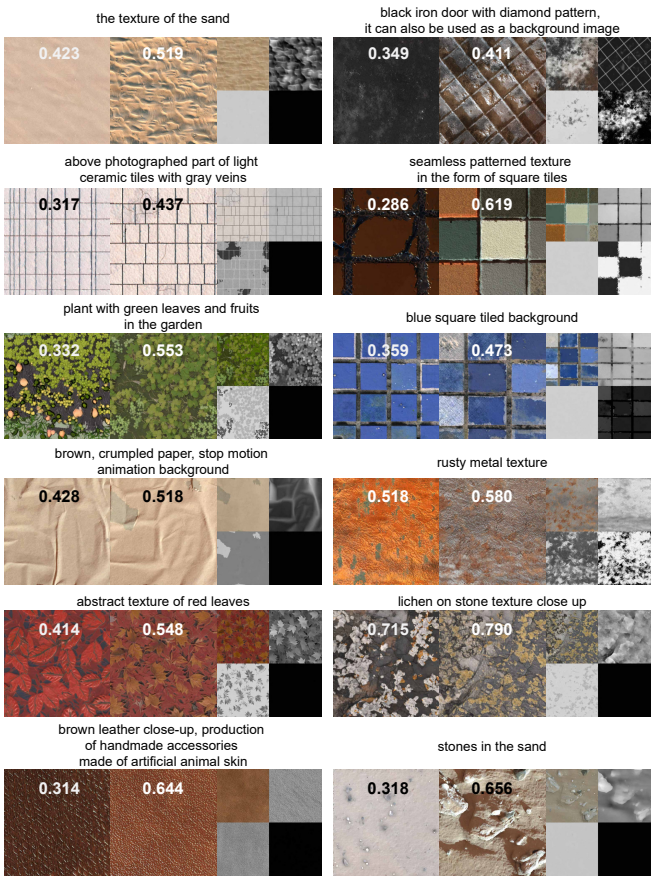

Fig. 3. We show a side-by-side comparison of renderings before and after RL finetuning, together with the corresponding material maps, given various text prompts. In addition, we show the realism rewards. This demonstrates the impact of our training strategy both numerically and visually.

In total, we collect 32 sub-survey responses, corresponding to 641 individual feedback points. On average, 46.8% of responses prefer the materials after realism finetuning, 28.9% of responses hold the opposite opinion, and 24.3% find the results "similar". Among the 421 different material pairs, 43.5% of pairs are marked as improved after finetuning, compared with 23.0% of pairs rated as degraded. The remaining 33.5% of pairs are considered as "similar" or "tie". Note that all the material pairs in the user study are randomly sampled without any data curation. This confirms that RL finetuning improves the realism and visual quality of sampled materials. We further compute the 95% confidence intervals for preference: after finetuning $[42.9\%, 50.7\%]$; before finetuning $[25.4\%, 32.4\%]$ and no preference $[21.0\%, 27.6\%]$.

6.3 Ablation Study

Effect of RL. We evaluate how the realism scores of generated materials evolve using both train and test prompts as our RL finetuning progresses in Fig. 7 (a). We can see that while materials generated using prompts from the training set benefit most from the RL finetuning, the benefits generalize well to unseen prompts.

Effect of training prompt set size. We analyze the effect of training prompt number during the reinforcement learning finetuning stage. Both DDPO [Black et al. 2023] and Carve3D [Xie et al. 2024] claim that finetuning on a limited set of prompts can generalize well to test prompts, which is consistent with our observation. However, in RealMat, we need to carefully distribute the training prompts across all material categories. In Fig. 7 (b), we show the effect of training prompt number on realism score. As shown in the figure, even trained on a prompt number as small as 5, the realism still improves as training progresses but appears to be worse than prompt numbers of 30, 100 and 150. Therefore, we choose a training prompt size of 100 in the final setting of RealMat as it achieves the highest realism score based on our analysis.

Effect of TV loss. TV loss term is added to Eq. 5 to regularize the training of the reward realism function, with the goal of ensuring that similar inputs yield similar realism scores. In Fig. 8, we show a set of real stones with similar visual appearance, and the estimated normalized score by realism function with and without TV regularization. These similar stones are expected to output similar scores based on human perceptual assessment. However, without TV regularization (the bottom row in Fig. 8), the maximum and minimum scores are 0.848 and 0.618 respectively, with a standard deviation of 0.074. In comparison, the standard deviation with TV regularization is reduced to 0.030. We conclude that the TV term is helpful to regularize our realism reward function.

6.4 Comparison with Previous Methods

We compare RealMat against ReflectanceFusion [Xue et al. 2024], MatFuse [Vecchio et al. 2024b], and PhotoMat [Zhou et al. 2023a]. We used the official code from MatFuse and PhotoMat, and asked the authors of ReflectanceFusion to sample their model with our text prompts.

General comparison. We show materials generated from text prompts rendered under a single point light to compare MatFuse, ReflectanceFusion, and RealMat in Fig. 11. As shown, given only a text prompt, MatFuse cannot generate diverse materials. We believe this is largely due to training the model from scratch on synthetic materials, lacking the important priors learned from images that ReflectanceFusion and our approach benefit from. ReflectanceFusion, on the other hand, generates materials that closely follow the input prompt and are overall realistic, but suffer from light-baking artifacts, since the method is based on a base (non-finetuned) SD model that aims to generate RGB images with highlights and shadows, which are non-trivial to undo by the second stage of their approach. In particular, light-baking can be observed in the albedo component (all renderings use a single, centered point light) of the ReflectanceFusion results in Fig. 11. In contrast, our approach generates materials with higher photorealism and fewer artifacts, while preserving good prompt alignment and diversity.

Competing methods user study. To further validate the perceptual quality of our results, we conduct a user study with competing methods (MatFuse, ReflectanceFusion, and RealMat) asking participants to assess the realism and prompt alignment of generated materials. We use a similar interface to the RL finetuning user study (right of Fig. 6), in which participants are asked to choose the most realistic material that also reasonably follows the accompanying text prompt by selecting "left", "middle", "right", or "similar". The randomized materials are rendered with the same animated lighting: a single point light moving along a circle parallel to the material plane. We use GIF animations with 60 frames displayed in a loop to better assess the reflectance behavior of the material as it is illuminated from different directions, and make any baking artifacts clear.

For this user study, we assemble a total of 110 sub-surveys, each containing 10 unique material comparisons. Similarly to the RL finetuning study, participants can take more than one sub-survey. In total, we collect 44 sub-survey responses, corresponding to 440 individual comparisons without any data curation. Here, we show the average followed by the 95% confidence interval range. 53% [48.3%-57.6%] of responses prefer the materials generated by RealMat, while 40% [35.6%-44.8%] opt for results from ReflectanceFusion. Only 3% [1.2%-4.3%] of responses choose MatFuse, and 4% [2.2%-5.9%] find the materials "similar". Among the 373 unique material prompts, RealMat is considered best for 52%, ReflectanceFusion comes second with 40%, MatFuse is chosen for 3%, and 5% are ties ("similar"). The study shows that our approach generates more realistic materials than competing methods. It also highlights the impact of leveraging image priors, strategy shared by RealMat and ReflectanceFusion, which combined constitute the overwhelming majority of realistic materials according to the participants.

PhotoMat comparison. We compare RealMat with PhotoMat, a material generator that is exclusively trained on real flash photos. Results are shown in Fig. 9. Note that PhotoMat is only designed for class-conditioned generation, without text control. Due to the limited scale of the PhotoMat training dataset, the sampled materials are fairly realistic but tend to lack diversity (we show sampled “stone” of PhotoMat in Fig. 9). In comparison, RealMat leverages existing realistic priors and produces realistic stones with greater diversity.

7 LIMITATIONS

We observe that detailed text control for material generation remains challenging for precise descriptions (e.g. separate control of the colors of each tile). We believe this may be difficult due to the lack of high-quality text prompts in material datasets. A second limitation of our proposed approach is spatial resolution, which is traded off for additional channels. While this could be mitigated using newer, high-resolution models, it makes the generalization of the grid approach beyond 4 maps more challenging.

8 CONCLUSION

We propose RealMat, a realistic material generator trained to preserve the prior knowledge in large text-to-image diffusion models and enhance generated material realism. We first adapt the SDXL model to generate materials using finetuning on 2×2 tiled SVBRDF grids. We further enhance the realism of generated materials by utilizing reinforcement learning (RL) with a realism reward function designed on a large-scale dataset of realistic images. We demonstrate the effectiveness of our realism reward function and the RL finetuning stage and show that RealMat can produce more diverse

and photo-realistic materials compared to state-of-the-art material generators.

REFERENCES

2024. PolyHaven HDR environments. https://polyhaven.com/hdris.

Adobe. 2023. Substance. https://substance3d.adobe.com/assets.

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

Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-image SVBRDF Capture with a Rendering-aware Deep Network. *ACM Trans. Graph.* 37, 4 (2018), 128:1–128:15.

Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2019. Flexible SVBRDF Capture with a Multi-Image Deep Network. *Computer Graphics Forum* 38, 4 (2019).

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

Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2019. Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. *ACM Trans. Graph.* 38, 4 (2019).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*. 2672–2680.

Dar'ya Guarnera, Giuseppe Claudio Guarnera, Abhijeet Ghosh, Cornelia Denk, and Mashhuda Glencross. 2016. BRDF Representation and Acquisition. *Computer Graphics Forum* (2016).

Paul Guerrero, Milos Hasan, Kalyan Sunkavalli, Radomir Mech, Tamy Boubekeur, and Niloy Mitra. 2022. MatFormer: A Generative Model for Procedural Materials. *ACM Trans. Graph.* 41, 4, Article 46 (2022). https://doi.org/10.1145/3528223.3530173

Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. 2021. Highlight-Aware Two-Stream Network for Single-Image SVBRDF Acquisition. *ACM Trans. Graph.* 40, 4, Article 123 (jul 2021), 14 pages. https://doi.org/10.1145/3450626.3459854

Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020. MaterialGAN: Reflectance Capture using a Generative SVBRDF Model. *ACM Trans. Graph.* 39, 6 (2020), 254:1–254:13.

Zhen He, Jie Guo, Yan Zhang, Qinghao Tu, Mufan Chen, Yanwen Guo, Pengyu Wang, and Wei Dai. 2023. Text2Mat: Generating Materials from Text. (2023).

Philipp Henzler, Valentin Deschaintre, Niloy J. Mitra, and Tobias Ritschel. 2021. Generative Modelling of BRDF Textures from Flash Images. *ACM Trans. Graph.* 40, 6, Article 284 (dec 2021), 13 pages.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

Yiwei Hu, Paul Guerrero, Milos Hasan, Holly Rushmeier, and Valentin Deschaintre. 2023. Generating Procedural Materials from Text or Image Prompts. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Yiwei Hu, Chengan He, Valentin Deschaintre, Julie Dorsey, and Holly Rushmeier. 2022. An Inverse Procedural Modeling Pipeline for SVBRDF Maps. *ACM Trans. Graph.* 41, 2, Article 18 (jan 2022), 17 pages. https://doi.org/10.1145/3502431

Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. 2022. *Mitsuba 3 renderer*. https://mitsuba-renderer.org.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 694–711.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192* (2023).

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2023. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214* (2023).

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling Surface Appearance from a Single Photograph Using Self-Augmented Convolutional Neural Networks. *ACM Trans. Graph.* 36, 4 (2017), 45:1–45:11.

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. 2018. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 11207)*. 74–90.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2024a. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 5404–5411.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2024b. Common Diffusion Noise Schedules and Sample Steps Are Flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 5404–5411.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

Xuejiao Luo, Leonardo Scandolo, Adrien Bousseau, and Elmar Eisemann. 2024. Single-Image SVBRDF Estimation with Learned Gradient Descent. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15018.

Rosalie Martin, Arthur Roullier, Romain Rouffet, Adrien Kaiser, and Tamy Boubekeur. 2022. MaterIA: Single Image High-Resolution Material Capture in the Wild. *Computer Graphics Forum* 41, 2 (2022), 163–177. https://doi.org/10.1111/cgf.14466 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14466

Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. 2003. A Data-Driven Reflectance Model. *ACM Trans. Graph.* 22, 3 (2003), 759–769.

Yongwei Nie, Jiaqi Yu, Chengjiang Long, Qing Zhang, Guiqing Li, and Hongmin Cai. 2024. Single-Image SVBRDF Estimation Using Auxiliary Renderings as Intermediate Targets. *IEEE Transactions on Visualization and Computer Graphics* (2024).

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Sam Sartor and Pieter Peers. 2023. Matfusion: a generative diffusion model for svbrdf capture. In *SIGGRAPH Asia 2023 Conference Papers*. 1–10.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

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

Giuseppe Vecchio and Valentin Deschaintre. 2024. MatSynth: A Modern PBR Materials Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22109–22118.

Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. 2024a. ControlMat: A Controlled Generative Approach to Material Capture. *ACM Trans. Graph.* 43, 5, Article 164 (sep 2024), 17 pages. https://doi.org/10.1145/3688830

Giuseppe Vecchio, Renato Sortino, Simone Palazzo, and Concetto Spampinato. 2024b. Matfuse: controllable material generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4429–4438.

Li Wang, Lianghao Zhang, Fangzhou Gao, and Jiawan Zhang. 2023. DeepBasis: Hand-Held Single-Image SVBRDF Capture via Two-Level Basis Material Model. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

Desai Xie, Jiahao Li, Hao Tan, Xin Sun, Zhixin Shu, Yi Zhou, Sai Bi, Sören Pirk, and Arie E Kaufman. 2024. Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6369–6379.

Bowen Xue, Claudio Guarnera, Shuang Zhao, and Zahra Montazeri. 2024. ReflectanceFusion: Diffusion-based text to SVBRDF Generation. In *Eurographics Symposium on Rendering*. Eurographics Association.

K. Yan, F. Luan, M. Hašan, T. Groueix, V. Deschaintre, and S. Zhao. 2023. PSDR-Room: Single Photo to Scene using Differentiable Rendering. In *ACM SIGGRAPH Asia 2023 Conference Proceedings*.

Lianghao Zhang, Fangzhou Gao, Li Wang, Minjing Yu, Jiamin Cheng, and Jiawan Zhang. 2023. Deep SVBRDF Estimation from Single Image under Learned Planar Lighting. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2023a. Photomat: A material generator learned from single flash photos. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2022. TileGen: Tileable, Controllable Material Generation and Capture. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea)

(SA '22). Association for Computing Machinery, New York, NY, USA, Article 34, 9 pages. https://doi.org/10.1145/3550469.3555403

Xilong Zhou, Miloš Hašan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2023b. A Semi-Procedural Convolutional Material Prior. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, e14781.

Xilong Zhou and Nima Khademi Kalantari. 2021. Adversarial Single-Image SVBRDF Estimation with Hybrid Training. *Computer Graphics Forum* (2021). https://doi.org/10.1111/cgf.142635

Xilong Zhou and Nima Khademi Kalantari. 2022. Look-Ahead Training with Learned Reflectance Loss for Single-Image SVBRDF Estimation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–12.

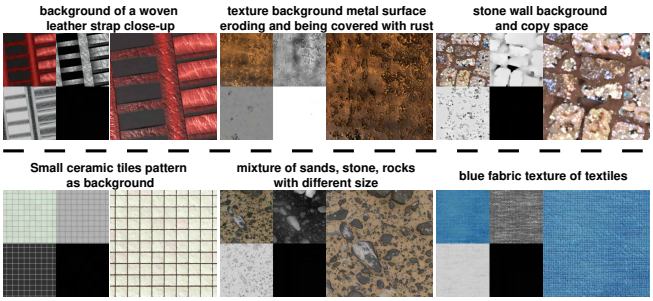

Fig. 4. The sampled material maps and renderings of RealMat at the first stage of fine-tuning with the synthetic dataset. Here, we partition examples as synthetic (top row) and realistic (bottom row) to motivate our reinforcement learning realism fine-tuning stage (Sec. 4.3). In all cases, the materials are high-quality, show consistent features across different material maps, and preserve good text alignment.

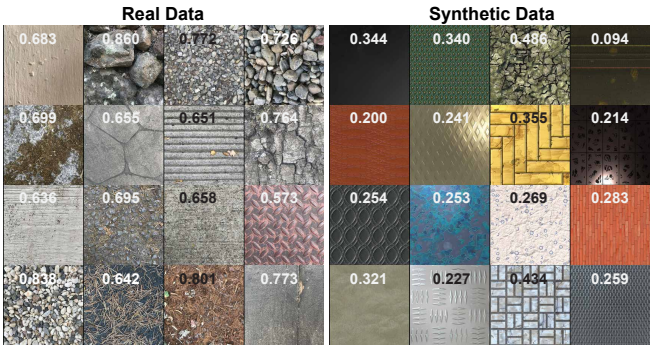

Fig. 5. The estimated normalized realism scores from our realism reward function. Left shows the scores of real materials and right shows the scores of rendered synthetic materials, demonstrating the effectiveness of the realism reward function.

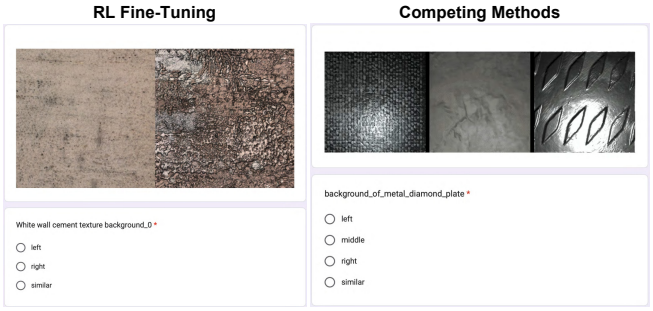

Fig. 6. The user interfaces of our user studies. For the RL fine-tuning study, text prompts and material pairs are shown to participants who need to select the most realistic option by selecting left, right, or similar. Similarly for the competing methods user study, participants select the most realistic material from three randomized options generated by MatFuse [Vecchio et al. 2024b], ReflectanceFusion [Xue et al. 2024], and RealMat by selecting left, middle, right, or similar.

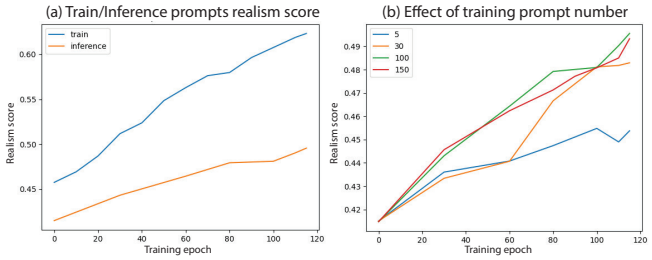

Fig. 7. Plots of realism score during training: (a) Normalized realism score plot for training and inference text prompts during the RL fine-tuning stage. (b) The effect of training prompt number on the reinforcement learning process. We show the realism score of models trained under 5, 30, 100 and 150 training prompts. Based on the plot, we choose $N = 100$ as final optimal training setting.

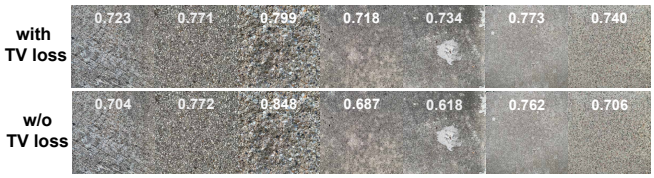

Fig. 8. We demonstrate the effect of TV regularization when training realism reward function. As demonstrated, using realism reward trained without TV loss, stones with similar appearance yield normalized scores with large variance.

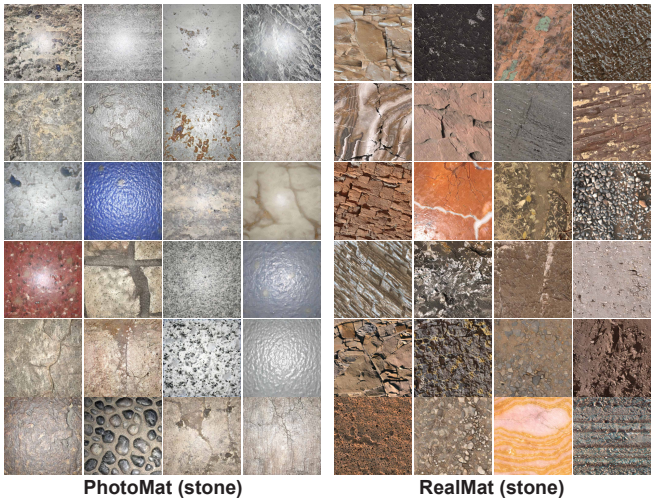

Fig. 9. Comparison of sampled results of PhotoMat and RealMat (we only focus on one specific category, "stone"). Although the PhotoMat results are realistic, the sampled materials show limited diversity due to a small-scale real dataset. In comparison, RealMat is trained with realistic priors and demonstrates both strong realism and diversity.

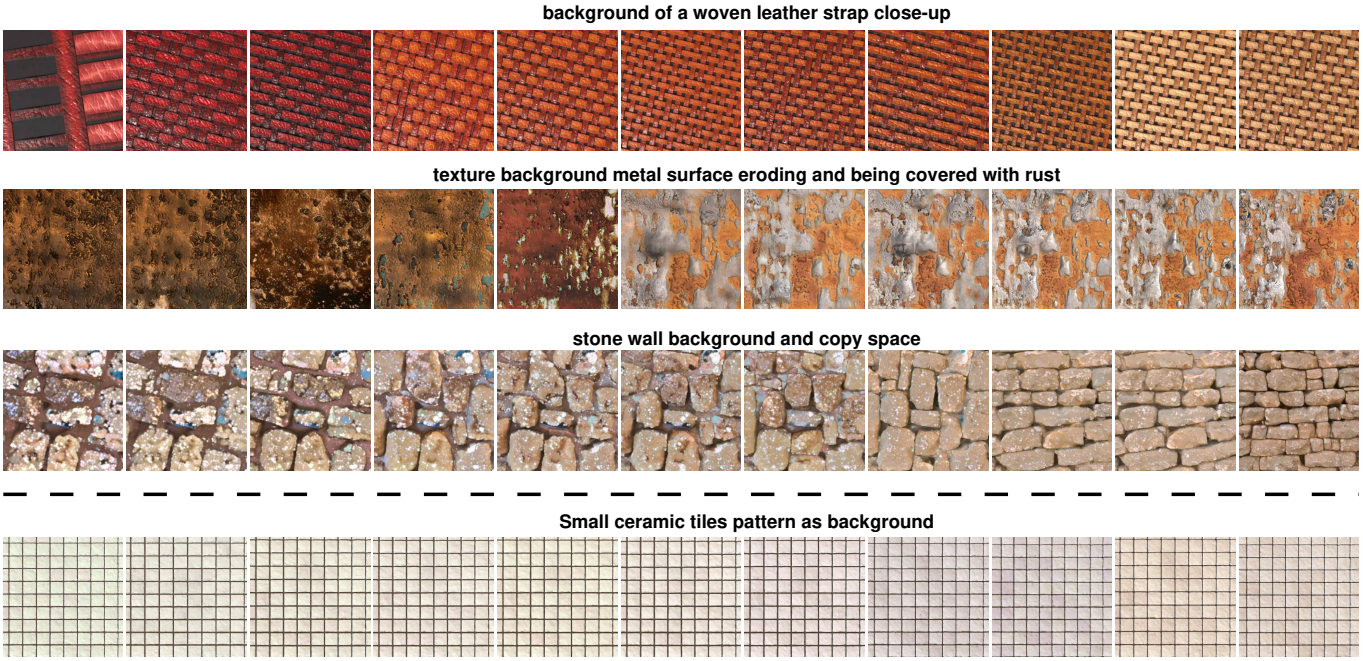

Fig. 10. This figure shows materials resulting from increasing levels of RL fine-tuning with realism reward, starting from no fine-tuning (left) to fine-tuning for 110 epochs (right). RL fine-tuning progressively improves synthetic materials (top three rows), while remaining consistent for realistic materials (last row). This demonstrates the effectiveness and robustness of the second stage of RealMat.

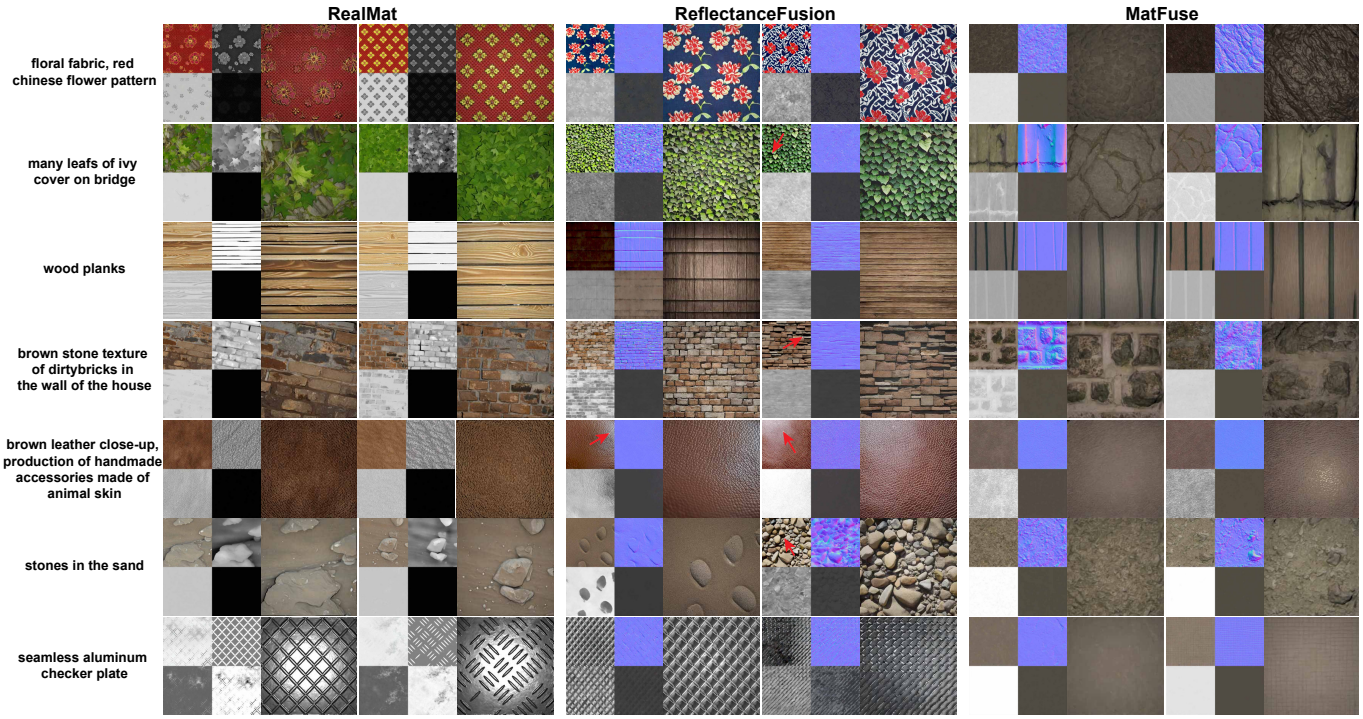

Fig. 11. Visual comparison between RealMat, ReflectanceFusion and MatFuse. The maps from ReflectanceFusion have light baking artifacts (marked by red arrows), produced by a base (non-finetuned) image model, which are difficult to undo by further processing. MatFuse produces less plausible and diverse materials, due to exclusively synthetic training data. In comparison, RealMat generates diverse, realistic and artifact-free material maps.