TexGarment: Consistent Garment UV Texture Generation via Efficient 3D Structure-Guided Diffusion Transformer

Supplementary Material

1. Preliminaries

Diffusion formulation. We review fundamental concepts essential for understanding diffusion models, specifically denoising diffusion probabilistic models (DDPMs). Gaussian diffusion models define a *forward noising process* that progressively corrupts real data x_0 by adding Gaussian noise. This process is expressed as $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$, where the constants $\bar{\alpha}_t$ are predefined hyperparameters. Using the reparameterization trick, samples can be generated as $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. The objective of diffusion models is to learn a reverse process that denoises x_t to reconstruct x_0. This reverse process is modeled as $p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t), \Sigma_\theta(x_t))$, where neural networks predict the mean μ_θ and covariance Σ_θ. The model is trained by maximizing the variational lower bound [2] on the log-likelihood of x_0. The training objective simplifies to $\mathcal{L}(\theta) = -p(x_0|x_1) + \sum_t \mathcal{D}_{KL}(q^*(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$ excluding a constant term irrelevant for training. Since both q^* and p_θ are Gaussian, the KL divergence $\mathcal{D}_{KL}$ can be efficiently computed using their means and covariances. To simplify training, μ_θ can be reparameterized as a noise prediction network ϵ_θ. In this formulation, the model minimizes the simple mean squared error (MSE) between the predicted noise $\epsilon_\theta(x_t)$ and the ground truth noise ϵ_t: $\mathcal{L}_{simple}(\theta) = ||\epsilon_\theta(x_t) - \epsilon_t||_2^2$ However, when learning the reverse process covariance Σ_θ, optimizing the full KL divergence term becomes necessary. We train ϵ_θ with $\mathcal{L}_{\text{simple}}$ and separately train Σ_θ using the full variational loss $\mathcal{L}$. Once the model p_θ is trained, new samples can be generated by initializing $x_{t_{\text{max}}} \sim \mathcal{N}(0, \mathbf{I})$ and iteratively sampling $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ using the reparameterization trick.

Classifier-free guidance. Conditional diffusion models incorporate additional input information, such as a text prompt c_{text}. In this scenario, the reverse process is expressed as $p_\theta(x_{t-1} \mid x_t, c_{text})$, with both ϵ_θ and Σ_θ conditioned on c_{text}. To guide the sampling process towards generating samples x that align strongly with c_{text}, *classifier-free guidance* can be employed [1]. Using Bayes' rule, we have: $\log p(c_{text}|x) \propto \log p(x|c_{text}) - \log p(x)$,. By interpreting the output of diffusion models as the score function, the DDPM sampling procedure can be guided to sample x with high $p(x \mid c_{text})$ using: $\hat{\epsilon}_\theta(x_t, c_{text}) = \epsilon_\theta(x_t, \emptyset) + s \cdot \nabla_x \log p(x|c_{text}) \propto \epsilon_\theta(x_t, \emptyset) + s \cdot (\epsilon_\theta(x_t, c_{text}) - \epsilon_\theta(x_t, \emptyset))$ where $s > 1$ represents the guidance scale (with $s = 1$ recovering standard sampling). To evaluate the diffusion model with $c_{text} = \emptyset$, we randomly drop out c_{text} during training and replace it with a learned "null" embedding $\emptyset$. Classifier-free guidance is well-known for producing significantly improved samples compared to generic sampling techniques [1, 4], and this trend is consistent with our models.

2. Analysis of Point Number

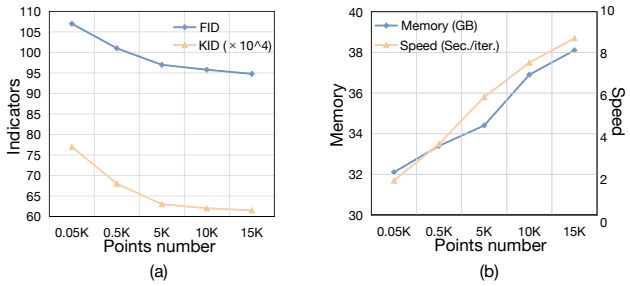


Figure 1. We analyze our method's average FID, KID($\times 10^4$), GPU memory and trainingg speed under various numbers of sampled points.

The number of point clouds significantly impacts the richness of the 3D information. Figure 1 analyzes how the number of sampled point clouds affects FID, KID, GPU memory, and training speed. Using the farthest point sampling method [3], we extract between 50 and 15,000 points from the mesh surface, leading to the following observations: 1) As the number of points increases, both FID and KID initially decrease, indicating enhanced texture quality due to richer 3D information. Beyond a certain threshold, these metrics stabilize, suggesting that further increases in point density provide diminishing returns and the performance approaches convergence. 2) GPU memory and training time per iteration consistently increase as the number of points increases, reflecting higher computational demands.

Balancing performance improvements against training costs, we determined that sampling 5,000 points per mesh achieves an optimal trade-off.

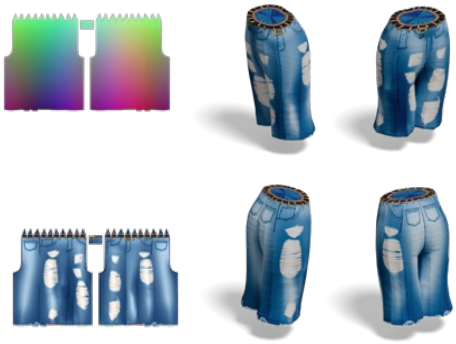
3. Additional Results

In order to further showcase the effectiveness of our method, we present additional results in Figure 2. The generated results further validate the effectiveness of our method. The texture produced by our approach is detailed and realistic, aligning closely with the text descriptions.

a dress with ethnic prints, incorporating colors from diverse cultures

the garment is a royal blue princess gown, adorned with delicate lace and embroidery

light blue denim shorts with distressed details and rolled cuffs



an ethereal, ivory lace and silk evening gown with delicate floral lace overlay on the bodice

a pair of black jeans

a casual denim dress with button-down front and rolled-up sleeves, for a laid-back style

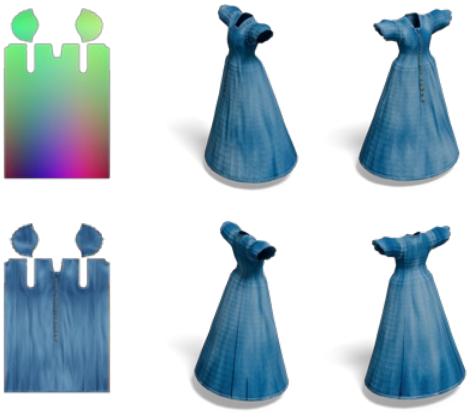


Figure 2. More garment textures generated by our method.

References

[1] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[2] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[3] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1

[4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 1