MATLAT: Material Latent Space for PBR Texture Generation

Anonymous CVPR workshop submission

Paper ID

Figure 1. **PBR Textures Generated by MATLAT.** Our method produces PBR textures that accurately represent rough materials (left), metallic surfaces (middle), and complex mixed materials (right).

Abstract

Text-guided PBR texture generation is essential for relightable 3D asset creation, but high-quality PBR supervision is scarce and pretrained RGB latent priors do not readily extend to roughness and metallic channels. We present MATLAT*, a framework that learns a material latent space by fine-tuning a pretrained latent diffusion model for multi-view PBR generation. Unlike prior methods that freeze the pretrained encoder and suffer from distribution shifts when encoding additional PBR channels,* MATVAE *fine-tunes the encoder to incorporate material channels through residual latent prediction while regularizing the resulting posterior toward the original RGB latent distribution. We further show that correspondence-aware attention alone is insufficient for cross-view consistency unless the latent-to-image mapping preserves spatial locality; to enforce this, we introduce a regularization that crops latent patches, decodes them, and aligns the results with corresponding image regions.*

On Objaverse-XL, MATLAT *improves the strongest multi-view baseline from 6.309 to 3.083 shaded FID and from 9.630 to 4.599 albedo FID while retaining competitive roughness and metallic accuracy. Ablations confirm that both distributional alignment and locality preservation are individually necessary for high-fidelity PBR texture synthesis.*

1. Introduction

Generative modeling has driven remarkable progress in 3D asset creation, spanning geometry [2, 12, 20, 31], textures [1, 13, 21, 30], motions [8, 19, 25], and complete scenes [11, 15, 34]. Among these, Physically Based Rendering (PBR) textures—comprising albedo, roughness, and metallic channels—are essential for production-ready assets, as they enable physically accurate relighting under arbitrary illumination. Recent PBR texture generation methods [1, 3, 21, 26] predict material maps for relightable assets, yet progress is constrained by the scarcity of large-

scale PBR datasets [4, 5]. A natural remedy is to leverage pretrained image generative models with strong priors from large-scale RGB data [22]. Among existing strategies, multi-view generation followed by mesh projection [9, 10] has proven most effective, as SDS-based approaches [16, 20, 27] often produce saturation artifacts. However, this multi-view approach still faces two challenges for PBR synthesis: (1) directly applying diffusion models trained on RGB latents to PBR generation is non-trivial, since simply mapping additional material channels to the pretrained encoder introduces a substantial domain gap; and (2) preserving multi-view consistency remains difficult, with failures causing blurring and artifacts in overlapping regions upon unprojection. To address both challenges, we introduce MATLAT, a framework that learns a **Mat**erial **Lat**ent space for high-quality PBR texture generation. Our two-stage pipeline first fine-tunes the pretrained VAE on PBR images to obtain Material VAE (MATVAE), then fine-tunes a diffusion model for multi-view material generation in the adapted latent space. Our first contribution is a latent-space adaptation module in MATVAE that extends the pretrained space to incorporate roughness and metallic channels. Unlike prior works [3, 7, 9, 10, 28, 36, 37] that freeze the encoder and use zero-channel padding—causing distributional mismatch and suboptimal diffusion fine-tuning—we fine-tune the encoder with a distributional regularization that constrains deviations from the original latent distribution. While similar extensions exist for transparent [33] or RGB-D synthesis [14], this is the *first* adaptation to PBR texture generation, and we identify crucial components for its practical effectiveness. Our second contribution is locality regularization in MATVAE, which enforces spatial alignment between latent tokens and image pixels. This alignment is subsequently leveraged by correspondence-aware attention (CAA) [24] in the diffusion model to enforce multi-view consistency. While spatial locality largely holds for pretrained RGB encoders [9], it often breaks for additional material channels. We address this by cropping latent patches, decoding them, and applying an ℓ_2 reconstruction loss against corresponding image regions. Experiments show that MATLAT outperforms baselines trained from scratch, SDS-based methods, and multi-view diffusion models on most quantitative metrics, achieving state-of-the-art performance. Ablation studies validate each proposed component.

2. Method

Given a mesh $\mathcal{M}$ and text prompt y, our goal is to generate multi-view PBR material images $\mathbf{x}_i = [\mathbf{a}_i, \mathbf{r}_i, \mathbf{m}_i]$ (albedo/roughness/metallic) for viewpoints $\{c_i\}_{i=1}^N$, and then project them to a consistent texture map. Our pipeline has two stages: (1) prior-preserving material latent learning with MATVAE, and (2) multi-view diffusion fine-tuning in the learned latent space, so that the first stage restores compatibility with pretrained RGB priors and the second stage exploits that compatible latent space for multi-view generation.

Prior-preserving material latent. Prior multi-view PBR pipelines often rely on a frozen RGB VAE [9], which encodes albedo reasonably well but places roughness and metallic maps off the pretrained latent manifold. We instead adapt the latent space while explicitly anchoring it to pretrained RGB statistics.

Let $q(\mathbf{z}_{\text{base}} \mid \mathbf{a}) = \mathcal{N}(\boldsymbol{\mu}_{\text{base}}, \boldsymbol{\sigma}^2_{\text{base}})$ be the pretrained latent distribution from albedo. We add a residual encoder $\mathcal{E}_{\text{res}}$ that predicts $(\boldsymbol{\mu}_{\text{res}}, \boldsymbol{\sigma}_{\text{res}})$ from full PBR input $\mathbf{x}$ and compose:

$$q(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\text{base}} + \boldsymbol{\mu}_{\text{res}},\ \boldsymbol{\sigma}^2_{\text{base}} \odot \boldsymbol{\sigma}^2_{\text{res}}). \tag{1}$$

To preserve compatibility with pretrained diffusion priors, we apply distribution alignment:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{reg}}\, \mathrm{KL}(q(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z}_{\text{base}} \mid \mathbf{a}))\,. \tag{2}$$

Combined with standard VAE reconstruction, KL, and discriminator losses, this objective gradually injects roughness and metallic information into an RGB-aligned latent manifold. Compared with residual-only prediction [33] or direct KL-aligned prediction [14], MATVAE combines residual initialization and distribution alignment in a single stochastic posterior.

Locality-preserved multi-view diffusion. We fine-tune a multi-view diffusion backbone in the learned material latent space and adopt correspondence-aware attention (CAA) [24] so that each latent token aggregates features only from geometrically corresponding regions in other views (Fig. 2(a)). However, CAA is only meaningful if the adapted latents remain locally aligned with decoded pixels. We therefore enforce locality during MATVAE training by matching cropped image patches before and after encode–decode (Fig. 2(b)):

$$\mathcal{L}_{\text{local}} = \lambda_{\text{local}}\, d(\mathcal{T}(\mathbf{x}),\ \mathcal{D}(\mathcal{T}(\mathcal{E}(\mathbf{x}))))\,. \tag{3}$$

This patch-level constraint keeps decoded pixels tied to spatially aligned latent tokens, reducing projection seams and making CAA translate into real pixel-space consistency.

Training and inference. We first train MATVAE on rendered PBR material images, where the pretrained encoder is kept frozen and only the residual encoder $\mathcal{E}_{\text{res}}$ and decoder $\mathcal{D}$ are optimized. The final convolution of $\mathcal{E}_{\text{res}}$ is zero-initialized so that training begins at zero residual, exactly reproducing the pretrained latent and allowing the material

channels to be incorporated gradually. We then fine-tune the multi-view diffusion model with CAA modules inserted alongside the original attention layers, using a conditional flow-matching objective [17] in the learned material latent space. At inference, the model generates six canonical-view PBR images that are projected and blended into texture space for relightable mesh rendering.

3. Experiments

Setup. We fine-tune the STABLE DIFFUSION 3.5-MEDIUM VAE and diffusion checkpoints. Training uses 40,723 Objaverse-XL assets [4] with captions from Bootstrap3D and Cap3D [18, 23]. For each training mesh we render 26 fixed views, and at inference we generate six canonical views (front, back, left, right, top, and bottom) following prior multi-view texturing methods. For evaluation, we follow PacTure [7] on 128 held-out meshes with 20 views per mesh: eight at $0°$ elevation, six at $30°$, four at $45°$, and two polar views at $\pm 90°$. For each view, we render shaded RGB images under matched HDR environment maps as well as the corresponding albedo, roughness, and metallic maps.

Evaluation metrics. We report shaded/albedo FID and KID in CLIP feature space, CLIP text–image similarity, and roughness/metallic RMSE. Because these metrics do not directly measure multi-view agreement, we also report c-PSNR, which computes PSNR between each pixel and its geometric correspondences across views. Runtime in seconds captures whether the method is practical for iterative text-guided asset creation.

Main comparison. We compare against three baseline families using official released checkpoints: trained-from-scratch methods (SC) [3, 29], SDS-based optimization pipelines (SDS) [6, 32, 35], and multi-view diffusion methods (MV) [9, 10]. As prior work differs substantially in pretraining, datasets, and optimization procedures, a strictly unified retraining protocol is impractical; we therefore render and evaluate all methods under the same pipeline and treat the comparison as a realistic positioning against existing systems.

Table 1 shows that MATLAT achieves the best shaded and albedo fidelity among all compared methods. Compared with MaterialMVP [9], shaded FID improves from 6.309 to 3.083 and albedo FID from 9.630 to 4.599, while retaining competitive roughness and metallic RMSE. SC methods generally exhibit suboptimal performance, indicating that limited PBR supervision alone is insufficient to learn both material realism and text alignment. SDS methods benefit from pretrained 2D priors but, aside from DreamMat [35] which incurs a prohibitive runtime of approximately 2400s, show only marginal improvements over SC baselines. Among multi-view diffusion methods, which are overall the strongest baseline family, MATLAT outperforms on most metrics, with the exception of the albedo CLIP score which is marginally behind MaterialAnything [10].

Qualitative results. Figure 3 presents representative outputs of MATLAT across diverse prompts and geometries, spanning glossy food, mixed-material footwear, patterned upholstery, and wood-and-metal props. Across these cases, MATLAT preserves sharp local boundaries and plausible albedo/roughness/metallic interactions, yielding coherent relighting and materially distinct parts within a single asset.

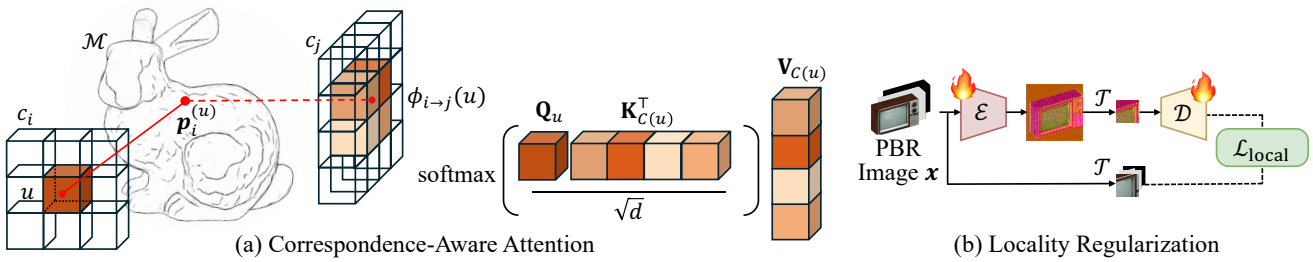
Ablation. We isolate the two core claims of the paper: (i) latent adaptation must preserve the pretrained RGB distribution, and (ii) CAA requires locality-preserved latents to be effective.

We first compare encoder-side adaptation strategies in Table 2. All variants share the same diffusion architecture with both CAA and locality regularization enabled; as Frozen VAE does not require encoder fine-tuning, we allocate additional diffusion training steps to match overall compute. Frozen VAE [9] degrades shaded FID ($3.083 \rightarrow 3.419$) and albedo FID ($4.599 \rightarrow 4.926$), confirming that encoding roughness and metallic channels with zero-padding introduces a distribution shift that hinders downstream generation. LayerDiffuse-style residual prediction with identity loss [33] and Orchid-style direct prediction with KL alignment [14] each recover part of the improvement but neither matches MATVAE, which combines residual initialization with distributional regularization in a single stochastic posterior.

We next isolate the multi-view consistency components. Removing locality regularization ($\mathcal{L}_{\text{local}}$) while keeping CAA hurts both appearance and consistency (albedo FID $4.599 \rightarrow 5.873$, c-PSNR $21.934 \rightarrow 19.437$): without spatially local latent–pixel mappings, CAA propagates features across unrelated regions. Conversely, removing CAA while keeping $\mathcal{L}_{\text{local}}$ preserves appearance metrics but still lowers c-PSNR to 18.687, as the model must infer correspondences implicitly via dense attention. MATLAT, which combines both, achieves the highest c-PSNR without degrading other metrics, confirming that geometry-aware attention and locality regularization are complementary.

4. Conclusion

We presented MATLAT, a two-stage pipeline that learns a material latent space for text-guided PBR texture generation by leveraging pretrained RGB diffusion priors. Our ap-



(a) Correspondence-Aware Attention (b) Locality Regularization

Figure 2. **Locality-preserved multi-view generation.** (a) Correspondence-aware attention (CAA) restricts feature exchange to geometrically corresponding tokens across views. (b) Our locality regularizer enforces patch-wise reconstruction so that decoded pixels depend on spatially aligned latent tokens. Together, they turn the adapted material latent space into a reliable representation for multi-view PBR texture synthesis.

Table 1. **Main quantitative comparison on held-out Objaverse-XL meshes.** Best scores are **bold**, and the runner-up scores are underlined. We highlight shaded/albedo FID (overall fidelity), roughness/metallic RMSE (material faithfulness), c-PSNR in Tab. 2 (cross-view consistency), and runtime as the primary metrics. MATLAT improves the strongest multi-view baseline MaterialMVP by over 50% in both shaded and albedo FID while keeping similar runtime and competitive material-map accuracy.

Method	Type	Shaded			Albedo			Rough.	Metal.	Time ↓
		FID_{CLIP} ↓	KID ↓	CLIP ↑	FID_{CLIP} ↓	KID ↓	CLIP ↑	RMSE ↓	RMSE ↓	
MeshGen [3]	SC	8.637	11.105	0.282	11.322	16.193	0.282	**0.143**	0.201	195s
TexGaussian [29]	SC	6.025	3.571	0.301	12.119	9.381	0.299	0.145	0.243	73s
Paint-it [32]	SDS	8.547	5.382	0.309	13.063	10.665	0.299	0.168	0.200	1260s
DreamMat [35]	SDS	5.422	2.668	0.311	9.621	6.002	0.311	0.167	0.165	2400s
FlashTex [6]	SDS	7.119	5.354	0.305	12.320	10.441	0.298	**0.143**	0.186	285s
MaterialAnything [10]	MV	6.582	5.287	0.312	12.691	9.325	**0.317**	0.233	0.200	500s
MaterialMVP [9]	MV	6.309	5.744	0.294	9.630	8.811	0.290	0.175	**0.133**	35s
MATLAT **(Ours)**	MV	**3.083**	**1.327**	**0.318**	**4.599**	**1.574**	0.314	0.158	0.134	**34s**

Figure 3. **Representative qualitative results of MATLAT.** We show generated shaded renderings for diverse prompts and geometries. MATLAT produces coherent PBR textures across glossy, metallic, wooden, and mixed-material assets while preserving fine local details and consistent relighting.

Table 2. **Detailed ablation summary.** The first block compares latent adaptation strategies, and the second isolates locality regularization and CAA. We report shaded FID, albedo FID, and c-PSNR; lower is better for FID, and higher is better for c-PSNR.

Variant	Shaded FID	Albedo FID	c-PSNR
Frozen VAE [9]	3.419	4.926	19.869
Res. Pred. + $\mathcal{L}_{\text{id}}$	3.210	4.871	20.977
Direct Pred. + $\mathcal{L}_{\text{reg}}$	3.192	4.768	21.468
w/o $\mathcal{L}_{\text{local}}$	3.419	5.873	19.437
w/o CAA	3.110	4.732	18.687
MATLAT	**3.083**	**4.599**	**21.934**

proach addresses the domain gap introduced by roughness and metallic channels through MATVAE, which incorporates PBR material information via residual posterior prediction while regularizing the learned distribution toward the pretrained latent space. Combined with locality regularization that maintains spatial alignment between latent tokens and decoded pixels, correspondence-aware attention reliably enforces multi-view consistency in pixel space. Experiments on Objaverse-XL show that MATLAT achieves superior fidelity and multi-view consistency over strong baselines at practical runtime, with ablations confirming that both distributional alignment and locality preservation are essential.

References

[1] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. In *ICCV*, 2023. 1

[2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 1

[3] Zilong Chen, Yikai Wang, Wenqiang Sun, Feng Wang, Yiwen Chen, and Huaping Liu. Meshgen: Generating pbr textured mesh with render-enhanced auto-encoder and generative data augmentation. In *CVPR*, 2025. 1, 2, 3, 4

[4] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. In *NeurIPS*, 2023. 2, 3

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023. 2

[6] Kangle Deng, Timothy Omernick, Alexander Weiss, Deva Ramanan, Jun-Yan Zhu, Tinghui Zhou, and Maneesh Agrawala. Flashtex: Fast relightable mesh texturing with lightcontrolnet. In *ECCV*, 2024. 3, 4

[7] Fan Fei, Jiajun Tang, Fei-Peng Tian, Boxin Shi, and Ping Tan. Pacture: Efficient pbr texture generation on packed views with visual autoregressive models. *arXiv preprint arXiv:2505.22394*, 2025. 2, 3

[8] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH*, 2023. 1

[9] Zebin He, Mingxin Yang, Shuhui Yang, Yixuan Tang, Tao Wang, Kaihao Zhang, Guanying Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, and Wenhan Luo. Materialmvp: Illumination-invariant material generation via multi-view pbr diffusion. In *ICCV*, 2025. 2, 3, 4

[10] Xin Huang, Tengfei Wang, Ziwei Liu, and Qing Wang. Material anything: Generating materials for any 3d object via diffusion. In *CVPR*, 2024. 2, 3, 4

[11] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, 2023. 1

[12] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1

[13] Jaihoon Kim, Juil Koo, Kyeongmin Yeo, and Minhyuk Sung. Synctweedies: A general generative framework based on synchronized diffusions. In *NeurIPS*, 2024. 1

[14] Akshay Krishnan, Xinchen Yan, Vincent Casser, and Abhijit Kundu. Orchid: Image latent diffusion for joint appearance and geometry generation. In *ICCV*, 2025. 2, 3

[15] Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-Hang Lee, and Peng Yuan Zhou. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In *ECCV*, 2024. 1

[16] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 2

[17] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. 3

[18] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. In *NeurIPS*, 2023. 3

[19] Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, 2023. 1

[20] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1, 2

[21] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *SIGGRAPH*, 2023. 1

[22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2

[23] Zeyi Sun, Tong Wu, Pan Zhang, Yuhang Zang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Bootstrap3d: Improving multi-view diffusion model with synthetic data. In *ICCV*, 2025. 3

[24] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023. 2

[25] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *ICLR*, 2023. 1

[26] Yitong Wang, Xudong Xu, Li Ma, Haoran Wang, and Bo Dai. Boosting 3d object generation through pbr materials. In *SIGGRAPH*, 2024. 1

[27] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 2

[28] Xiaokang Wei, Bowen Zhang, Xianghui Yang, Yuxuan Wang, Chunchao Guo, Xi Zhao, and Yan Luximon. Pbr3dgen: A vlm-guided mesh generation with high-quality pbr texture. *arXiv preprint arXiv:2503.11368*, 2025. 2

[29] Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao, Errui Ding, and Zhouhui Lian. Texgaussian: Generating high-quality pbr material via octree-based 3d gaussian splatting. In *CVPR*, 2025. 3, 4

[30] Kyeongmin Yeo, Jaihoon Kim, and Minhyuk Sung. Stochsync: Stochastic diffusion synchronization for image generation in arbitrary spaces. In *ICLR*, 2025. 1

[31] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang

Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024. 1

[32] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *CVPR*, 2024. 3, 4

[33] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. In *SIGGRAPH*, 2024. 2, 3

[34] Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, and Hsin-Ying Lee. Scenewiz3d: Towards text-guided 3d scene composition. In *CVPR*, 2024. 1

[35] Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, and Xiaogang Jin. Dreammat: High-quality pbr material generation with geometry- and light-aware diffusion models. In *SIGGRAPH*, 2024. 3, 4

[36] Lingting Zhu, Jingrui Ye, Runze Zhang, Zeyu Hu, Yingda Yin, Lanjiong Li, Jinnan Chen, Shengju Qian, Xin Wang, Qingmin Liao, and Lequan Yu. Muma: 3d pbr texturing via multi-channel multi-view generation and agentic post-processing. *arXiv preprint arXiv:2503.18461*, 2025. 2

[37] Shenhao Zhu, Lingteng Qiu, Xiaodong Gu, Zhengyi Zhao, Chao Xu, Yuxiao He, Zhe Li, Xiaoguang Han, Yao Yao, Xun Cao, Siyu Zhu, Weihao Yuan, Zilong Dong, and Hao Zhu. Mcmat: Multiview-consistent and physically accurate pbr material generation. *arXiv preprint arXiv:2412.14148*, 2024. 2