VideoMatGen: PBR Materials through Joint Generative Modeling

Supplementary Material

Jon Hasselgren NVIDIA
Miloš Hašan NVIDIA
Zheng Zeng NVIDIA
Jacob Munkberg NVIDIA

Figure 1. Our text-to-PBR material generations (renderings in Blender) on a collection of TRELLIS.2-generated base geometry.

1. Supplemental material

1.1. Results on AI generated geometry

To illustrate that our method can complement an image→3D pipeline with high quality PBR material generation, we create 3D objects using TRELLIS.2 [6] (conditioned on images from our test set), and use the unmodified GLTF meshes in our pipeline. There are no significant mapping issues as shown in Fig. 1. Our method does not rely on the UV mapping except in the final projection to UVs, which is optional.

1.2. Frame concatenation vs. Compressed VAE

Implementation Following UniRelight [1], in our experiment with frame concatenation, we concatenate two encoded latent videos along the *frame* dimension of the input tensor of the DiT. The first video represents sixteen views of base color, the second video represents the corresponding sixteen views of height, roughness, and metallicity, packed into RGB images. Both videos are encoded with the Cosmos Tokenizer, $\mathcal{E}$, using the image (keyframe) mode. To distinguish the two video segments, we leverage the *view* encoding used in the multi-view post training example of Cosmos [5]. Note that frame concatenation doubles the number of tokens (and inference time), which limits us to train with examples with 16 frames in a resolution of 768×768 pixels. In contrast, our proposed architecture using VAE_{pbr} is more memory-efficient, and we can train with videos with a spatial resolution of 1024×1024 pixels. We implemented both variants of joint prediction for our material prediction task to evaluate their quality. For positional encoding in the frame concatenation version, we leverage the *view* encoding used in the multi-view post training example of Cosmos [5], to distinguish the two video segments.

Results In Fig. 2 we show examples of the generated materials for a frame-concatenation variant (which doubles the number of tokens and inference time) vs. our proposed VAE_{pbr}. In Tab. 1 we present metrics comparing the two series, using the same evaluation protocol from Section 4.1 of the main paper.

Table 1. Quantitative metrics for text-to-material generation. We compare two variants of joint prediction, frame concatenation (FCat) which doubles the number of video frames/latent tokens, and our version with VAE_{pbr}.

Method	CLIP-FID (↓)	CMMD (↓)	LPIPS (↓)
VideoMatGen (VAE_{pbr})	**5.638**	**0.035**	**0.126**
VideoMatGen (FCat)	5.712	0.039	0.129

1.3. Text-alignment metric

Our prompts (generated by Qwen2.5-VL-7B) exceeds CLIP's limitation of 77 tokens. Therefore, we report BLIP scores [3] below (Using the `Salesforce/blip-itm-base-coco` model). We use 16 views of each of our 32 test examples. The score is a binary classification probability ($\in [0, 1]$). Here, the reference means views of the test set materials with captions from Qwen2.5-VL-7B.

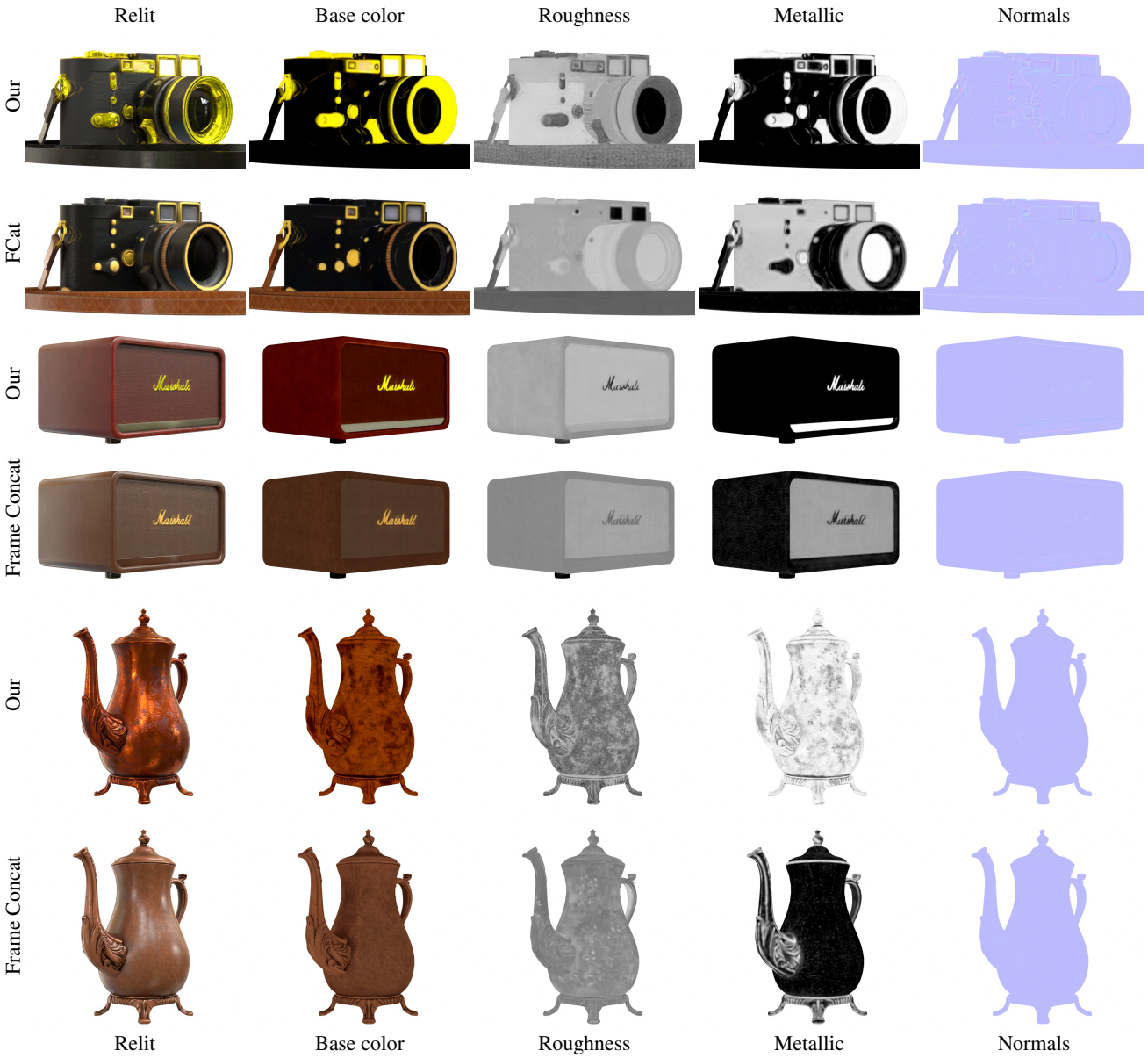

Figure 2. We compare two versions of joint prediction. Our version uses the latent space of VAE_{pbr} and predicts a video with 16 frames. Frame concatenation doubles the number of tokens by treating the material prediction as diffusing a video with twice the number of frames [16× base color, 16× HRM]. The resulting quality is similar, arguably with more coherent material predictions in our approach. Please zoom to see the details.

Method	BLIP Mean Image-Text Matching scores
VideoMat	0.9015 ± 0.2474
VideoMatGen	0.9271 ± 0.2025
Reference	0.9339 ± 0.1944

1.4. Rendering loss experiment

We experimented with including a *rendering loss* when finetuning the diffusion model, where we leverage a *differentiable* version of split sum shading [4]. In each training iteration, we load a random HDR probe, and evaluate image-based shading for each view using a Lambertian term and a Cook-Torrance microfacet specular shading approximated by split sum. However, we did not see improved results compared to only training with the denoising score matching loss. In our setting, given that our predicted latent includes all material modalities, a rendering loss is only representing a different weighting factor of each modality. This is in contrast to methods which predict each material modality in separate networks [2], where a rendering loss

is critical to align the modalities. Note also that an image space loss term requires more memory during training, as it is computed *after* VAE decoding, and hence, requires backpropagation through the VAE decoder, $\mathcal{D}_{\text{pbr}}$, to update the DiT weights.

1.5. Prompts

In this subsection, we include the text prompts for the examples shown in the main paper. For the full set of 32 prompts used in our test set, please refer to the image viewer.

Diver: "A vintage diving helmet with a worn, copper-colored finish rotates against a black background. The helmet features multiple circular windows for visibility, with one prominently positioned on the front. It has a sturdy, metallic construction with visible bolts and rivets, giving it a rugged and industrial appearance. The helmet's design includes a curved neck guard and a handle on top, suggesting it was used for deep-sea exploration. The surface shows signs of age and wear, with patches of rust and discoloration, adding to its historical charm. A close-up shot from various angles highlights the intricate details and craftsmanship of this classic diving gear."

Robot: "A quirky, retro-futuristic robot with a boxy head and a small screen displaying green code. It has two large, striped arms and legs, each ending in simple, rounded feet. The robot's body is adorned with various mechanical components, including a circular antenna on top and a small, round sensor on one side. It moves slowly, swaying slightly as it walks, giving off a playful and endearing vibe. The background is plain black, emphasizing the robot's unique design and movements. A medium shot capturing the robot's full body as it navigates through space."

Shed: "A small wooden house model rotates against a bright background. Light, bright, vivid colors. The structure is made of weathered wooden planks, with a sloped roof covered in corrugated metal sheets. The house features two small windows, one on each side, and a small door with a window above it. A small awning with a striped pattern hangs over the entrance. The model is detailed with visible joints and supports, giving it a rustic and handmade appearance. The camera pans around the house, showcasing its various angles and features."

Lantern: "A vintage-style lantern rotates against a black background. The lantern is made of metal with a weathered, rustic appearance, featuring a white glass globe protected by a wire cage. The handle is coiled and attached to the side, and the base has a textured, ribbed design. The lantern's intricate details and sturdy construction suggest it is designed for practical use, possibly for camping or outdoor activities. A medium shot captures the lantern from various angles as it spins slowly."

Motorcycle: "A sleek black motorcycle rotates against a bright background, Light, bright, vivid colors, showcasing its intricate design and polished chrome details. The bike features a classic retro aesthetic with a rounded front fender, a prominent headlight, and a comfortable-looking black seat. The handlebars are equipped with round mirrors, and the engine is exposed, revealing a robust and powerful build. The wheels have spoked rims, adding to its vintage charm. The motorcycle's design is highlighted from multiple angles as it spins, emphasizing its elegant lines and craftsmanship. A close-up shot from various perspectives."

References

[1] Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. UniRelight: Learning Joint Decomposition and Synthesis for Video Relighting, 2025. 1

[2] Peter Kocsis, Lukas Höllein, and Matthias Nießner. IntrinsiX: High-Quality PBR Generation using Image Priors, 2025. 2

[3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[4] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8280–8290, 2022. 2

[5] NVIDIA. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 1

[6] Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and Compact Structured Latents for 3D Generation. *Tech report*, 2025. 1