MeSS: City Mesh-Guided Outdoor Scene Generation with Cross-View Consistent Diffusion — Supplementary Material —

Xuyang Chen[1,2], Zhijun Zhai[3], Kaixuan Zhou[1], Zengmao Wang[3], Jianan He[1], Dong Wang[1], Yanfeng Zhang[1], Mingwei Sun[1,3], Xuqin Wang[1,2], Rüdiger Westermann[2], Tao Wu[1], Konrad Schindler[4], and Liqiu Meng[2]

[1] Huawei Hilbert Research Center
[2] Technical University of Munich, Munich, Germany
[3] Wuhan University, Wuhan, China
[4] ETH Zurich, Zurich, Switzerland

This supplementary material provides additional evidence and details for the main paper. Secs. 1 to 5 consolidate the additional experiments and analyses; the remaining sections give architecture, texturing-baseline, and stylization details.

1 Compute Budget

Tab. 1 reports per-stage wall-clock and peak VRAM on a single RTX PRO 6000 for a 200-frame (200 m) trajectory. Both stages are autoregressive, so cost scales roughly linearly with trajectory length and remains dominated by Stage II; peak memory is unchanged for longer paths since sub-sequences are processed independently. Training each ControlNet takes ≈10 h on eight RTX A6000 GPUs.

Table 1. Per-stage wall-clock and peak VRAM on a single RTX PRO 6000 for a 200-frame / 200 m trajectory.

Stage	**Wall time**	**Peak VRAM**
Stage I	3 min 00 s	5.6 GB
Stage II	≈17 min 10 s	≈24.7 GB
End-to-end	≈20 min 10 s	≈24.7 GB

2 Cross-View Consistency and Novel-View Synthesis

Cross-view LPIPS alone does not verify whether the *same* surface keeps a consistent appearance across overlapping viewpoints. We therefore evaluate 12 shared trajectories against VistaDream[†], the stronger matched baseline (Tab. 2), using three additional measures. MEt3R [2] quantifies multi-view consistency directly

on the rendered surface; KPM counts LoFTR keypoint inliers retained after RANSAC across cross-view offsets; and held-out novel-view PSNR, SSIM, and LPIPS are measured against UE5 ground truth. MeSS improves on all five metrics. We report held-out novel-view synthesis in place of FVD because MeSS emits a re-renderable 3DGS field rather than a video, so geometry-consistent novel-view rendering—not video temporal statistics—is the operative test of consistency.

Table 2. Cross-view consistency and held-out novel-view synthesis against the matched baseline VistaDream† (12 shared trajectories).

	MEt3R↓	KPM↑	PSNR↑	SSIM↑	LPIPS↓
VistaDream†	0.155	1310	12.49	0.327	0.568
MeSS (Ours)	**0.129**	**1382**	**13.92**	**0.495**	**0.371**

3 Real-World LoD3: Comparison with a Video-Diffusion ControlNet

The main paper shows that MeSS transfers zero-shot to the real LoD3 mesh TUM2Twin [21]. Here we contrast it against Cosmos-Transfer2.5 [16], a strong video-diffusion ControlNet, conditioned on its own depth prior. MeSS generates facades that follow the LoD3 geometry and stay consistent across views, whereas Cosmos-Transfer2.5 preserves only a coarse layout with visibly wrong facade detail. This is consistent with a *control-distribution mismatch*: Cosmos's depth ControlNet is trained on estimated depth from real video and is out-of-distribution on clean mesh-rendered depth. It does not by itself prove that retraining Cosmos's ControlNet on mesh-rendered priors could not close the gap—an experiment beyond the scope of this paper—so we position MeSS and video diffusion as complementary rather than as substitutes, and we state the corresponding claim in the abstract and Sec. 1 of the main paper in terms of this control-distribution rationale rather than as a categorical image-vs-video position.

4 AGInpaint Efficiency: N_g Early Stopping

We also probed an NMSE-thresholded early stop on the inner gradient-guidance loop of Alg. 1, in the spirit of exact-inversion samplers [10]. With a fixed threshold τ=10^{-3} across LCM denoising steps, the stop fires inconsistently—never at early steps (the residual stays above τ for all iterations) and immediately at late steps (halving the guidance budget)—so a single threshold does not yield a clean, monotone speedup. Searching nearby, we found instead that an SDEdit-style [14] 50% noise initialization with a reduced N_g=64 at the same 5 LCM

steps matches the FID/KID of the original N_g=100 setting at ≈30% lower wall-clock. We therefore adopt this setting as the default; *all numbers in the main paper are reported with it.*

5 Appearance Guided Sampling Algorithm

As noted in the main paper, intermediate novel views rendered from the coarse Gaussian field exhibit *silhouettes*—regions of low rendered opacity that AGInpaint is tasked to fill. Fig. 1 illustrates these regions on two example novel views.

Fig. 1. Silhouettes on novel views. The ellipses highlight the silhouette regions—areas of low rendered opacity in the Gaussian field—that AGInpaint is tasked to fill.

Alg. 1 gives the complete AGInpaint procedure described in the main paper. To adaptively regulate the step size s_t, we scale the gradient proportionally to the ratio of the mean absolute magnitudes of the current noise estimate and the gradient, accelerating convergence.

Algorithm 1 Appearance Guided Sampling

Require: Target signal $\mathbf{y}_{\text{guide}}$, mask $\mathbf{M}_{\text{guide}}$, condition $\mathbf{c}$, steps N_g, learning rate lr

$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  $\hat{\boldsymbol{\epsilon}}_t^{(0)} \leftarrow \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{c}, t)$
  for $k = 0, \dots, N_g - 1$ **do**
    $\hat{\mathbf{y}}_t \leftarrow \mathcal{D}\left(\frac{\mathbf{z}_t - \sigma_t\, \hat{\boldsymbol{\epsilon}}_t^{(k)}}{\alpha_t}\right)$ ▷ Tweedie estimate
    $\mathbf{g}_t \leftarrow \nabla_{\mathbf{z}_t} \left\| \hat{\mathbf{y}}_t \odot \mathbf{M}_{\text{guide}} - \mathbf{y}_{\text{guide}} \odot \mathbf{M}_{\text{guide}} \right\|^2$
    $s_t \leftarrow lr \times \frac{\text{mean}(|\hat{\boldsymbol{\epsilon}}_t^{(k)}|)}{\text{mean}(|\mathbf{g}_t|)}$
    $\hat{\boldsymbol{\epsilon}}_t^{(k+1)} \leftarrow \hat{\boldsymbol{\epsilon}}_t^{(k)} - s_t\, \mathbf{g}_t$
  end for
  $\mathbf{z}_{t-1} \leftarrow \text{Denoise}(\mathbf{z}_t, t, \hat{\boldsymbol{\epsilon}}_t^{(N_g)})$
end for
return $\mathcal{D}(\mathbf{z}_0)$

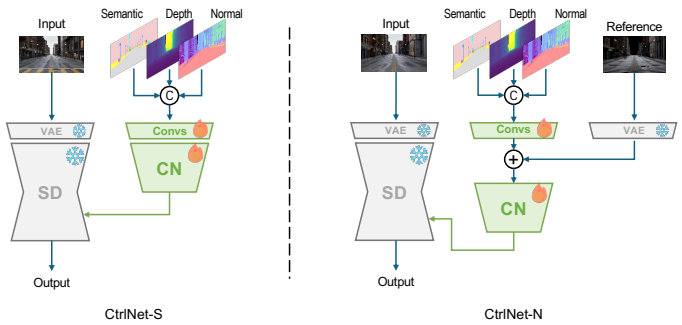

Fig. 2. Architecture of our two cascaded CtrlNets. **Left:** *CtrlNet-S* takes three geometric conditions (semantic, depth, normal) concatenated along the channel dimension. **Right:** *CtrlNet-N* additionally takes a reference image encoded by the frozen VAE; the latent features are element-wise added to the geometric condition features before injection into the ControlNet. Snowflake and flame icons denote frozen and trainable components, respectively.

6 Architecture of the Cascaded CtrlNets

The architecture of our two cascaded CtrlNets is illustrated in Fig. 2. Both build on the standard ControlNet [24] but are extended to accept multiple control signals simultaneously. The key difference of *CtrlNet-N* from *CtrlNet-S* lies in the processing of the reference image. Unlike the geometric control signals (depth, semantic and normal map), the reference image is first processed by the VAE encoder and then passed through several convolutional layers to align its feature dimension with the other control signal features before injection into the ControlNet.

7 Trial on Mesh Texturing Method – Text2Tex

Several authors have adopted diffusion-based generative models [18,3,11,20] for mesh texturing. TEXTure [19] and Text2Tex [5] use depth-to-image diffusion models to texture meshes through inpainting, but can suffer from visible seams and gradual amplification of texture artifacts. Other methods, including Fantasia3D [6], Decorate3D [9], SceneTex [4], and Latent Paint [15], use Score Distillation Sampling (SDS) [17,22]. Their costly test-time optimization limits them mainly to single objects or low-polygon indoor scenes, and they require training views that cover the relevant view field.

Text2Tex [5] textures a mesh with a depth-to-image model by generating and inpainting the visible texels from predefined camera views. The final texture is represented as a UV-mapped RGB image. We adapt it to outdoor scene texturing, where City Sample Project instances are substantially more complex than Objaverse objects [7]. To make a 100-meter city block tractable, we apply

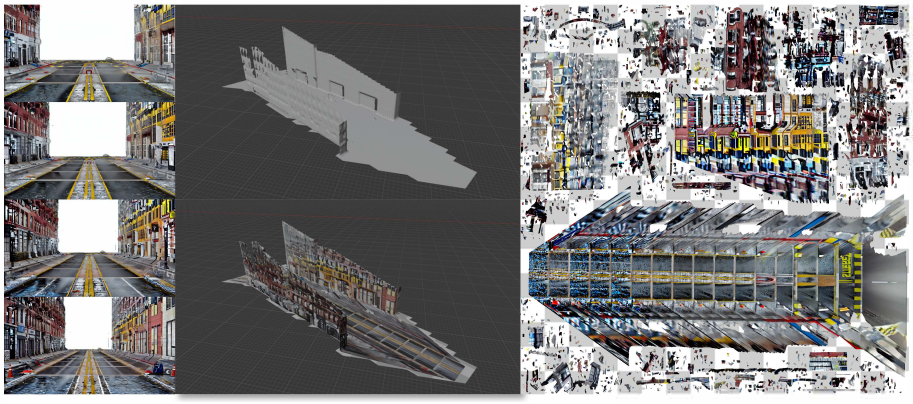

Fig. 3. Text2Tex trial on a crossroad mesh from the City Sample Project.

Quadric Edge Collapse Decimation [8] and retain only mesh faces visible from the camera views (the second column in Fig. 3). We replace the original inpainting module with ours, extend the texture while moving the camera backwards, and increase the texture resolution from 3000×3000 to 4096×4096. Several randomly sampled views are then used to refine the texture. As shown in Fig. 3, seams remain between successively inpainted regions, indicating limited scalability to large outdoor meshes.

8 Distributing Flattened Gaussian Surfels On Mesh Surfaces

Our approach for surfel generation draws conceptual inspiration from WonderWorld [23], where depth and surface normals are predicted for novel views. However, since we directly utilize mesh geometry, we bypass this prediction step by retrieving accurate depth and normal information through rasterization.

Here, Gaussian surfels are parameterized with position $\mathbf{p}$, orientation in quaternion form $\mathbf{q}$, scales in orthogonal directions $\mathbf{s} = [s_x, s_y, \epsilon]$, opacity o and RGB color $\mathbf{c}$. The Gaussian kernel at any spatial position $\mathbf{x}$ is given by:

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{p})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{p})\right), \tag{1}$$

where the covariance matrix $\mathbf{\Sigma}$ encodes the shape and orientation of the surfel and is defined as:

$$\mathbf{\Sigma} = \mathbf{Q} \cdot \mathrm{diag}(s_x^2, s_y^2, \epsilon^2) \cdot \mathbf{Q}^T, \tag{2}$$

with $\mathbf{Q}$ derived from the quaternion $\mathbf{q}$. Our renderer uses the rasterization and alpha compositing process of 3D Gaussian Splatting (3DGS) [12].

Given an image $\mathbf{I}$ of resolution $H \times W$, we construct $H \times W$ Gaussian surfels. For each surfel, the position $\mathbf{p}$ is directly derived from its corresponding 3D

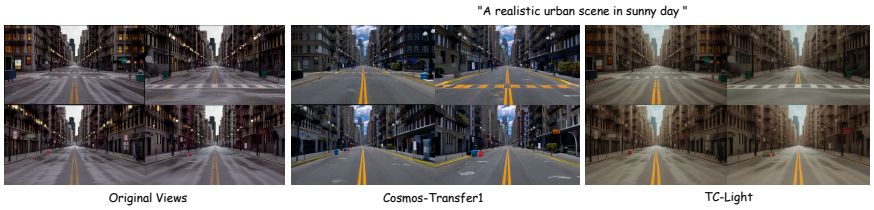

Fig. 4. Stylized rendering results with Cosmos-Transfer1 and TC-Light. Please zoom in to check for details.

position on mesh, while it inherits color $\mathbf{c}$ from the pixel's RGB value. We assume that all surfaces are Lambertian, so $\mathbf{c}$ is treated as view-invariant. To avoid rendering artifacts such as undersampling or holes when zooming in, the scales along x and y axes are set as $d/\sqrt{2}f_x$, $d/\sqrt{2}f_y$, respectively. And ϵ is kept small enough but still larger than zero to avoid numerical errors. The opacity property is set to a constant 0.9. Similar to position, we align the surfel's normal with the mesh surface normal at the corresponding pixel. Then we receive the rotation matrix:

$$\mathbf{Q}_z = \mathbf{n}, \quad \mathbf{Q}_x = \frac{\mathbf{u} \times \mathbf{n}}{\|\mathbf{u} \times \mathbf{n}\|}, \quad \mathbf{Q}_y = \frac{\mathbf{n} \times \mathbf{Q}_x}{\|\mathbf{n} \times \mathbf{Q}_x\|}, \tag{3}$$

with the global up-direction $\mathbf{u} = [0, 1, 0]^T$ used to ensure a consistent coordinate frame.

9 Stylized Rendering

The appearance of our generated scenes is strongly branded with the City Sample style, and simply tweaking the text prompt during view generation does not achieve stylization. We therefore customize the scene appearance through stylized rendering: given a camera path in the scene, we render a video and transfer it to a different style, either with the video relighting method TC-Light [13] or via SDEdit [14] with Cosmos-Transfer1 [1]. As depicted in Fig. 4, both complete this task by taking the original views as prior, with Cosmos-Transfer1 outperforming TC-Light in visual fidelity; the styled videos can be projected back to the Gaussian scene if desired.

The following prompts are leveraged for the inference of Cosmos-Transfer1 and TC-Light to achieve these results.

Full prompt for Cosmos-Transfer1:

```
A photorealistic scene of a quiet, empty urban street on a
sunny day with some clouds in the sky. The wide road with
yellow center lines stretches into the distance, flanked by
tall buildings with classic architecture and street lamps.
Some building fronts feature small potted shrubs, while a
```

few windowsills display potted flowers or trailing leafy plants, adding touches of greenery and charm to the scene. The view is centered and symmetrical, creating a peaceful and cinematic atmosphere. Cool moonlight and warm streetlamp glows softly illuminate the buildings and pavement, casting gentle shadows. There are no people or vehicles, enhancing the stillness. Urban details like trash bins and cones add realism. The composition draws the eye toward a distant vanishing point.

Full prompt for TC-Light:
A photorealistic depiction of a calm, empty urban street at noon on a sunny day.

Further style variants (night, rain, snow) can be obtained with the following prompts:

Cosmos-Transfer1 night time prompt:
A photorealistic scene of a quiet, empty urban street at night under a clear sky with scattered clouds. The wide road with yellow center lines stretches into the distance, flanked by tall buildings with classic architecture and street lamps. Some building fronts feature small potted shrubs, adding a touch of greenery to the scene. The view is centered and symmetrical, creating a peaceful and cinematic atmosphere. Cool moonlight and warm streetlamp glows softly illuminate the buildings and pavement, casting gentle shadows. Urban details like trash bins and cones add realism. The composition draws the eye toward a distant vanishing point.

TC-Light night time prompt:
A photorealistic depiction of a calm, empty urban street at night under a moonlit sky.

Cosmos-Transfer1 rainy day prompt:
A photorealistic scene of a quiet, empty urban street during daytime under a cloudy, rainy sky. The wide road with yellow center lines stretches into the distance, flanked by tall buildings with classic architecture. Rows of street trees line both sides of the road, their wet trunks and leaves glistening slightly under the diffused daylight. The rain-slick asphalt reflects the buildings, trees, and urban elements, with scattered puddles creating mirror-like surfaces. Soft daylight filters through the overcast sky, producing muted shadows and emphasizing the textures of wet pavement and facades. Urban details like

`trash bins and traffic cones add realism. The composition is centered and symmetrical, guiding the eye toward a distant vanishing point softened by mist and rain.`

Cosmos-Transfer1 snowy day prompt:
`A photorealistic scene of a quiet, empty urban street during daytime in snowy weather. The wide road stretches into the distance, flanked by tall buildings with classic architecture and street lamps dusted with snow. Trees lining the sidewalks are covered with snow-laden branches, adding to the serene winter atmosphere. Some building fronts feature small potted shrubs, adding subtle touches of greenery amid the white landscape. The view is centered and symmetrical, creating a peaceful and cinematic atmosphere. Soft daylight reflects off the snow-covered pavement and buildings, casting diffuse shadows. Urban details like trash bins and cones partially covered in snow add realism. The composition draws the eye toward a distant vanishing point.`

References

1. Alhaija, H.A., Alvarez, J., Bala, M., Cai, T., Cao, T., Cha, L., Chen, J., Chen, M., Ferroni, F., Fidler, S., et al.: Cosmos-transfer1: Conditional world generation with adaptive multimodal control. arXiv preprint arXiv:2503.14492 (2025)
2. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: MEt3R: Measuring multi-view consistency in generated images. In: CVPR (2025)
3. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
4. Chen, D.Z., Li, H., Lee, H.Y., Tulyakov, S., Nießner, M.: Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21081–21091 (2024)
5. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18558–18568 (2023)
6. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22246–22256 (2023)
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
8. Garland, M., Heckbert, P.S.: Surface simplification using quadric error metrics. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques. pp. 209–216 (1997)
9. Guo, Y., Zuo, X., Dai, P., Lu, J., Wu, X., Yan, Y., Xu, S., Wu, X., et al.: Decorate3d: text-driven high-quality texture generation for mesh decoration in the wild. Advances in Neural Information Processing Systems **36**, 36664–36676 (2023)

10. Hong, S., Lee, K., Jeon, S.Y., Bae, H., Chun, S.Y.: On exact inversion of DPM-solvers. In: CVPR (2024)
11. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9492–9502 (2024)
12. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
13. Liu, Y., Luo, C., Tang, Z., Li, Y., Yang, Y., Ning, Y., Fan, L., Zhang, Z., Peng, J.: Tc-light: Temporally coherent generative rendering for realistic world transfer. CoRR (2025)
14. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
15. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023)
16. NVIDIA: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)
17. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
18. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
19. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
20. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4713–4726 (2022)
21. TUM2Twin Project: TUM2Twin: Digital twin of TU Munich (lod2/lod3) (2025)
22. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems **36**, 8406–8441 (2023)
23. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: Wonderworld: Interactive 3d scene generation from a single image. arXiv preprint arXiv:2406.09394 (2024)
24. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)