Appendix

A. Datasets

From Tanks and Temples, we use images with the following names: Auditorium: 00007.png, Ballroom: 00001.png, Courtroom: 00001.png, Museum: 00015.png, Palace: 00019.png, Temple: 00001.png. The images were selected such that there are no humans in the images and such that they show a larger scene, instead of close-up captures.

B. Implementation Details

Our ControlNet follows [55], i.e., creates trainable copies of the transformer blocks and adds back features from zero-initialized layers. We did not tune this design specifically for our setting. The ControlNet trains for 5k iterations within a few hours on 8 A100 GPUs with a batch size of 8. We do not add any LoRAs. Inference takes ∼20s per image. 3DGS converges within minutes on a single A100 GPU. For the equirectangular to projective projection, we add a slight blur to avoid sharp masking edges. We found that the inpainting models can be highly sensitive to such edge-artifacts, and an overall higher inpainting quality when ensuring smooth transitions.

DimensionX. We use the official implementation for our experiments. Specifically, we generate two trajectories. Both trajectories start from the input image. One rotates left and the other one right, moving along a circular trajectory and facing inwards. We tuned the pipeline to the best of our abilities but could not extend it to perform a matching full rotation, since the generated sequences do not produce overlapping content when meeting opposite of the input image.

WonderJourney. We use the official code release for our experiments. Similarly to DimensionX, we found it best to generate two camera rotations on an inward-facing circle, left and right, starting from the input image. Since there is no metric scale, we tune the hyperparamters to the best of our ability to make the radius of the circle reasonable.
For both approaches, we extract metric poses after generating the trajectories using CUT3R [42] as it robustly works with long input sequences. We use the resulting images and poses with the same 3D reconstruction settings as for our approach.

B.1. 3D Reconstruction - implementation details

Gaussian Splatting Training. We train our Gaussian Splatting models using the Splatfacto implementation from NerfStudio [38]. Most training hyper-parameters are kept at their default values, matching the original GS implementation from [16], with the following changes:

- we reduce the training schedule from 30k to 5k iterations;
- we disable periodic opacity reset, and keep adaptive density control active from iteration 500 to iteration 2500;
- we set the degree of the spherical harmonic functions that model view-dependent colors to 1;
- we increase batch size from 1 to 2.

	BRISQUE↓		NIQE↓		Q-Align↑	
	metric3D	MoGE	metric3D	MoGE	metric3D	MoGE
WL	45.0	**41.1**	5.9	**5.6**	3.2	**3.5**
TaT	41.3	**39.5**	5.4	**5.3**	3.2	**3.4**

Table 4. **Depth Estimator Ablation:** Quality metrics for images rendered from the 3DGS representation at resolution 1024×1024 pixels, using a field of view of 60 degrees.

Trainable Distortion. As mentioned in Sec. 3.3 in the main document, the function $f(\mathbf{p}, \mathbf{c}_I; \theta)$ that models the point sampling offsets in our trainable distortion model, is implemented as a small MLP. Specifically, we use three linear layers with 128 hidden dimensions and ReLU activations, except for the last one that uses tanh. Before feeding them to the MLP, the input positions $\mathbf{p}$ are encoded into 32-dimensional harmonic embeddings akin to NeRF [22]. The per-image codes $\mathbf{c}_I$ have 32 channels, and the grid at which we evaluate f has a resolution of 128×128. We provide qualitative results with and without adding grid distortion in Fig. 10.

C. Qualitative Ablation Studies

As our heuristic for panorama synthesis is difficult to quantify, we instead provide extensive qualitative results in Fig. 7 and Fig. 8. Similar to the results from the main paper, Ad-hoc synthesis yields pleasantly looking, yet geometrically incorrect results. Sequential synthesis often generates reasonable panorama images, too. However, we observe in multiple instances that the floor does not match the scene well. The anchored approach overall results in the best panorama images. We further compare results for using the image caption as prompt, finding that this often simply repeats characteristics of the input image in all directions. While the coarse prompt from a vision-language model improves on the duplications, the sky and ground-specific prompts used in the anchored approach further improve the results qualitatively. In Fig. 10, we qualitatively show the effect of adding a trainable image distortion. While both approaches generate good results overall, grid distortion preserves more fine-grained details, e.g., for the branches of the trees.

Ablation on Depth Estimator. Tab. 4 compares results for our pipeline using Metric3D against MoGE for estimating the depth. Our approach is robust towards the depth estimation method but MoGE consistently achieves slightly better results. We hence choose MoGE as default depth estimator.

Text-to-World Generation and Comparison to DreamScene360. Our approach can be extended to text-to-world

Figure 7. **Qualitative Ablation For Panorama Synthesis:** We compare different heuristics for progressive panorama synthesis and prompt generation. From left to right: Ad-Hoc, Sequential, Anchored, Prompt is caption generated from the input image, Non-specific prompt from vision-language model.

synthesis by generating an image from a given text prompt with a pre-trained T2I generator before passing the result to our pipeline. We compare our results against DreamScene360 [56], which does not support image input but only text input. The qualitative comparison in Fig. 9 suggests that our approach achieves higher fidelity of the 3D scene and adheres more closely to the prompt, e.g., DreamScene360 misses the river and moon (first column), and the spaghetti (third column).

Discussion About WorldLabs.ai Results. WorldLabs.ai recently introduced a solution for the single image to 3D generation task and presented their results on their blog. Their approach is neither publicly disclosed nor are their resulting models available for a direct, side-by-side compar-

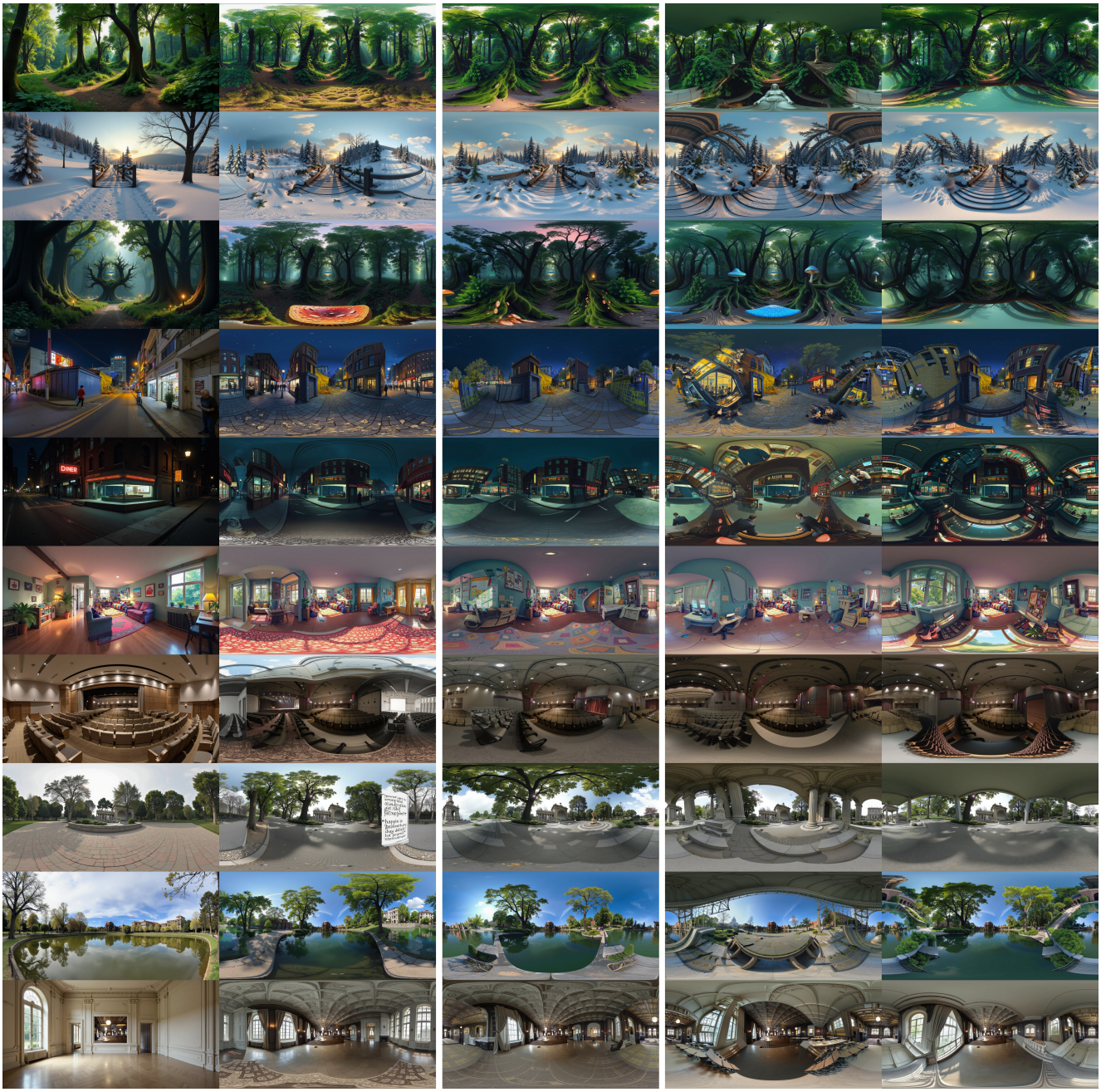
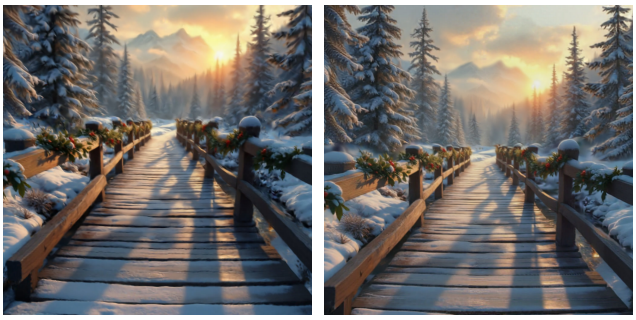


Figure 8. **Qualitative Ablation For Panorama Synthesis:** We compare different heuristics for progressive panorama synthesis and prompt generation. From left to right: Ad-Hoc, Sequential, Anchored, Prompt is caption generated from the input image, Non-specific prompt from vision-language model.

ison. We observe that our results look sharper and provide a better sense of continuity. We also observe that our generated back side views are better aligned with respect to the style of the input image.

Failure Cases. For challenging input images, e.g. artworks with a distinct style, we observe that the initial outpainting can fail to faithfully extend the image w.r.t. the style, see Fig. 11. This results in visible border artifacts around the input image. Further, we observe that occasionally the spatial layout of the panorama can be imperfect, even when using our anchored heuristic. While all synthesis results in this work were generated using a single seed, such failure cases could also be addressed with re-sampling or user edits on the prompt.

Figure 9. **Comparison to DreamScene360:** Rendered images from the generated 3DGS comparing Ours (upper row) and DreamScene360 (lower row). The prompts are below the images. Best viewed with zoom.



(a) w/o distortion (b) with distortion

Figure 10. **Trainable Image Distortion:** Integrating a trainable image distortion results in sharper reconstructions, compare e.g. the details on the tree branches.

Figure 11. **Failure Case For Panorama Synthesis:** For challenging input images, e.g. artworks with a distinct style, the inpainting model can struggle to adapt the global style of the panorama, resulting in visible border artifacts around the input image. Further, even with our anchored heuristic, the spatial layout of the panorama can sometimes be imperfect.