A Appendix

A.1 Experimental Settings

A.1.1 Training datasets and details

1) For perception tasks, we use two synthetic datasets covering both indoor and outdoor scenes: **A)** Hypersim (Roberts et al., 2021) is a photorealistic synthetic dataset featuring 461 indoor scenes. We use the official training split, which contains approximately 54K samples. After filtering out incomplete samples, around 39K samples remain, all resized to 576×768 for training. **B)** Virtual KITTI 2 (Cabon et al., 2020) is a synthetic street-scene dataset with five urban scenes under various imaging and weather conditions. We utilize four of these scenes for training, comprising about 20K samples. We set the far plane at 80m. **2)** For generation tasks, we filter and randomly select 1M samples from MultiGen-20M (Qin et al., 2024), ensuring a minimum edge length of 768. We utilize the existing annotations in MultiGen-20M, including *normal, canny, sketch, human pose, and blur*, while re-annotating *depth* using GeoWizard (Fu et al., 2025). This re-annotation is necessary because the original *depth* annotations are not suitable for unified generation and perception tasks in diffusion-based models. Discriminative tools like Depth Anything Series (Yang et al., 2024a;b) and MiDaS (Ranftl et al., 2020) typically output inverse normalized depth, whereas diffusion-based depth estimation models output non-inverse normalized depth. Therefore, we employ GeoWizard, a diffusion-based depth estimation model, to ensure compatibility and consistency for the *depth* annotations.

During training: **1)** For perception dataset, we use only the *original image* as the condition. **2)** For generation dataset, we select *normal, canny, sketch, human pose, blur, or NONE* as the condition with equal probability, where NONE represents an empty input for joint-generation tasks.

A.1.2 Evaluation datasets and metrics

1) For perception tasks: **A)** For zero-shot affine-invariant depth estimation, we evaluate UniGP on NYUv2 (Silberman et al., 2012), ScanNet (Dai et al., 2017), KITTI (Geiger et al., 2013) and ETH3D (Schops et al., 2017) using absolute mean relative error (*AbsRel*), and also report $\delta 1$ and $\delta 2$ values. **B)** For surface normal estimation, we employ NYUv2, ScanNet, iBims-1 (Koch et al., 2018) and Sintel (Butler et al., 2012) datasets, reporting mean angular error (*m.*) as well as the percentage of pixels with an angular error below 11.25° and 30°. **2)** For generation tasks: **A)** We evalute UniGP on 5K samples from the MSCOCO (Lin et al., 2014) validation set using Fréchet Inception Distance (FID) (Heusel et al., 2017) and CLIP(Radford et al., 2021) Text-Image Similarity. **B)** We also perform task-specific metric comparisons for controllable generation and joint-generation. Specifically, for joint-generation task, we recompute the depth and normal from the generated RGB image and then calculate the RMSE against the generated depth and normal. For the depth (or normal) to image task, we calculate the RMSE between the generated depth (or normal) and the given depth (or normal).

A.1.3 How to handle multiple input conditions

For multiple conditions, we duplicate DUGP, apply each DUGP to each condition, and perform the interaction with the backbone in the same way as discussed in Sec. 3.2.1, achieved by a single forward pass for all conditions.

A.2 Limitations and Social Impacts

Limitations. Although our model can simultaneously output *image*, *depth*, and *normal*, its size increases from 2B to 3B. During inference with 20 steps at FP16 precision on an NVIDIA V100, the runtime increases from 5.07s for the base model to 9.54s. It is worth noting that the increase in inference time and parameters is not linearly related, as we added extra computations.

Social Impacts. Our model equips the SOTA DiT with unified perception and generation capabilities but carries potential misuse risks, similar to other generative models. We emphasize the importance of responsible use and transparency.

A.3 More Results

A.3.1 Generalization to other perception tasks

UniGP builds on diffusion model, transforming vision tasks into conditional generation problems, enabling arbitrary perception tasks. We trained UniGP on additional perception tasks, including semantic segmentation, albedo estimation, and shading estimation, with qualitative results in Fig. 7.

Figure 7: UniGP can be adapted to more perception tasks.

A.3.2 More qualitative results

More qualitative generation results shown in Fig. 9 and more qualitative controllable perception results shown in Fig. 8.

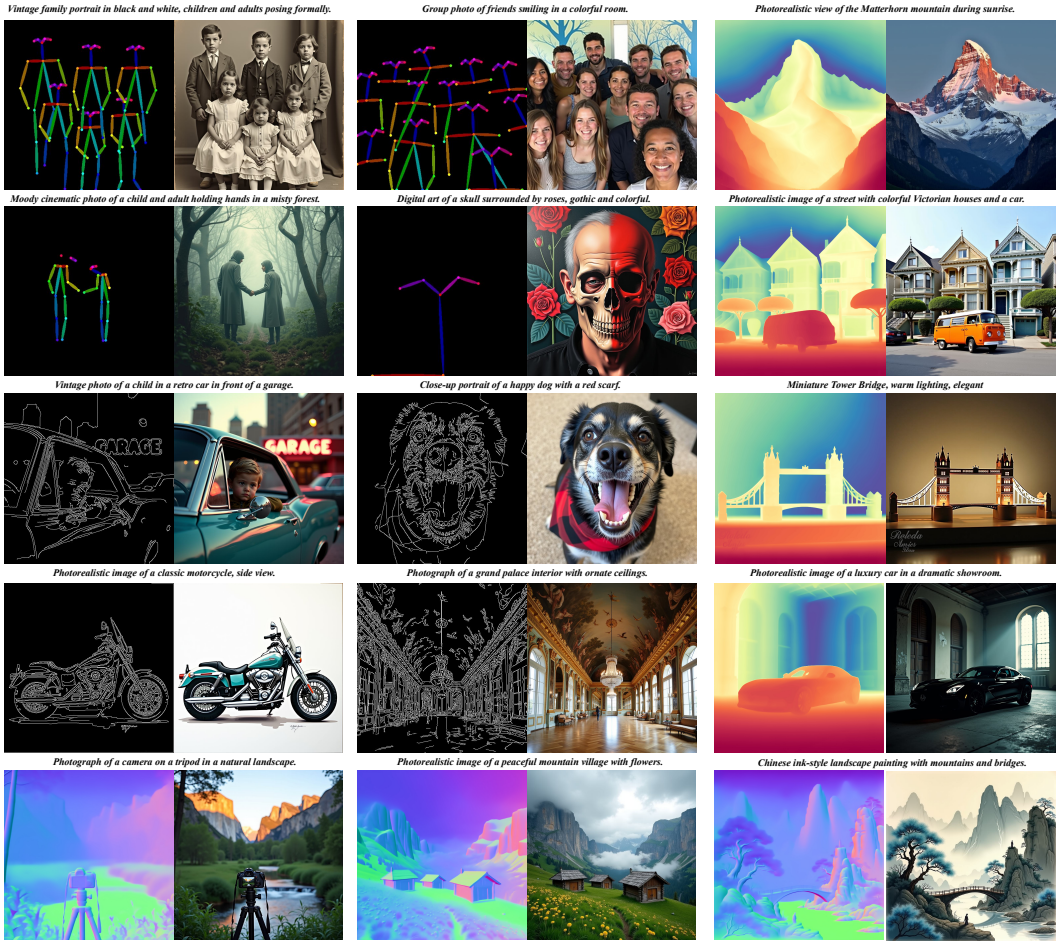

Figure 8: **Additional Qualitative Generation Results**.

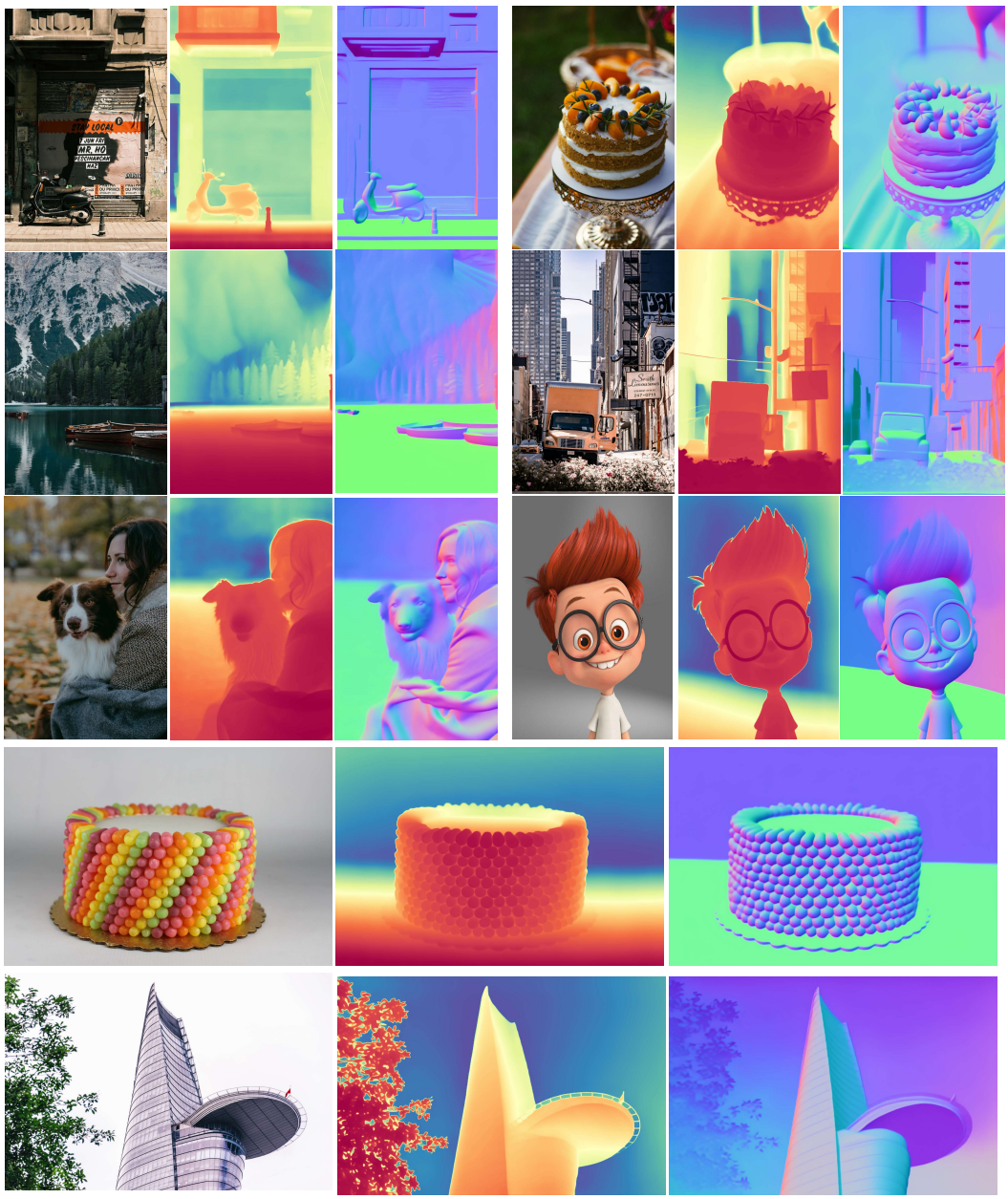

Figure 9: **Additional Qualitative Perception Results**.