# Generative Depth Completion

Jose Cuaran, Asher Mai, Kulbir Ahluwalia, Junzhe Wu
University of Illinois at Urbana-Champaign

## Motivation

- Affordable Low-Cost RGB-D Sensors: Applications in robotic navigation, perception, 3D reconstruction
- Challenges in Raw Depth Data: Noise, Occlusions, Missing values, Sensor or Object specific artifacts
- Limitations of Monocular Depth Estimation: While learning-based monocular methods yield smooth, visually appealing depth maps, they lack metric accuracy and scale due to the absence of real depth signals.
- Integrating Sensor Data with Generative Models: Combine visual features obtained from pretrained models with the geometric guidance from noisy but real sensor depth for conditioning a diffusion model for more accurate depth maps.

## Method

**Training.** Following Marigold's training pipeline [1], we adapt the pre-trained U-Net from Stable Diffusion V2 as the latent denoiser, and take the frozen VAE to encode input image, input sensor depth map, and ground-truth depth map. We concatenate all three latent codes into a single input along the feature dimension. The input channels of the latent denoiser are tripled. To preserve the pretrained weights, we duplicate the weight tensor of the input layer 3 times and divide the values by three, similar to [1]. The latent denoiser is trained to predict the added Gaussian noise with uniformly sampled noise level $t$.

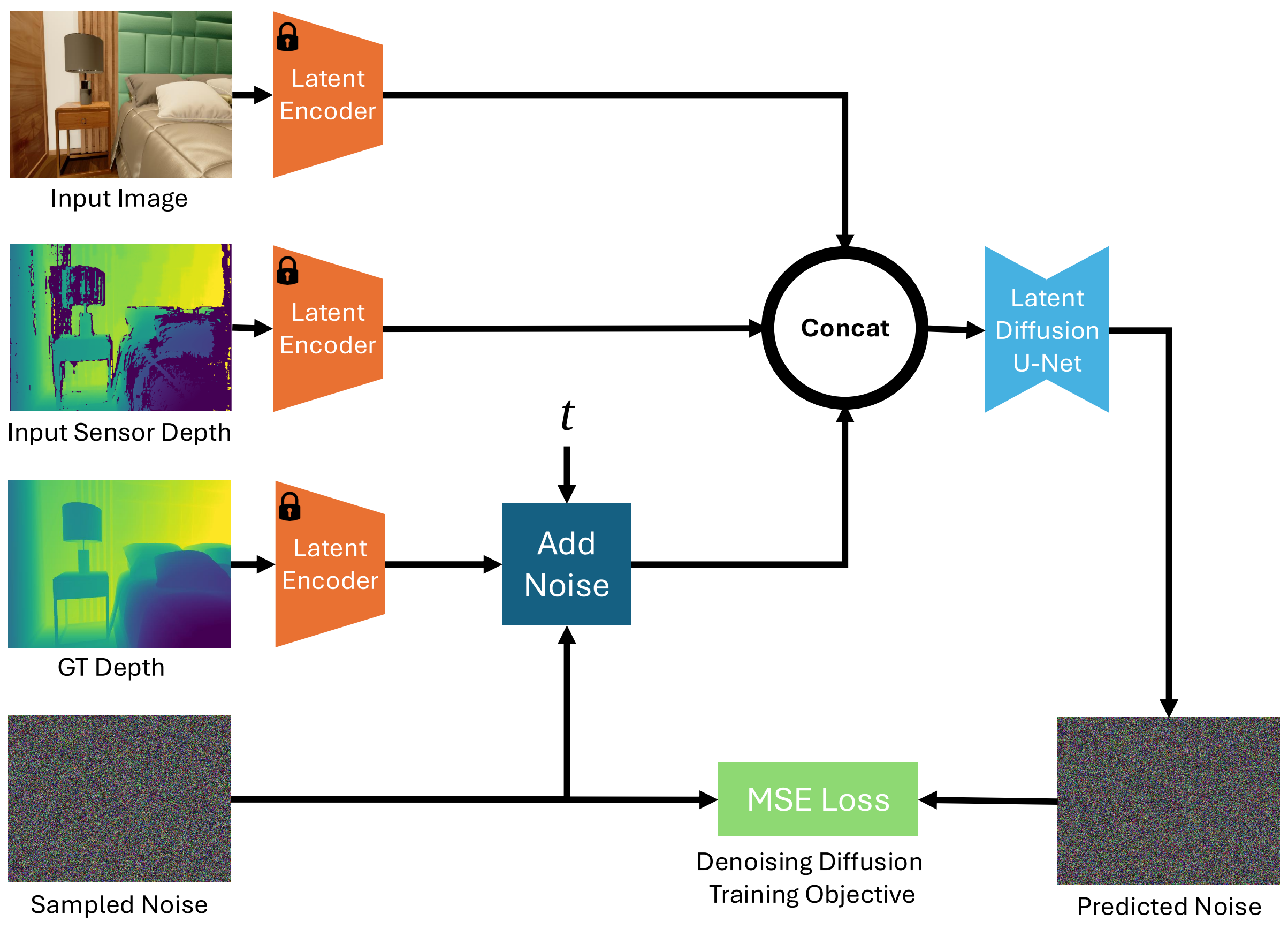


Fig. 1: System overview during training.

**Inference.** At inference time, the latent denoiser reconstructs the output depth map by iteratively denoising an initially normally-distributed Gaussian noise.

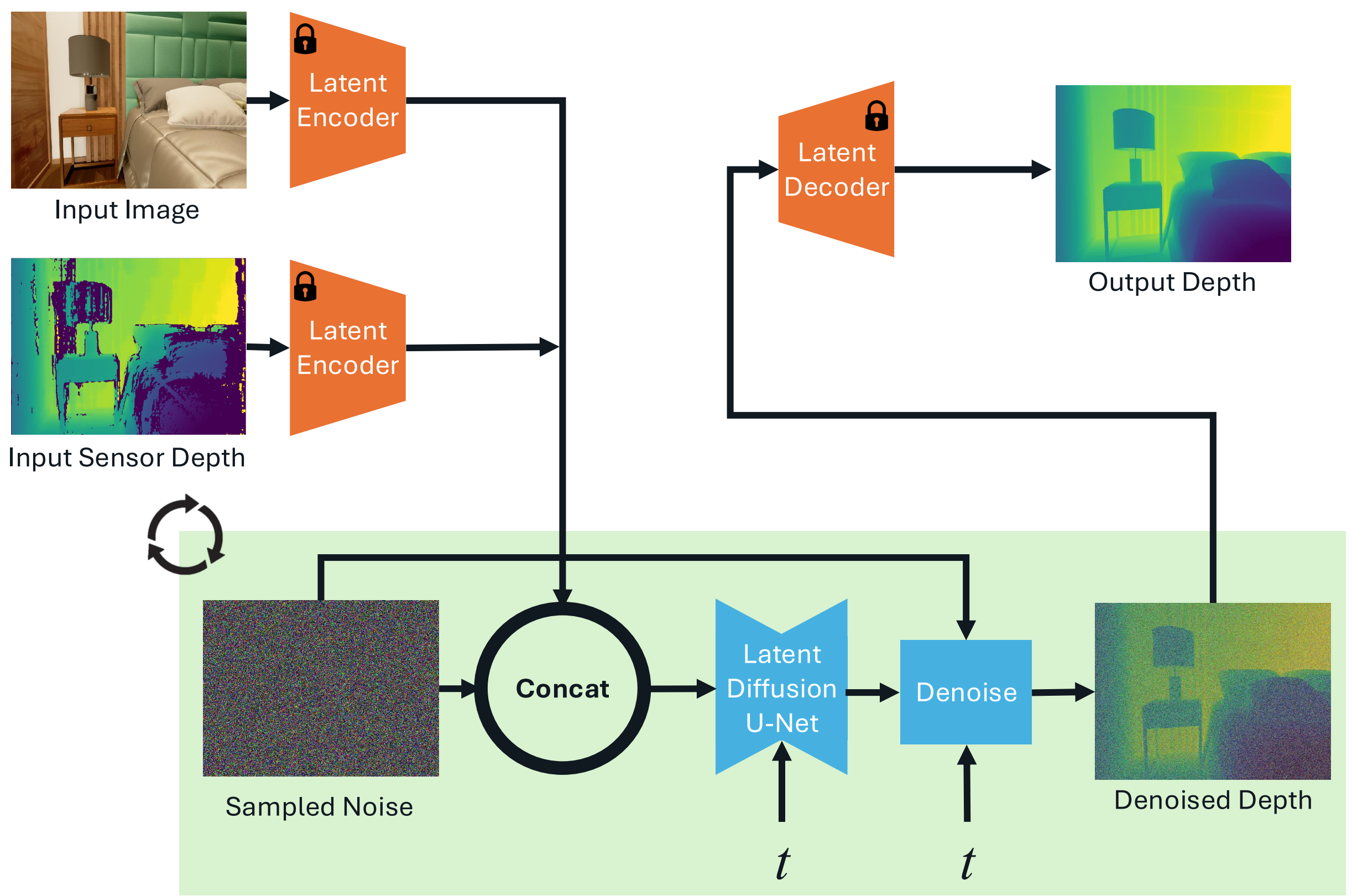


Fig. 2: System overview during inference.

**Sensor Depth Simulation.** We use SimSense [3], a depth sensor simulator, to generate sensor depth maps using stereo images. This ensures that our input sensor depth maps closely align with real-world depth sensors.

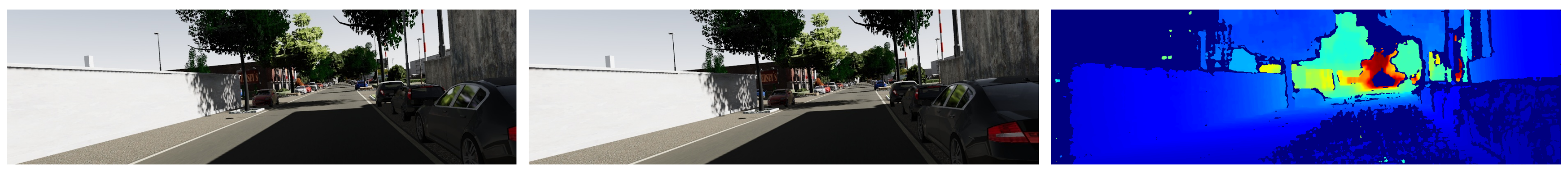


Fig. 3: Depth maps generated by SimSense using the stereo images exhibit the noise and incompleteness typical of real-world depth maps which helps to reduce the sim-to-real gap.

## Results

**Baselines**

- **Marigold [1]**. A diffusion-based monocular depth estimation method.
- **Depth Anything V2 [2](small)**. A discriminative method for monocular depth estimation.
  The predicted depth maps are aligned to the ground truth using least squares method before evaluation.

**Datasets**

- **IRS and VKITTI**. Seen during training.
- **NYUv2**. Unseen during training.

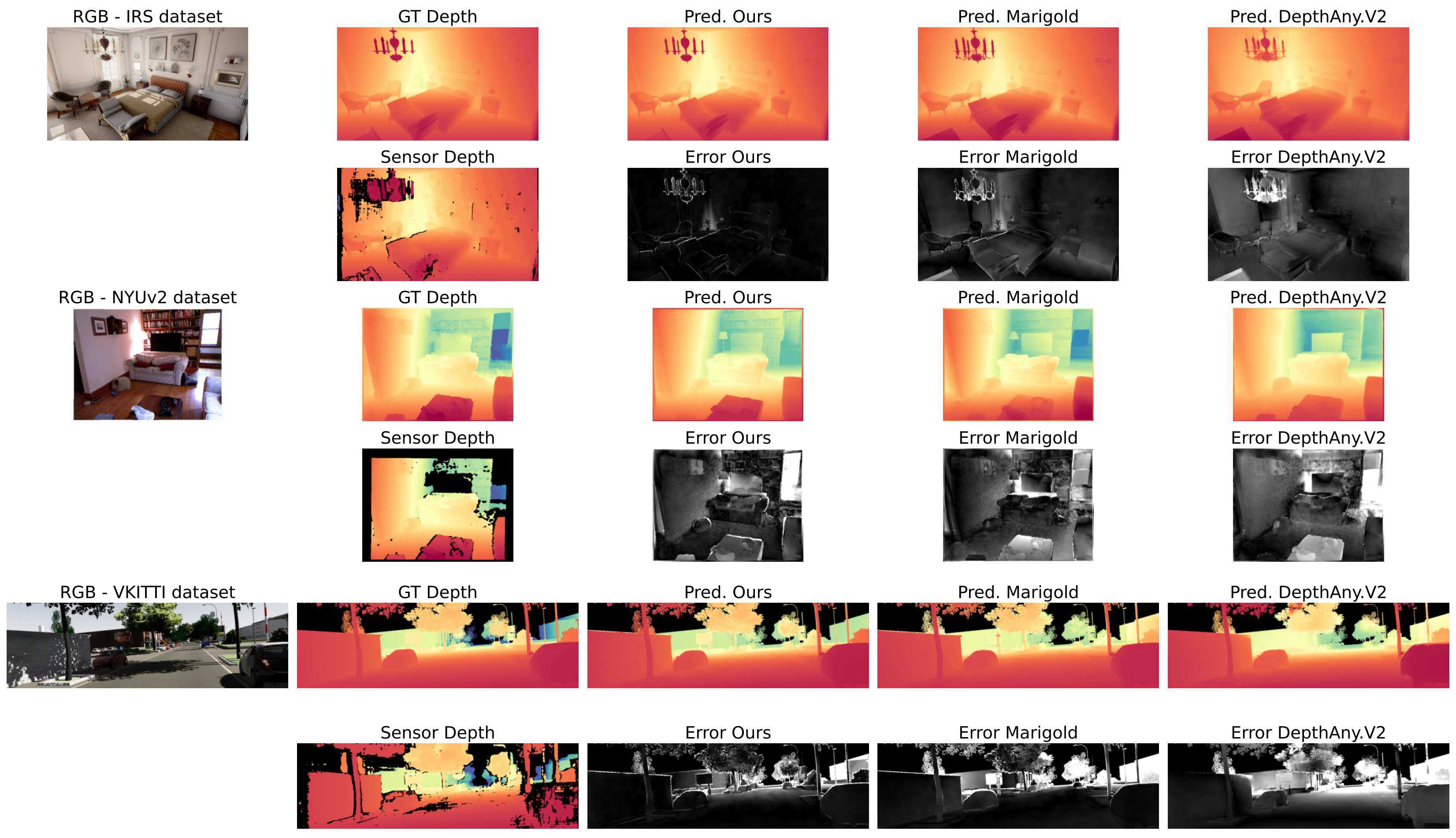
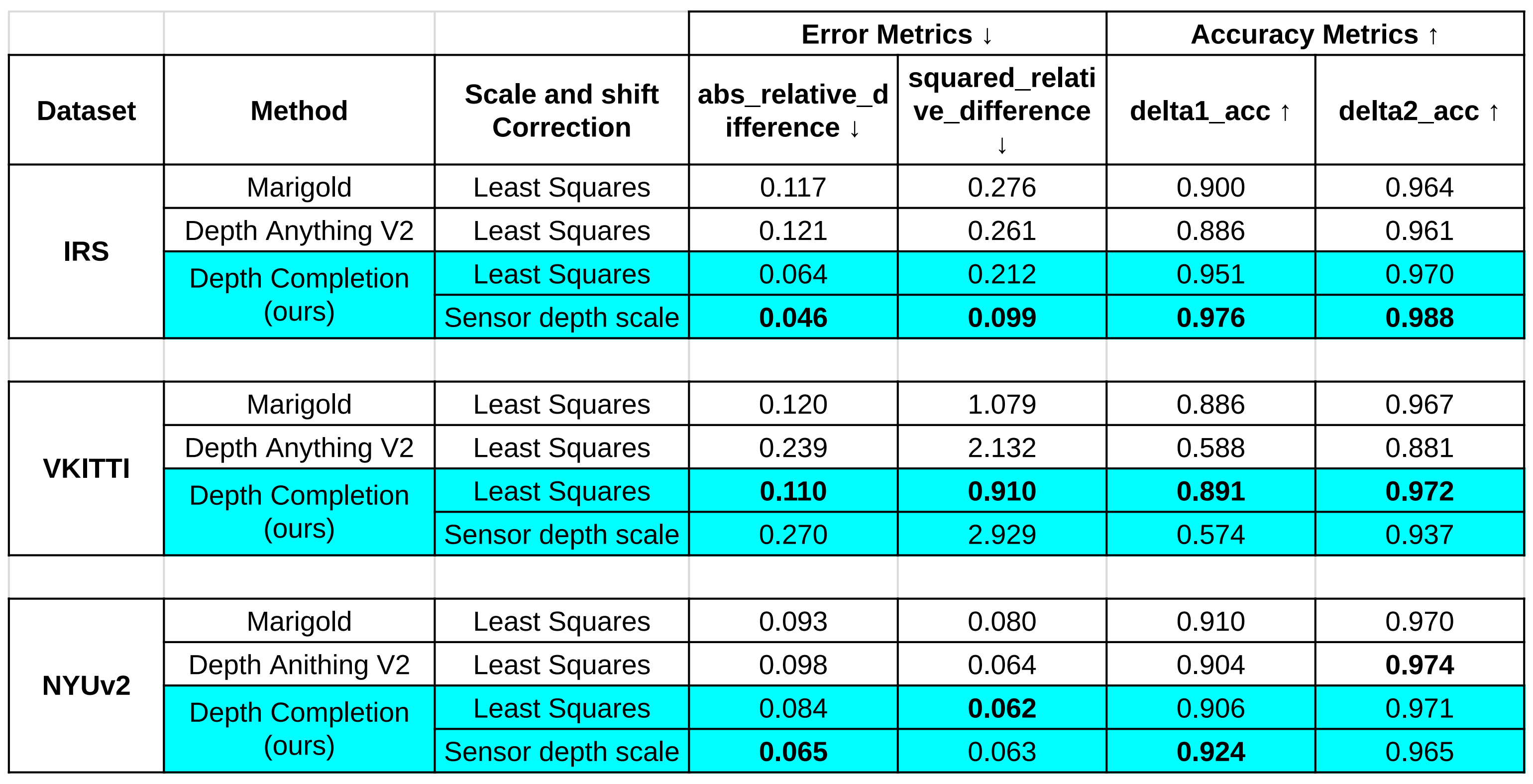


Fig. 4: Qualitative results



| | | | Error Metrics ↓ | | Accuracy Metrics ↑ | |
|---|---|---|---|---|---|---|
| **Dataset** | **Method** | **Scale and shift Correction** | **abs_relative_difference ↓** | **squared_relative_difference ↓** | **delta1_acc ↑** | **delta2_acc ↑** |
| **IRS** | Marigold | Least Squares | 0.117 | 0.276 | 0.900 | 0.964 |
| | Depth Anything V2 | Least Squares | 0.121 | 0.261 | 0.886 | 0.961 |
| | Depth Completion (ours) | Least Squares | 0.064 | 0.212 | 0.951 | 0.970 |
| | | Sensor depth scale | **0.046** | **0.099** | **0.976** | **0.988** |
| | | | | | | |
| **VKITTI** | Marigold | Least Squares | 0.120 | 1.079 | 0.886 | 0.967 |
| | Depth Anything V2 | Least Squares | 0.239 | 2.132 | 0.588 | 0.881 |
| | Depth Completion (ours) | Least Squares | **0.110** | **0.910** | **0.891** | **0.972** |
| | | Sensor depth scale | 0.270 | 2.929 | 0.574 | 0.937 |
| | | | | | | |
| **NYUv2** | Marigold | Least Squares | 0.093 | 0.080 | 0.910 | 0.970 |
| | Depth Anithing V2 | Least Squares | 0.098 | 0.064 | 0.904 | **0.974** |
| | Depth Completion (ours) | Least Squares | 0.084 | **0.062** | 0.906 | 0.971 |
| | | Sensor depth scale | **0.065** | 0.063 | **0.924** | 0.965 |

Fig. 5: Quantitative results

## Conclusion and Future Work

- Diffusion models conditioned on RGB and noisy sensor depth exhibit a slight improvement in accurate depth estimation compared to pure monocular depth estimation methods. This suggests that sensor depth provides valuable guidance to the model for estimating accurate depth.
- Our approach demonstrates generalization to real-world indoor scenes, despite being fine-tuned on synthetic data only. This can be attributed to the fact that the original Stable Diffusion model was trained on large datasets, capturing strong semantic and geometric priors from diverse domains.
- Future work aims to expand the conditioning inputs beyond RGB and standard depth by incorporating additional modalities such as thermal (infrared), monochrome, near-infrared (NIR) images, semantic and instance masks to capture both local and global features at the object part level for improving generalization.

## References

[1] Bingxin Ke et al. "Repurposing diffusion-based image generators for monocular depth estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9492–9502.

[2] Lihe Yang et al. "Depth Anything V2". In: *arXiv preprint arXiv:2406.09414* (2024).

[3] Xiaoshuai Zhang et al. "Close the Optical Sensing Domain Gap by Physics-Grounded Active Stereo Sensor Simulation". In: *IEEE Transactions on Robotics* (2023), pp. 1–19. DOI: 10.1109/TRO.2023.3235591.