LDM3D: Latent Diffusion Model for 3D

Gabriela Ben Melech Stan[1], Diana Wofk[1], Scottie Fox[2], Alex Redden[2], Will Saxton[2], Jean Yu[3], Estelle Aflalo[1], Shao-Yen Tseng[1], Fabio Nonato[3], Matthias Müller[1], Vasudev Lal[1]

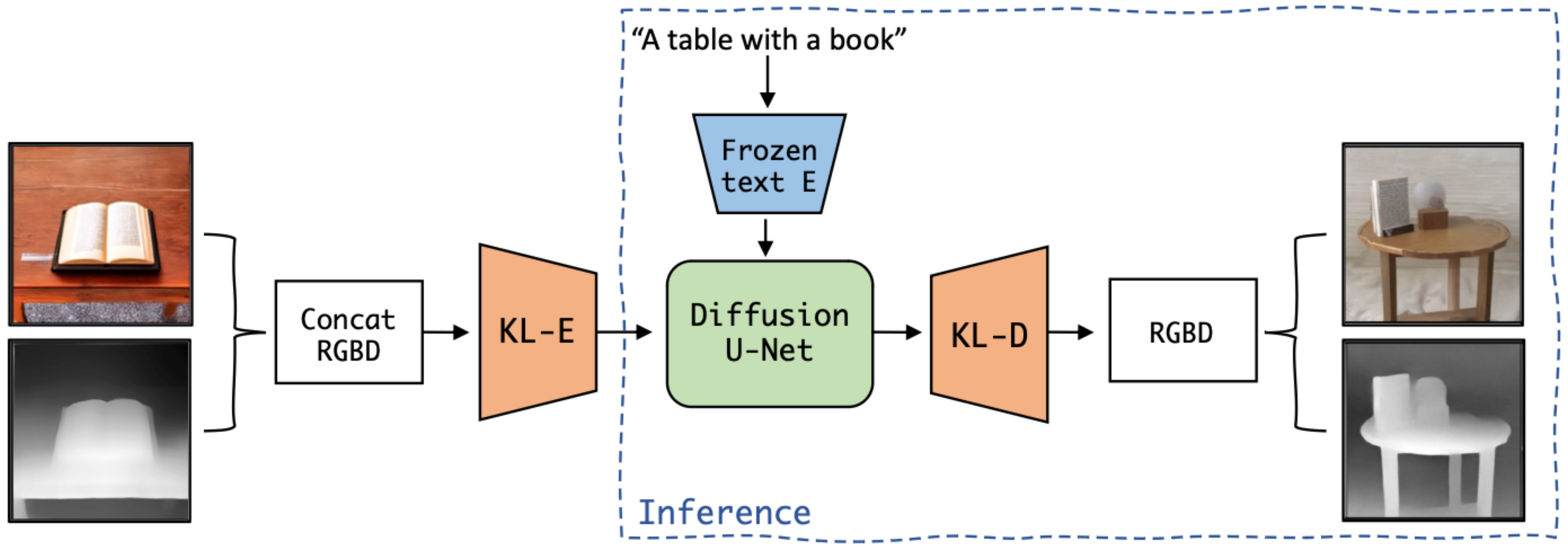
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

1. Introduction

- Our Latent Diffusion Model for 3D (LDM3D) generates RGB image and depth map pairs for given text prompts, allowing users to generate RGBD outputs from text inputs.
- We demonstrate integration of LDM3D into an application called DepthFusion, which uses diffused images and depth maps to create immersive and interactive 360°-view experiences with TouchDesigner.

2. Methodology

- 6-channel RGBD input: 16b grayscale depth is packed into 3-chn 8b depth, concatenated with the RGB image
- Input is passed through **modified KL-encoder** and mapped to the latent space
- Noise is added to the latent representation, which is then iteratively denoised by the U-Net
- **Text prompt** is passed through a frozen CLIP-text encoder and mapped to U-Net layers via cross-attention
- Denoised latent representation is passed through **modified KL-decoder** and mapped back to pixel space as a 6-channel RGBD output. This is then separated into an RGB image and a 16b grayscale depth map
- LDM3D was trained on Intel AI Supercomputing Cluster with Intel Xeon and Habana Gaudi AI accelerators



Our model is on par with Stable Diffusion with nearly the same number of parameters (1.06B). We finetune on a subset of ~10k samples from LAION-400M. Depth labels for supervised training are produced using DPT-Large.

3. Evaluation

We evaluate text-conditional image synthesis on 30k samples of the MS-COCO validation dataset.

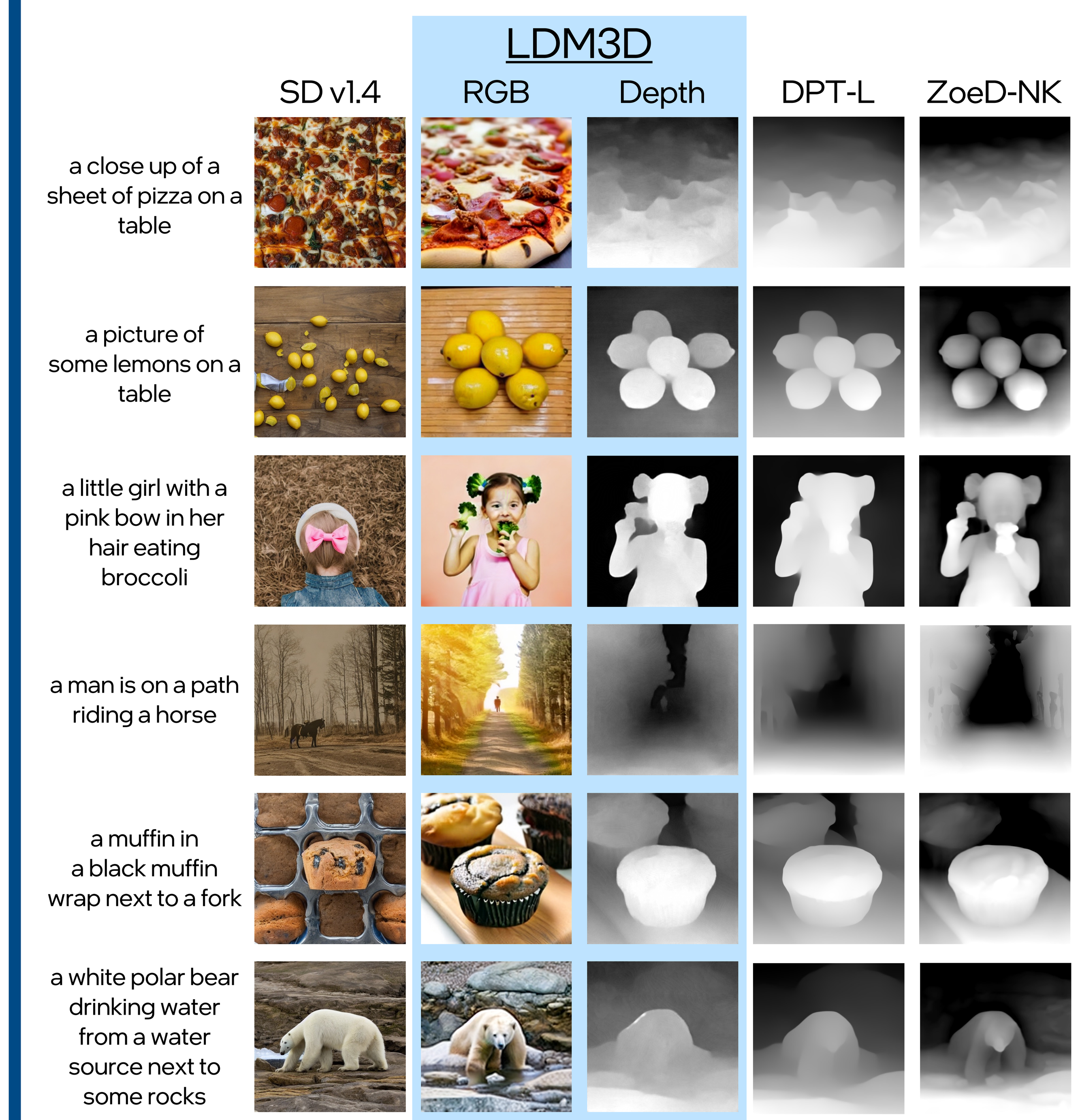


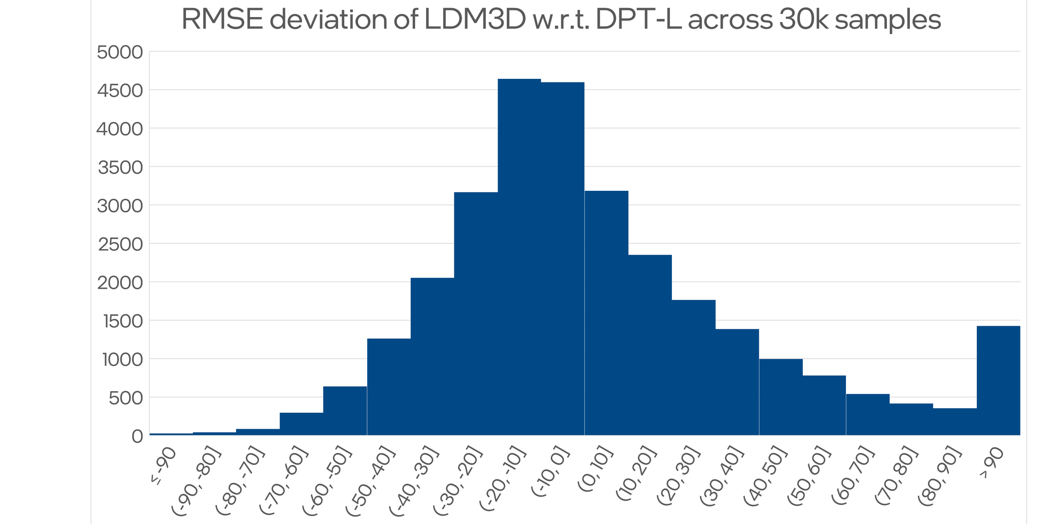
Image Analysis Metrics

Method	FID↓	IS↑	CLIP↑
SD v1.4	28.08	**34.17** ± 0.76	26.13 ± 2.81
SD v1.5	**27.39**	34.02 ± 0.79	26.13 ± 2.79
LDM3D (ours)	27.82	28.79 ± 0.49	**26.61** ± 2.92

Depth Error Metrics
(using depth maps from ZoeDepth-NK as reference/GT)

Method	AbsRel	RMSE
DPT-Large	**0.098**	1.57 [m]
LDM3D (ours)	0.109	**1.51** [m]

valid depth defined above 0m, with unbounded maximum



For ~50% of test samples, LDM3D achieves depth error within ±20% of DPT-Large.

4. Application: DepthFusion

LDM3D is integrated into DepthFusion:

1. Image-to-image inference with LDM3D: an RGBD input consisting of a panoramic image and depth map is passed through LDM3D to generate a new transformed image and depth map, guided by a given text prompt.
2. Generated images are projected onto a sphere and manipulated based on diffused depth, followed by meshing.
3. Different viewpoints are assembled.