Turbo3D: Ultra-fast Text-to-3D Generation

Supplementary Material

7. Details of Multi-step Multi-view Generation Model

We directly fine-tune an internal DiT [31] based text-to-image model into a text-to-multiview model. We fine-tune the model on the Objaverse dataset [3]. For the generation task, we render the dataset at a fixed elevation (20 degrees) and 16 equidistant azimuths. We empirically find that training with random views performs better than fixed views. In particular, during training, we randomly sample f views from the rendered 16 views for each instance, where f can be 4 or 8. Each view is conditioned on the corresponding Plücker embedding. For inference, we only infer 4 views for efficiency.

8. Experiments on 512 resolution

Some of the previous methods (Instant3D and SV3D) generate results with a higher resolution of 512. For a fair comparison, we also perform experiments on 512 resolution. Tab. 4 presents the quantitative comparisons with several state-of-the-art methods, where inference time is all measured under the resolution of 512. We can see our Turbo3D-512 version performs slightly better than our Turbo3D (256 resolution) with longer inference time, while outperforming other state-of-the-art methods by a large margin in terms of Clip score, VQA score, and inference speed.

Tab. 5 displays the effectiveness of latent GS-LRM. Under a higher resolution, the speed-up gain for latent GS-LRM gets larger. Overall, the latent GS-LRM archives a speed-up of 0.34s for the whole text-to-3D process.

9. Details of User Study

The interface example is shown in Fig. 7. For each question, we show two rendered videos from two different methods and ask the user to pick their preferred one. The two methods are randomly chosen from the total 4 methods: LGM [47], Instant3D [17], our multi-step multi-view model and our Turbo3D.

Method	CLIP Score ↑	VQA Score ↑	Inference Time ↓
TripoSR [49]	23.85	0.57	**1.28s**
SV3D [51]	24.92	0.64	35.96s
Instant3D [17]	26.23	0.65	20.00s
LGM [47]	24.73	0.58	6.56s
Turbo3D-512	**27.66**	**0.78**	**1.28s**

Table 4. **Comparison against state-of-the-art 3D generation methods.** Our Turbo3D-512 generates 3D assets with highest CLIP and VQA scores while using the least amount of time (benchmarked on a A100 GPU).

Ablation	CLIP Score ↑	VQA Score ↑	Inference Time ↓
Pixel GS-LRM [62]	**27.68**	**0.78**	1.62s
Latent GS-LRM	27.66	**0.78**	**1.28s**

Table 5. **Comparison between pixel and latent GS-LRM.** We report the CLIP score, VQA score, and overall text-to-3D inference time for comparison. Our latent GS-LRM achieves similar image quality while enabling better efficiency (∼21% speedup).

Figure 7. Interface example for user study.