1 REVISED QUALITATIVE COMPARISON

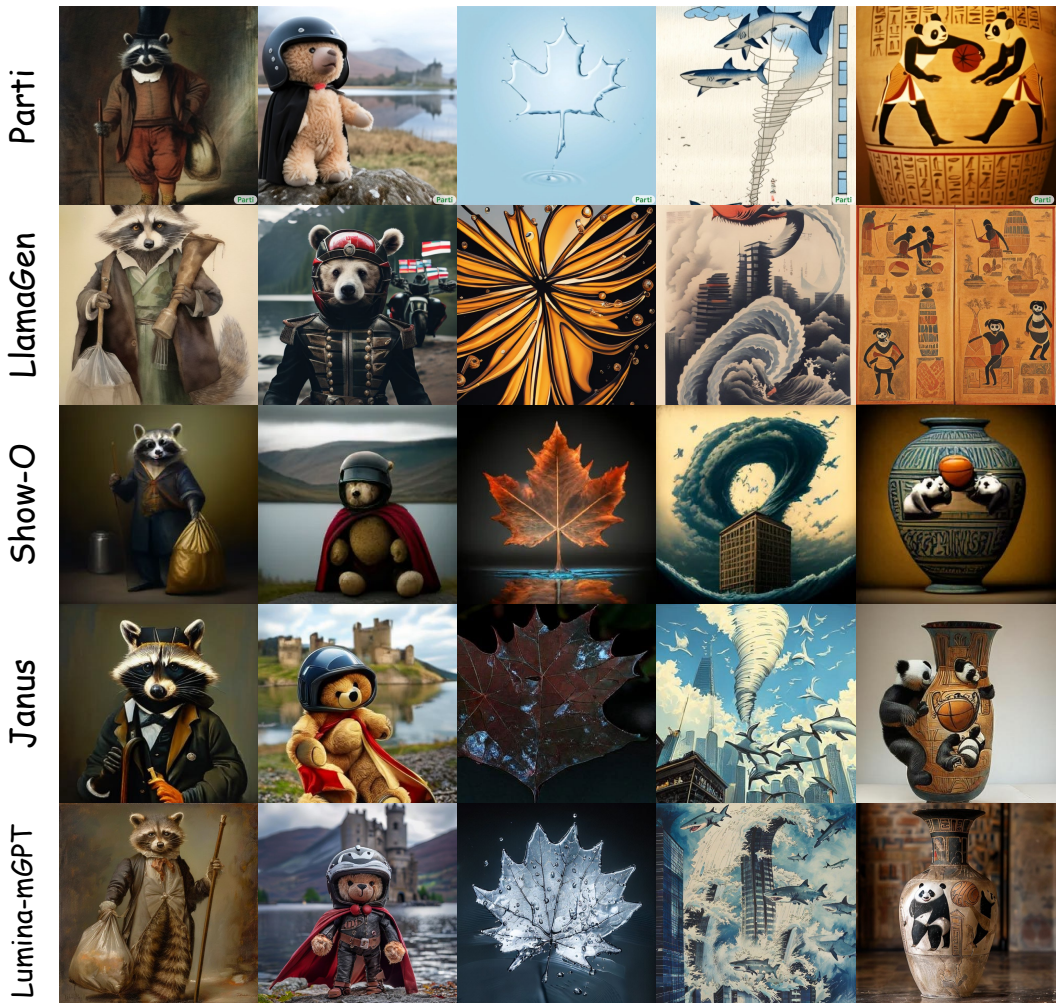

Figure 1: Text-to-image qualitative comparison. Prompts: (1) A raccoon wearing formal clothes, wearing a tophap and holding a cane. The raccoon is holding a garbage bag. Oil painting in the style of Rembrandt. (2) A teddy bear wearing a motorcycle helmet and cape is standing in front of Loch Awe with Kilchurn Castle behind him, dslr photo. (3) A photo of a maple leaf made of water. (4) A tornado made of sharks crashing into a skyscraper. Painting in the style of Hokusai. (5) A photo of an Athenian vase with a painting of pandas playing basketball in the style of Egyptian hieroglyphics.

2 EXTENDED T2I RESULTS

Methods	GenEval ↑				DPG ↑				T2I-CompBench ↑		
	Two Obj.	Counting	Color Attri.	**Overall**	Entity	Relation	Attribute	**Overall**	Color	Shape	Texture
Diffusion Models											
SDv1.5	-	-	-	0.40	74.23	73.49	75.39	63.18	0.3730	0.3646	0.4219
Lumina-Next	0.49	0.38	0.15	0.46	83.78	89.78	82.67	75.66	0.5088	0.3386	0.4239
SDv2.1	0.51	0.44	0.50	0.47	-	-	-	68.09	0.5694	0.4495	0.4982
SDXL	0.74	0.39	0.23	0.55	82.43	86.76	80.91	74.65	0.6369	0.5408	0.5637
SD3-medium	0.74	0.63	0.36	0.62	91.01	80.70	88.83	84.08	-	-	-
DALL-E3	0.87	0.47	0.45	0.67	89.61	90.58	88.39	83.50	0.8110	0.6750	0.8070
FLUX-dev	0.81	0.79	0.47	0.67	89.5	91.1	88.7	84.0	-	-	-
FLUX-schnell	0.92	0.73	0.54	0.71	91.3	86.5	89.7	84.8	-	-	-
AutoRegressive-Diffusion Mixed Models											
Transfusion†	-	-	-	0.63	-	-	-	-	-	-	-
Show-o †	0.52	0.49	0.28	0.53	-	-	-	67.48	-	-	-
AutoRegressive Models											
LlamaGen	0.34	0.21	0.04	0.32	-	-	-	65.16	-	-	-
Chameleon	-	-	-	0.39	-	-	-	-	-	-	-
Janus†	0.68	0.30	0.42	0.61	87.38	85.46	87.70	79.68	-	-	-
Lumina-mGPT (Ours)	0.77	0.27	0.32	0.56	86.60	91.29	84.61	79.68	0.6371	0.4727	0.6034

Table 1: **Quantitative comparison on text-to-image benchmarks.** † indicates concurrent or later work.

Method	**Non-Spatial**	**2D-Spatial**
Lumina-mGPT (Ours)	0.3176	0.2394

Table 2: Performance on T2I-Compbench across spatial and non-spatial dimensions. Metrics are reported independently as these were not provided by prior works.

3 Additional Downstream Results

3.1 Surface Normal Estimation

Table 3: NYUv2 Surface Normal Benchmark

Method	**11.25° ↑**	**22.5° ↑**	**30° ↑**	**mean↓**	**median↓**	**RMS_normal↓**
Deep3D	0.420	0.612	0.682	20.9	13.2	-
SURGE	0.473	0.689	0.766	20.6	12.2	-
SkipNet	0.479	0.700	0.778	19.8	12.0	28.2
GeoNet	0.484	0.715	0.795	19.0	11.8	26.9
PAP	0.488	0.722	0.798	18.6	11.7	25.5
GeoNet++	0.502	0.732	0.807	18.5	11.2	26.7
Lumina-mGPT(Ours)	0.497	0.717	0.782	20.4	11.3	30.2

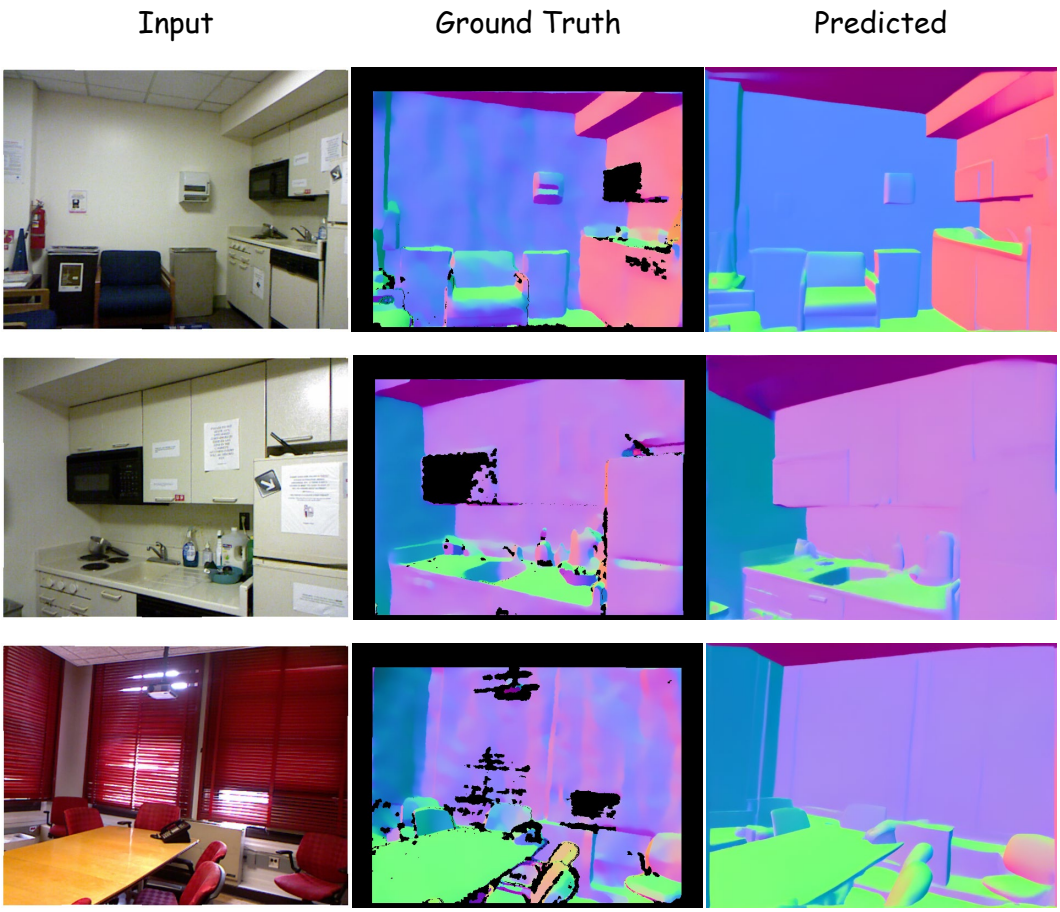

Figure 2: Visualization of Surface Normal Estimation

3.2 CONDITIONAL GENERATION

Table 4: Conditional Generation Comparison (CLIP-S↑) on MultiGen-20M

Methods	Canny-to-Image	Depth-to-Image
ControlNet-SD1.5	32.15	32.45
T2I-Adapter-SD1.5	31.71	31.46
Lumina-mGPT(Ours)	32.41	32.67