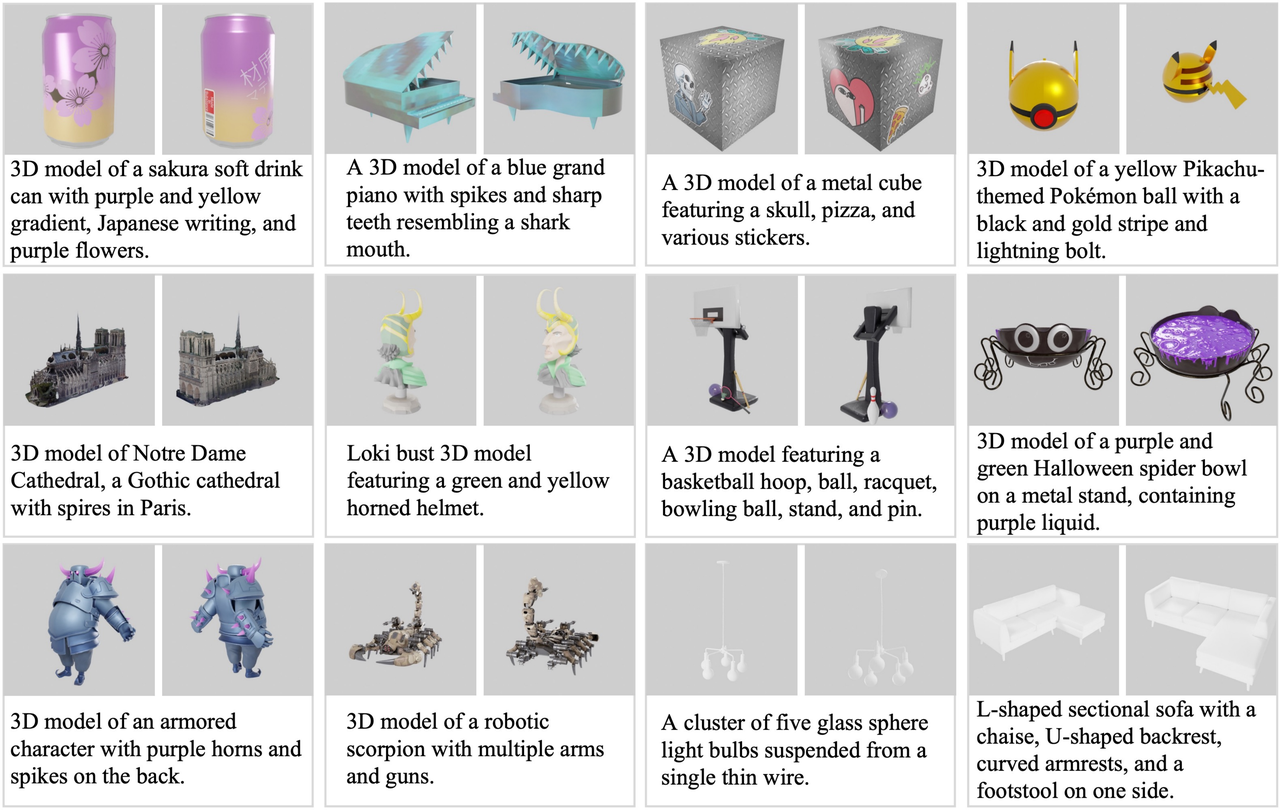
Scalable 3D Captioning with Pretrained Models

Tiange Luo*, Chris Rockwell*, Honglak Lee†, Justin Johnson†
University of Michigan, LG AI Research

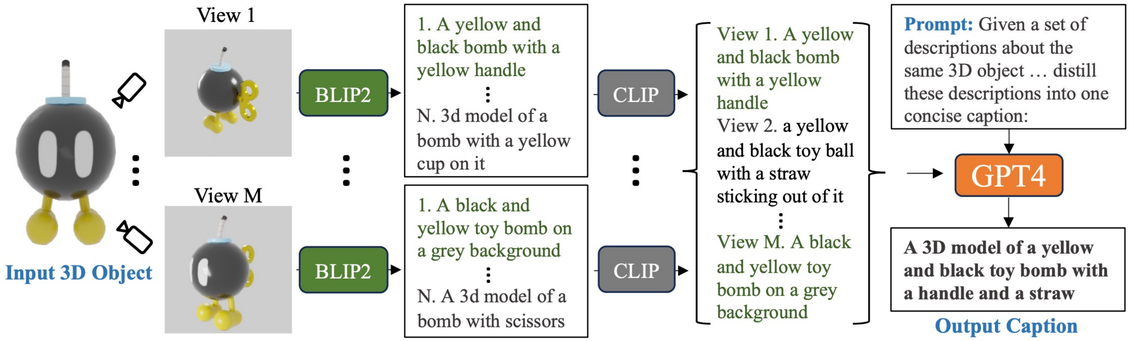
Project Page | Dataset

ICCV23 PARIS

Task: 3D Captioning



Method: Cap3D



Cap3D is better, cheaper, and faster than crowdsourced annotation

Method	A/B Human Testing Win % (Tie %)	Cost per 1k Objects	Annotation Speed
Human	37.8% ± 0.5% (9.5%)	$87.18	1.4k / day
Cap3D	**52.3% ± 0.5% (9.5%)**	**$8.35**	**65k / day**

All Cap3D captions and annotated objects have commercial-friendly licenses.

1k Objects Cost Breakdown	
BLIP2	$3.79
CLIP	$0.38
GPT4	$4.18
Cap3D Total Cost	**$8.35**

Apply Cap3D on Objaverse and finetune Text-to-3D models

	Pretrained					Finetuned on Cap3D				
	FID↓	CLIP Score	CLIP R-Precision (2k) R@1	R@5	R@10	FID↓	CLIP Score	CLIP R-Precision (2k) R@1	R@5	R@10
Ground Truth Images	-	81.6	32.7	55.1	64.3	-	81.6	32.7	55.1	64.3
Point-E (Text-to-3D) [88]	36.1	72.4	6.0	16.2	22.4	**32.8**	**75.6**	**12.4**	**28.1**	**36.9**
S. Diff. [22] (CNet) [63]+ [88](Im-to-3D)	54.7	73.6	11.0	23.4	30.0	**53.3**	**74.6**	**12.4**	**26.2**	**33.8**
S. Diff. [22] (LoRA) [103]+ [88](Im-to-3D)	54.7	73.6	11.0	23.4	30.0	**53.7**	**74.4**	**11.6**	**24.6**	**31.4**