Textfield3d: Towards enhancing open-vocabulary 3d generation with noisy text fields

Tianyu Huang[1,3], Yihan Zeng[2], Bowen Dong[1], Hang Xu[2], Songcen Xu[2], Rynson W.H. Lau[3], Wangmeng Zuo[1*]

[1] Harbin Institute of Technology [2] Huawei Noah's Ark Lab [3] City University of Hong Kong

ICLR

Motivation:

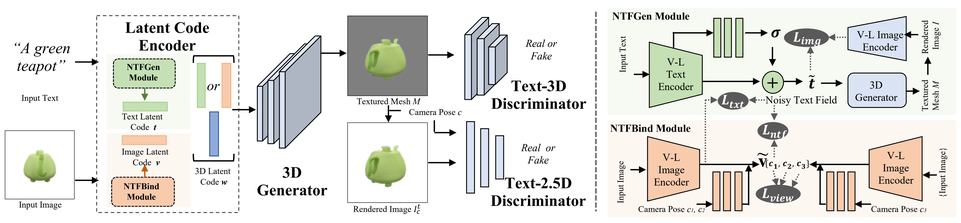
- **Limited Vocabulary of 3D Native Generation**
  - Remaining under development due to limited training data, *e.g.*, Objaverse (largest available dataset, 800k objects)
- **One Text Prompt Can Match Multiple Possible Scenes**
  - A training instance is represented as a point in latent space
  - pre-trained concepts are overfitted to limited points
  - fail to synthesis complicated objects
- **Noisy Text Fields (NTFs) for Enlarging Expression Range**
  - Inject dynamic noise to the latent code $\tilde{\mathbf{t}} = \mathbf{t} + N(0, \boldsymbol{\sigma}),\ where\ \boldsymbol{\sigma} = f_{txt}(\mathbf{t}), \boldsymbol{\sigma} \in (\epsilon_1, \epsilon_2)$
  - Expect that simpler descriptions have larger text space

Methodology:

- **GAN-Based Supervision with 3D and 2.5D Guidance**
  - 3D Discriminator: sampled points
  - 2.5D Discriminator: rendered images with camera pose
  - Further align the noisy text fields with generated results

$$\mathcal{L}_{img} = \mathcal{L}_{contrast}(\tilde{\mathbf{t}}, I) = -\log \frac{\exp(\tilde{\mathbf{t}} \cdot [f_{vis}(I_c^t)]^T/\tau)}{\exp(\tilde{\mathbf{t}} \cdot [f_{vis}(I_c^t)]^T/\tau) + \sum_{\bar{t} \neq t} \exp(\tilde{\mathbf{t}} \cdot [f_{vis}(I_c^{\bar{t}})]^T/\tau)},$$

- **Open-Vocabulary 3D Training Data**
  - Filter out 3D assets from Objaverse
  - Caption multi-view renderings with BLIP-2 and MiniGPT-4



Experiments:

• Quantitative Results

Table 1: CLIP R-Precision on COCO evaluation prompts.

Method	ViT-B/32	ViT-L/14	Latency	Data Scale
DreamFields	78.6	82.9	$\sim$ 200 V100-hr	-
DreamFusion	75.1	79.1	$\sim$ 12 V100-hr	-
Point-E (300M, text-only)	33.6	35.5	24 V100-sec	$>$ 1M
Shap-E (300M, text-only)	37.8	40.9	13 V100-sec	$>$ 1M
Point-E (300M)	40.3	45.6	1.2 V100-min	$>$ 1M
Point-E (1B)	41.1	46.8	1.5 V100-min	$>$ 1M
Shap-E (300M)	41.1	46.4	1.0 V100-min	$>$ 1M
TextField3D (ours, 1-shot)	45.8	49.0	$\sim$ 6.9 V100-sec	$\sim$ 175K
TextField3D (ours, 9-shot)	63.4	67.3	$\sim$ 1.0 V100-min	$\sim$ 175K

• Ablation Study

(a) directly use CLIP embeddings; (b) add L_{img} to CLIP embeddings; (c) use noisy text fields; (d) add 3D discriminator

Table 2: Quantitative Analysis. FID (Heusel et al., 2017) and CLIP-score (Hessel et al., 2021) are adopted to evaluate the generation quality and text consistency of different solutions.

Method	BLIP-2 (Li et al., 2023)			MiniGPT-4 (Zhu et al., 2023)		
	FID↓	CLIP-Score↑		FID↓	CLIP-Score↑	
		ViT-B/32	ViT-L/14		ViT-B/32	ViT-L/14
GET3D (Gao et al., 2022)	41.66	-	-	41.66	-	-
(a)	31.67	28.33	22.79	34.67	28.59	23.44
(b)	29.02	29.80	24.39	29.91	30.03	24.05
(c)	29.73	**30.36**	25.43	28.69	30.68	25.17
(d)	**26.94**	30.35	**25.57**	**25.46**	**30.89**	**25.77**

• Qualitative Results

• Complicated Text Descriptions

• Data Captioning

	TAPS3D	BLIP-2	MiniGPT-4
	"A red hatchback"	"A red sports car"	"The image is a red sports car with a sleek and shiny body, designed with a streamlined style"
	"A brown chair"	"A wooden chair"	"The wooden chair has a rectangular shape with a curved back and a flat seat"
	"A wooden desk"	"A wooden table"	"The image is a wooden table with a rectangular top and four legs"

Personal Web: