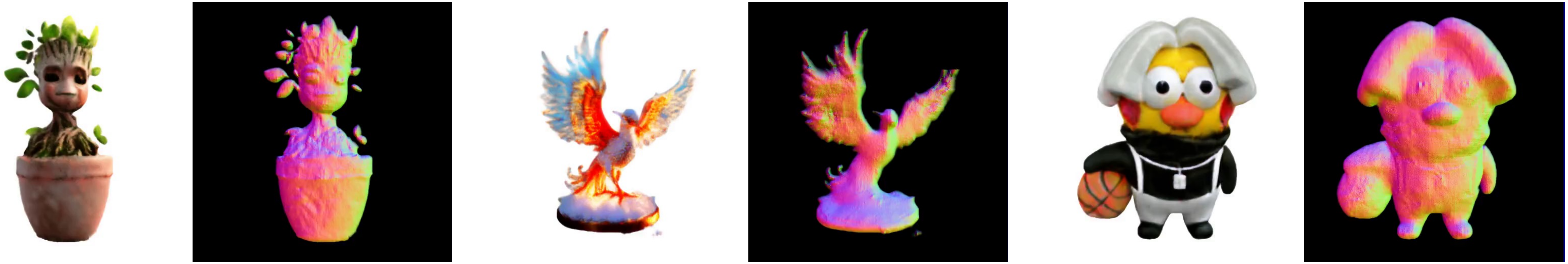
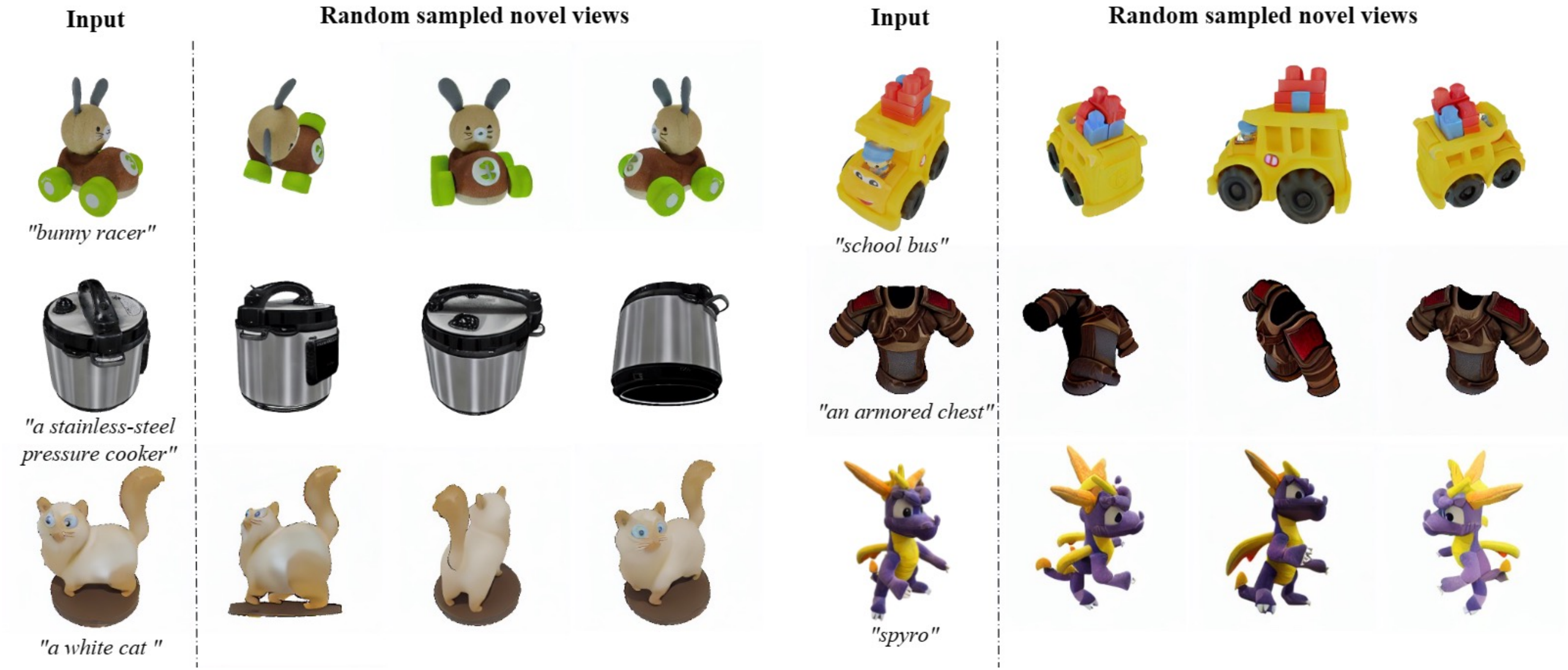
TOSS: High-quality Text-guided Novel View Synthesis from a Single Image

Yukai Shi[1,3]*† Jianan Wang[3]* He Cao[2,3]*†
Boshi Tang[1,3] Xianbiao Qi[3] Tianyu Yang[3] Yukun Huang[3] Shilong Liu[1,3]
Lei Zhang[3] Heung-Yeung Shum[1,3]

[1] Tsinghua University [2] Hong Kong University of Science and Technology
[3] International Digital Economy Academy (IDEA)

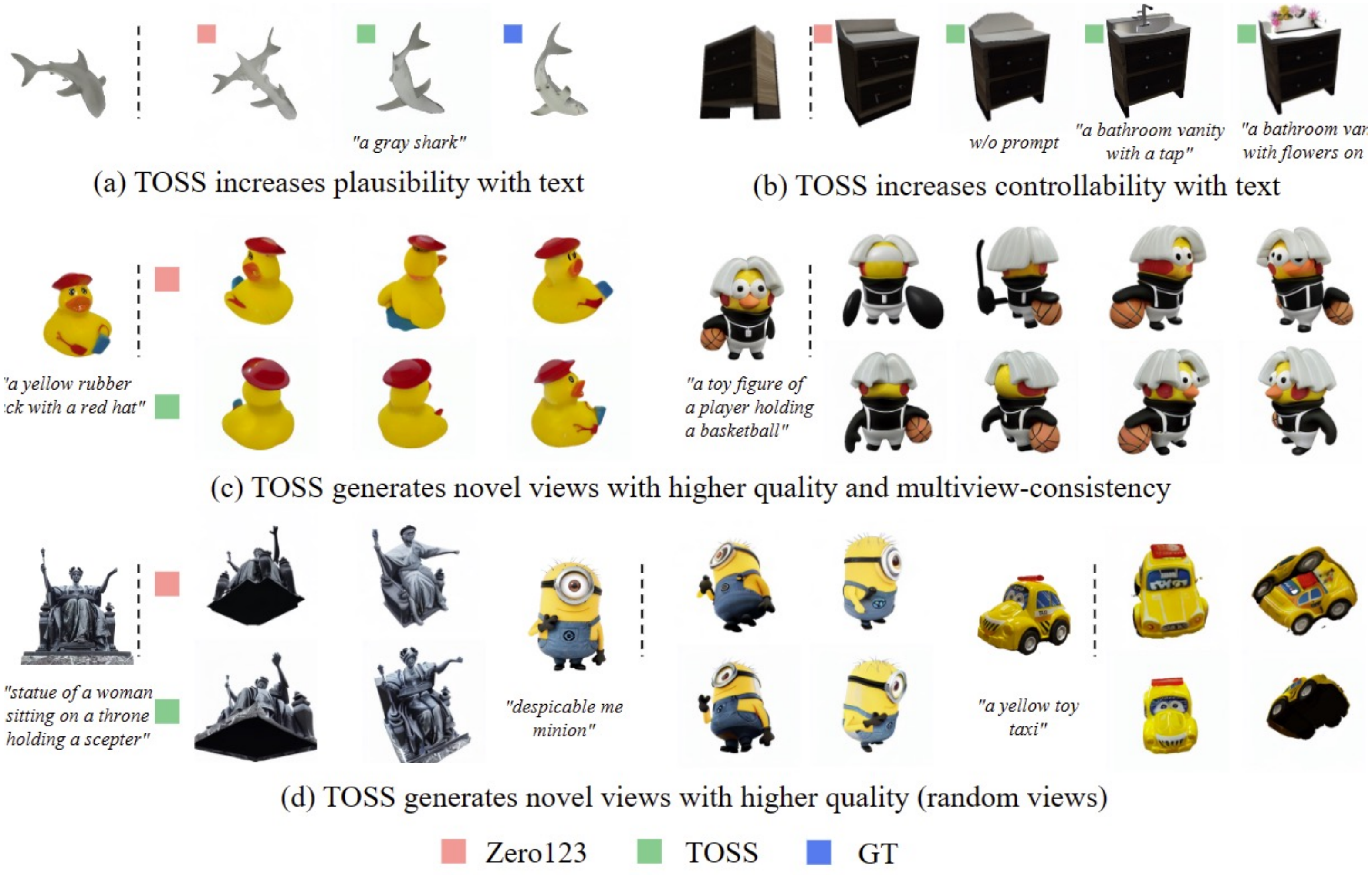


What can TOSS do


Input
Random sampled novel views
"bunny racer"
"a stainless-steel pressure cooker"
"a white cat "
Input
Random sampled novel views
"school bus"
"an armored chest"
"spyro"

Generate high-quality images from arbitrary camera poses based on a single image of arbitrary objects

Related works


"a gray shark"
w/o prompt
"a bathroom vanity with a tap"
"a bathroom var with flowers on
'a yellow rubber ck with a red hat"
"a toy figure of a player holding a basketball"
"statue of a woman sitting on a throne holding a scepter"
"despicable me minion"
"a yellow toy taxi"
🟥 Zero123
🟩 TOSS
🟦 GT

(a) TOSS increases plausibility with text

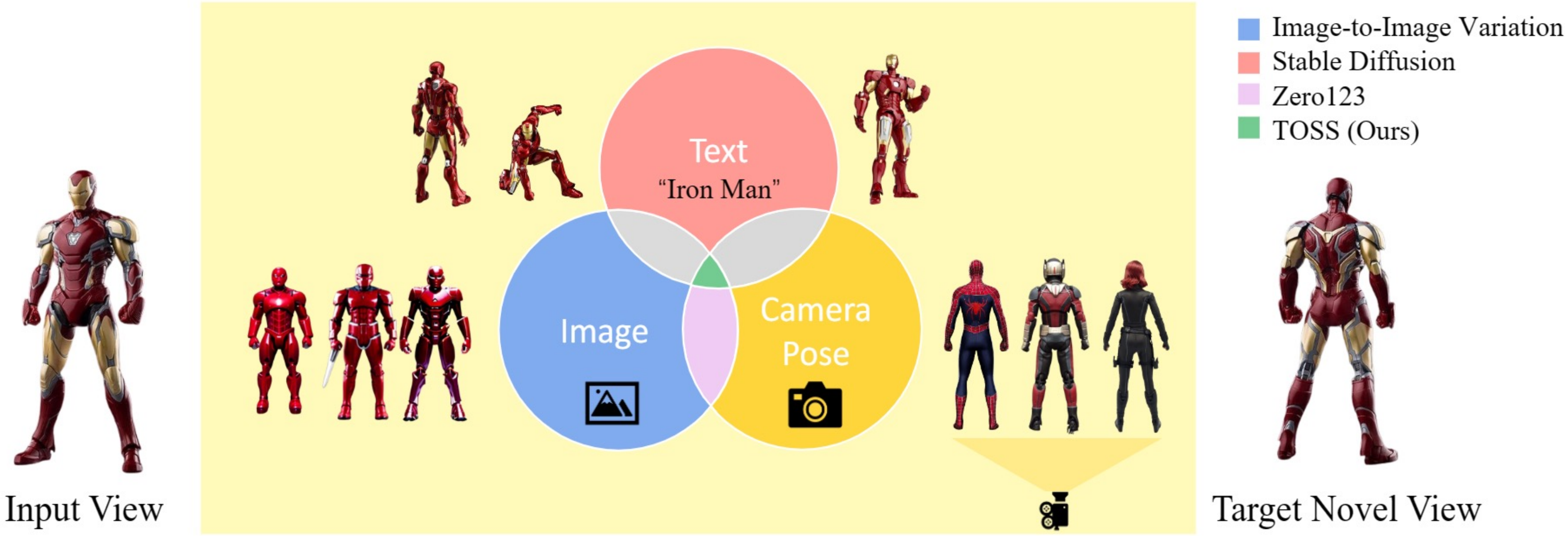
(b) TOSS increases controllability with text

(c) TOSS generates novel views with higher quality and multiview-consistency

(d) TOSS generates novel views with higher quality (random views)

Impalusible, uncontrollable and inconsistent results due to the ill-posed nature of single-view NVS

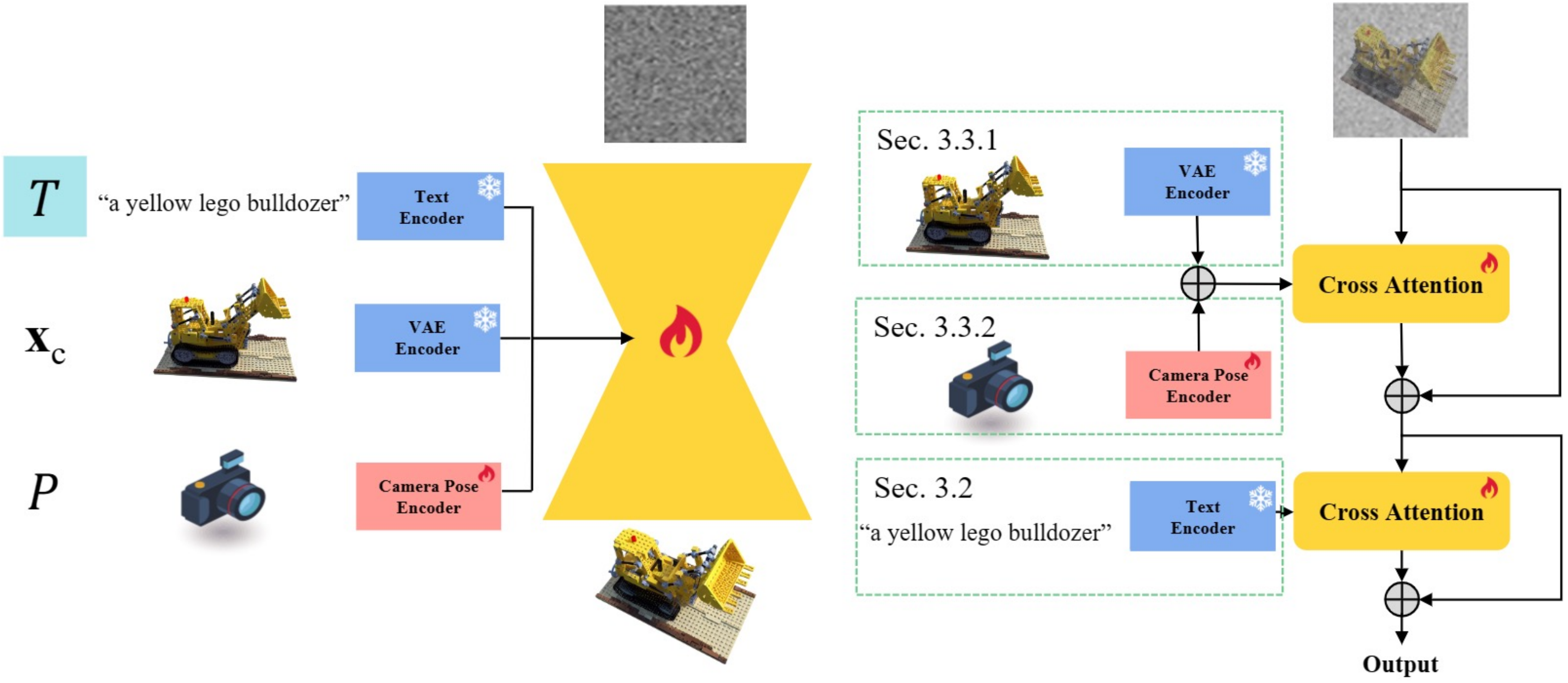
Compare with related works



Introduce text as high-level sementic information
to constraint the NVS solution space
for more controllable and more plausible results

Method

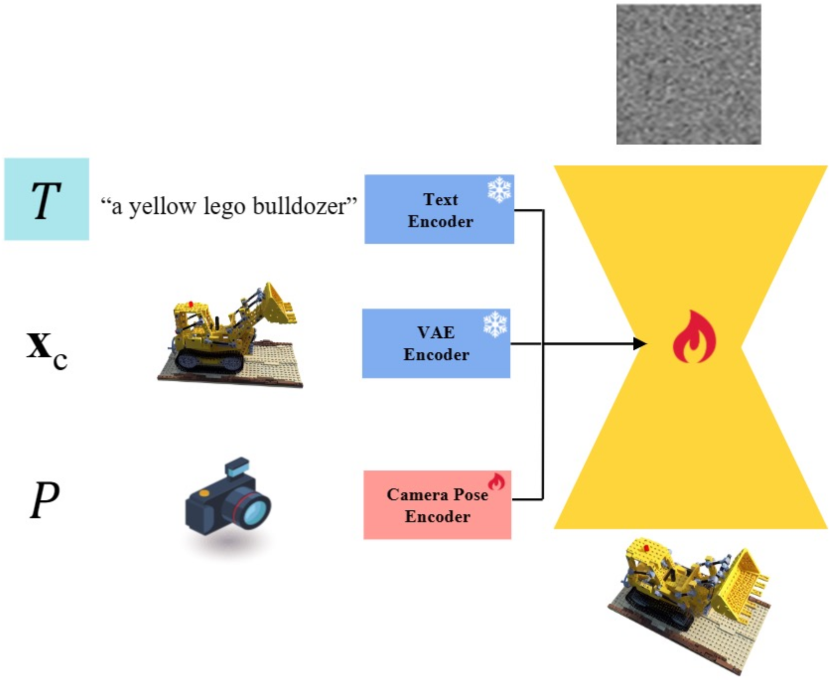
- **Totally geometry-free generation task**
- **Texts constraint:** finetune pretrained t2i stable diffusion model
- **Image constraint:** dense cross attention
- **Camera pose constraint:** key-value pair in dense cross attention



Method

Texts constraint

- finetune pretrained t2i stable diffusion model
- objective function
- guidance scale settings

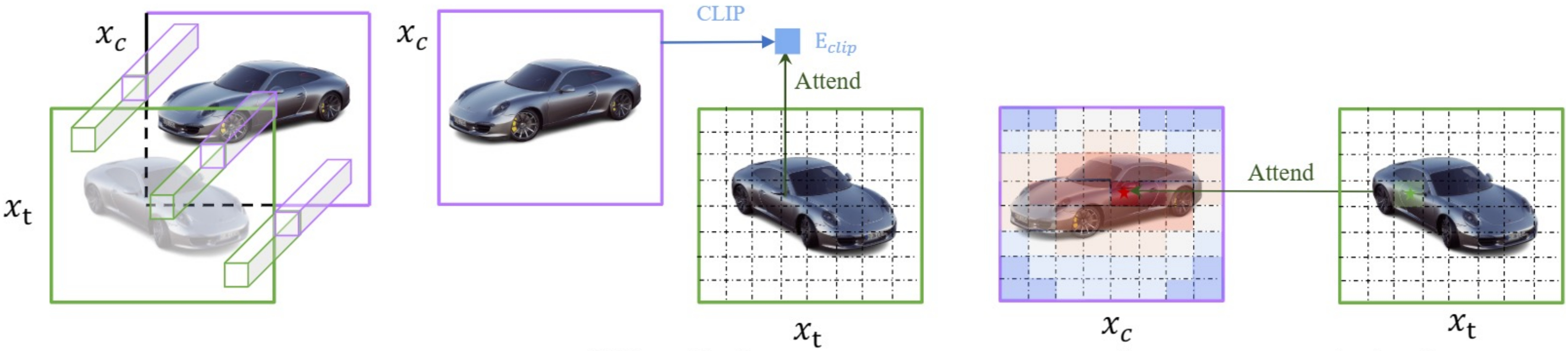


$$\min_{\theta} \mathbb{E}_{t,\mathbf{x},\boldsymbol{\epsilon},\mathbf{x}_c,P}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{x}_c, P, T)\right\|_2^2\right].$$

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, \mathbf{x}_c, P, T) = {} & \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \emptyset, P, \emptyset) \\ & + \underline{\alpha}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{x}_c, P, \emptyset) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \emptyset, P, \emptyset)] \\ & + \underline{\beta}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{x}_c, P, T) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{x}_c, P, \emptyset)],\end{aligned}$$

Method

- **Image constraint**:
  - dense cross attention
  - Channel concatenation: Information misalignment
  - CLIP embedding: excessive information compression
- **Camera pose constraint**: key-value pair in dense cross attention



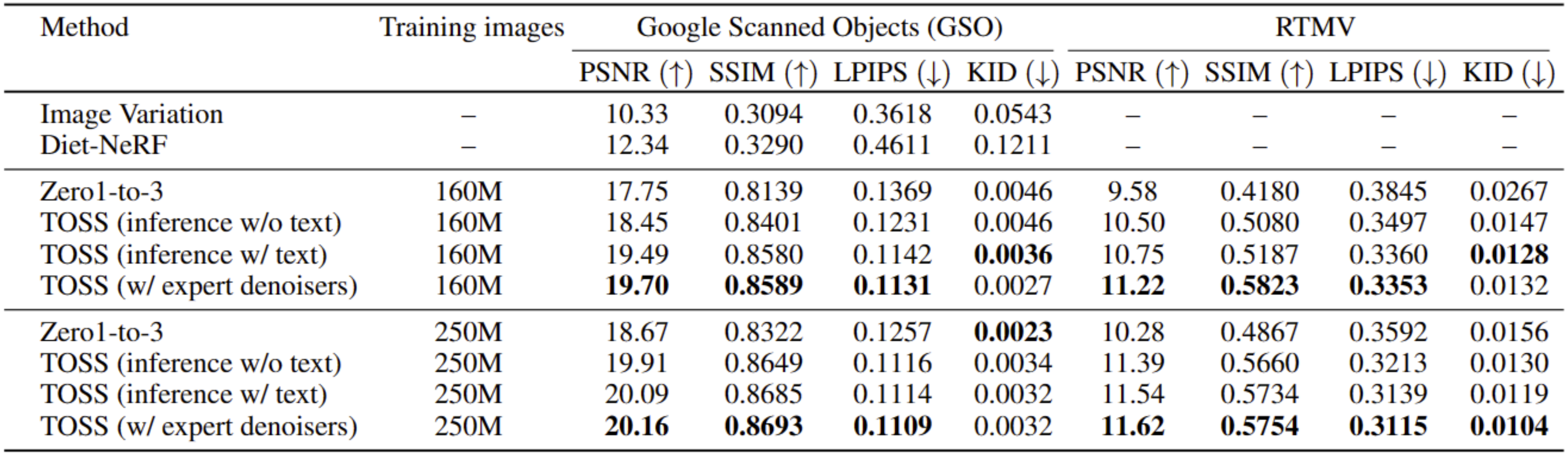
(a) Channel Concatenation

(b) Cross Attention w/ CLIP embedding

(c) Dense Cross Attention (Ours) w/ VAE Image Embedding

Results: Quantitative comparison

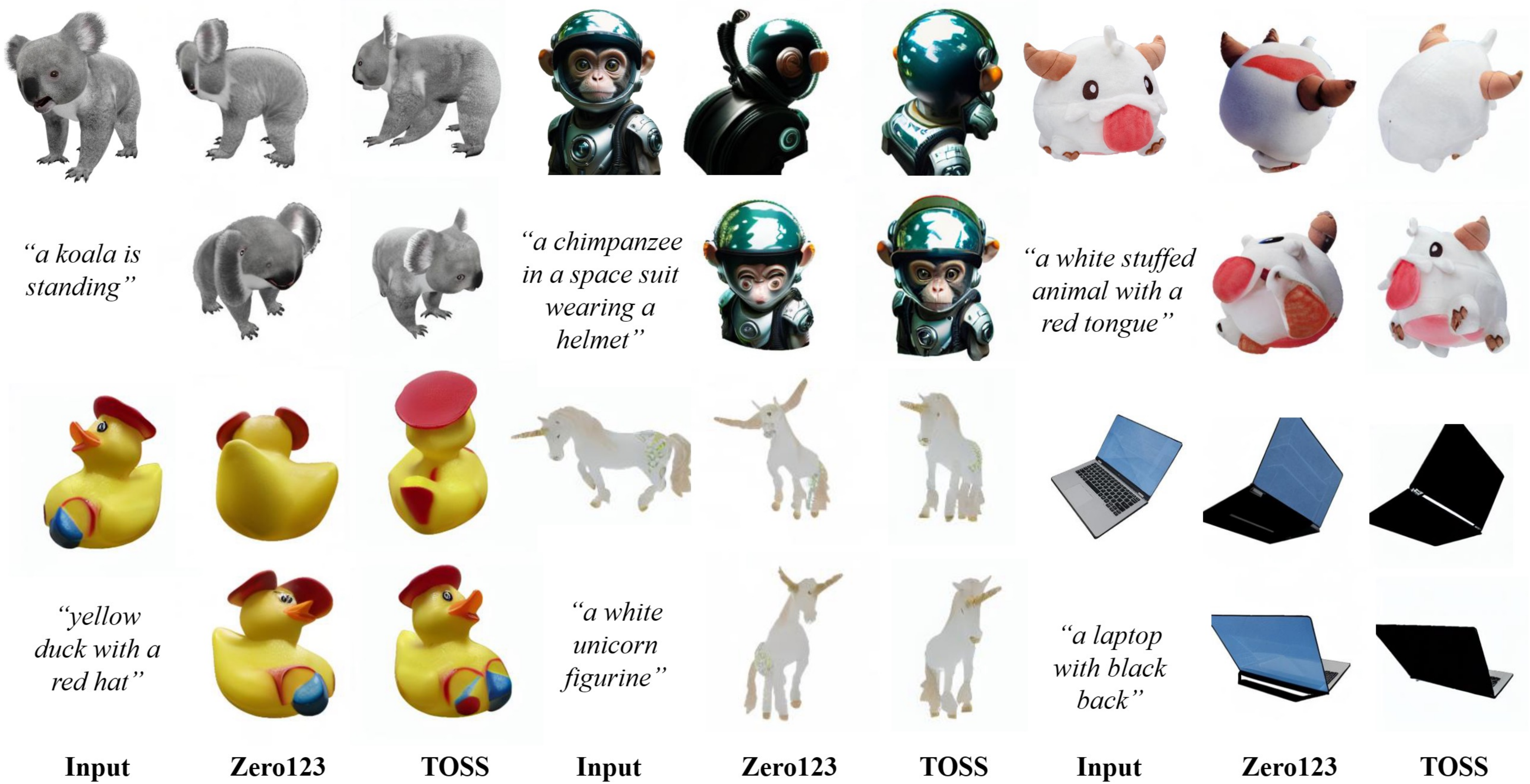
- **Training:** objaverse 800k, cap3d + clip rank, 5-6 days for 8*A100 (fp16)
- **Evaluation:** GSO/RTMV datasets, random views in the whole sphere



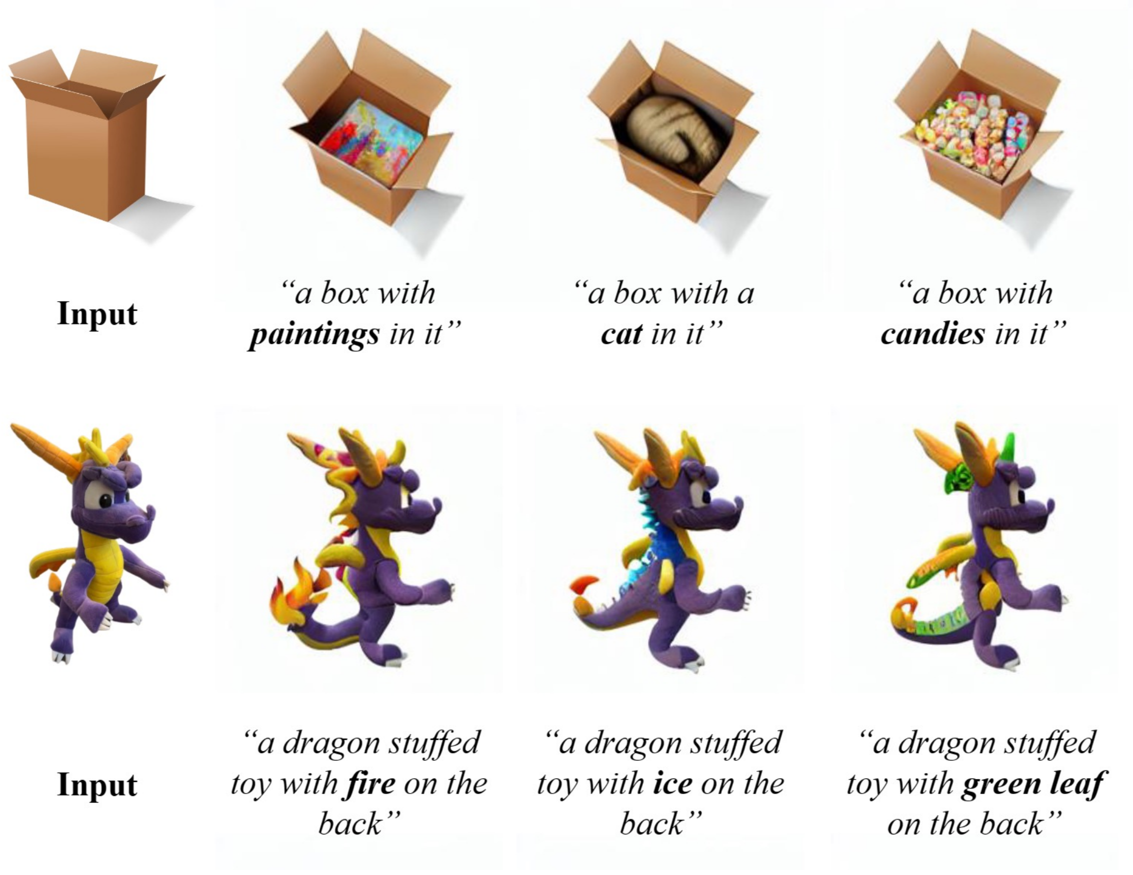
Method	Training images	Google Scanned Objects (GSO)				RTMV			
		PSNR (↑)	SSIM (↑)	LPIPS (↓)	KID (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	KID (↓)
Image Variation	–	10.33	0.3094	0.3618	0.0543	–	–	–	–
Diet-NeRF	–	12.34	0.3290	0.4611	0.1211	–	–	–	–
Zero1-to-3	160M	17.75	0.8139	0.1369	0.0046	9.58	0.4180	0.3845	0.0267
TOSS (inference w/o text)	160M	18.45	0.8401	0.1231	0.0046	10.50	0.5080	0.3497	0.0147
TOSS (inference w/ text)	160M	19.49	0.8580	0.1142	**0.0036**	10.75	0.5187	0.3360	**0.0128**
TOSS (w/ expert denoisers)	160M	**19.70**	**0.8589**	**0.1131**	0.0027	**11.22**	**0.5823**	**0.3353**	0.0132
Zero1-to-3	250M	18.67	0.8322	0.1257	**0.0023**	10.28	0.4867	0.3592	0.0156
TOSS (inference w/o text)	250M	19.91	0.8649	0.1116	0.0034	11.39	0.5660	0.3213	0.0130
TOSS (inference w/ text)	250M	20.09	0.8685	0.1114	0.0032	11.54	0.5734	0.3139	0.0119
TOSS (w/ expert denoisers)	250M	**20.16**	**0.8693**	**0.1109**	0.0032	**11.62**	**0.5754**	**0.3115**	**0.0104**

Table 1: **Quantitative comparison of single-view novel view synthesis on GSO and RTMV.**

Results: more plausible than baselines


"a koala is standing"
"a chimpanzee in a space suit wearing a helmet"
"a white stuffed animal with a red tongue"
"yellow duck with a red hat"
"a white unicorn figurine"
"a laptop with black back"
Input
Zero123
TOSS
Input
Zero123
TOSS
Input
Zero123
TOSS

Results: more controllable than baselines



Input

*"a box with **paintings** in it"*

*"a box with a **cat** in it"*

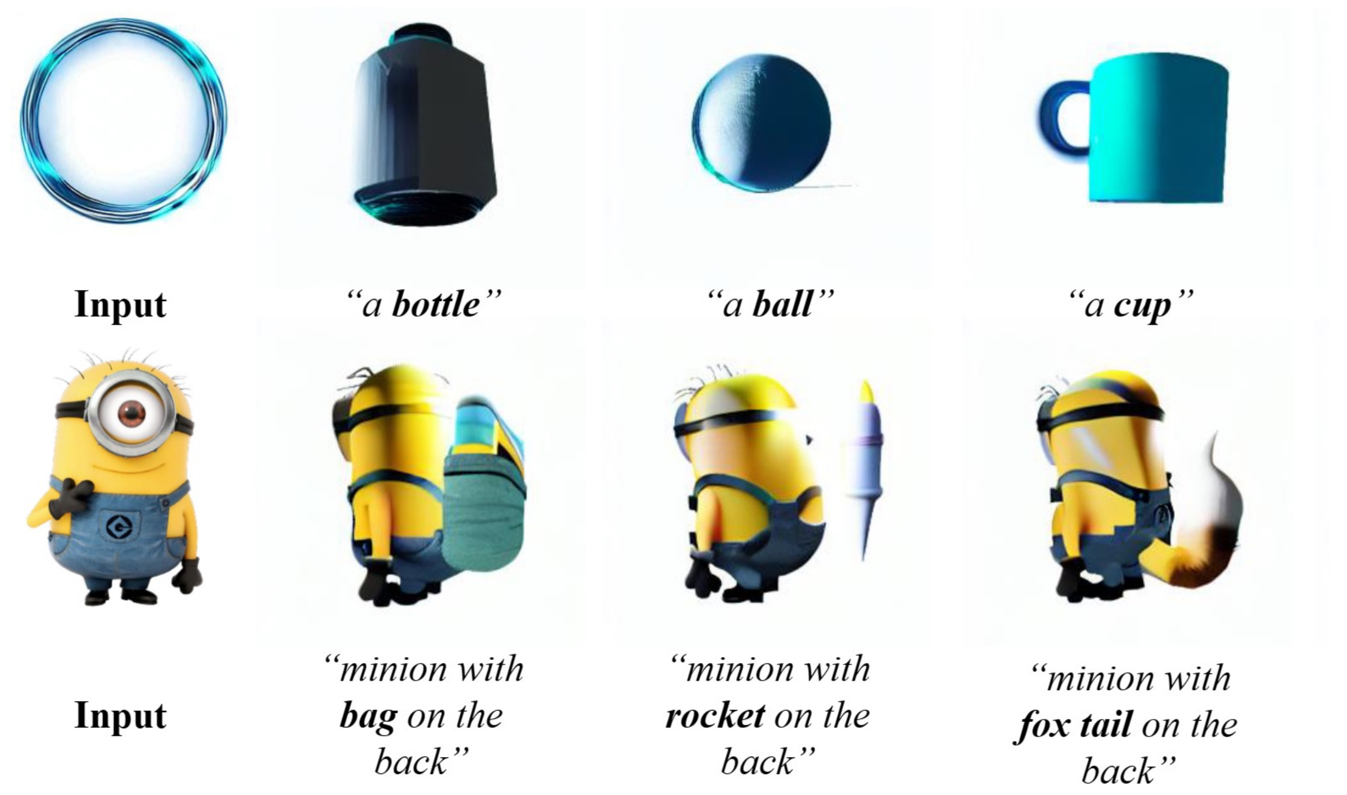
*"a box with **candies** in it"*

Input

*"a dragon stuffed toy with **fire** on the back"*

*"a dragon stuffed toy with **ice** on the back"*

*"a dragon stuffed toy with **green leaf** on the back"*



Input

*"a **bottle**"*

*"a **ball**"*

*"a **cup**"*

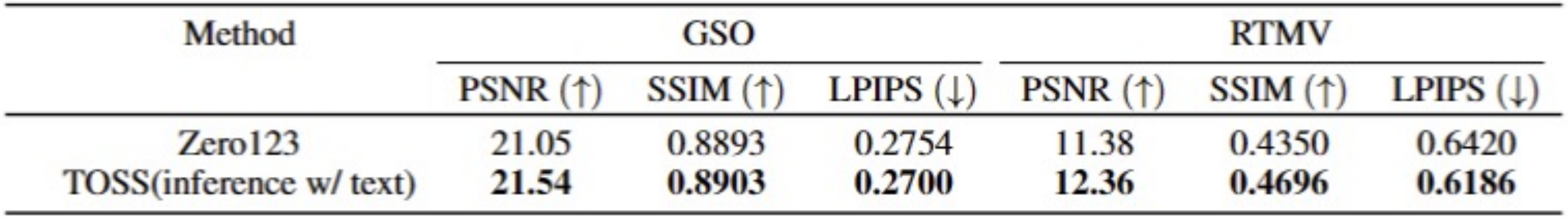
Input

*"minion with **bag** on the back"*

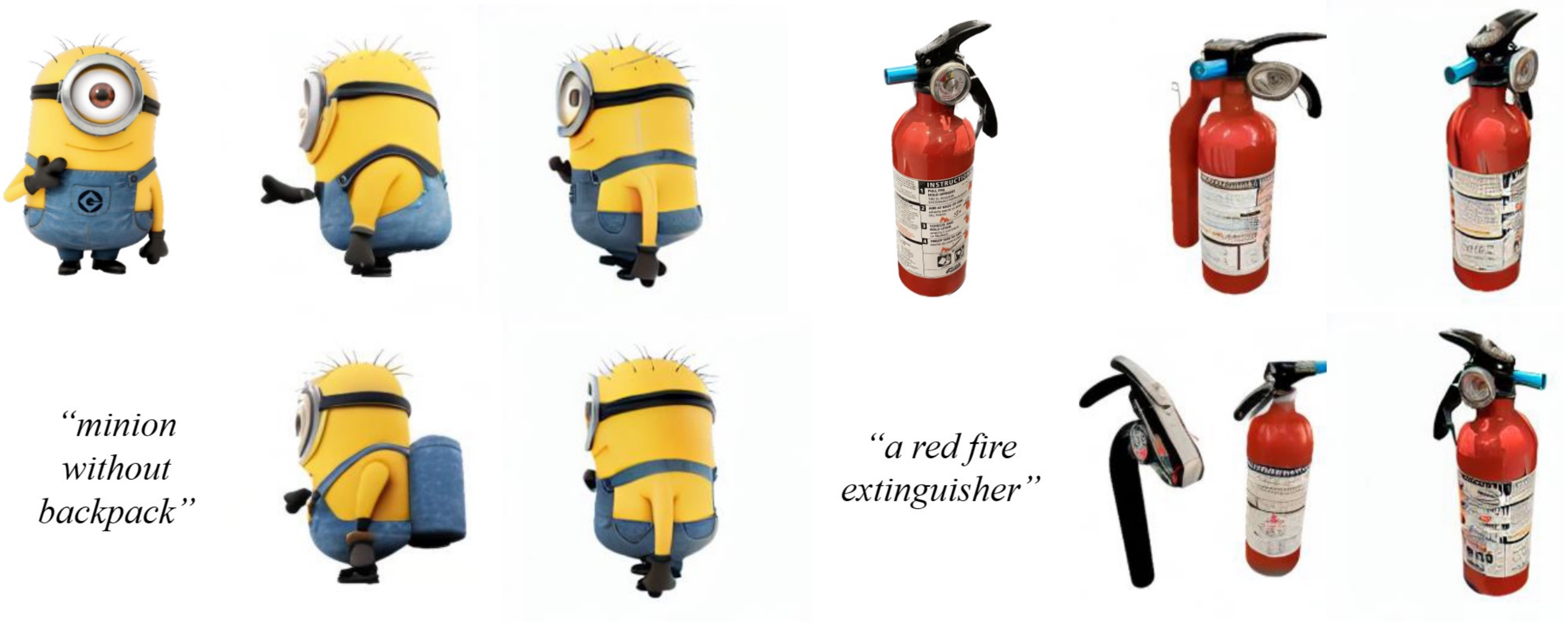
*"minion with **rocket** on the back"*

*"minion with **fox tail** on the back"*

Results: better multi-view consistency

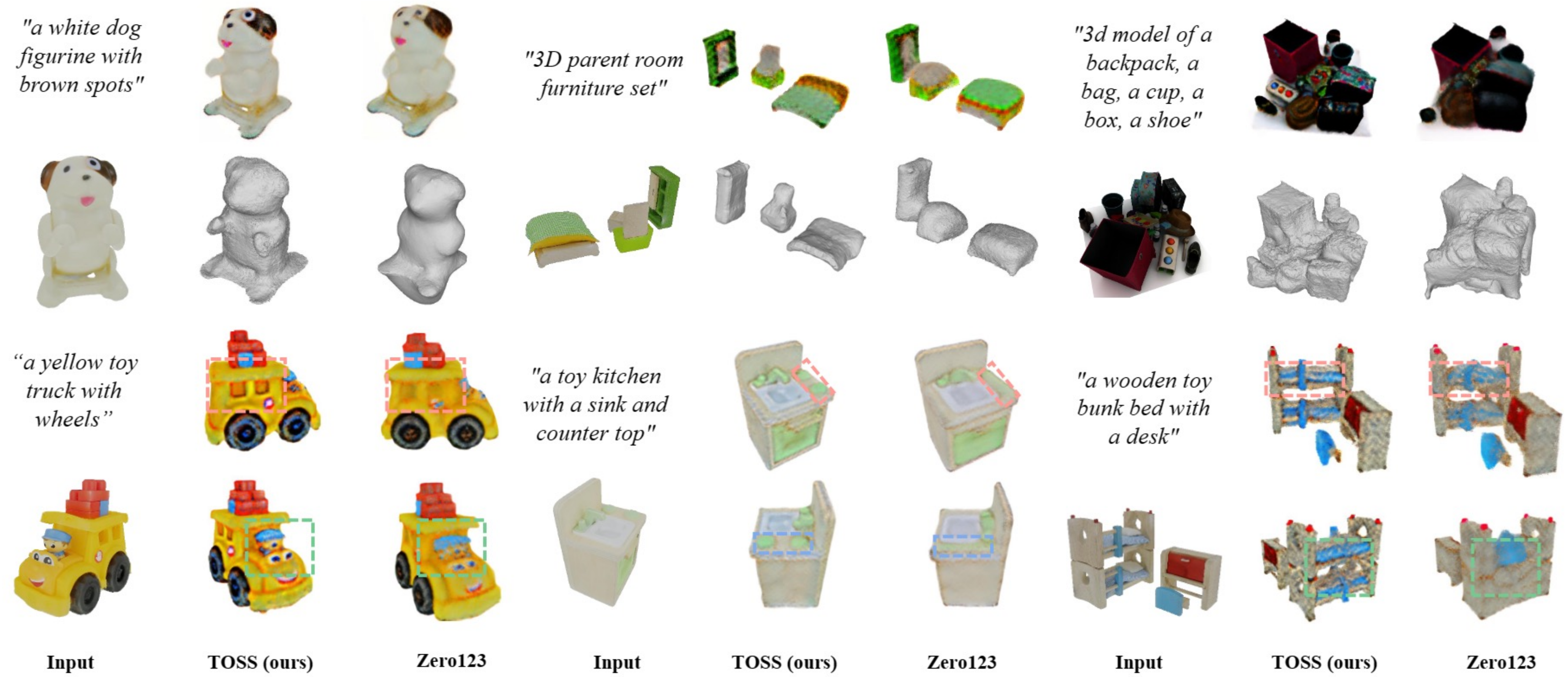


Method	GSO			RTMV		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Zero123	21.05	0.8893	0.2754	11.38	0.4350	0.6420
TOSS(inference w/ text)	**21.54**	**0.8903**	**0.2700**	**12.36**	**0.4696**	**0.6186**

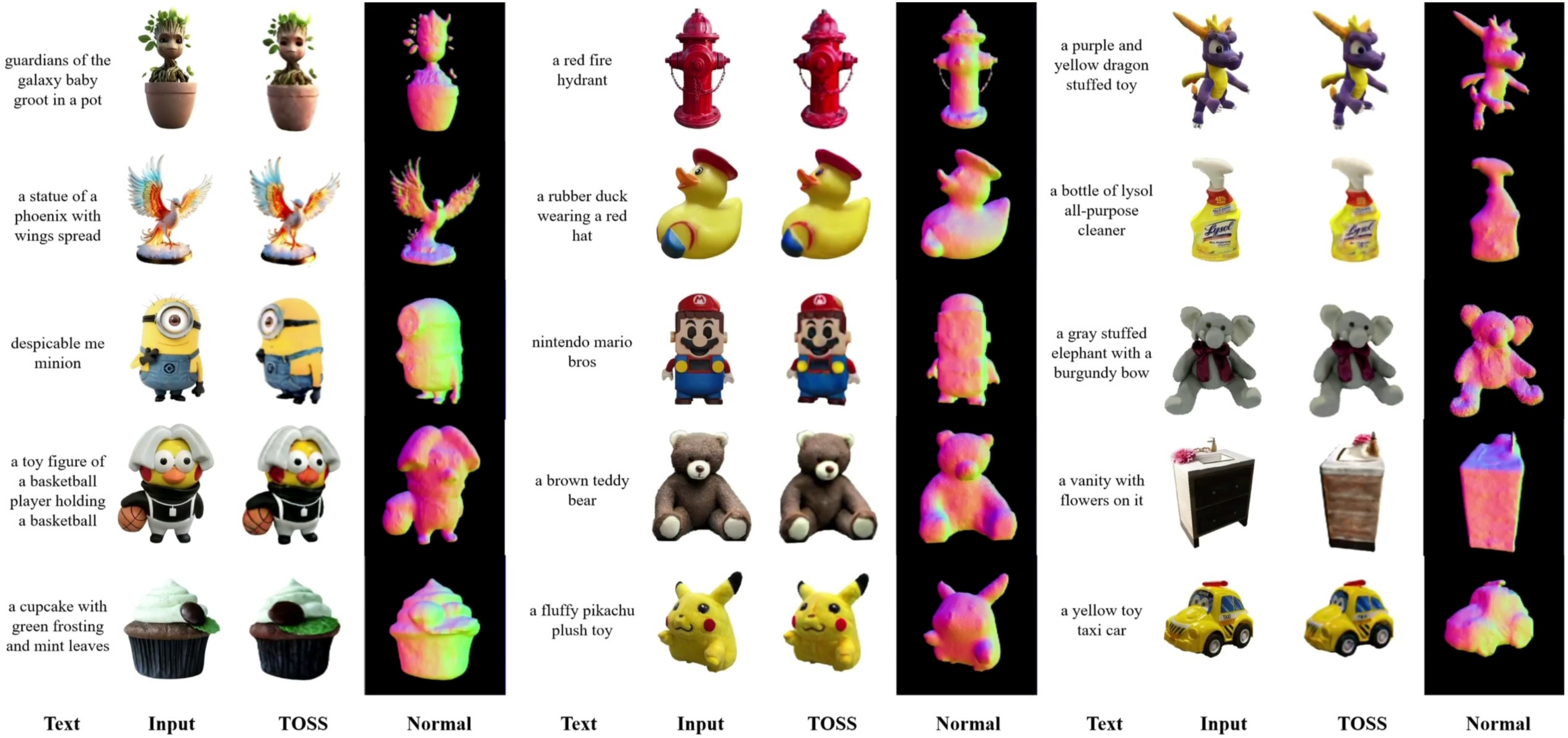


Input **Zero123** **TOSS** **Input** **Zero123** **TOSS**

Results: better 3D generation results


"a white dog figurine with brown spots"
"3D parent room furniture set"
"3d model of a backpack, a bag, a cup, a box, a shoe"
"a yellow toy truck with wheels"
"a toy kitchen with a sink and counter top"
"a wooden toy bunk bed with a desk"
Input
TOSS (ours)
Zero123
Input
TOSS (ours)
Zero123
Input
TOSS (ours)
Zero123

More 3D results


guardians of the galaxy baby groot in a pot
a statue of a phoenix with wings spread
despicable me minion
a toy figure of a basketball player holding a basketball
a cupcake with green frosting and mint leaves
a red fire hydrant
a rubber duck wearing a red hat
nintendo mario bros
a brown teddy bear
a fluffy pikachu plush toy
a purple and yellow dragon stuffed toy
a bottle of lysol all-purpose cleaner
a gray stuffed elephant with a burgundy bow
a vanity with flowers on it
a yellow toy taxi car
Text
Input
TOSS
Normal
Text
Input
TOSS
Normal
Text
Input
TOSS
Normal

Conclusion

- **What do we do?**
  - Introduce text constraints to NVS task for more controllable and more plausible (high-quality) results
- **Why we focus on NVS task?**
  - NVS from a single image is a significant proxy task of 3D generation, combining both 2D diversity and 3D geometry priors
- **Why we introduce texts to NVS task?**
  - As a high-level constraint, text information greatly improve the controllibility and plausibility of NVS results
- **How do we model the task?**
  - Geometry-free generation task
  - Texts constraint: finetune pretrained t2i stable diffusion model
  - Image constraint: dense cross attention
  - Camera pose constraint: key-value pair in dense cross attention
- **What conclusion can we get?**
  - T2i diffusion model deserves more attention in NVS task

Twitter

Thank you!

shiyukai22@gmail.com

Zhihu

Paper

Page

Code