Multimodal Semantic Bias Mitigation for Diverse Text-To-3D Generation

Supplementary Material

Figure 8. More qualitative comparisons with TRELLIS-text. Our method gain a better performance in more complex multi-target or fantastic object generation scenes, offering better semantic-aligned results, marked in blue.

Figure 9. Comparison results of different generate conditions, Refined denotes the prompt is refined following the format of 3D-FUTURE. Image-guided denotes the results of TRELLIS-image conditioned on CFD intermediate images.

Table 5. Comparisons of different reference images for TRELLIS-image. w/ SD denotes we directly use the Stable Diffusion (SD) model to generate image for TRELLIS-image. A&E denotes the 2D debiased method Attend-and-Excite [1] for SD model of CFD. * denotes only alter the words with the last 50% grad score.

Method	Prompts	Single Object				Multiple Object			
		Basic	Refined	Complex	Fantastic	Grouped	Action	Spatial	Imaginative
w/ SD	0	7.43	7.34	6.48	6.27	5.77	5.31	6.36	5.73
w/ CFD	0	7.67	7.42	6.52	6.33	5.81	5.34	6.40	6.07
w/ A&E	0	7.52	7.46	6.51	6.34	5.82	5.32	6.37	6.09
w/ A&E*	0	7.62	7.51	6.55	6.37	5.84	5.38	6.43	6.11
w/ CFD	1	7.82	7.57	6.54	6.51	5.93	5.38	6.47	6.14
w/ A&E*	1	7.94	7.66	6.57	6.59	5.93	5.42	6.43	6.14

A. More Qualitative Results

We present more results for our method using TRELLIS-text in the Fig. 8. Although TRELLIS-text could produce multi-view consistent 3D meshes, its performance declines when processing long or complex prompts due to its biased perception of the prompt text semantics. Specifically, as shown in rows 1 and 2, TRELLIS-text overemphasizes the words "teapot" and "tree", neglecting the overall semantics of the prompt. Our method refines the TRELLIS-text to generate robust, semantically consistent 3D meshes across varying prompt lengths and complexities to construct more diverse results.

B. Analysis of the Reference Images

As shown in the Table 5 and Fig. 9, we have the following findings: First, due to the limited training data, TRELLIS lacks some 3D semantic concepts. Thus directly constructing reference images using a 2D generation model or relying on a single intermediate reference image of CFD does not yield good generation results. Therefore, we fine-tuned TRELLIS by generating 3D meshes using the text-based 3D method CFD. Second, based on quantized word-level bias, we can combine existing 2D image debiasing methods to accurately enrich the semantics of reference images, thus promoting diverse 3D generation.

References

[1] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2