

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

Yukuan Min¹, Muli Yang^{†2}, Jinhao Zhang¹, Yuxuan Wang¹, Yihang Zhu¹, Jiexi Yan¹, Cheng Deng^{†1}

¹Xidian University, China

²A*STAR, Singapore

ykmin@stu.xidian.edu.cn, yangml@i2r.a-star.edu.sg, {zhangjinhao, wangyuxuan}@stu.xidian.edu.cn

zhuyihang@stu.xidian.edu.cn, yanjiexi@xidian.edu.cn, chdeng.xd@gmail.com *

Abstract

The latest progress in text-to-3D generative models makes it possible to generate high-quality 3D content. Recent text-to-3D large model have achieved remarkable breakthroughs in multi-view consistency. However, their effectiveness is often affected by inherent biases, resulting in sensitivity to design settings such as prompt format, leading to difficulty understanding complex prompts. To help text-to-3D generative models understand more diverse prompts, we propose a framework to localize and mitigate the bias in the current text-to-3D large model. Specifically, we first use the existing model to generate 3D content and use the quality evaluation model to identify the cross-modality bias. Then, we use the predicted quality score to quantify the contribution of the prompt text to the bias. Finally, in order to reduce these biases, we construct diverse pairwise examples to help the current text-to-3D large model construct unbiased visual-text connections. The experiment shows that our method has achieved competitive results and can provide higher quality, more diverse 3D content compared to existing methods.

1. Introduction

3D content generation plays a pivotal role across a wide range of domains, including product design, medical imaging, and scientific visualization [16]. Despite its broad applicability, creating high-quality 3D assets requires significant expertise. To alleviate this burden, text-to-3D generation [5, 25, 32] is proposed to create 3D assets that accurately align with text descriptions. A general solution involves distilling 3D representations directly from pre-trained 2D text-to-image diffusion models (DMs) via Score Distillation Sampling (SDS) [25, 31]. SDS leverages the generative capacity of DMs as powerful 3D priors to opti-

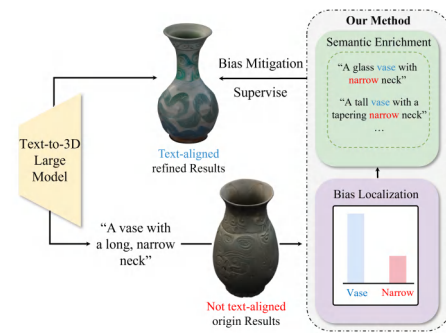


Figure 1. Our method localize and mitigate biases in text-to-3D large models and generate both cross-view consistent and text-aligned results.

mize differentiable 3D representations such as Neural Radiance Fields (NeRF) [23] and 3D Gaussian Splatting (3DGS) [15].

Despite having impressive quantitative results through SDS, qualitative analysis shows that such text-to-3D generation often results in cross-view inconsistency or ambiguous texture details due to the intrinsic bias of 2D diffusion priors [4, 17, 22, 32], and the learnt 3D assets with implicit 3D scene fail to be directly integrated into the graphics pipeline. Recently, a text-to-3D large model TRELLIS [35] was proposed. By training on a large amount of 3D asset datasets, TRELLIS can generate cross-view consistent 3D assets. However, since paired text-3D data is relatively scarce, TRELLIS fails to achieve diverse text-to-3D generation. We find this is due to inherent biases in the model. As shown in Fig. 2, TRELLIS gain significantly varying performance across different prompt types in the text-to-3D benchmark, and overfit specific prompt words. Therefore, TRELLIS tends to favor only a particular word (e.g., "vase") when provided with a textual prompt (e.g., "A ceramic vase with a long, narrow neck."), resulting in generated results that are inconsistent with semantics (see Fig.

*† The corresponding authors.

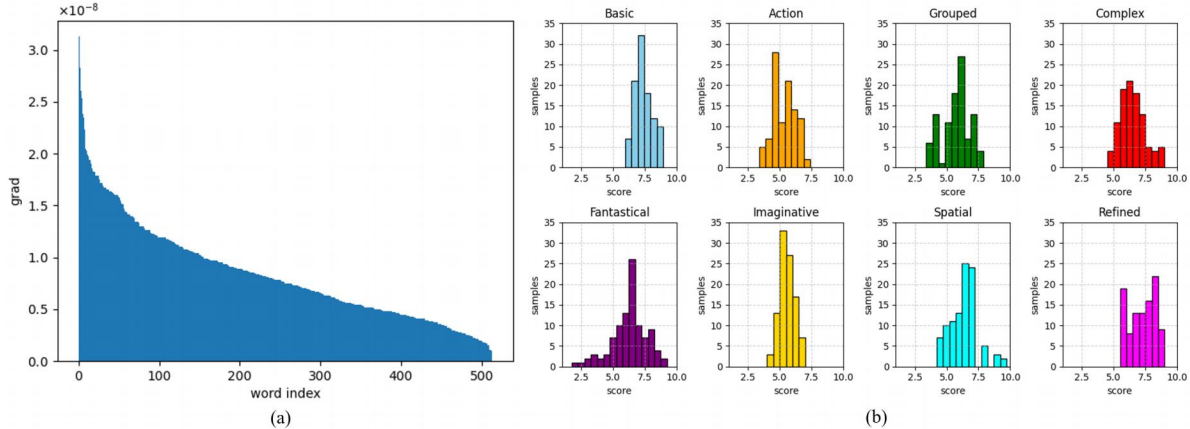


Figure 2. Quantitative results of TRELIS on the text-to-3D benchmark MATE-3D [43]. As shown, TRELIS has inherent bias. (a) Average grad of different prompt words on the MATE-3D benchmark. TRELIS over-focuses on few certain prompt words. (b) Distribution histogram of the quality scores of TRELIS for 8 different categories of text prompts on the MATE-3D benchmark. TRELIS can understand the common “Basic” prompts, but has a poor understanding of complex prompts such as “Fantastical” and “Grouped”.

4). Issues that reduce the applicability and adaptability of the current model for more diverse 3D generation.

To address the aforementioned limitations, as shown in Fig. 1, we present a pipeline for localizing and mitigating biases in the current TRELIS model and constructing a diverse and cross-view consistent text-to-3D generative model. Specifically, we first use the text-to-3D large model to generate 3D assets and analyze biases at the word level via a text-to-3D evaluation model. Then, with the word-level bias, we construct semantically stable and richer text-3D pairs with its assistance. Finally, the generated 3D pairs can be used to fine-tune the text-to-3D large model to improve the diversity of 3D generation results. The contributions of this work are:

- To the best of our knowledge, we are the first to address the problem of bias detection and mitigation in the text-to-3D generative model.
- We present a novel framework specifically designed for word-level bias localization and mitigation. By leveraging a text-to-3D evaluation model, our framework could extend more semantically stable and richer text-3D pairs.
- Extensive results have shown our framework can be a new benchmark for text-to-3D generation. Our methods outperform 8 state-of-the-art methods and delivering highly impressive and controllable qualitative results.

2. Related Work

2.1. Text-to-3D Generation

While 2D image generation has achieved remarkable success, extending these advances to 3D remains a major challenge, largely due to the difficulty in acquiring high-quality 3D data. The emergence of Neural Radiance Fields (NeRF) [3, 23] offered a significant breakthrough, enabling

3D scene reconstruction using only multi-view 2D images. Leveraging the strong priors from powerful text-to-image diffusion models, researchers have introduced several pioneering approaches—such as DreamField [13], DreamFusion [25], and SJC [31]—that aim to synthesize 3D content guided by textual prompts. Nevertheless, early methods relying on vanilla Score Distillation Sampling (SDS) [25] are plagued by limitations including over-saturation, over-smoothing, and the Janus problem. To overcome these issues, more recent works [41, 45, 46] propose various enhancements to improve generation fidelity. Concurrently, the growing availability of large-scale 3D datasets [6, 7] has enabled the development of methods [19–21, 28] that better utilize multi-view supervision to mitigate view-inconsistency issues like the Janus effect. Unlike previous methods, we aim to analyze and address the bias of text-to-3D large models at the data level to improve their generation diversity.

2.2. Bias Detection

Bias mitigation in generative models [36–39] has been a long-standing area of research. For instance, [14] proposes gradient clipping to promote fairer representations across sensitive groups, while [29] focuses on adjusting the semantic distribution of the latent space during inference to enhance fairness. More recently, research has increasingly turned to addressing biases in text-to-image (T2I) diffusion models. FairDiffusion [8] introduces fairness constraints via user-defined instructions, incorporating a fairness-aware term into the classifier-free guidance mechanism [11] to steer Stable Diffusion [27] toward more equitable outputs, particularly in job-related contexts. Similarly, ITI-GEN [42] adopts prompt tuning strategies to mitigate bias in digit generation tasks.

3. Method

We localize and mitigate cross-modality bias in the text-to-3D generation model as in Fig. 3. First of all, in the localization stage, we use real prompt captions to generate 3D contents with the text-to-3D generation model. Then, we adopt a text-to-3D evaluation model to evaluate the potential biases. Finally, these biases will be used to guide the generation of semantically stable and diverse text-3D pairs, providing enhanced supervision information for the text-to-3D large model.

3.1. Bias Localization

Formulation of the Bias. In the evaluation of text-to-3D generative models, we expect the model’s ability to be an ideal fixed quantity ϕ . Given the evaluation data $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$, the evaluation model constructs an estimator $\hat{\phi} = \mathcal{E}(\mathcal{X})$ as a function of the data $\mathcal{X}$ to provide an approximation of ϕ . Then, the difference between the ideal expected value ϕ and $\hat{\phi}$ can be defined as the evaluation bias:

$$\epsilon(\hat{\phi}) = \hat{\phi} - \phi, \quad (1)$$

where ϕ is the target expected value of $\mathcal{X}$, so a bias of 0 means that the estimator is unbiased and can provide perfect estimates.

Let G be the text-to-3D generative model. Our objective is to evaluate the bias of G over the evaluation data $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$. Given a generated mesh x_n and its corresponding prompt t_n , traditional single-dimensional evaluators directly map them to the subjective score q [10]. The procedure can be described as $\hat{q} = \psi(F(x, t))$, where $\hat{q}$ is the predicted score, $F(\cdot)$ denotes a feature extractor and $\psi(\cdot)$ represents a mapping function. Recently, some researchers have proposed a multi-task network with a shared feature extractor and multiple mapping heads for multi-dimensional evaluation,

$$\hat{q}_i = \psi(F(x, t) | \pi(f_c^i)), \quad (2)$$

where $\pi(\cdot)$ denotes a hypernetwork to generate θ_i for a unified mapping head; f_c^i denotes a condition feature related to the i -th evaluation dimension. Specifically, for a given prompt caption $t_n \in \mathcal{T}$, we generate a set of D meshes x_n^d , defined as:

$$x_n^D = \{G(t_n)\} \in \mathbb{R}^D, \quad (3)$$

thus we could evaluate the bias $\epsilon(\hat{\phi})$ over $\mathcal{X}$ by querying a state-of-the-art text-to-3D evaluation model. For each 3D mesh x_n^d , the text-to-3D evaluation model infer the quality score q according to the corresponding prompt t_n , hence, we define the predicted score q as:

$$q = \psi(x_n^d, t_n). \quad (4)$$

By applying the text-to-3D evaluation model to the whole dataset $\mathcal{X}$, we then explore two different scenarios: local bias, where we examine bias under the same context prompt t_n . And global bias, where we consider the entire set of generated 3D meshes associated with the bias.

For the local bias, we capture the bias by considering the prompt caption context t_n . Since the factors within the certain prompt may influence the direction and intensity of the bias. For instance, the captions “A ceramic vase with swirling patterns etched into the surface, a chipped rim, and a long, narrow neck” and “A blue vase” both describe a vase, but the former focuses on shape while the latter focuses on texture. Specifically, we assess the bias by analyzing the set of 3D meshes x_n^D generated from a specific prompt t_n , effectively collecting statistics at the local level. Formally, the local quality score q_l is computed as:

$$q_l = \frac{1}{|x_n^D|} \sum_{x_n^d} q, \quad (5)$$

where $q = \mathcal{E}(x_n^d, t_n)$ is the prediction of the evaluation 3D model as defined in Eq. 4.

In addition to local bias, we also need to understand the overall behavior of model G to understand the direction toward which the bias tends and the overall intensity of the bias. Therefore, we propose to take the average quality score for q_g in all prompt captions T to describe the overall bias of the model:

$$q_g = \frac{1}{|\mathcal{X}|} \sum_{(x,t) \in \mathcal{X}} q. \quad (6)$$

Note that the local bias is a special case of the global bias, where x is a certain instance, i.e., $x = x_n$.

Quantification of the Bias. After determining the form of bias in text-to-3D models, we need to quantify the cross-modality bias between the visual 3D meshes and text prompts. Specifically, our target is to explore the influence of the prompt text information on the bias in 3D visual information. To achieve this, as in Eq. 1, since $\epsilon(\hat{\phi}) = \hat{\phi} - \phi \propto \hat{\phi}$, we first compute the $\hat{\phi}$, i.e., the predicted score q as in Eq. 4. To analyze which parts of the prompt contribute most to the bias, we perform backpropagation using the predicted score to compute the gradients with respect to the text token embeddings in the text-to-3D evaluation model. Given a prompt caption t_n , we denote the sequence of token embeddings as $\{e_0, e_1, \dots, e_n\}$. For each embedding e_i , we calculate the absolute value of the scalar gradients as:

$$e'_i = \sum \left| \frac{\partial q}{\partial e_i} \right|, q \in \{q_l, q_g\}. \quad (7)$$

Therefore, we define e'_i as an estimate of the contribution of token e_i to the observed bias. To quantify the influence

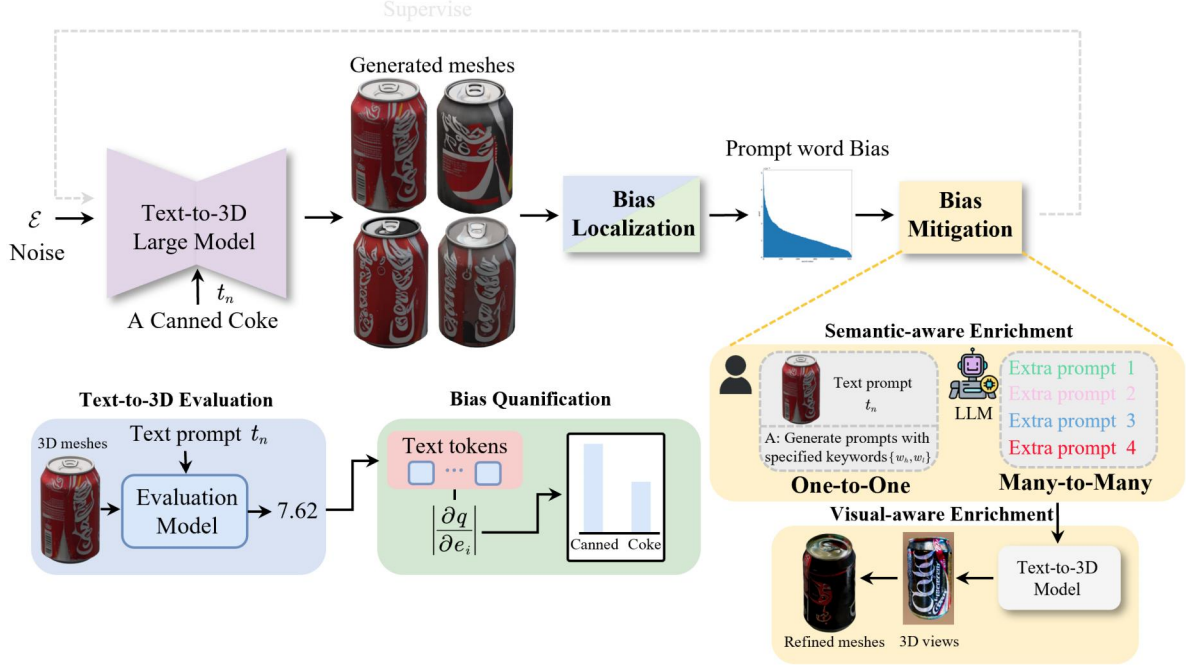


Figure 3. Overview of our method. Our method first uses the text-to-3D large model to generate examples, then localizes the bias with a 3D evaluation model. Finally, we quantify the bias and use it to guide the generation of text-3D pairs with diverse semantics to fine-tune the text-to-3D large model.

Algorithm 1 Text-3D pairs enrichment for Bias Mitigation

Input: Prompt t_n .

- 1: Extract meshes x_n^D via the Text-to-3D Large Model.
- 2: **for** x_n^d in $x_n^D = \{x_n^1, x_n^2, \dots\}$ **do**
- 3: Calculate quality score q with Eq. 4.
- 4: Update local bias e_l and global bias e_g with grad e_i according to Eq. 5 and Eq. 6.
- 5: Calculate the grad e'_i for each token in t_n with Eq. 7.
- 6: Enrich prompt text t_n based on e'_i of each token.
- 7: Enrich 3D meshes x_n^d according to rank model.
- 8: **end for**

Output: Enriched samples $\{x_{nC}, t_{nC}\}$

of each word $w \in t_n$ in the prompt captions, we aggregate the gradients of the tokens that compose that word. The intuition is that the magnitude of these gradients reflects the sensitivity of the model’s output to changes in specific input words. Thus they can serve as a proxy for measuring how much each word contributes to the quality score of the text-to-3D evaluation model, and consequently indicate the contribution of each word in influencing the generative model towards biased outcomes.

3.2. Bias Mitigation

To solve the imbalances ϵ in the 3D generative model, our idea is to train a text-to-3D generative model that con-

tains wider semantic information while maintaining its general knowledge. To implement this idea, we leverage the quantified bias for controllable generation of enhanced text-3D pairs, including: 1) semantic-aware enrichment and 2) visual-aware enrichment.

Semantic-aware Enrichment. To enrich the semantic information, motivated by the theory of Invariant Risk Minimization [2], we construct diverse context environments $\mathcal{C} = \{C_1, C_2, \dots\}$ with different context prompt captions. Specifically, as previously described, we compute the gradient of the predicted score to obtain the most important word w_h and the least important word w_l in the prompt caption t_n . Then, we generate prompts with different contexts with the assistance of GPT-4 [1] while preserving w_h and w_l . Through this operation, we enrich the semantic information of the model by expanding its understanding of uncommon words w_l while preserving the base semantics w_h . Based on the instruction, we utilize GPT-4 to create a large pool of candidate prompts tailored to the identified challenges. Finally, we manually filter out any duplicate or inappropriate prompts to ensure consistency and diversity.

Visual-aware Enrichment. After the prompt generation, we adopt a recent text-to-3D generative method CFD [40] to generate corresponding text-3D aligned representations by using open source code, which are then transformed into textured meshes. Then, we adopt a rank model following [44] to select semantic faithful 3D meshes.

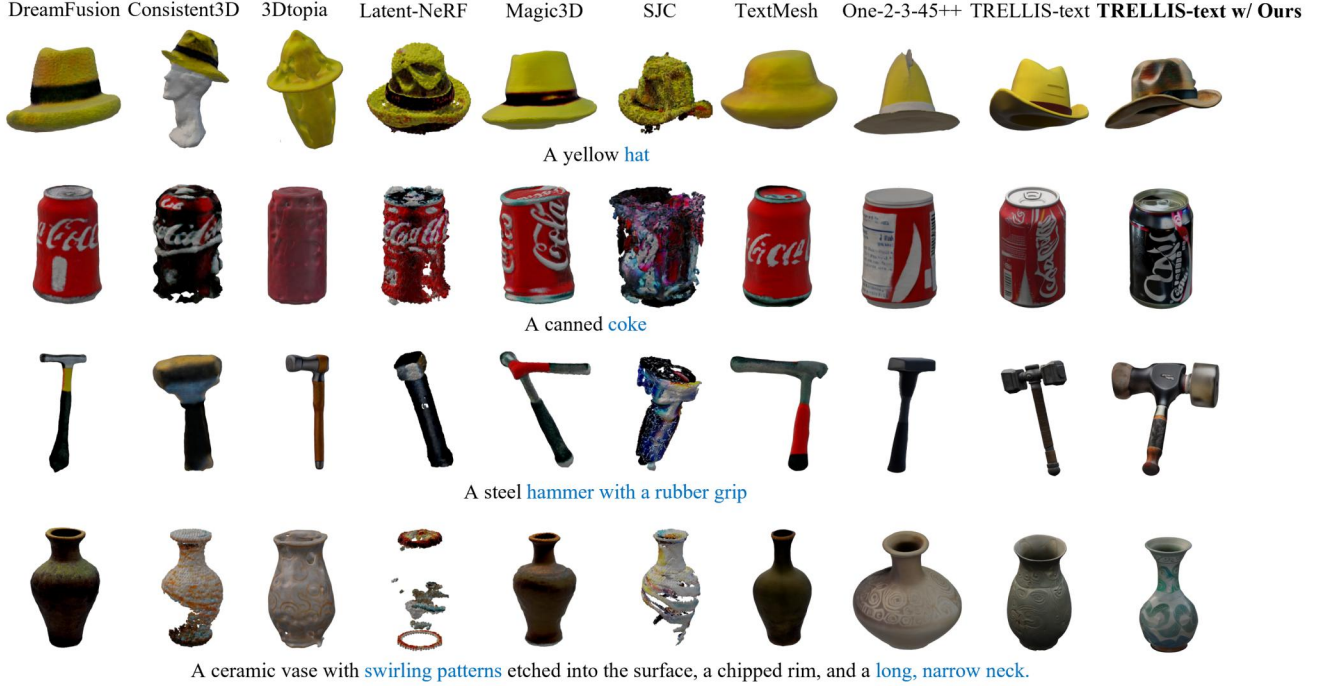


Figure 4. Qualitative comparisons on the benchmark of MATE-3D. Existing methods struggle with text matching, as marked in blue. Our method improves both text matching and geometric/texture details.

Table 1. Qualitative comparisons on 160 prompts generated by MATE-3D. We calculate the average scores (in terms of overall quality) of each generative method on MATE-3D for comprehensive 3D quality evaluation. The best performance in each case is shown in bold.

Method	Single Object				Multiple Object			
	Basic	Refined	Complex	Fantastic	Grouped	Action	Spatial	Imaginative
DreamFusion [25]	4.99	4.96	4.64	3.24	3.45	3.73	4.47	3.45
Magic3D [17]	5.26	4.89	5.06	4.95	3.92	4.73	5.06	4.93
SJC [31]	3.28	3.48	2.95	2.92	2.50	2.77	2.92	3.25
TextMesh [30]	4.08	4.56	4.79	4.22	3.22	3.68	4.17	4.26
3DTopia [12]	5.06	4.91	4.83	3.89	4.05	2.80	3.89	2.95
Consistent3D [34]	4.15	4.92	4.40	4.32	3.37	3.64	4.39	3.47
LatentNeRF [22]	3.20	3.76	3.47	4.16	2.94	3.48	3.46	4.01
One-2-3-45++ [18]	7.79	7.69	6.50	6.60	6.49	5.65	6.81	6.13
TRELLIS-text [35]	7.39	7.27	6.50	6.24	5.73	5.30	6.32	5.57
TRELLIS-text w/ ours	8.19	7.83	6.72	6.95	6.64	5.70	6.92	6.25

Specifically, given the origin text-3D pair $\{x_n, t_n\}$ and the generated relative prompts $t_{n_C} = \{t_{n_1}, t_{n_2}, \dots\}$, we compare the generated 3D mesh x_{n_C} and the original 3D mesh x_n yields a preferred prediction $\mathbf{x}^{\text{win}}$ and a less preferred one $\mathbf{x}^{\text{lose}}$. We only select generated 3D meshes that belong to $\mathbf{x}^{\text{win}}$ to build new text-3D pairs $\{x_{n_C}, t_n\}$. By construct a many-to-many text-to-3D assignment, we could enrich the supervision signals from the text modality. Algorithm. 1 shows details of enrich the text-3D pairs for bias mitigation. To reduce computation cost, we directly calculate and count the average gradient instead of calculating the average score first and then calculating its gradient.

4. Experiments

4.1. Experimental Setups

Datasets and measurements. We here evaluate the proposed method with 160 prompts from MATE-3D [43] and 300 prompts from T^3 Bench [9]. Specifically, in MATE-3D, we perform pairwise comparisons with baselines under different prompt categories. For single object generation, four sub-categories are defined including “Basic”, “Refined”, “Complex” and “Fantastical”. For multiple object generation, the text prompt is composed of at least two objects with specified relationships between the objects, which are

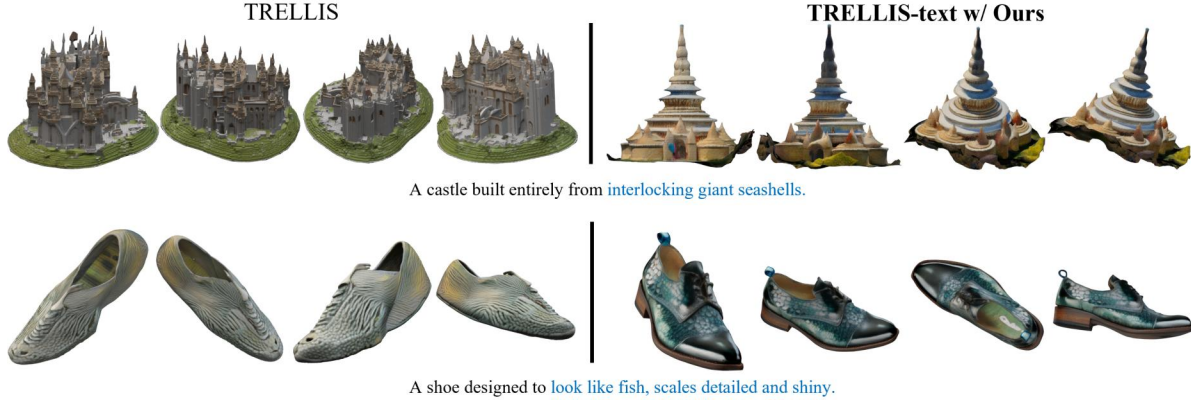


Figure 5. Qualitative comparisons with TRELIS-text. Our method gain a better performance in fantastic object generation scenes, offering better semantic-aligned results, marked in blue.

Table 2. Qualitative comparisons on 300 prompts generated by T^3 Bench. The best performance in each case is shown in bold.

Method	Single Object			Single Object with Surroundings			Multiple Objects		
	Quality	Alignment	Average	Quality	Alignment	Average	Quality	Alignment	Average
DreamFusion [25]	24.9	24.0	24.4	19.3	29.8	24.6	17.3	14.8	16.1
Magic3D [17]	38.7	35.3	37.0	29.8	41.0	35.4	26.6	24.8	25.7
SJC [31]	26.3	23.0	24.7	17.3	22.3	19.8	17.7	5.8	11.7
LatentNeRF [22]	34.2	32.0	33.1	23.7	37.5	30.6	21.7	19.5	20.6
RichDreamer [26]	57.3	40.0	48.6	43.9	42.3	43.1	34.8	22.0	28.4
ProlificDreamer [32]	51.1	47.8	49.4	42.5	47.0	44.8	45.7	25.8	35.8
TRELIS-text [35]	45.3	44.2	44.8	40.6	46.2	43.4	39.2	17.8	28.5
TRELIS-text w/ ours	52.1	48.3	50.2	44.4	51.2	47.8	46.7	28.3	37.5

denoted by “Grouped”, “Action”, “Spatial”, and “Imaginative” according to the type of the relationships. We calculate average overall quality scores of each generative method on eight prompt categories; In T^3 Bench, we perform comparisons for single object, single object with surroundings, and multiple objects generation. We normalize the original scores on quality and alignment assessment from the range $[-2.5, 2.5]$, $[1, 5]$ to $[0, 100]$.

Baselines. Following MATE-3D [43], we benchmark our method against 8 recent text-to-3D generative methods, including DreamFusion [25], Magic3D [17], Score Jacobian Chaining (SJC) [31], TextMesh [30], 3DTopia [12], Consistent3D [34], LatentNeRF [22], and One-2-3-45++ [18]. As for T^3 Bench, we benchmark our method against recent 5 text-to-3D generative methods, including DreamFusion [25], Magic3D [17], Score Jacobian Chaining (SJC) [31], LatentNeRF [22], RichDreamer [26], and ProlificDreamer [32].

Implementation. We conduct experiments using PyTorch [24] with TRELIS [35] as the backbone of our method. We follow the training strategy of TRELIS and use HPSv2 [33] as the rank model for visual-aware enrichment. In T^3 Bench, we set the five focal lengths used for multi-focal capturing to 3.0, 4.0, 5.0, 6.0, 7.5, and set the

resolution of the rendered image to 512×512 . All experiments are conducted on an NVIDIA A6000-48GB GPU.

4.2. Comparison with Prior Methods

Qualitative Comparisons. We conduct two qualitative evaluations, which include comparing benchmarks of MATE-3D [43] and TRELIS [35]. As illustrated in Fig. 4 and Fig. 5, the evaluation results demonstrate the consistent improvements in text alignment, generation stability, and fine-grained texture-geometry consistency. First, compared with previous methods, our method can generate more realistic textures and more reasonable shapes. For example, in the prompt “A yellow hat”, our result takes into account both the wrinkles on the hat and the interior space of the hat, and has a more realistic texture. Second, although baseline methods produce high-fidelity outputs, they frequently fail to align with the input text, as indicated by blue highlights. For example, in the prompt “A ceramic vase with swirling patterns etched into the surface, a chipped rim, and a long, narrow neck”, most baselines omit key elements such as the “long, narrow neck” or “swirling patterns” (see the last row in Fig. 4). In contrast, our approach accurately captures all described components, highlighting its effectiveness in maintaining semantic alignment.

Table 3. Ablation study for prompt generation on MATE-3D.

Method	Single Object				Multiple Objects			
	Basic	Refined	Complex	Fantastic	Grouped	Action	Spatial	Imaginative
One-2-3-45++ [18]	7.79	7.69	6.50	6.60	6.49	5.65	6.81	6.11
One-2-3-45++ [18] w/ prompt	7.86	7.83	6.52	6.64	6.50	5.67	6.83	6.14
TRELLIS-text [35]	7.39	7.27	6.50	6.24	5.73	5.30	6.32	5.57
w/o grad guide	4.15	4.57	3.62	3.24	3.37	2.61	4.27	3.11
only w_l	4.39	4.96	5.14	5.23	3.11	2.84	4.11	4.21
only w_h	7.68	7.44	6.57	6.43	5.81	5.53	6.41	5.77
Ours	8.19	7.83	6.72	6.95	6.64	5.70	6.92	6.25

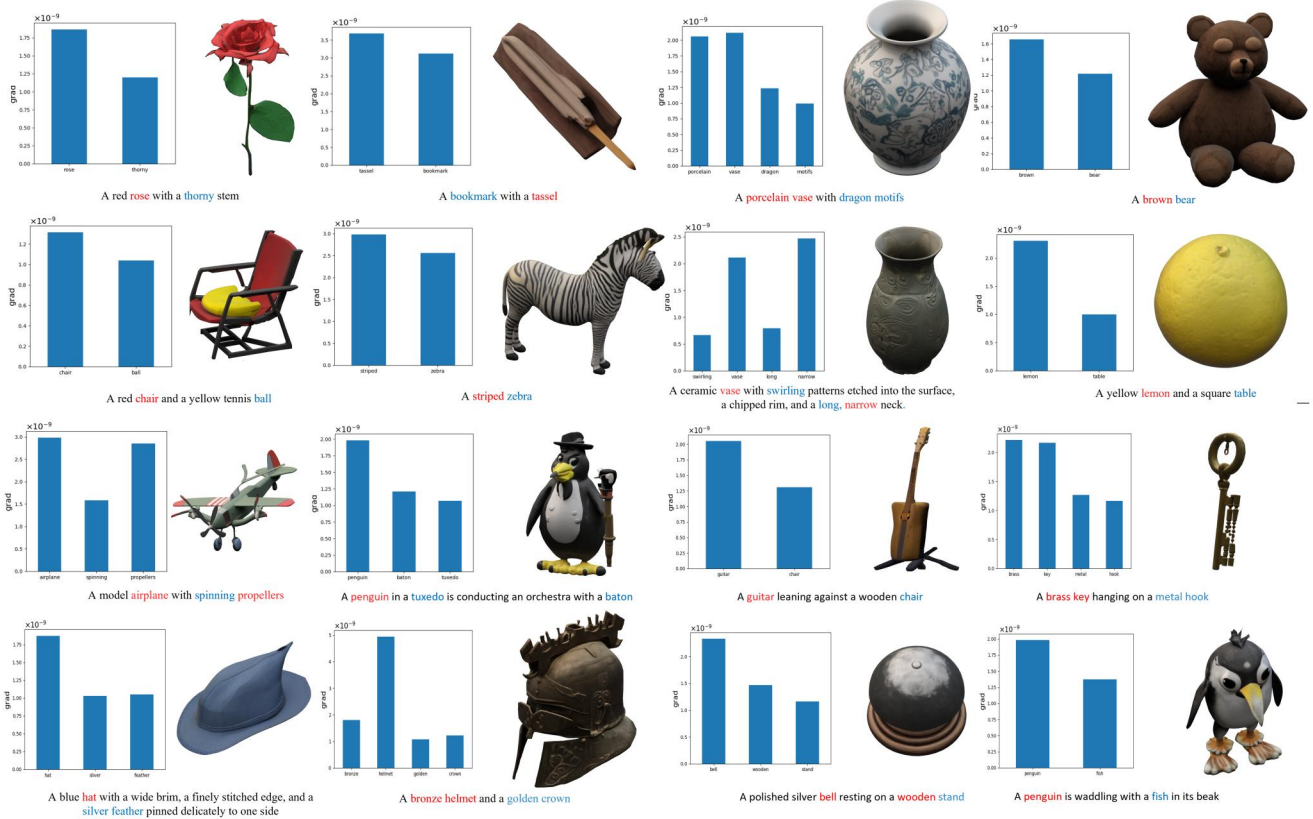


Figure 6. We explored the potential of our method to leverage 3D evaluation models (e.g., MATE-3D) to identify word-level bias, specifically the cause of poor 3D asset generation. The figure shows the gradients of relative prompt words and the corresponding generation results. We highlight in red the most influential word identified by our method. The model’s overemphasis on influential words leads to insufficient understanding of other words.

Quantitative Comparisons. We present quantitative comparisons on the MATE-3D [43] and T^3 Bench [9] benchmark. The results in Table 1 and Table 2 demonstrate that our method surpasses previous state-of-the-art (SOTA) approaches, ranking first across all metrics. Compared with the prior best method, One-2-3-45++, our approach achieves a minimum score improvement of 0.12, a maximum of 0.4, and an average improvement of 0.19. These results highlight the effectiveness of our method in enhancing both text alignment and geometric fidelity, while

preserving 3D consistency.

4.3. More Analyses and Justifications

Verification for Localizing Bias. We further explore the potential of our method for explicit bias localization in 3D content evaluation. As illustrated in Fig. 6, the gradient-based statistical analysis could localize bias at the token level, linking specific linguistic concepts to geometric or appearance-related distortions in 3D meshes. The visualization results demonstrate that biased prompt words corre-

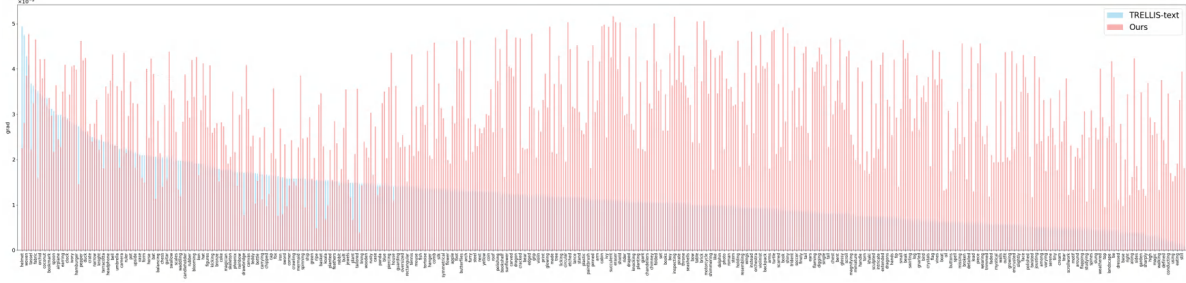


Figure 7. Comparison of TRELLIS and our method for the gradient of all prompt words in the MATE-3D benchmark. Our method alleviates the bias of TRELLIS for certain prompt words and expands the semantic information.

Table 4. Analysis for computation cost for enriched prompt numbers and training samples on MATE-3D. w/o training denotes we input the intermediate images generated by CFD to TRELLIS-image.

Method	Time(h)	Training	Enriched Prompts	Samples(%)	Single Object		Multiple Object	
					Basic	Refined	Spatial	Imaginative
TRELLIS-text [35]	-	w/o	-	-	7.39	7.27	6.32	5.57
TRELLIS-image [35]	4	w/o	0	-	7.67	7.42	6.40	6.07
TRELLIS-image [35]	8	w/o	1	-	7.74	7.53	6.43	6.12
TRELLIS-text [35]	10.5	w/	1	Last 50% score	7.87	7.63	6.44	6.15
TRELLIS-text [35]	10.5	w/	1	Last 50% grad	7.92	7.67	6.57	6.19
TRELLIS-text [35]	13	w/	1	100%	8.07	7.81	6.63	6.22
TRELLIS-text [35]	25	w/	4	100%	8.19	7.83	6.92	6.25
CFD [40]	4	-	0	-	8.02	7.65	6.51	6.24
CFD [40]	8	-	1	-	8.21	7.78	6.68	6.37
CFD [40]	20	-	4	-	8.34	7.92	7.08	6.67

spond to inconsistent content in the meshes, revealing the potential of our method for explaining connections between semantics and 3D meshes.

Evaluation on Mitigating Bias. To provide a more detailed analysis of our method’s effectiveness in mitigating the bias in the text-to-3D model, Fig. 7 show the detailed results of comparing TRELLIS and TRELLIS w. ours with respect to the average grad (importance) of prompt words in the MATE-3D task. We can observe that our method effectively prevents the model from overfitting the general pattern on the few head prompt words and improves the model’s comprehension of tail prompt words.

Ablation study for prompt generation. Ablation study for prompt generation. As shown in Table 3, vanilla TRELLIS still performs worse than the state-of-the-art method One-2-3-45++ due to semantic bias, and simply augmenting prompts without considering semantic bias may not significantly improve the performance. For a more comprehensive evaluation, we list the results of the existing method One-2-3-45++ with the augmented prompt. Our method significantly improves the performance of vanilla TRELLIS by alleviating semantic bias. We investigate the impact of using w_h and w_l for generating relative prompts. Specifically, we provide results in Table 3, which show that only using w_h can also generate diverse text-3D pairs while preserving

semantics, but adding w_l can further enhance the semantic diversity. To generate diverse text-3D pairs, we suggest using both w_h and w_l .

Analysis for Computation cost. We analyzed the computational cost of our method in Table 4. Due to the need for model fine-tuning and enriched sample generation, our method takes about 25 hours. To fit scenarios with limited computational resources, our method is compatible with a TRELLIS training-free pipeline by using intermediate images to guide the image-to-3D version of TRELLIS-image. Furthermore, we could reduce the computational cost by sampling hard samples based on word-level bias. As in Table 4, this approach reduces the generation time by 50% (to around 10.5 hours) while maintaining improved performance and achieving better performance than using the evaluation scores for hard sampling.

5. Conclusion

In this work, we propose a framework to localize and mitigate the bias in the current text-to-3D model. Specifically, we leverage the identified biases to guide the creation of semantically consistent and diverse text-3D pairs to finetune the text-to-3D large model. Comprehensive experimental results demonstrate our method surpasses previous works in both output quality and controllability.

Acknowledgement. Our work is supported by the National Natural Science Foundation of China (U25B2048, 62132016).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4
- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 4
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2
- [4] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 1
- [5] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 1
- [6] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023. 2
- [7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 2
- [8] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 2
- [9] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T3 bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023. 5, 7
- [10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3
- [11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2
- [12] Fangzhou Hong, Jiaxiang Tang, Ziang Cao, Min Shi, Tong Wu, Zhaoxi Chen, Shuai Yang, Tengfei Wang, Liang Pan, Dahua Lin, et al. 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*, 2024. 5, 6
- [13] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 867–876, 2022. 2
- [14] Patrik Joslin Kenfack, Kamil Sabbagh, Adín Ramírez Rivera, and Adil Khan. Repfair-gan: Mitigating representation bias in gans using gradient clipping. *arXiv preprint arXiv:2207.10653*, 2022. 2
- [15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1
- [16] Chenghao Li, Chaoning Zhang, Joseph Cho, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv preprint arXiv:2305.06131*, 2023. 1
- [17] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023. 1, 5, 6
- [18] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10072–10083, 2024. 5, 6, 7
- [19] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2
- [20] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.
- [21] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 2
- [22] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12663–12673, 2023. 1, 5, 6
- [23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

- thesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2
- [24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 5, 6
- [26] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9914–9925, 2024. 6
- [27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [28] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2
- [29] Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020. 2
- [30] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. In *3DV*, 2024. 5, 6
- [31] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12619–12629, 2023. 1, 2, 5, 6
- [32] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36: 8406–8441, 2023. 1, 6
- [33] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 6
- [34] Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, and Hanwang Zhang. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9892–9902, 2024. 5, 6
- [35] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 1, 5, 6, 7, 8
- [36] Chenghao Xu, Jiexi Yan, Yanhua Yang, and Cheng Deng. Implicit compositional generative network for length-variable co-speech gesture synthesis. *IEEE Transactions on Multimedia*, 26:6325–6335, 2023. 2
- [37] Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. Llm knows body language, too: Translating speech voices into human gestures. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5004–5013, 2024.
- [38] Chenghao Xu, Jiexi Yan, Muli Yang, and Cheng Deng. Rethinking noise sampling in class-imbalanced diffusion models. *IEEE Transactions on Image Processing*, 33:6298–6308, 2024.
- [39] Chenghao Xu, Jiexi Yan, and Cheng Deng. Keep and extent: Unified knowledge embedding for few-shot image generation. *IEEE Transactions on Image Processing*, 2025. 2
- [40] Runjie Yan, Yinbo Chen, and Xiaolong Wang. Consistent flow distillation for text-to-3d generation. In *The Thirteenth International Conference on Learning Representations*, 2024. 4, 8
- [41] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023. 2
- [42] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023. 2
- [43] Yujie Zhang, Bingyang Cui, Qi Yang, Zhu Li, and Yiling Xu. Benchmarking and learning multi-dimensional quality evaluator for text-to-3d generation. *arXiv preprint arXiv:2412.11170*, 2024. 2, 5, 6, 7
- [44] Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. In *Forty-second International Conference on Machine Learning*, 2025. 4
- [45] Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 2
- [46] Wenjie Zhuo, Fan Ma, Hehe Fan, and Yi Yang. Vividreamer: invariant score distillation for hyper-realistic text-to-3d generation. In *European Conference on Computer Vision*, pages 122–139. Springer, 2024. 2