Debiasing Scores and Prompts of 2D Diffusion for View-consistent Text-to-3D Generation Supplemental Material

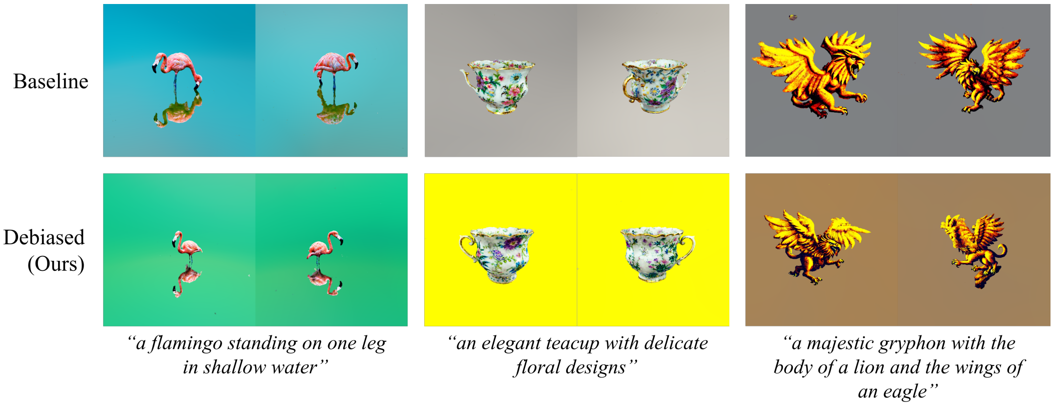

Baseline
Debiased (Ours)
"a flamingo standing on one leg in shallow water"
"an elegant teacup with delicate floral designs"
"a majestic gryphon with the body of a lion and the wings of an eagle"

Figure 1: **Results of the debiased ProlificDreamer (VSD) [11] framework.** We utilize the VSD implementation, introduced in ProlificDreamer, of threestudio [2]. In the baseline examples, we observe additional necks, handles, and faces. These artifacts are mitigated in our debiased examples.

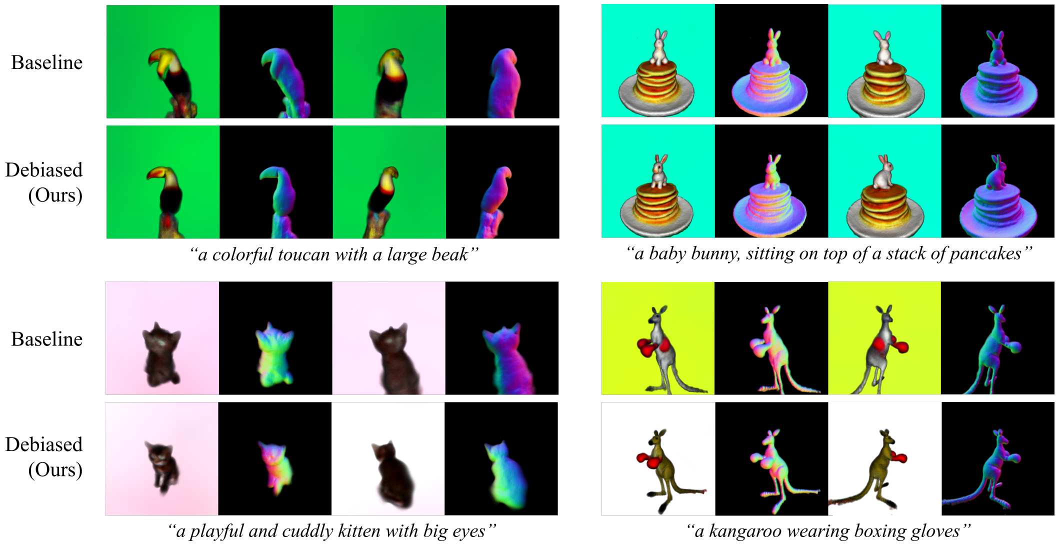

Baseline
Debiased (Ours)
"a colorful toucan with a large beak"
"a baby bunny, sitting on top of a stack of pancakes"
Baseline
Debiased (Ours)
"a playful and cuddly kitten with big eyes"
"a kangaroo wearing boxing gloves"

Figure 2: **Results of the debiased DreamFusion [6] framework.** We utilize the DreamFusion implementation of threestudio [2], which leverages DeepFloyd-IF.

A More Results

A.1 Qualitative results

We present additional qualitative results in Figs. 4. These results clearly show that our methods alleviate the Janus problem, also known as view inconsistency.

In certain instances, even though the Janus problem is present, the images from each angle still display reasonable appearances due to smooth transitions between angles. To illustrate this, we present a series of 10 sequential images arranged in order of the camera angles, right, back, and left of the object, in Fig. 4. In the baseline images, the front view appears in the back view or side

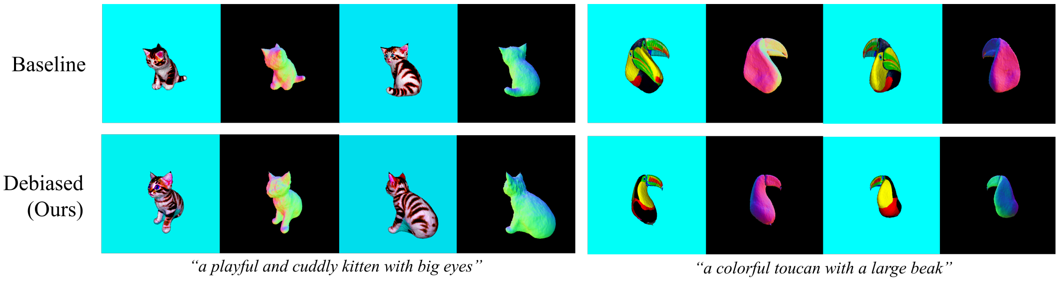

Baseline
Debiased (Ours)
"a playful and cuddly kitten with big eyes"
"a colorful toucan with a large beak"

Figure 3: **Results of the debiased Magic3D [3] framework.** We utilize the Magic3D implementation of threestudio [2].

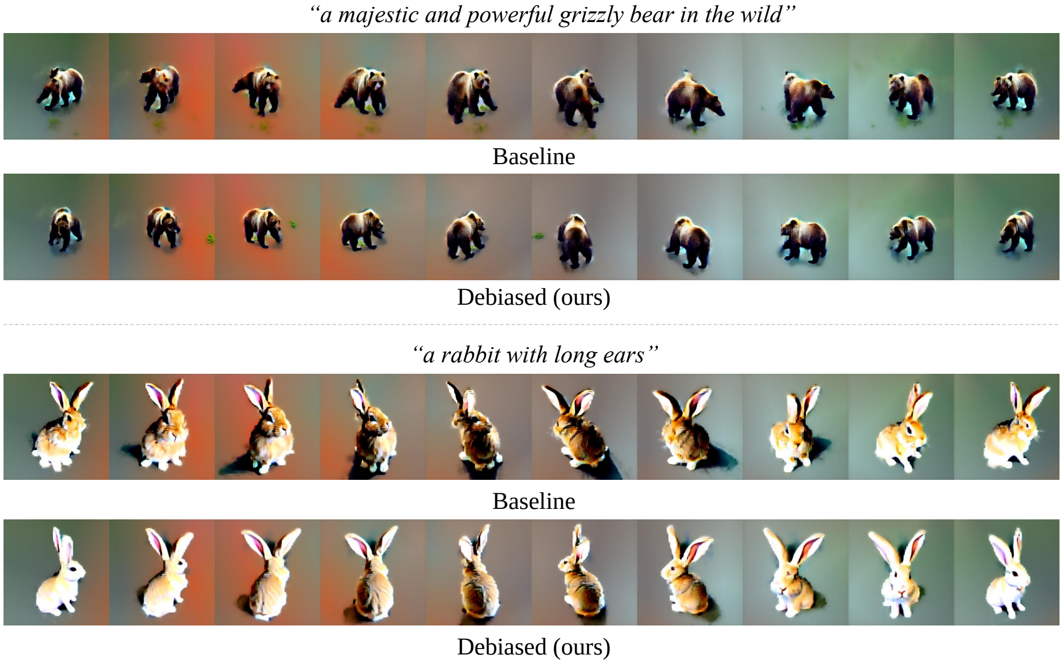

"a majestic and powerful grizzly bear in the wild"
Baseline
Debiased (ours)
"a rabbit with long ears"
Baseline
Debiased (ours)

Figure 4: **Comparison of our method with baseline (SJC [10]) in 360°.** In both cases, the baseline [10] exhibits the Janus problem, where the face appears in every view. Our debiased methods ensure proper view consistency in the 360° images.

view. However, after applying our debiasing methods, we observe a significant improvement in view consistency, resulting in more realistic representations.

Furthermore, we provide another example of ablation study on our methods in Fig. 5. This analysis clearly demonstrates that both prompt debiasing and score debiasing techniques significantly contribute to improved realism, reduction of artifacts, and achievement of view consistency.

A.2 Dynamic clipping of 2D-to-3D scores

We provide an additional example where we examine the outcomes of dynamic clipping in comparison to static clipping and the absence of clipping, as shown in Fig. 6. In the case of no clipping (row (a)), several artifacts appear in certain views. Using a high threshold for static clipping yields a similar outcome (row (b)). A low threshold successfully removes artifacts, but also makes necessary objects, like icebergs, appear transparent (row (c)). Gradually reducing the threshold from high to low preserves the main object while eliminating artifacts (row (d)). Overall, this demonstrates that dynamic clipping reduces artifacts and enhances realism.

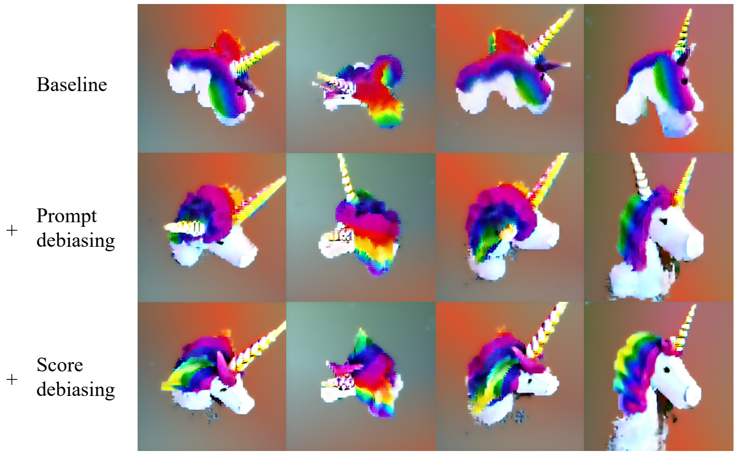
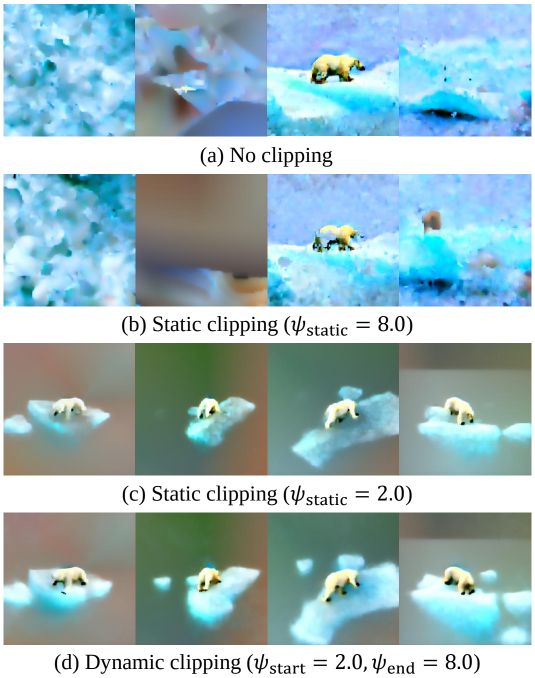


Figure 5: **Improvement of view consistency through prompt and score debiasing.** The baseline is original SJC [10], and *Prompt* and *Score* denote prompt and score debiasing, respectively. The given user prompt is *"an unicorn with a rainbow horn."*


(a) No clipping

(b) Static clipping ($\psi_{\text{static}} = 8.0$)

(c) Static clipping ($\psi_{\text{static}} = 2.0$)

(d) Dynamic clipping ($\psi_{\text{start}} = 2.0, \psi_{\text{end}} = 8.0$)

Figure 6: **Dynamic clipping of 2D-to-3D scores.** The given user prompt is *"a polar bear on an iceberg".*

A.3 User study

We conducted a user study to evaluate the view-consistency, faithfulness, and overall quality of the baseline and our debiased results. The results are presented in Table 2 in the main paper. According to the study, our method surpassed the baseline in all human evaluation criteria. We tested 75 participants anonymously, and the format of instructions provided to the users was as follows:

1. Which one has a more realistic 3D form? (above/below)
2. Which one is more consistent with the prompt? Prompt: {prompt} (above/below)
3. Which one has better overall quality? (above/below)

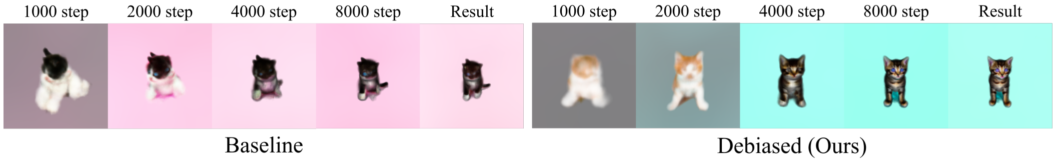


Figure 7: **Evaluation of rendered images during optimization.** Note that the geometry of the object is mostly formed within the first 4,000 optimization steps, which is when the problem in the geometry is clearly identified.

A.4 Results on other text-to-3D frameworks

Our method is designed to enhance 2D score-based text-to-3D generation methods. While it has been mainly claimed to be applicable to the SJC [10] and DreamFusion [6] frameworks, the applicability of our approach extends beyond these models. This approach can be adapted for any text-to-3D generation method that relies on a score generated by a text-to-image diffusion model and incorporates view-augmented prompting [6, 3, 11]. These methods, including contemporary works such as Magic3D [3] and ProlificDreamer [11], are to some extent susceptible to the Janus problem. With a recent implementation of the text-to-3D frameworks, threestudio [2], we have provided results that demonstrate the applicability of our method to recent frameworks such as Magic3D (Fig. 3), DreamFusion (Fig. 2), and ProlificDreamer (Fig. 1). We use the same seed for a fair comparison and only apply score debiasing for this experiment. Notably, even in instances with complex geometries that are susceptible to challenges like the Janus problem (e.g., "a majestic griffon with a lion's body and eagle's wings" or "an elegant teacup with delicate floral patterns"), the results show clear improvement when our method is applied.

A.5 Visualization of optimization process and convergence speed

We present Fig. 7 to demonstrate how the rendered image evolves at each step during the first stage of Magic3D [3]. This experiment underscores the motivation for dynamic clipping of 2D-to-3D scores, as the geometry is determined in the early stages.

In addition, the 3D scenes for both the baseline and ours evolve similarly in terms of optimization steps, with ours being debiased. It indeed shows that the impact of gradient clipping on convergence speed is quite marginal; the number of optimization steps is comparable to that of the baseline models, and the convergence speed is nearly unchanged by our approach (approximately 20 minutes for both SJC and ours).

B Limitations and Broader Impact

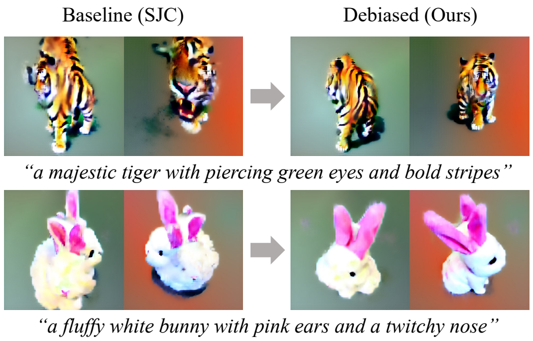
B.1 Limitations

Although our debiasing methods effectively tackle the Janus problem, the results produced by some prompts remain less than perfect. This is primarily due to the Stable Diffusion's limited comprehension of view-conditioned prompts. Despite the application of our debiasing methods, these inherent limitations result in constrained outputs for specific user prompts. Fig. 8 presents examples of such failure cases.

B.2 Broader impact

Our strategy pioneers the realm of debiasing. It possesses the capability to be integrated into any Score Distillation Sampling (SDS) technique currently under development, given that these methods universally deploy view prompts [6, 3, 10, 5, 4, 8].

Artificial Intelligence Generated Content (AIGC) has paved the way for numerous opportunities while simultaneously casting certain negative implications. However, it is important to note that our procedure is not identified as having any deleterious impact since it is exclusively designed for the purpose of debiasing the existing framework.



“a majestic tiger with piercing green eyes and bold stripes”

“a fluffy white bunny with pink ears and a twitchy nose”

Figure 8: **Failure cases.** In some prompts where Stable Diffusion has a severely limited ability to generate view-conditioned images, the view consistency of the result is constrained.

C Implementation Details

C.1 Common settings

We base our debiasing techniques on the publicly available repository of SJC [10]. To ensure consistency, we conduct 10,000 optimization steps for both SJC and our methods to enhance the 3D fields. The hyperparameters of SJC are set to fixed values and remain unchanged throughout our experiments. For future research, we present the prompts we used in our experiments in Table 1, where some prompts are taken from DreamFusion [6], Magic3D and SJC [10]. When comparing the use of these prompts, we intentionally omit Stable Dreamfusion [9, 6] because its occasional tendency to fail to generate an object and only generate backgrounds can significantly skew our evaluation metrics.

C.2 Score debiasing

In terms of score debiasing, we gradually increase the truncation threshold from one-fourth of the pre-defined threshold to the pre-defined threshold, according to the optimization step. Specifically, we linearly increase the threshold from 2.0 to 8.0 for all experiments using Stable Diffusion [7] that leverage dynamic clipping of 2D-to-3D scores. When using Deepfloyd-IF, we adopt a linearly increasing schedule from 0.5 to 2.0, considering the lower scale of the scores.

C.3 Prompt debiasing

To compute the pointwise mutual information (PMI), we use the uncased model of BERT [1] to obtain the conditional probability. Additionally, we set $P(u) = 1$ for words that should not be erroneously omitted. Otherwise, we set $P(u) = 1/2$. To use a general language model for the image-related task, we concatenated "This image is depicting a" when evaluating the PMI between the view prompt and user prompt. We first get u, v pairs such that $\frac{P(v,u)}{P(v)P(u)} < 1$. Then, given a view prompt, we remove words whose PMI for that view prompt, normalized across all view prompts, is below 0.95.

For the view prompt augmentation, we typically follow the view prompt assignment rule of DreamFusion [6] and SJC [10]. However, we slightly modify the view prompts and azimuth ranges for each prompt as mentioned in Sec. 4.2. For example, we assign an azimuth range of $[-22.5°, 22.5°]$ for the "front view." Also, we empirically find that using a view prompt augmentation $v \in \{“front\ view”, “back\ view”, “side\ view”, “top\ view”\}$ without $“of”$ depending on a viewpoint gives us improved results for Stable Diffusion v1.5 [7].

French fries
Mcdonald's fries
a 3D modeling of a lion
a baby bunny, sitting on top of a stack of pancakes
a blue bird with a hard beak
a cherry
a clam with pearl
a collie dog, high resolution
a crocodile
a dolphin swimming in the ocean
a dragon with a long tail
a flamingo standing on one leg in shallow water
a frog wearing a sweater
a kangaroo wearing boxing gloves
a monkey swinging through the trees
a mug with a big handle
a penguin
a piece of strawberry cake
a polar bear
a polar bear on an iceberg
a rabbit with long ears
a beautiful mermaid with shimmering scales and flowing hair
a brightly colored tree frog
a colorful parrot with a curved beak and vibrant feathers
a colorful toucan with a large beak
a colorful and exotic parrot with a curved beak
a colorful and vibrant chameleon
a colorful and vibrant poison dart frog perched on a leaf
a confident businesswoman in a sharp suit with briefcase in hand
a cozy coffee mug with steam rising from the top
a creepy spider with long, spindly legs
a curious meerkat standing on its hind legs
a curious and mischievous raccoon with a striped tail
a cute and chubby panda munching on bamboo
a cute and cuddly sugar glider jumping from tree to tree
a delicious sushi on a plate
a fearsome dragon with iridescent scales
a fluffy white bunny with pink ears and a twitchy nose
a graceful ballerina mid-dance, with flowing tutu and pointed toes
a majestic eagle
a majestic elephant with big ears and a long trunk
a majestic giraffe with a long neck
a majestic gryphon with the body of a lion and the wings of an eagle
a majestic lioness
a majestic tiger with piercing green eyes and bold stripes
a majestic and powerful grizzly bear in the wild
a photo of a comfortable bed
a pirate collie dog, high resolution
a playful baby elephant spraying water with its trunk
a playful dolphin leaping out of the water
a playful and cuddly kitten with big eyes
a quirky and mischievous raccoon with a striped tail
a refreshing and tangy grapefruit with juicy segments
a regal queen in a crown and elegant gown
a relaxed yoga practitioner in a serene pose
a ripe strawberry
a sleek and graceful shark swimming in the deep sea
a sleek and speedy cheetah in mid-stride
a sleek and speedy cheetah racing across the savannah
a sleek and speedy falcon soaring through the sky
a sleek and stealthy panther
a small kitten
a smiling cat
a sneaky fox
a toy helicopter
a toy sports car
an octopus in the ocean
an unicorn with a rainbow horn
an bitten apple
an elegant teacup with delicate floral designs

Table 1: **Example prompts.**

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Yuan-Chen Guo, Ying-Tian Liu, Chen Wang, Zi-Xin Zou, Guan Luo, Chia-Hao Chen, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023.

[3] 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. *arXiv preprint arXiv:2211.10440*, 2022.

[4] 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. *arXiv preprint arXiv:2303.11328*, 2023.

[5] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022.

[6] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[8] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023.

[9] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion. https://github.com/ashawkey/stable-dreamfusion, 2022.

[10] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022.

[11] 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. *arXiv preprint arXiv:2305.16213*, 2023.