
SWEETDREAMER: ALIGNING GEOMETRIC PRIORS IN 2D DIFFUSION FOR CONSISTENT TEXT-TO-3D

Weiyu Li¹, Rui Chen^{2,3*}, Xuelin Chen³, Ping Tan¹

¹ Hong Kong University of Science and Technology

² South China University of Technology

³ Tencent AI Lab

{weiyuli.cn, xuelin.chen.3d}@gmail.com, riorui@foxmail.com, pingtan@ust.hk

ABSTRACT

It is inherently ambiguous to lift 2D results from pre-trained diffusion models to a 3D world for text-to-3D generation. 2D diffusion models solely learn view-agnostic priors and thus lack 3D knowledge during the lifting, leading to the multi-view inconsistency problem. We find that this problem primarily stems from geometric inconsistency, and avoiding misplaced geometric structures substantially mitigates the problem in the final outputs. Therefore, we improve the consistency by aligning the 2D geometric priors in diffusion models with well-defined 3D shapes during the lifting, addressing the vast majority of the problem. This is achieved by fine-tuning the 2D diffusion model to be viewpoint-aware and to produce view-specific coordinate maps of canonically oriented 3D objects. In our process, only coarse 3D information is used for aligning. This “coarse” alignment not only resolves the multi-view inconsistency in geometries but also retains the ability in 2D diffusion models to generate detailed and diversified high-quality objects unseen in the 3D datasets. Furthermore, our aligned geometric priors (AGP) are generic and can be seamlessly integrated into various state-of-the-art pipelines, obtaining high generalizability in terms of unseen shapes and visual appearance while greatly alleviating the multi-view inconsistency problem. Our method represents a new state-of-the-art performance with a 85+% consistency rate by human evaluation, while many previous methods are around 30%. Our project page is <https://sweetdreamer3d.github.io/>

1 INTRODUCTION

Generative models have achieved diverse and high-quality image generation, in a highly controllable way with input text prompts (Nichol et al., 2022; Ramesh et al., 2021; Saharia et al., 2022b; Rombach et al., 2022). This remarkable achievement has been attained by training scalable generative models, particularly diffusion models, on an extensive corpus of paired text-image data. To replicate such success in 3D, a substantial endeavor is obviously necessary to gather a vast amount of high-quality text-3D pairs, which is currently receiving commendable attention (Deitke et al., 2023; Wu et al., 2023; Shrestha et al., 2022). However, it is evident that the effort required to collect a comprehensive 3D dataset covering highly varied subjects is considerably more significant, given the high cost associated with acquiring high-quality 3D content.

On the other end, attempts to achieve text-controlled 3D generative models have taken several routes, among which the 2D-lifting technique has emerged as a particularly promising direction and is increasingly gaining momentum in the field (Poole et al., 2022). This technique lifts 2D results into a 3D world and features an optimization framework, wherein a 3D representation is updated in differentiable parameterizations with the Score Distillation Sampling (SDS) loss derived from a pre-trained 2D diffusion model. By combining SDS with various suitable 3D representations (Lin et al., 2023; Chen et al., 2023; Wang et al., 2023b; Shi et al., 2023), this technique can generate high-fidelity 3D objects and scenes for a diverse set of user-provided text prompts.

*Work done during internship at Tencent AI Lab.



Figure 1: Our work can generate high-fidelity and highly diversified 3D results from various text prompts, free from the notorious multi-view inconsistency problem. We highly recommend referring to the supplementary materials for a more immersive viewing experience of the 3D results.

Yet lifting 2D observations into 3D is inherently ambiguous. 2D diffusion models solely learn 2D priors from individual images and therefore lack 3D knowledge for disambiguation during the lifting, leading to notorious multi-view inconsistency problems, e.g., the multi-face Janus problem. While learning robust 3D priors from extensive and comprehensive 3D datasets is seemingly the very answer, in reality, we are only presented with 3D data that is rather scarce compared to plentifully available images. Hence, a currently compelling direction is to incorporate 3D priors learned from relatively limited 3D data into 2D diffusion priors that possess high generalizability, thus obtaining *the best of both worlds*.

In particular, the issues related to multi-view inconsistency can be primarily categorized into two types: i) geometric inconsistency issues, that are caused by the ambiguity in the spatial arrangement of geometric structures, i.e., a geometric structure can position and orient differently in 3D. Importantly, geometry inconsistency is further exacerbated during the lifting by the supervision imposed by 2D priors that *lack 3D awareness*, where many irrational 3D structures resulting in identical 2D projections can deceive 2D priors; ii) appearance inconsistency issues, that arise due to the ambiguity in the mapping from geometric structures to corresponding appearance, and again, is exacerbated by the lack of 3D awareness in 2D diffusion for disambiguation. Empirically, we found the geometry inconsistency issue is the primary cause contributing to most multi-view inconsistent results within various existing methods, whereas the appearance inconsistency issue manifests itself alone in only extreme cases and thereby holds lesser significance. This is evidenced by the fact that the majority of 3D inconsistent results exhibit repetitive geometric structures, typically multiple hands, or faces, that are generated under the guidance of 2D diffusion. It is worth noting that addressing these misplaced structures plays a significant role in mitigating 3D inconsistency in the final outcomes, as the inclusion of geometric hints from 3D consistent geometries greatly aids the appearance modeling. This holds true for both one-stage text-to-3D pipelines (Poole et al., 2022), where geometry and appearance are updated simultaneously, as well as pipelines that model geometry and appearance separately (Chen et al., 2023; Richardson et al., 2023). However, it should also be acknowledged that there may still be exceptional circumstances where appearance inconsistency can manifest with 3D consistent geometric structures.

These findings have motivated us to prioritize addressing geometric inconsistency issues in text-to-3D, by equipping the 2D priors with the capability to produce 3D consistent geometric structures while retaining their generalizability. In a way analogous to (Leike & Sutskever, 2023), we enforce the 2D geometric priors¹ in diffusion models act in a way that *aligns with well-defined 3D geometries* as depicted in 3D datasets during the lifting, addressing the vast majority of the inconsistency problem from the origin. We refer to the resulting geometric priors as “AGP”, for Aligned Geometric Priors. Specifically, we align the geometric priors by fine-tuning the 2D diffusion model to produce coordinate maps of objects in canonical space, thereby *disambiguating the geometry distribution* in 3D for ease of learning, and further conditioning it on additional camera specifications, thereby conferring *3D awareness* eventually. Notably, in stark contrast to methods that hinge heavily upon the geometric and appearance information in 3D datasets, we only capitalize on *coarse geometries*, avoiding over-reliance on geometric and visual appearance details that may further introduce undesired inductive bias. This “coarse” alignment of geometric priors not only enables the generation of 3D objects without the multi-view inconsistency problem but also retains the ability in 2D diffusion models to generate vivid and diversified objects unseen in 3D datasets.

Finally, yet importantly, our AGP possesses high compatibility that is generally lacking in competing methods. We show that AGP is highly generic and can be seamlessly integrated into various state-of-the-art pipelines using different 3D representations, obtaining high generalizability in terms of unseen geometries and appearances while significantly alleviating multi-view inconsistency. It represents a new state-of-the-art performance with a 85+% consistency rate in human evaluation (see qualitative results gallery in Figure 1 and quantitative results in Table 1).

2 RELATED WORK

In the following, we mainly review related literature that exploit 2D priors learned in text-conditioned generative image models for text-to-3D, and refer readers to Zhang et al. (2023) for a more in-depth survey of text-to-image diffusion models.

¹Please refer to Section A.1 for ours findings about geometric priors in 2D diffusion.

Text-to-3D using 2D Diffusion. Following successful text-to-image diffusion models, there has been a surge of studies that lift 2D observations in diffusion models to perform text-to-3D synthesis, bypassing the need for large-scale text-3D datasets for training scalable 3D diffusion models. In particular, the pioneer work by Poole et al. (2022) introduces a key technique – Score Distillation Sampling (SDS), where diffusion priors are used as score functions to supervise the optimization of a 3D representation. Concurrent with Poole et al. (2022), a similar technique is proposed in Wang et al. (2023a), which applies the chain rule on the learned gradients of a diffusion model and back-propagate the score of a diffusion model through the Jacobian of a differentiable renderer to optimize a 3D world. An explosion of text-to-3D techniques occurred in the community since then that improves the text-to-3D in various aspects, such as improved sampling schedules (Huang et al., 2023), adopting various 3D representations (Lin et al., 2023; Tsalicoglou et al., 2023b; Chen et al., 2023), new score distillation (Wang et al., 2023b), etc. Although these methods have shown the capability to circumvent the necessity for text-3D data and generate photo-realistic 3D samples of arbitrary subjects with user-provided textual prompts, more often than not, they are prone to the notorious 3D inconsistency issues. Hence, prior works have attempted to address the inconsistency with improved score functions and/or text prompts (Armandpour et al., 2023; Hong et al., 2023). Nonetheless, these methods cannot guarantee 3D consistency and tend to fail on consistent text-to-3D synthesis.

Concurrently, MVDream (Shi et al., 2023) addresses the multi-view inconsistency problem via training a dedicated multi-view diffusion model, which simultaneously generates multi-view images that are consistent across a set of sparse views. Particularly, they jointly fine-tune the diffusion model on real images and synthetic multi-view images, in order to inherit the generalizability in 2D diffusion and to obtain the multi-view consistency in 3D datasets. Instead of relying on computationally intensive renderings and fine-tuning on both synthetic and real images, our method uses only low-resolution and low-cost geometry maps, and hence the “coarse” alignment of geometric priors is computationally efficient. Besides, our AGP is generic and can be seamlessly integrated into existing pipelines to confer 3D consistency. Such compatibility is obviously lacking in MVDream.

Generative Novel Views with Diffusion Models Another route of 3D generation is to model it as a view-conditioned image-to-image translation task with diffusion models, and directly generate novel views of the 3D, detouring the need for optimizing a 3D world. Watson et al. (2022) train a pose-conditional image-to-image diffusion model to take a source view and its pose as inputs and generate a novel view for a target pose. This method has only been demonstrated on synthetic data in the ShapeNet dataset (Chang et al., 2015). Similarly, Zhou & Tulsiani (2023) build a view-conditioned novel view diffusion model in the latent space and demonstrate its utility in sparse view 3D reconstruction. Chan et al. (2023) improve the view consistency of the diffusion model by reprojecting the latent features of the input view prior to diffusion denoising. More recently, Liu et al. (2023) propose a variant of this technique for performing text-to-3D with the generative ability of novel views enabled by fine-tuning language-guided priors on renderings from 3D datasets.

Generally, models trained in these methods are unable to accurately capture the view specifications, resulting in the generation of multiple views that are only *approximately* 3D consistent. Although our AGP also takes as input camera specifications to be viewpoint-aware, its purpose is merely to generate coarse geometries that will evolve subsequently into a 3D consistent object. Furthermore, aside from the differences in the formulation of these methods and ours on the text-to-3D task, our approach diverges significantly in that it does not capitalize on the appearance information, i.e., synthetic renderings, residing in 3D datasets, which is at the risk of compromising the visual priors learned in pre-trained diffusion models and thus may result in degraded visual quality.

3 METHOD

As aforementioned, the issues of multi-view or 3D inconsistency can be categorized from two perspectives: geometric inconsistency issues, which pertain to misplaced geometric structures in 3D, and appearance inconsistency issues, which relate to incorrect visual appearance modelling on the 3D geometric structures. Realizing that geometric inconsistency is the main reason for most 3D inconsistent results, our goal is to equip the 2D priors with the capability to produce 3D consistent geometric structures while retaining their generalizability. As a result, the generated consistent geometric structures play a continuous role in contributing to the modeling of intricate geometric details and visual appearance in pipelines that perform text-to-3D.

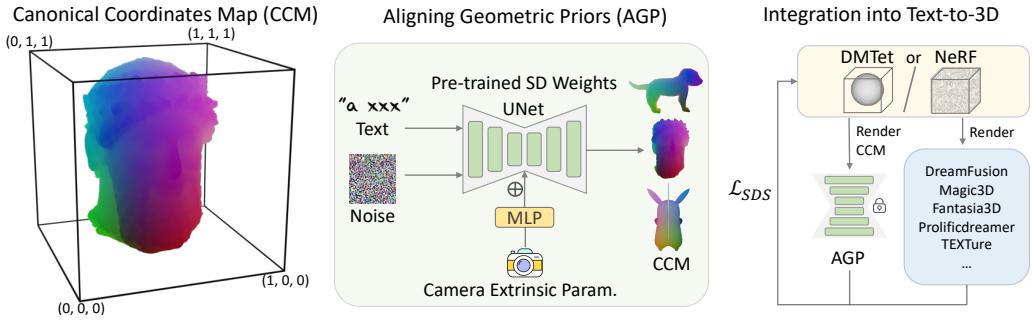


Figure 2: Method overview. We fine-tune the 2D diffusion model (middle) to generate viewpoint-conditioned canonical coordinates maps, which are rendered from canonically oriented 3D assets (left), thereby aligning the geometric priors in the 2D diffusion. The aligned geometric priors can then be seamlessly integrated into existing text-to-3D pipelines to confer 3D consistency (right), while retaining their generalizability to obtain high-fidelity and highly varied 3D content.

To this end, we propose to ensure the geometric priors in 2D diffusion act in a way that aligns with well-defined 3D geometries as depicted in 3D dataset (Section 3.1). Specifically, we assume to have access to a 3D dataset, which comprises extensive and diverse 3D models that are canonically oriented and normalized. We then render depth maps from random views, and convert them into canonical coordinates maps. Note we only render from rather coarse geometries, as the goal is merely to use 3D data for aligning rather than for generating geometric details. The benefits of using such 3D data are two-fold: i) all geometries are well-defined in 3D, so there is no ambiguity in their spatial arrangement; ii) by further injecting the viewpoint into the model, we can confer viewpoint awareness and eventually 3D awareness. Then, we fine-tune the 2D diffusion model to generate the canonical coordinates map under a specified view, eventually aligning the geometric priors in 2D diffusion. Finally, the aligned geometric priors can be seamlessly integrated into various text-to-3D pipelines (Section 3.2), significantly mitigating the inconsistency issues, resulting in the generation of high-quality and diverse 3D content. Figure 2 presents an overview.

3.1 ALIGNNING GEOMETRIC PRIORS IN 2D DIFFUSION

Then, we elaborate technical details involved in aligning the geometric priors in 2D diffusion with well-defined 3D geometries during the lifting, while retaining its generalizability.

Canonical Coordinates Map (CCM) To eliminate the distraction caused by the gauge freedom and thereby ease the modeling, we assume that all objects within the same category adhere to a canonical orientation in the training data, a common practice in various publicly accessible datasets (Deitke et al., 2023; Chang et al., 2015). Note that, while the object orientation is assumed to be canonicalized per category, our objective is not to learn category-specific data priors. Instead, our aim is to extract general knowledge from a variety of objects in the 3D datasets, which will aid in aligning the 2D geometric priors. Analogous to (Wang et al., 2019; Shotton et al., 2013), the canonical object space is defined as a 3D space contained within a unit cube $\{x, y, z\} \in [0, 1]$. Specifically, given an object, we normalize its size by uniformly scaling the object such that the max extent of its tight bounding box has a length of 1 and is centered at the origin. While we can render coordinates maps at random views from these canonically oriented and uniformly normalized objects for training, we further propose to anisotropically scale the three components in the coordinate maps rendered from an object, such that the value of each component is within the range from 0 to 1. This anisotropic normalization amplifies the discrepancy of spatial coordinates on thin structures at different views, easing the perception of the 3D structures and thereby improving the 3D-awareness in the subsequent learning.

Camera Condition Although the canonical coordinates maps contain rough viewpoint information, we found that the diffusion model has difficulties in exploiting it. Therefore, we inject the camera information into the model to improve viewpoint-awareness, following MVDream (Shi et al.,

2023). Specifically, we explicitly input the corresponding camera extrinsic parameters to the diffusion model, which is passed through an MLP before being fed to the middle layer of the diffusion model. Note that, in contrast to other models that rely on accurate viewpoint-awareness for generating consistent 3D, the use of camera conditions in our model is only to roughly generate coarse geometries that will evolve subsequently into a 3D consistent object.

Fine-tuning 2D Diffusion for Alignment Given the pairs of the canonical coordinates map and its corresponding camera specification, we keep the architecture of the 2D diffusion model while slightly adapting it to be conditioned on camera extrinsic parameters. This enables us to leverage the pre-trained 2D diffusion model for transfer learning, thereby inheriting their generalizability in terms of highly varied subjects unseen in the 3D dataset. Finally, we fine-tune the diffusion model, originally intended for generating raw RGB or latent images, to generate the canonical coordinates map under a viewpoint condition, eventually aligning the geometric priors in 2D diffusion.

Implementation Details By default, we conduct experiments based on the Stable Diffusion model (we use v2.1), which is a commonly used public large pre-trained text-to-image diffusion model.

3D dataset. We use a public 3D dataset – Objaverse (Deitke et al., 2023), which contains around 800k models created by artists, to generate the data for fine-tuning. Notably, while there is no explicit specification for the coordinate system, many artists still adhere to a convention regarding the orientation when creating 3D assets. Hence, by having a significant portion of the 3D objects in canonical orientation and only a few remaining misoriented, we are able to achieve satisfactory outcomes without the necessity of manually correcting those misoriented ones. On the other hand, due to the presence of considerable noise in the textual annotations, we employ a 3D captioning model (Luo et al., 2023) to augment the textual description of each 3D asset and randomly switch between the augmented caption and its original textual annotation (typically names and tags) during the training. In addition, to ensure relevance, we apply a filtering process based on tags to eliminate 3D assets such as point clouds and low poly models, resulting in approximately 270k objects.

Camera sampling. We render canonical coordinates maps from the 3D objects. The camera is randomly positioned at a distance ranging from 0.9 to 1.1 units, with a field of view set at 45 degrees. Additionally, the camera’s elevation is randomly varied between -10° and 45°. As we do not rely on visual appearance information, we are able to utilize a fast rasterization renderer for generating the training data, avoiding the computational intensity associated with ray tracing renderers.

Training. We fine-tune our model on the latent space of Stable Diffusion using Diffusers (von Platen et al., 2022). Note that the canonical coordinates map is directly treated as a latent image for the latent diffusion model to produce. This leads to an attractive feature that our aligned geometric priors can be trained fast, without involving the encoding and decoding process of the VAE. We keep the default optimizer setting, as well as the ϵ -prediction. Since we input the camera extrinsic parameters as a condition to the diffusion model, the training objective is now formulated as follows:

$$\mathcal{L}_{LDM} := \mathbb{E}_{c,y,z,t,\epsilon \in \sim N(0,1)} \left[\|\epsilon - \epsilon_\theta(c, \tau_\theta(y), z_t, t)\|_2^2 \right], \text{ where } c \text{ is the camera extrinsic parameters, } y \text{ the input text prompt and } \tau_\theta(y) \text{ its embedded feature using tokenizer, and } z_t \text{ the noisy latent image generated by adding noise } \epsilon \text{ to a clean latent image } z \text{ at a diffusion timestep } t.$$

3.2 INTEGRATION INTO TEXT-TO-3D

Finally, we elaborate on how to integrate our aligned geometric priors into existing pipelines using different 3D representations, significantly mitigating their inconsistency issues and achieving state-of-the-art text-to-3D performance. To showcase such compatibility, we provide demonstrations of two state-of-the-art text-to-3D methods that utilize different 3D representations, namely, Fantasia3D (Chen et al., 2023), which explicitly disentangles the geometry and appearance modeling and uses a hybrid representation – DMTet (Shen et al., 2021) – for the underlying 3D geometry, and DreamFusion (Poole et al., 2022), which employs the neural radiance field (NeRF) (Mildenhall et al., 2020) as the 3D representation. Please see Figure 3 for the system pipelines of integrating our aligned geometric priors into these two methods. For more technical details, we refer readers to the original papers, as we do not elaborate on them further in this context.

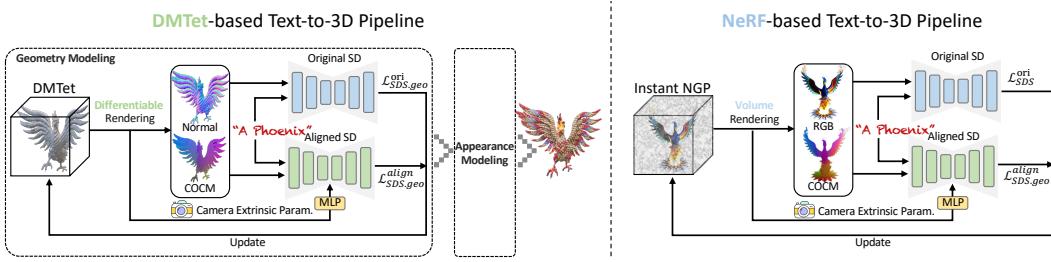


Figure 3: Seamless integration of our AGP in various text-to-3D pipelines.

DMTet-based Pipeline For the sake of clarity, we refer to the variant obtained by integrating our aligned geometric priors into Fantasia3D as the DMTet-based pipeline. All that is required is an additional parallel branch, to incorporate our aligned geometric priors for supervising the geometry modeling in the original pipeline. With this seamless integration of our aligned geometric prior, high-quality and view-consistent results can be easily achieved, without the need for carefully designed initialization shapes as in the original pipeline.

Optimization. Specifically, we add this additional supervision imposed by our aligned geometric priors in both the coarse and fine geometry modeling stages. Simply, our aligned diffusion model takes as input the canonical coordinates map and produces the SDS loss to update the 3D representation. Then, the final loss function in the geometry modeling can be written as $\mathcal{L}_{SDS,geo} = \lambda^{ori}\mathcal{L}_{SDS,geo}^{ori} + \lambda^{align}\mathcal{L}_{SDS,geo}^{align}$, where the first term is the geometry SDS loss derived from the original diffusion model, the latter is the SDS loss derived from our aligned geometric priors. Here, λ^{ori} and λ^{align} are the weights to balance their effects. The revised system pipeline is shown on the left of Figure 3. Note that this integration is implemented only in the coarse and fine geometry stages, while the appearance modeling stage is untouched.

NeRF-based Pipeline NeRF is another common choice for the 3D representation in text-to-3D, as it is more friendly for optimization compared to traditional discrete meshes, and can also be combined with volume rendering for great photo-realism. Specifically, we base on a popular implementation (Guo et al., 2023) of the pioneer – DreamFusion (Poole et al., 2022), which uses NeRF as the 3D representation, and refer to it as the NeRF-based pipeline. Particularly, the 3D scene is represented by Instant-NGP with an extra MLP for modeling the environment map, allowing the modeling of rich details with low computing cost. Then we can volume-render the 3D object/scene to obtain the RGB images and feed them into the Stable Diffusion to calculate SDS loss.

Optimization. During the lifting optimization, we render the canonical coordinates map and feed it to our aligned geometric priors to calculate the geometry SDS loss $\mathcal{L}_{SDS,geo}^{align}$ to help update the geometry branch of the NeRF, in addition to the origin SDS loss \mathcal{L}_{SDS} calculated with the RGB image. Similar to the previous integration, the final loss is the weighted combination of the original SDS loss and our aligned geometric SDS loss: $\mathcal{L}_{SDS} = \lambda^{ori}\mathcal{L}_{SDS}^{ori} + \lambda^{align}\mathcal{L}_{SDS,geo}^{align}$, where λ^{ori} and λ^{align} are the weights balancing these two terms. Note our AGP continues to model 3D consistent coarse geometries in this pipeline, while again leaving the appearance modeling untouched.

4 TEXT-TO-3D GENERATION

We present the qualitative and quantitative evaluation of the text-to-3D pipelines as described in Section 3.2, as well as comparison results against other text-to-3D baseline methods. For convenience and clarity, we refer to the DMTet-based pipeline and NeRF-based pipeline as *Ours (DMTet-based)* and *Ours (NeRF-based)*, respectively. Furthermore, depending on the different pre-trained diffusion models used in its *original* pipeline, we developed two versions of *Ours (NeRF-based)*, namely, *Ours (NeRF-based IF)* using Deepfloyd IF, and *Ours (NeRF-based full)* using DeeFloyd IF first and then Stable Diffusion. Please refer to the supplementary for more details.

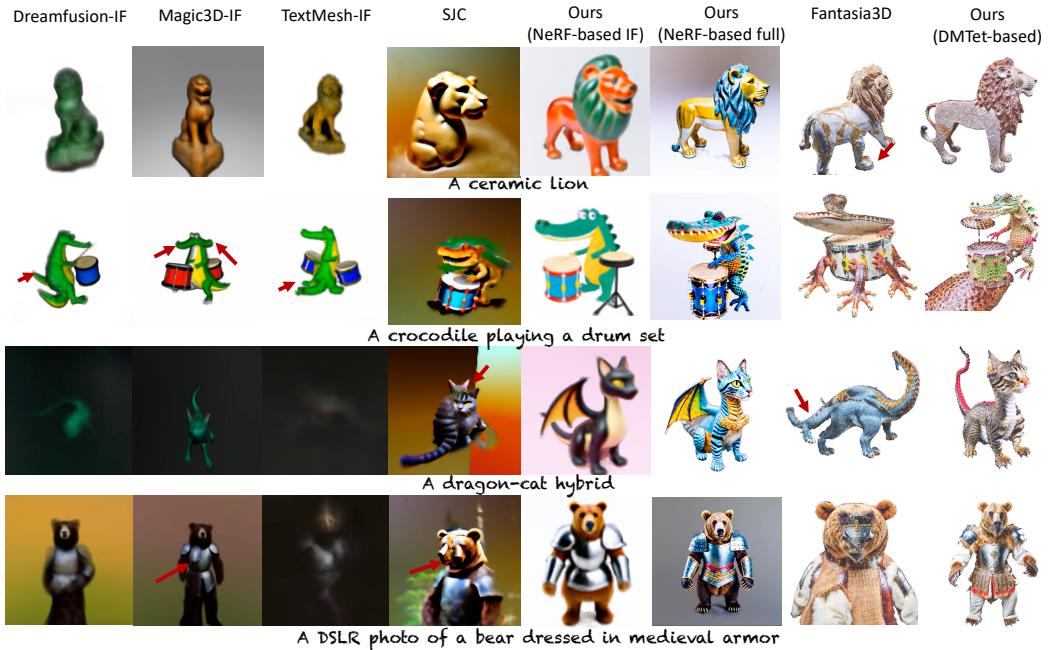


Figure 4: Visual comparisons. Compared to other competing methods, our text-to-3D pipelines can generate high-fidelity 3D content with high 3D consistency. See 3D inconsistency issues in the baseline results (highlighted by red arrows).

Baselines We extensively compare to baselines as follows: i) *Fantasia3D*, based on which our DMTet-based pipeline is implemented. We compare our DMTet-based pipeline against it to show specifically the effectiveness of our AGP; ii) *DreamFusion-IF*, which replaces the unreleased Imagen (Saharia et al., 2022a) with – DeepFloyd IF (IF, 2023). We compare our NeRF-based against it to validate our AGP again; And more other baselines implemented in (Guo et al., 2023), including iii) *SJC* (Wang et al., 2023a), that applies the chain rule on the learned gradients of a diffusion model and backpropagates the score of a diffusion model through the Jacobian of a differentiable renderer to optimize a 3D world; iv) *Magic3D-IF* (Lin et al., 2023) is a hybrid pipeline that uses NeRF at the coarse stage and then converts it into a mesh for fine details. We also adapt it to use DeepFloyd IF for SDS calculation; v) *TextMesh-IF* (Tsalicoglou et al., 2023a), which is also a DeepFloyd IF-based implementation, is similar to Magic3D but uses an SDF-based representation at the coarse stage. vi) *MVDream* (Shi et al., 2023), which is a concurrent work to us. Note that, since their official implementation is unavailable by the time of our submission, we use the same prompts as listed on their website for side-by-side comparisons.

Due to various reasons, we were not able to obtain the original implementation of most baselines. Therefore, except for Fantasia3D and MVDream, we use the implementation from Guo et al. (2023) for all baselines. We consider these implementations to be the most reliable and comprehensive open-source option available in the field. By default, we use the Stable Diffusion model as the prior, except those with the name suffixed “IF” use DeepFloyd IF within the pipeline.

Quantitative Evaluation It is important to acknowledge that currently, there is a lack of well-established metrics that can quantitatively and comprehensively evaluate the text-to-3D results from various perspectives. In this work, our primary focus lies in generating multi-view consistent 3D content, rather than placing specific emphasis on enhancing the appearance or texture quality of existing pipelines. So we focus on quantitatively evaluating the multi-view consistency of the 3D results. Specifically, we randomly select 80 text prompts from the DreamFusion gallery (dre, 2023), and perform text-to-3D synthesis to generate 80 results using each method. We then manually check and count the number of occurrences of 3D inconsistencies (e.g., multiple heads, hands, or legs) and report the success rate, i.e., the number of 3D consistent objects divided by the total number of

generated results. As shown in Table 1, our method outperforms other methods by a large margin. Our success rates are over 85% in both pipelines, while the previous methods are only around 30%.

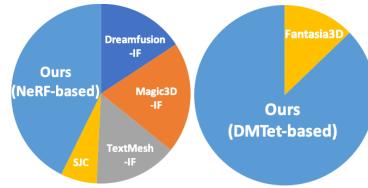
	Dreamfusion -IF	Magic3D -IF	TextMesh -IF	SJC	Fantasia3D	Ours (DMTet-based)	Ours (NeRF-based IF)
Cons. Rate ↑	30.0%	35.0%	23.8%	7.5%	32.5%	87.5%	88.8%

Table 1: Quantitative comparison results for the 3D consistency rate.

Qualitative Evaluation As shown in Figure 4, By integrating our AGP into Fantasia3D, i.e., *Ours (DMTet-based)*, the results have been significantly improved. The original Fantasia3D only produced coarse and inaccurate results without hand-crafted initial geometries. We believe this is due to the domain gap between the rendered normal map and the geometric information extracted in the latent space of the Stable Diffusion, resulting in optimization difficulties in converging to a reasonable 3D shape. As for Ours (NeRF-based), the generated results clearly have high 3D consistency and possess a more realistic appearance. This is because our aligned geometric priors only contribute to the geometry modeling during the lifting, while they do not compromise the appearance modeling guided by the powerful visual priors learned by Stable Diffusion from billions of real images. In general, the 3D results generated by most remaining baselines, even when equipped with the more powerful Deepfloyd IF, suffer from multi-view inconsistency (easier to see from spinning views in the supplementary file). Note that, when evaluating the results, we focus on assessing the 3D consistency, and hence do not heavily penalize blurry images, as they can be caused by the use of DeepFloyd IF with limited computing resources.

Last, although the concurrent work, MVDream, can also resolve the multi-view inconsistency problem, we observe that it is prone to overfit the limited 3D data, consequently resulting in a compromise of the generalizability in the original powerful 2D diffusion model. Specifically, as shown in the supplementary (A.4), MVDream misses the “backpack” in its generated result presented with the prompt “an image of a pig carrying a backpack”. Additionally, since they use synthetic multi-view renderings for fine-tuning their multi-view diffusion model, the appearance of the generated results lacks the desired level of photo-realism.

User Study We also conducted a user study on 30 generated 3D results of relevant methods. The study engaged 36 participants to assess 3D results in 30 rounds. In each round, each participant was presented with videos rendered from the 3D models obtained by various methods based on one of the 30 text prompts. Then, they were requested to choose a 3D model they favored the most, considering only the 3D consistency. We report the rate of preference for each method in the inset pie chart. As shown, our method outperforms the competing methods by a large margin, showing the robustness of our method in generating results of high 3D consistency.



5 CONCLUSION

We introduced Aligned Geometric Priors (AGP), which is obtained by fine-tuning a pre-trained 2D diffusion model to generate viewpoint-conditioned coarse geometric maps of canonically oriented objects, thereby conferring 3D awareness. AGP is generic and can be seamlessly integrated into various existing pipelines to generate 3D objects of high consistency. Most importantly, AGP improves the geometry modeling and does not compromise the appearance modeling guided by strong priors learned from billions of real images.

Discussion While AGP has shown state-of-the-art performance in text-to-3D, we also note a few limitations. Our work does not directly consider appearance modeling, where inconsistency may still arise *rarely* due to the remaining ambiguity in the mapping from the geometric structure to its associated appearance. Early in our development, we attempted to incorporate an appearance generator by fine-tuning the 2D diffusion model to generate the appearance image conditioned on a given canonical coordinates map. Unfortunately, this approach resulted in overfitting to the renderings derived from 3D data, leading to 3D results that lacked the desired level of photorealism. We leave

the study in this direction for future work. All that being said, we believe the work opens up a novel “less is more” direction of utilizing relatively limited 3D data to enhance 2D diffusion priors for text-to-3D synthesis.

REFERENCES

- Dreamfusion project webpage, 2023. URL <https://dreamfusion3d.github.io/>.
- Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023.
- Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Genvs: Generative novel view synthesis with 3d-aware diffusion models, 2023.
- Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.
- Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *International Conference on Computer Vision (ICCV)*, October 2023.
- 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13142–13153, 2023.
- Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. <https://github.com/threestudio-project/threestudio>, 2023.
- Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. *arXiv preprint arXiv:2303.15413*, 2023.
- Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023.
- DeepFloyd IF. Deepfloyd if, 2023. URL <https://huggingface.co/DeepFloyd>.
- Jan Leike and Ilya Sutskever. Introducing superalignment, 2023. URL <https://openai.com/blog/introducing-superalignment>.
- 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- 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.
- Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pre-trained models. *arXiv preprint arXiv:2306.07279*, 2023.
- B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (SIGGRAPH)*, 41(4):102:1–102:15, July 2022. doi: 10.1145/3528223.3530127. URL <https://doi.org/10.1145/3528223.3530127>.

-
- Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, pp. 16784–16804. PMLR, 2022.
- Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2022.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pp. 8821–8831. PMLR, 2021.
- Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022a.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022b.
- Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- 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.
- Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgbd images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2937, 2013.
- Rakesh Shrestha, Siqi Hu, Minghao Gou, Ziyuan Liu, and Ping Tan. A real-world dataset for multi-view 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, pp. 56–73, 2022.
- Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. 2023a.
- Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439*, 2023b.
- Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022.
- Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12619–12629, 2023a.
- He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651, 2019.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolific-dreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023b.

Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 803–814, 2023.

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12588–12597, 2023.

A APPENDIX

A.1 GEOMETRIC PRIORS IN 2D DIFFUSION

The development our method is started by the findings about the geometric priors learned in a publicly available 2D diffusion model – Stable Diffusion (Rombach et al., 2022), which is a latent diffusion model that generates images from text. Instead of operating in the high-dimensional image space, it first compresses the image into a latent space by training a variational autoencoder (VAE). As shown in Figure 5, We can observe that the latent images produced by the VAE exhibit clear geometric structures, e.g., the contour of the head, flat surface of cheeks, sharp edges around the eyes, dense dots on the whiskers, etc. It is evident that during the training of the diffusion model to generate these images, certain geometric priors are learned as a by-product. Moreover, this finding is indirectly supported by the success of the coarse geometry modeling stage as described in (Chen et al., 2023), where coarse geometries are effectively modeled under the supervision of the diffusion model. Nonetheless, as mentioned earlier, such geometric priors in 2D diffusion are not 3D-aware, leading to multi-view inconsistency issues during the 2D-to-3D lifting.



Figure 5: We render the 3D head from two views, encode the renderings into the latent space where a latent diffusion model is trained, and obtain latent images that exhibit clear geometric structures.

A.2 MORE TEXT-TO-3D RESULTS

We present more text-to-3D synthesis results obtained with our methods (Figure 6, Figure 7, Figure 8, and Figure 9).

A.3 MORE IMPLEMENTATION DETAILS OF OUR PIPELINES

Ours (DMTet-based) We integrate our Aligned Geometric Priors in the official repository of Fantasia3D (Chen et al., 2023) as described in Section 3.2. We follow the same parameters as the original paper. We also disentangle the learning of geometry and appearance. It takes about 12 and 8 minutes to generate a fine geometry and its corresponding Physically-Based Rendering (PBR) materials, respectively, for each object. For the time step range of SDS loss, We adopt a uniform sampling strategy of annealing from [0.5, 0.98] to [0.05, 0.5]

Ours (NeRF-based full) We implement it in the threestudio (Guo et al., 2023), which implemented a diverse set of state-of-the-art text-to-3D generation pipelines. Specifically, we use the Instant-NGP (Müller et al., 2022) as the 3D representation to optimize, which uses a multi-resolution hash-grid to predict the RGB and the density of the sampled ray points. The sampled camera views follow the same protocol as the render dataset to finetune the UNet. In addition, we also use time annealing, negative prompts, and CFG rescaling tricks from the open-source implementation for improved performance. For SDS, the maximum time step is decreased from 0.98 to 0.5 linearly and the maximum time step is kept to 0.02. We use a rescale factor of 0.7 for the CFG rescale. The whole process takes about 1 hour to generate each object with 10,000 steps using 4 V100 GPUs.

A.4 MORE COMPARISON RESULTS USING PROMPTS FROM MVDREAM

Note that, since MVDream’s official implementation is unavailable by the time of our submission, we use the same prompts as listed on their website for side-by-side comparisons. We present the visual comparisons in Figure 10. Although the concurrent work, MVDream, can also resolve the multi-view inconsistency problem, we observe that it is prone to overfit the limited 3D data, consequently resulting in a compromise of the generalizability in the original powerful 2D diffusion model. Specifically, as shown in the results, MVDream misses the “backpack” in its generated result presented with the prompt “an image of a pig carrying a backpack”. Additionally, since they use synthetic multi-view renderings for fine-tuning their multi-view diffusion model, the appearance of the generated results lacks the desired level of photorealism.

A.5 MORE DISCUSSION

When training AGP, we do not introduce any regularization constraint. Other fine-tuned models are usually trained with additional objectives to preserve their original capabilities. Theoretically, there is a potential risk to the integrity of the geometric priors learned in the original pre-trained diffusion model, resulting in the degradation of the generalizability in terms of highly diverse geometries. Nevertheless, we found our model still possesses strong generalizability, as evidenced by the successful results.

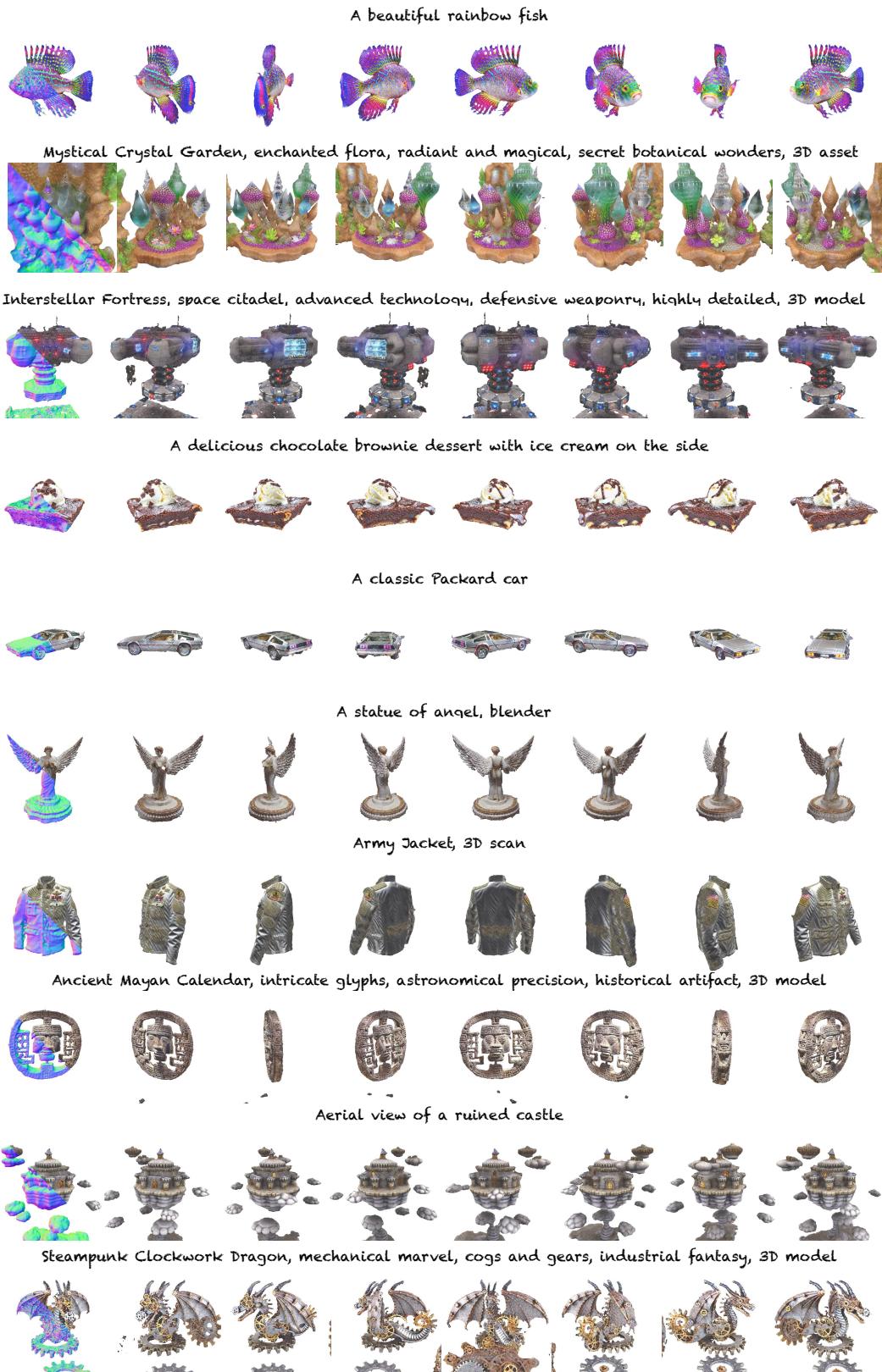


Figure 6: More generated results using our proposed DMTet-based model.

Ancient Roman Colosseum, historic arena, architectural wonder, gladiators and spectacles, 3D render



mini China town, highly detailed, blender 3d



Enchanted Elven Citadel, ethereal fortress, magical spires, elven stronghold, 3D asset



A crab, low poly



A 3D model of A Darth Vader helmet, highly detailed



A bulldog wearing a black pirate hat



A 3D model of Flying Dragon, highly detailed, breathing fire



Mystical Elven Bow, ethereal craftsmanship, enchanted arrows, forest protector, 3D asset



Space Explorer's Exosuit, advanced astronaut armor, HUD visor, interstellar adventure, 3D asset



A bear dressed in medieval armor



Figure 7: More generated results using our proposed DMTet-based model.

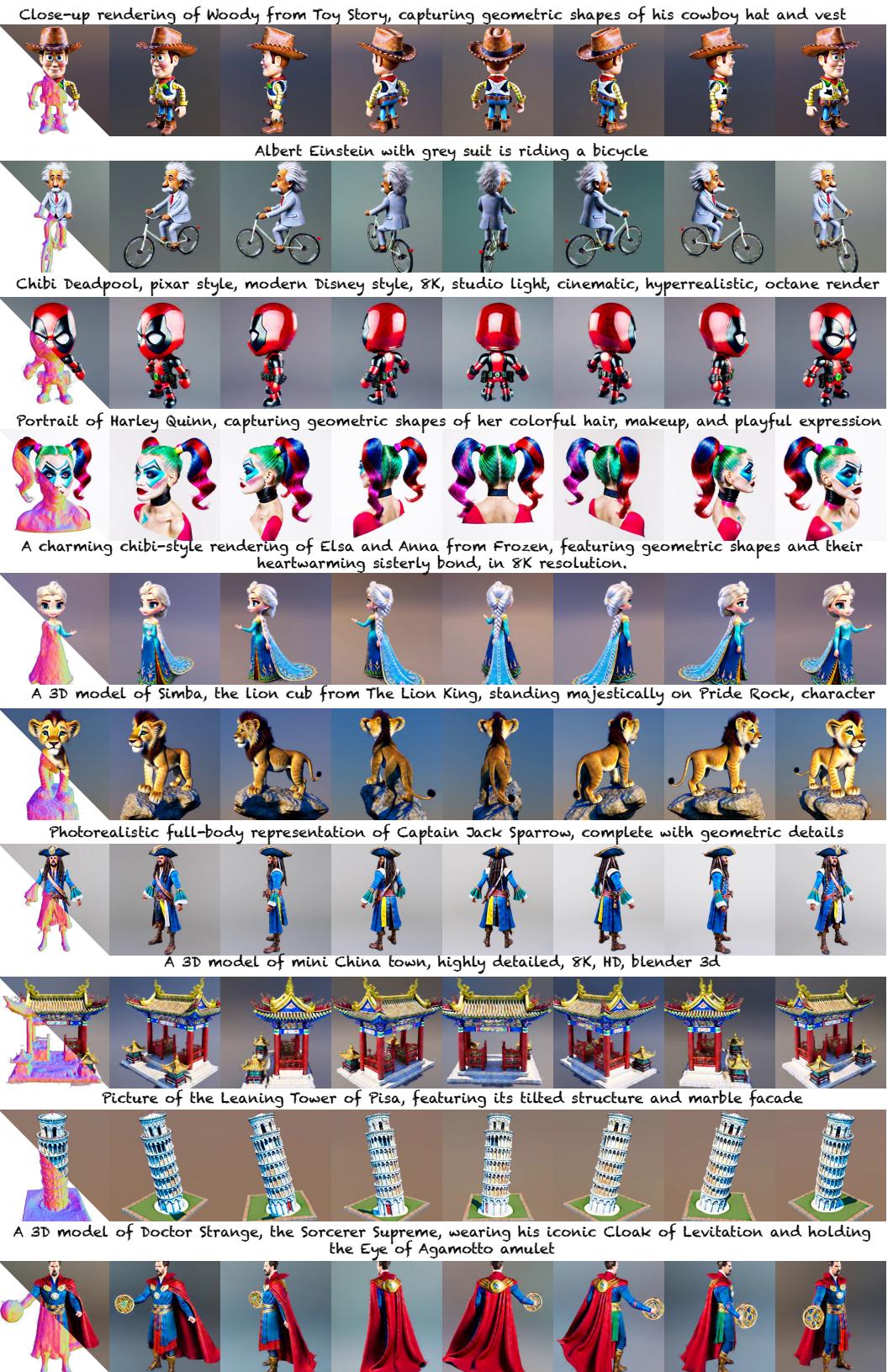


Figure 8: More generated results using our proposed NeRF-based model.

The Hulk smashing through a wall, showcasing his muscular physique and powerful pose in photorealistic 4K detail



Fisherman House, cute, cartoon, blender, stylized



Mini Paris, highly detailed, 8K, HD



Image of Michael Jackson, showcasing his signature dance moves, fedora hat, and stylish wardrobe



Scene of the Temple of Heaven in Beijing, displaying its circular architecture and ornate details



Fire-breathing Phoenix, mythical bird, engulfed in flames, rebirth and renewal, 3D render, 8K, HD



View of Sydney Opera House, showcasing its unique sail-like design and waterfront location



Detailed headshot of Thor, the God of Thunder, emphasizing geometric shapes of his majestic beard and intense gaze



Highly detailed, majestic royal tall ship, realistic painting



Floating Steampunk City, gears and balloons, Victorian-era airship metropolis, 3D render, 4K, HD



Figure 9: More generated results using our proposed NeRF-based model.



Figure 10: Side-by-side visual comparisons using prompts from MVDream. Note that some key concepts in the prompts are missing in MVDream results, such as the rocket, backpack, and squirrel missing in their results.