View Selection for 3D Captioning via Diffusion Ranking

Tiange Luo[1], Justin Johnson[1,†], and Honglak Lee[1,2,†]

[1] University of Michigan
[2] LG AI Research
https://huggingface.co/datasets/tiange/Cap3D

Abstract. Scalable annotation approaches are crucial for constructing extensive 3D-text datasets, facilitating a broader range of applications. However, existing methods sometimes lead to the generation of hallucinated captions, compromising caption quality. This paper explores the issue of hallucination in 3D object captioning, with a focus on Cap3D [35] method, which renders 3D objects into 2D views for captioning using pre-trained models. We pinpoint a major challenge: certain rendered views of 3D objects are atypical, deviating from the training data of standard image captioning models and causing hallucinations. To tackle this, we present DiffuRank, a method that leverages a pre-trained text-to-3D model to assess the alignment between 3D objects and their 2D rendered views, where the view with high alignment closely represent the object's characteristics. By ranking all rendered views and feeding the top-ranked ones into GPT4-Vision, we enhance the accuracy and detail of captions, enabling the correction of 200k captions in the Cap3D dataset and extending it to 1 million captions across the entire Objaverse dataset and a portion of the Objaverse-XL high-quality subset. Additionally, our dataset includes 20 rendered images per caption, providing both intrinsic and extrinsic camera details, depth data, and masks, resulting in a total of 60 million PNG images. Beyond datasets, we showcase the adaptability of DiffuRank by applying it to pre-trained text-to-image models for a Visual Question Answering task, where it outperforms the CLIP model.

1 Introduction

Recent advancements in generative models have shown remarkable performance in both image [2,50] and video [4] domains, driven by the availability of extensive captioned datasets. Despite these successes, extending generative modeling to 3D domains has been challenging due to the scarcity of high-quality 3D-text pairs. This gap has been partially bridged by Cap3D [35], which generates captions for 3D objects by rendering them into 2D images and employing image-based captioning models, further refined by Large Language Models (LLMs) to synthesize

† equal advising

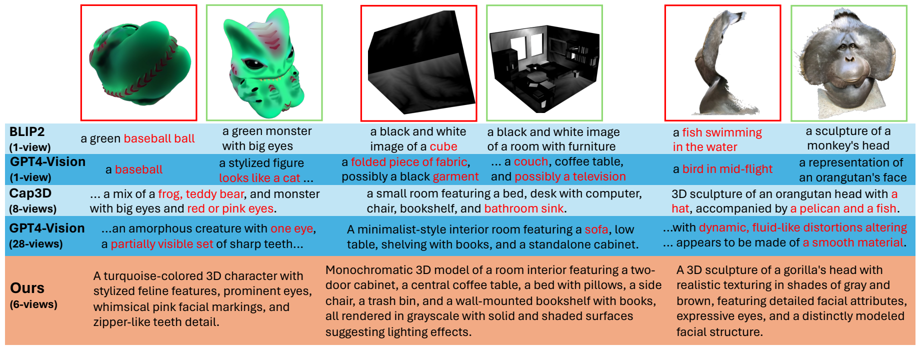

Fig. 1: DiffuRank enhances caption accuracy and reduces hallucinations by prioritizing key rendered views (green box), contrasting the atypical views (red box) that cause errors. Surprisingly, using fewer views (6 vs. 28) not only saves computational resources but also may yield more accurate and detailed outcomes (the middle example) by countering the uncertainty caused by excessive views.

captions. Cap3D has contributed 660k captions for the Objaverse dataset [13], facilitating developments in Text-to-3D [23, 65], Image-to-3D [64, 68], robot simulator [61] and learning [46], and the pre-training of 3D LLMs [43, 63, 70].

Despite the utility of Cap3D, our analysis reveals that a significant portion of Cap3D captions includes inaccurate and hallucinated information, potentially compromising model training [57]. Upon inspection, we found that the key is the rendered view: as Cap3D adheres to the Objaverse's default orientation for 3D objects, it positions the rendering cameras horizontally based on heuristic hyperparameters. Some of the renderings are hard to distinguish even for humans, which existing captioning models cannot handle [24]. Consequently, when these challenging views are included, even advanced captioning models like GPT4-Vision [1] may generate erroneous information, as illustrated in Figure 1.

To address this, we introduce DiffuRank, an approach for ranking rendered views with pre-trained diffusion models. By leveraging a pre-trained text-to-3D diffusion model [20], DiffuRank evaluates the alignment between the captions of each view and the corresponding 3D object's information. The underlying premise is that captions generated from rendered views that closely match the object's 3D information will exhibit a higher alignment, indicating these views are more representative of the object. Consequently, DiffuRank promotes the preferable views (Figure 2) for captioning as those that better reflect the true essence of the 3D objects, leading to more accurate and truthful captions.

Specifically, we first employ image-based captioning models to caption all candidate rendered views, and then perform multiple iterations over diffusion model objective to obtain average score estimation for all captions conditional on the same 3D object feature, Gaussian noise, and timestamps. This score gauges the alignment between the captions and the corresponding 3D object feature. Following this, we rank the views based on their scores and forward

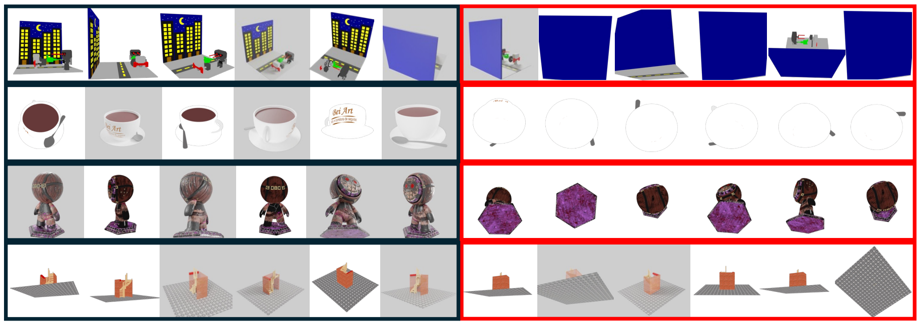

Fig. 2: The left row features the top-6 views as ranked by DiffuRank, while the right row displays the bottom-6. Comparative analysis shows that the top-6 views generally uncover more characteristics of the object compared to the bottom-6. This finding underscores DiffuRank's capability to identify views that more accurately represent the features of the 3D object. More randomly sampled results are included in Appendix B.5.

the top-N rendered views to GPT4-Vision for the final caption generation. Our evaluations through human studies indicate that captions produced with DiffuRank, in conjunction with GPT4-Vision, are of significantly higher quality and exhibit fewer inaccuracies compared to those generated by Cap3D. Moreover, our captions are usually richer in detail and fewer hallucinations when using only 6 rendered views than those produced using GPT4-Vision alone across all 28 rendered views or views selected based on default object orientations.

We extend DiffuRank to the 2D domain, demonstrating its effectiveness in the challenging Visual Question Answering task [59] when combined with text-to-2D diffusion models [49], and surpassing the zero-shot performance of CLIP [48].

Our contributions are as follows:

– We identify and alleviate the systematic hallucinations in Cap3D captions, revising approximately 200k entries with the help of DiffuRank and GPT4-Vision. The corrected captions consistently improve the finetuned performance of text-to-3D models (Point·E, Shap·E); note that Shap·E models fine-tuned with Cap3D captions show decreased performance.
– We extend the Cap3D caption dataset [35] from 660k to 1M across the entire Objaverse [13] and a portion of the Objavere-XL high-quality subset [12]. Each caption is complemented with point clouds containing 16,384 colorful points and 20 rendered images, including camera, depth, and MatAlpha details. This results in a total of 1 million point clouds and 60 million PNG images. All data is released under the ODC-By 1.0 license and is available at https://huggingface.co/datasets/tiange/Cap3D.
– We proposed DiffuRank which shown ability to model the alignment between 3D object and its 2D rendered views via a pre-trained Text-to-3D model and a captioning model. Additionally, we extend DiffuRank to 2D domain, and demonstrate DiffuRank beats CLIP on the VQA task [59] with the help of a pre-trained text-to-2D diffusion model [49].

2 Related Work

2.1 3D-Text

Recent advancements introduced by Objaverse have significantly enriched the field of 3D object research. By integrating a comprehensive set of 3D objects with descriptive captions from Cap3D, a wide array of 3D applications has been enabled. These include Text-to-3D methods [17, 23, 25, 37, 65], Image-to-3D conversion techniques [64, 68], enhancements in robot learning [46, 61], the pre-training of 3D language models [8, 27, 47, 63, 70], and the development of language models capable of processing diverse modalities [5, 16, 43].

Despite these advancements, we identified issues with hallucination contents in the captions provided by Cap3D. This discovery aligns with findings from concurrent research [21,32,57], pinpointing inaccuracies in Cap3D captions. Our investigation reveals that the root cause of these inaccuracies is attributed to atypical rendered views, which lead to failures in captioning models. These failures are exacerbated as text summarization models (GPT4) are unable to rectify these errors. To address this challenge, we introduce DiffuRank that selects rendered views capturing the essential characteristics of 3D objects. Furthermore, we utilize the recent advancements in vision-language models, specifically GPT4-Vision, to provide holistic captions for 3D objects. We release our dataset under ODC-By 1.0 license to enable research and commercial usage, and hope facilitate related 3D-Text research [6, 7, 9, 10, 14, 19, 23, 25, 26, 29, 30, 33, 34, 36, 38, 41, 45, 51, 53, 53, 56, 60, 62, 64, 66, 72].

2.2 Diffusion Model

Our proposed DiffuRank leverages denoising diffusion objective [18, 54, 55] to model the alignment between the input and output modalities. By using pre-trained text-to-3D [20, 41] and text-to-2D [2, 44, 50] diffusion models, we can model the alignment between given 3D object/image for a set of possible captions (text descriptions) as detailed in Section 3.2. In our listed algorithm 1, we adopt the objects $L_{3D} = E_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,\mathbf{I}), t \sim U[1,T]} \left\| x_\theta\left(x_t, t\right) - x_0 \right\|_2^2$ as used in Shap·E [20], where x_0 is data sampled from data distribution $q(x_0)$, ϵ is Gaussian noise, and t is timestamp. We also adopt the alternative but equivalent objective, $L_{\mathrm{2D}} = E_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,\mathbf{I}), t \sim U[1,T]} \left\| \epsilon - \epsilon_\theta\left(x_t, t\right) \right\|_2^2$, when we adopt the text-to-2D model, stable-diffusion, in Section 5.3.

DiffuRank is related to score sampling distillation proposed in [45], while we do not compute gradients but sampling loss to accumulate scores estimation for ranking. Our findings also relate to works which leverage pre-trained diffusion models for downstream tasks, such as image classification [22, 39], semantic segmentation [69], visual grounding [31], depth prediction [52, 67], and other low-level computer vision tasks [15].

When applying our method to the 2D domain, we discovered that our algorithm aligns closely with the insights of the approach presented in [22]. Consequently, our method can be considered an expansion of the findings from [22],

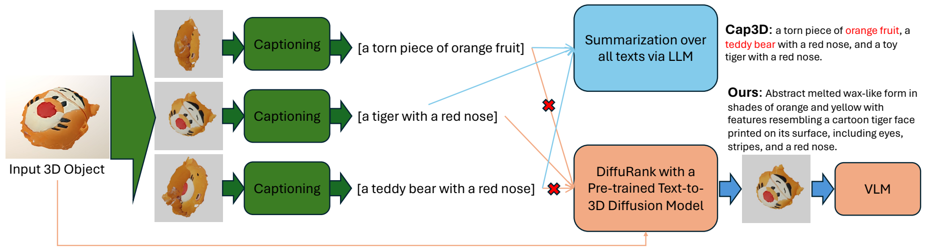

Fig. 3: Methods overview. Both Cap3D and our method render input 3D objects into multiple views for caption generation (green steps). However, while Cap3D consolidates these captions into a final description (blue steps), our method employs a pre-trained text-to-3D diffusion model to identify views that better match the input object's characteristics. These selected views are then processed by a Vision-Language Model (VLM) for captioning (orange steps).

extending its applicability from 2D classification to broader domains and tasks, including the use of a pre-trained text-to-3D diffusion model and a 2D-image-based captioning model to estimate the alignment between 3D objects and their 2D rendered views, as well as the application of a pre-trained text-to-2D diffusion model to solve Visual Question Answering tasks.

3 Method

In this section, we analyze the issues with atypical rendered views leading to hallucinations in Cap3D captions, motivating our proposed DiffuRank, a approach for selecting informative rendered views with 3D priors learned from a diffusion model. We then detail DiffuRank's formulation and describe our novel 3D captioning framework that integrates GPT4-Vision.

3.1 Issues in Cap3D

Firstly, we revisit the Cap3D pipeline, which unfolds across four stages. Initially, it renders a set of 2D views for each 3D object. Subsequently, image captioning is applied to generate preliminary descriptions (5 captions for each image). Then, the CLIP model is utilized in the third stage to select the best-aligned caption for each image, filtering out inaccuracies. The process culminates with an LLM synthesizing captions from various perspectives into a comprehensive caption.

However, the captioning of rendered views (the combined second and third stages) for given 3D objects can falter with atypical views, producing captions that diverge significantly from the actual 3D object. In the worst-case scenarios, each rendering view might correspond to an incorrect object, leading to compounded errors when these captions are summarized by GPT4. One example is shown in Figure 3. Since GPT4 operates solely on text, it cannot correct these inaccuracies, resulting in captions riddled with hallucinated details.

Due to the versatility of 3D object geometries, determining which rendered views best reflect a 3D object's characteristics is non-trivial. While measuring the geometric properties of 3D objects and computing their principal directions is feasible, positioning the camera orthogonally, as shown in the bottom-left example of Figure 2, is often suboptimal. Hence, we propose DiffuRank, which learns 3D priors from data to filter preferable rendered views by leveraging a pre-trained text-to-3D model. Our experiments demonstrate that DiffuRank efficiently enhances caption quality and reduces hallucinations with fewer renderings compared to using all available views.

3.2 DiffuRank Formulation

DiffuRank leverages a pre-trained text-to-3D diffusion model $D_{text-to-3D}$ to rank rendered views based on their alignment with both captions and the corresponding 3D information.

For a given 3D object $\mathcal{O}$, assuming a set of candidate captions c_i and the pre-trained model $D_{text-to-3D}$, the training objective of this pre-trained diffusion model is predicting a 3D object $\mathcal{O}$ based on a text description c, i.e., modeling the score function $\nabla_{\mathcal{O},c} p(\mathcal{O}|c)$ of the data distribution $p(\mathcal{O}_i|c)$. Specially, the diffusion model aims to minimize

$$\mathcal{L}_c = \|D_{text-to-3D}(\mathcal{O}_t|c) - \mathcal{O}_0\|$$

based on a text description c, where the noised input $\mathcal{O}_t = \sqrt{\bar{\alpha}_t}\mathcal{O} + \sqrt{1-\bar{\alpha}_t}\epsilon$, for timestamp t, and randomly sampled Gaussian noises $\epsilon \sim \mathcal{N}(0, I)$, with $\bar{\alpha}$ being a hyper-parameters defined by the noise schedule [18]. Our tuition here is simple: a caption closely aligned with the given 3D object in terms of characteristics (e.g. structure, colors, textures, etc), should aid the diffusion model in making accurate predictions starting from the same noised input $\mathcal{O}_i^t$, resulting in a lower score matching loss. By sampling multiple sets of t_j, ϵ_j for the same set of captions c_i, we can measure the alignment $Cor(\mathcal{O}, c_i)$ between the 3D object and captions via the average loss.

Initially, we generate candidate captions for $\mathcal{O}$ by rendering it into multiple views I_i and generating captions c_i^j with a captioning model $Dcap$. This captioning procedure aims to maximize the joint likelihood of the model distribution $p(c_i^j, I_i)$ over the image I_i and generated captions c_i^j. Thus, we estimate the alignment between the 3D object and all captions of the same rendering $Cor(\mathcal{O}, \mathbb{E}_j c_i^j)$, which is proportional to $Cor(\mathcal{O}, \mathbb{E}_j p(c_i^j, I_i)) \propto Cor(\mathcal{O}, I_i)$. Then, we write down the whole pipeline in Algorithm 1.

Specifically, we adopted shap-E as the text-to-3D diffusion model in our paper, and the above $\mathcal{O}_i$ should be $E_{encoder}(\mathcal{O}_i)$, where $E_{encoder}$ is the encoder (transmitter in [20]) to extract feature embeddings from given 3D object.

Furthermore, DiffuRank's application is not confined to 3D captioning; because it is a general framework for measuring the alignment between two modalities received and output by a diffusion model. It can be seamlessly extended to other domains, such as 2D images. In section 5.3, we show an example where we apply DiffuRank to perform 2D VQA and beat CLIP model [48].

Algorithm 1 DiffuRank for modeling the alignments between 3D object and its rendered views

Require: Given 3D object $\mathcal{O}$, pre-trained text-to-3D model $D_{\text{text-to-3D}}$, captioning model D_{cap}
# 1. rendered views $\{I_i\}_{i=1,\cdots,M}$ for $\mathcal{O}$ with rendering program (e.g., Blender).
# 2. Generate candidate captions for $\mathcal{O}$.
for each view I_i of $\mathcal{O}$ **do**
  Generate captions $\{c_i^j\}_{j=1,\cdots,N}$ with captioning model D_{cap}.
end for
# 3. Compute average alignment scores
for $k \leftarrow 1$ to num_samples **do**
  Sample timestamp $t_k \sim \text{Uniform}(0, 1)$.
  Sample noise $\epsilon_k \sim \mathcal{N}(0, I)$.
end for
for each rendering view I_i **do**
  for $k \leftarrow 1$ to num_samples **do**
    Compute noised input $\mathcal{O}_{t_k} = \sqrt{\bar{\alpha}_{t_k}}\mathcal{O}_0 + \sqrt{1-\bar{\alpha}_{t_k}}\epsilon_k$.
    for $j \leftarrow 1$ to N **do**
      Compute loss $\mathcal{L}_{c_i^j,k} = \|D_{\text{text-to-3D}}(\mathcal{O}_{t_k}|c_i^j) - \mathcal{O}_0\|$.
    end for
  end for
  Compute average loss for all captions of I_i, $Cor(I_i, \mathcal{O}) = -\mathbb{E}_{j,k}\mathcal{L}_{c_i^j,k}$.
end for
return Top-P($\{Cor(I_i, \mathcal{O})\}_{i=1,\cdots,M}$)

3.3 New 3D Captioning Framework

With the proposed DiffuRank, we establish a new 3D captioning pipeline, as shown in Figure 3. For given 3D object, we render it into 28 images, which are then captioned into 5 descriptions using an image-based captioning model. Following captioning, DiffuRank ranks the rendered views using a pre-trained text-to-3D model. This ranking enables the selection of the Top-6 rendered views for processing by a vision-language model, resulting in holistic captions that describe structure, form, color, and texture, with enhanced accuracy and detail.

To elaborate, our methodology integrates two distinct rendering strategies, as illustrated in Figure 4. The first strategy, derived from Cap3D [35], renders objects into 8 views against a uniform grey background, arranged horizontally around the object's default orientation, with Blender ray-tracing render engine 'CYCLES'. Concurrently, we apply a second technique from Shap·E [20], where 20 views are generated through randomized sampling after object normalization, set against a transparent background, with Blender real-time engine 'EEVEE'. These 20 views, created following the Shap·E methodology, are instrumental in forming Shap·E latent codes, i.e. $E_{encoder}(\mathcal{O}_i)$ in Section 3.2. Altogether, this approach results in 28 distinct views for each object. Additionally, as grey and transparent backgrounds may accentuate or obscure details variably across objects, we observed that DiffuRank adeptly selects the views with the proper

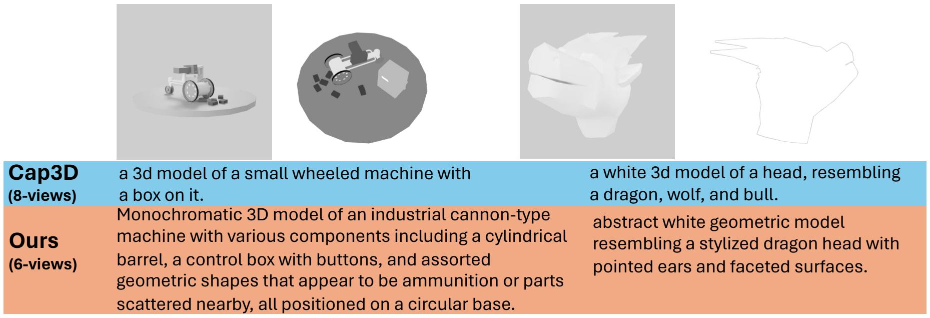

Fig. 4: We utilized both grey background + ray-tracing render engine (left images) and transparent background + real-time render engine (right images) for rendering, discovering that the effectiveness of each varies. We noticed DiffuRank can select the views with the appropriate rendering that highlight object features.

background that most effectively highlight object features, without manual intervention. Some examples are included in Appendix B.

Following this, the captioning model, BLIP2 [24], is employed to generate five captions for each view. These captions, alongside the pre-trained text-to-3D diffusion model, Shap·E [20], and the previously derived 3D latent code $E_{encoder}(\mathcal{O}_i)$, undergo analysis in the DiffuRank process, as detailed in Algorithm 1. Subsequent to DiffuRank, the six views that demonstrate the highest alignment scores are chosen to input into GPT4-Vision for caption generation.

4 Dataset

This section details our process for correcting the Cap3D captions, expanding the dataset with high-quality 3D objects from Objaverse-XL, and ethical filtering. More detailed hyper-parameters and comparisons are included in Appendix B.

4.1 Correction of Cap3D Captions

As Cap3D contains a lot of good quality captions, our first objective is to identify erroneous Cap3D captions, which might contain incorrect information or hallucinations. We tried three strategies as outlines the below.

Image-Text Alignment Method: We discovered that utilizing the maximum and average CLIP scores effectively filters out inaccurate captions. Most of erroneous captions, like those depicted in Figure 1, described improbable combinations of objects (e.g., “a mix of a frog, teddy bear, and monster" or “an orangutan accompanied by a pelican and a fish") in scenarios where only one entity was present in the given 3D object. Such discrepancies arise when different views of the same 3D object receive varied entity captions from BLIP2, which GPT4 then erroneously combines, shown in Figure 3. To detect this kind of case, we computed both the average and maximum CLIP scores between the

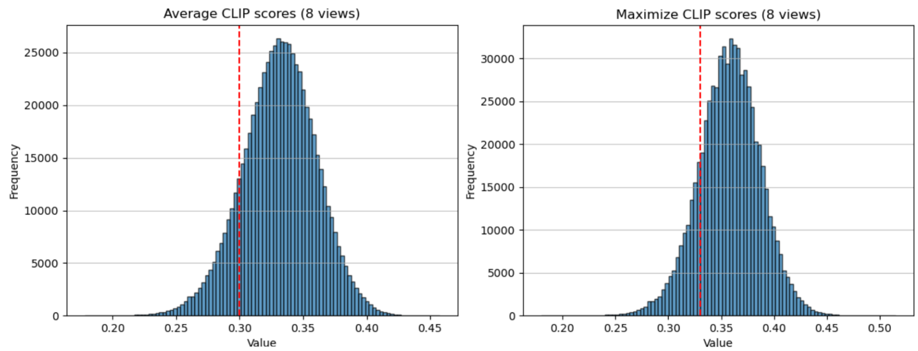

Fig. 5: Mean and Max clip score distribution for Cap3D captions and their 8 rendering images. Selected thresholds are the two red dash lines via our annotated validation set.

final caption and all eight rendered views used in Cap3D. A validation set of $\sim 7k$ objects with inaccurate captions was annotated and used to determine two thresholds (mean & max as shown in Figure 5), with the goal of encompassing all objects in this set. We then use the two selected thresholds to filter out $\sim 167k$ possible issued objects out of a total of $660k$.

Image-Based Method: Approximately $10k$ renderings in Cap3D dataset were identified as having all-grey images, likely due to rendering issues within the Cap3D process. We addressed this by re-rendering these objects and updating their captions with descriptions generated by our method (Section 3.3).

Text-Based Method: Attempting to identify errors solely based on captions proved challenging due to the diverse and complex nature of objects within Objaverse, making it difficult to detect hallucinations based on text alone. This complexity arises because some 3D objects genuinely comprise multiple or unusual components. Despite this, we developed a technique for identifying the misuse of terms related to "image" and "rendering", as these are directly associated with the rendering process rather than the 3D objects themselves. Through this method, we identified approximately 23,000 objects requiring correction.

4.2 Dataset Expansion and Ethical Filtering

Our expansion includes adopting the remaining objects of Objaverse, where Cap3D did not include, and high-quality 3D objects from Objaverse-XL's curated subset (Section 4.1 of [12], selected through human evaluation and heuristics from a pool of 10 million objects. This extension enhances the diversity and quality of our dataset.

Moreover, we apply ethical filtering to both the rendered images and generated captions to remove potentially NSFW content and identifiable human faces, following Cap3D's protocol. We also leverage GPT4-Vision's internal detection capabilities for identifying images with potential ethical issues. It returns 'content_policy_violation' once their model detection the image possibly against

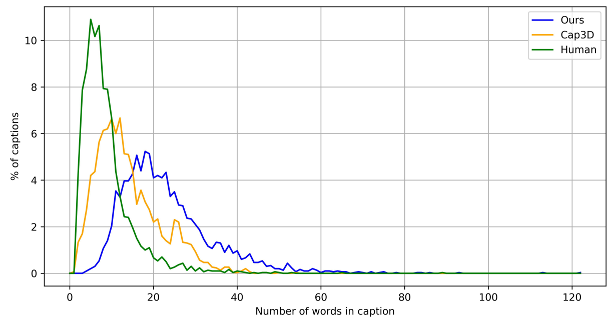

Fig. 6: Number of words in caption.

	Human	Cap3D	Ours
Unigrams	2,876	2,767	**5,600**
Bigrams	11,374	12,293	**29,521**
Trigrams	16,535	23,062	**52,457**

Fig. 7: Number of n-grams for captions generated by different methods.

their safety policy. These comprehensive measures have allowed us to detect a list of $\sim 35k$ objects.

We compared caption length and n-grams [3] of captions among Human, Cap3D, and our captions in a 5k common set. As shown in Figure 6, our captions usually contain longer length indicating more details than Cap3D and human-authored captions. Table 7 demonstrates we have the largest vocabulary size.

5 Experiments

In this section, we compare our captions against Cap3D captions and human-authored captions in terms of quality and hallucination degrees through human studies. We also ablate our methods to verify the effectiveness of the proposed DiffuRank. Then, we compare text-to-3D models finetuned on Cap3D and our updated Captions on the same set to measure the improvements of caption alignment at scale. Finally, we further verify the effectiveness of our propose DiffuRank by examining it on a VQA task. For the sake of space, we list quantitative results here and include qualitative comparisons in Appendix B and C.

5.1 Captioning Evaluation

Settings. We first evaluate the quality of captions generated by our method. Our captioning process involves selecting the top 6 captions out of a total of 28, as determined by DiffuPick, and then feeding these captions into GPT4-Vision (for further details, see Section 4). We evaluate the generated captions by comparing them to those produced by Cap3D, as well as to the human-authored captions that Cap3D provides. Our goal is to determine whether our method can produce captions of higher quality and with fewer inaccuracies or hallucinations.

Furthermore, we conduct ablation studies to assess the effectiveness of another component of our method, DiffuRank. We compare various approaches to highlight the benefits of DiffuRank: (1) **Allviews 28-views:** using all 28 rendered views as input to GPT4-Vision (details in Section 3.3), (2) **Horizontal**

Table 1: Objaverse Captions Evaluations. All A/B testing represents captions from other methods vs. ours. We tested on 5k objects.

Method	Quality A/B test			Hallucination A/B test			CLIP			
	Score(1-5)	Win %	Lose %	Score(1-5)	Win %	Lose %	Score	R@1	R@5	R@10
Human	2.57	31.9	62.1	2.88	39.9	46.4	66.2	8.9	21.0	27.8
Cap3D	2.62	32.7	60.2	2.43	25.8	63.9	71.2	20.5	40.8	51.9
Ours	-	-	-	-	-	-	**74.6**	**26.7**	**48.2**	**57.5**
Allviews 28-views	2.91	37.9	43.6	2.85	35.1	47.2	73.5	24.9	46.7	55.7
Horizontal 6-views	2.84	35.2	44.5	2.90	36.2	40.9	73.8	25.8	46.7	55.9
Bottom 6-views	2.74	31.1	52.0	2.61	30.1	57.0	72.8	24.6	45.1	55.2

6-views: selecting 6 rendered views that place the camera horizontally across the object's default orientation, applying the same up and down positioning heuristics as Cap3D, and (3) **Bottom 6-views:** using the bottom-6 captions, defined as those with the worst alignment scores according to our DiffuRank algorithm (see Alg. 1), as input to GPT4-Vision. Through these comparisons, we aim to demonstrate the impact of DiffuRank's selection process on the quality of the generated captions.

Metrics. Our primary evaluation method utilizes A/B testing with human judgment, where participants evaluate a pair of captions on a 1-5 scale, with 3 representing a neutral preference (i.e., tie). Our approach includes two distinct assessments: (a) evaluating which caption more accurately describes the object's type, appearance, and structure, and (b) determining which caption is less prone to presenting incorrect information or hallucinations. Each assessment involves over 10,000 ratings across 4,000 objects to ensure statistical reliability. We calculate and report the average scores and the frequency each option is preferred (i.e., excluding neutral (tie) responses). More human evaluation details are included in Appendix B.7. Additionally, we follow Cap3D [35] and employ automated metrics, including CLIP score, measuring the cosine similarity between CLIP encodings and input images, and CLIP R percision [45], assessing the match between a rendered image and all potential texts.

Results. The evaluation results, presented in Table 1, highlight the effectiveness of our captioning approach. According to scores from human evaluators on quality and hallucination metrics, our captions feature more accurate details with fewer instances of hallucination, compared to Cap3D and human-authored captions. Supporting qualitative findings are detailed in Appendix B.3, reinforcing these conclusions.

A comparison of our method, which selects the top-6 views, with alternatives—the bottom-6 views and horizontally placed 6-views—demonstrates the impact of DiffuRank on performance. Specifically, as depicted in Figure 2, bottom-6 views often relate less to the 3D object as they may capture only the back or bottom. This issue highlights the difficulties arising from Objaverse's random default orientation, positioning cameras 'horizontally' does not always ensure they are actually horizontal. More qualitative compairsons between the three

Table 2: Text-to-3D Finetuning experiments.

	FID↓	CLIP Score	CLIP R-Precision (2k) R@1	R@5	R@10
Ground Truth Images	-	81.6	32.7	55.1	64.3
Point·E	36.1	61.5	3.4	10.4	15.3
Point·E + Cap3D	32.8	65	7.1	19.4	26.4
Point·E + Ours (330k)	32.4	66.2	8.1	20.3	28.5
Point·E + Ours (825k)	**31.2**	**66.5**	**10.1**	**21.9**	**29.8**
Shap·E (STF)	37.2	68.8	12.7	29.0	37.9
Shap·E (STF) + Cap3D	35.5	68.2	11.9	28.8	37.4
Shap·E (STF) + Ours (330k)	35.6	69.4	13.4	29.7	39.3
Shap·E (STF) + Ours (825k)	**34.3**	**69.8**	**14.9**	**33.7**	**42.8**
Shap·E (NeRF)	48.7	68.3	12.2	27.9	36.2
Shap·E (NeRF) + Cap3D	48.2	68.0	11.7	27.1	35.1
Shap·E (NeRF) + Ours (330k)	48	68.4	13.2	29.3	38.4
Shap·E (NeRF) + Ours (825k)	**47.9**	**69.3**	**14.3**	**31.7**	**40.4**

types of view selection are included in Appendix B.5. Furthermore, DiffuRank does not consistently achieve optimal performance, as illustrated by the selection of the 6th image in the first row (referenced in Figure 2) captioned 'a blue laptop.' Enhancements could be achieved through using an improved text-to-3D diffusion models, a topic explored in detail in Appendix E.

Furthermore, our approach outperforms the variant using 24 views, delivering captions with greater detail and fewer hallucinations (See qualitative comparisons at Appendix B.4). Interestingly, providing a larger number of views (24) does not necessarily improve details; it appears to complicate the model's ability to access precise information due to the variance in detail across different perspectives. This observation contradicts expectations, suggesting an optimal balance of view selection is crucial for accurate 3D object captioning.

5.2 Text-to-3D Generation with New Captions

Settings. This section we finetune Text-to-3D models to check if our updated captions can bring more improvements compared to Cap3D captions. For this purpose, we would mainly conduct experiments over point-E [41] and shap-E [40] as they are used in Cap3D. We follow the same setting as Cap3D, including learning rate, batch size, optimizer, and steps. We adopted the same 330k training split and test split used in [35], and we have updated $72k$ captions in this 330k set ($\sim$20%). Additionally, we scale our experiment up, and train models with 825k ($2.5 \times 330k$) data from our full 3D-text pairs. More details and qualitative results are included in Appendix C.

Metrics. We incorporated the use of CLIP Score and CLIP R-Precision [35, 45] in our evaluation process. CLIP R-Precision involves ranking a rendered image among all text pairs within the test set based on the cosine similarity as measured by CLIP, then determining the precision based on accurate text-image matches. Given the availability of ground truth images, we employed the FID metric to compare the fidelity of 3D rendered images with these true images.

Results. Results are showcased in Table 2. Considering we updated nearly 20% captions of the 330k training set for Cap3D 3D-text pairs, we anticipated some improvement, albeit modest. However, the improvements exceeded our expectations. Our enhanced model ('model + Ours' with 330K data points) consistently outperformed both the 'model + Cap3D' (with 330K data points) version and pre-trained Shap·E model. Surpassing the pre-trained Shap·E model is non-trivial, as the 'model + Cap3D' version generally showed declining performance when compared to the pre-trained Shap·E model, indicating that fine-tuning on Cap3D data actually harms the performance. The performance enhancement achieved by correcting 20% of the data underscores the effectiveness of addressing misalignments in the 3D-text of Cap3D by locating the potential errors and refining with our new captioning approach. Furthermore, by expanding our dataset by 2.5 times, we have boosted performance across multiple metrics and models. Given that Shap·E model was trained on proprietary data, our findings suggest that our proposed 3D-text dataset could be a competitive open-source alternative.

5.3 DiffuRank on VQA

Settings. We extend our DiffuRank to solve Visual Question Answering task, with the help of a pre-trained text-to-2D diffusion model [49]. We list our detailed settings and the updated algorithm in Appendix D. We mainly compare to CLIP [48] in terms of zero-shot VQA performance and test on the Multimodal Visual Patterns (MMVP) benchmark [59], comprising nine fundamental visual patterns across 150 images pairs. Each pair of images (Figure 8), despite having clear visual distinctions, are perceived similarly by the CLIP model. Each pair is associated with a question that has two divergent answers. Numerous Vision-Language Models (VLMs) have been shown to underperform on this challenging benchmark.

Given that the task involves Visual Question Answering (VQA), neither our approach nor the CLIP model is inherently designed to generate textual responses directly. To address this, we employed GPT-4 to transform each question and its corresponding answers into declarative statements. Consequently, for each image pair, we obtained two distinct statements. For DiffuRank, we executed multiple iterations of alignment estimation for the statements corresponding to each image, selecting the statement with the highest alignment estimate as the correct answer/statement. For CLIP model, we determined the appropriate answer by calculating the cosine similarity between an image and each statement, choosing the statement with the greatest similarity as the response. We used "ViT-B/32" CLIP here for evaluation.

Model	Accuracy (%)
Human	95.7
Gemini [58]	40.7
GPT4-Vision [42]	38.7
Ours	30.7
Random Guess	25.0
LLaVA-1.5 [28]	24.7
Bard	19.0
Bing Chat	17.3
InstructBLIP [11]	16.7
CLIP [48]	13.3
mini-GPT4 [71]	12.7
LLaVA [28]	6.0

Table 3: Accuracy comparison among various VLMs, CLIP, and our method.

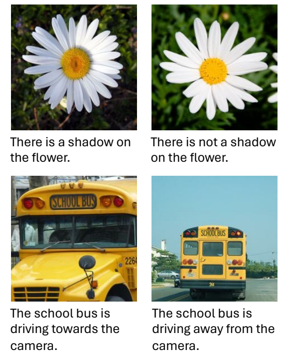

Fig. 8: Each row represents a matched pair, and the accompanying text beneath it is the description.

Metrics. Our evaluation metrics are aligned with those proposed by [59]. A model's response is deemed accurate only if it correctly identifies the appropriate statements for both images in a pair. Hence, if a model accurately selects the correct statement for only one image within the pair, its attempt is marked as incorrect. It is important to note that both DiffuRank and CLIP may occasionally select identical statements for different images within the same pair.

Results. Table 3 shows the quantitative results which demonstrate DiffuRank significantly outperforms CLIP in the MMVP benchmark with the help of pre-trained stable diffusion model. Also, for the example pairs shown on the Figure 8, our method is able to select the correct corresponding image-statement pairs. In contrast, the CLIP model incorrectly selects There is not a shadow on the flower' and The school bus is driving towards the camera' for both images in each pair.

6 Conclusion

This paper help alleviate inaccuracies and hallucinations in Cap3D captions (a 3D-Text dataset for Objaverse), attributed to suboptimal render views based on default object orientations. We introduced DiffuRank to address this issue, a method that ranks rendered views by their alignment with 3D object information using pre-trained text-to-3D diffusion models. Combining DiffuRank and GPT4, our new captioning approach improved caption quality, reduced inaccuracies, and enhanced detail richness with fewer views. Our efforts have not only improved the quality of existing Cap3D captions but also expanded the dataset to cover a total of 1M 3D-text pairs (whole Objaverse and a subset of Objaver-XL highquality set). We also extended DiffuRank's application to the 2D domain, demonstrating its effectiveness in Visual Question Answering tasks.

Acknowledgement

This work has been made possible through the generous support of the "Efficient and Scalable Text-to-3D Generation" grant from LG AI Research. We also thank the OpenAI Researcher Access Program for partially supporting our use of the GPT-4 API. We greatly appreciate Chris Rockwell for his invaluable technical support in caption evaluation, and Mohamed El Banani for his insightful feedback to our initial draft. Tiange thanks Minghua Liu and Jiaming Song for their insightful discussions back at NeurIPS 2023 in NOLA.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
3. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
4. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), `https://openai.com/research/video-generation-models-as-world-simulators`
5. Chen, C., Du, Y., Fang, Z., Wang, Z., Luo, F., Li, P., Yan, M., Zhang, J., Huang, F., Sun, M., et al.: Model composition for multimodal large language models. arXiv preprint arXiv:2402.12750 (2024)
6. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. arXiv (2023)
7. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
8. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. arXiv preprint arXiv:2311.18651 (2023)
9. Chen, Y., Fang, J., Huang, Y., Yi, T., Zhang, X., Xie, L., Wang, X., Dai, W., Xiong, H., Tian, Q.: Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. arXiv preprint arXiv:2312.04424 (2023)
10. Chen, Y., Pan, Y., Li, Y., Yao, T., Mei, T.: Control3d: Towards controllable text-to-3d generation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1148–1156 (2023)
11. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
12. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)

13. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects (2023)
14. Ding, L., Dong, S., Huang, Z., Wang, Z., Zhang, Y., Gong, K., Xu, D., Xue, T.: Text-to-3d generation with bidirectional diffusion using both 2d and 3d priors. arXiv preprint arXiv:2312.04963 (2023)
15. Du, X., Kolkin, N., Shakhnarovich, G., Bhattad, A.: Generative models: What do they know? do they know things? let's find out! arXiv preprint arXiv:2311.17137 (2023)
16. Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: Onellm: One framework to align all modalities with language. arXiv preprint arXiv:2312.03700 (2023)
17. He, Y., Bai, Y., Lin, M., Zhao, W., Hu, Y., Sheng, J., Yi, R., Li, J., Liu, Y.J.: T3 bench: Benchmarking current progress in text-to-3d generation. arXiv preprint arXiv:2310.02977 (2023)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33** (2020)
19. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. arXiv preprint arXiv:2112.01455 (2021)
20. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
21. Kabra, R., Matthey, L., Lerchner, A., Mitra, N.J.: Evaluating vlms for score-based, multi-probe annotation of 3d objects. arXiv preprint arXiv:2311.17851 (2023)
22. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. arXiv preprint arXiv:2303.16203 (2023)
23. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. arXiv preprint arXiv:2311.06214 (2023)
24. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv (2023)
25. Li, W., Chen, R., Chen, X., Tan, P.: Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. arXiv preprint arXiv:2310.02596 (2023)
26. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. arXiv preprint arXiv:2211.10440 (2022)
27. Liu, D., Huang, X., Hou, Y., Wang, Z., Yin, Z., Gong, Y., Gao, P., Ouyang, W.: Uni3d-llm: Unifying point cloud perception, generation and editing with large language models. arXiv preprint arXiv:2402.03327 (2024)
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
29. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., et al.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928 (2023)
30. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. arXiv (2023)
31. Liu, X., Huang, S., Kang, Y., Chen, H., Wang, D.: Vgdiffzero: Text-to-image diffusion models can be zero-shot visual grounders. arXiv preprint arXiv:2309.01141 (2023)
32. Liu, Y.T., Luo, G., Sun, H., Yin, W., Guo, Y.C., Zhang, S.H.: Pi3d: Efficient text-to-3d generation with pseudo-image diffusion. arXiv preprint arXiv:2312.09069 (2023)

33. Lorraine, J., Xie, K., Zeng, X., Lin, C.H., Takikawa, T., Sharp, N., Lin, T.Y., Liu, M.Y., Fidler, S., Lucas, J.: Att3d: Amortized text-to-3d object synthesis. arXiv preprint arXiv:2306.07349 (2023)
34. Luo, T., Lee, H., Johnson, J.: Neural shape compiler: A unified framework for transforming between text, point cloud, and program. Transactions on Machine Learning Research (2023), `https://openreview.net/forum?id=gR9UVgH8PZ`
35. Luo, T., Rockwell, C., Lee, H., Johnson, J.: Scalable 3d captioning with pretrained models. arXiv preprint arXiv:2306.07279 (2023)
36. Melas-Kyriazi, L., Laina, I., Rupprecht, C., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8446–8455 (2023)
37. Mercier, A., Nakhli, R., Reddy, M., Yasarla, R., Cai, H., Porikli, F., Berger, G.: Hexagen3d: Stablediffusion is just one step away from fast and diverse text-to-3d generation. arXiv preprint arXiv:2401.07727 (2024)
38. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13492–13502 (2022)
39. Mukhopadhyay, S., Gwilliam, M., Agarwal, V., Padmanabhan, N., Swaminathan, A., Hegde, S., Zhou, T., Shrivastava, A.: Diffusion models beat gans on image classification. arXiv preprint arXiv:2307.08702 (2023)
40. Nichol, A., Jun, H.: Shap-e: Generating conditional 3d implicit functions. arXiv (2023)
41. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv (2022)
42. OpenAI: Gpt-4 technical report. arXiv (2023)
43. Panagopoulou, A., Xue, L., Yu, N., Li, J., Li, D., Joty, S., Xu, R., Savarese, S., Xiong, C., Niebles, J.C.: X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. arXiv preprint arXiv:2311.18799 (2023)
44. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
45. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv (2022)
46. Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., Ma, K.: Shapellm: Universal 3d object understanding for embodied interaction. arXiv preprint arXiv:2402.17766 (2024)
47. Qi, Z., Fang, Y., Sun, Z., Wu, X., Wu, T., Wang, J., Lin, D., Zhao, H.: Gpt4point: A unified framework for point-language understanding and generation. arXiv preprint arXiv:2312.02980 (2023)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
50. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
51. Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C.Y., Fumero, M., Malekshan, K.R.: Clip-forge: Towards zero-shot text-to-shape generation. In: CVPR (2022)

52. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
53. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
54. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
55. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. NeurIPS (2019)
56. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. arXiv preprint arXiv:2303.14184 (2023)
57. Tang, Z., Gu, S., Wang, C., Zhang, T., Bao, J., Chen, D., Guo, B.: Volumediffusion: Flexible text-to-3d generation with efficient volumetric encoder. arXiv preprint arXiv:2312.11459 (2023)
58. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
59. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. arXiv preprint arXiv:2401.06209 (2024)
60. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
61. Wang, Y., Xian, Z., Chen, F., Wang, T.H., Wang, Y., Fragkiadaki, K., Erickson, Z., Held, D., Gan, C.: Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. arXiv preprint arXiv:2311.01455 (2023)
62. Wei, J., Wang, H., Feng, J., Lin, G., Yap, K.H.: Taps3d: Text-guided 3d textured shape generation from pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16805–16815 (2023)
63. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. arXiv preprint arXiv:2308.16911 (2023)
64. Xu, Y., Tan, H., Luan, F., Bi, S., Wang, P., Li, J., Shi, Z., Sunkavalli, K., Wetzstein, G., Xu, Z., et al.: Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. arXiv preprint arXiv:2311.09217 (2023)
65. Yariv, L., Puny, O., Neverova, N., Gafni, O., Lipman, Y.: Mosaic-sdf for 3d generative models. arXiv preprint arXiv:2312.09222 (2023)
66. Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. arXiv preprint arXiv:2310.08529 (2023)
67. Zhang, J., Herrmann, C., Hur, J., Polania Cabrera, L., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. Advances in Neural Information Processing Systems **36** (2024)
68. Zhao, M., Zhao, C., Liang, X., Li, L., Zhao, Z., Hu, Z., Fan, C., Yu, X.: Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion prior. arXiv preprint arXiv:2308.13223 (2023)
69. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. arXiv preprint arXiv:2303.02153 (2023)

70. Zhou, Q., Yu, C., Zhang, S., Wu, S., Wang, Z., Wang, F.: Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. arXiv preprint arXiv:2308.02299 (2023)
71. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
72. Zhu, J., Zhuang, P.: Hifa: High-fidelity text-to-3d with advanced diffusion guidance. arXiv preprint arXiv:2305.18766 (2023)