Table of Contents

A Broader Impact

By enhancing the accuracy and richness of captions for 3D objects, this work facilitates advancements in 3D modeling and prompote related applications in educational tools, interactive learning environments, and assistive technologies, making digital content more accessible and informative. Moreover, by addressing inaccuracies and hallucinations in captions which could be used in AI-content generations, our work underscores the pursuit of more reliable and trustworthy AI systems. During the process, we undertaken with a commitment to ethical considerations to filter out potential ethical issued 3D objects. We recognize the wide-reaching effects of our work on society and maintain that it chiefly offers positive contributions towards the progress of generative modeling and its implementation in diverse fields.

B Dataset: more details & results

B.1 Extra dataset details

In Section 4, we addressed approximately $200k$ caption corrections for the Cap3D dataset, significantly reducing its hallucinations. Our efforts also expand the dataset to include over 1 million 3D-text pairs, encapsulating the entirety of the Objaverse [13] and portions of the Objaverse-XL high-quality set [12]. The objects with updated captions are cataloged in a CSV file within the supplementary material, accessible via “uid" or cryptographic hash values ("sha256"). These identifiers correspond to the ones provided in the Objaverse and Objaverse-XL datasets.

As mentioned in the Introduction, we are excited to also provide access to rendered images associated with each object. These images include detailed camera information (both intrinsic fov and extrinsic RT matrix), depth map, and

MatAlpha, in addition to point clouds that complement the textual captions. Alongside these resources, we are releasing the source code for our DiffuRank methodology, which facilitates the replication of our findings. The distribution also includes pre-trained models, further aiding in the exploration and utilization of our dataset. This comprehensive package aims to empower researchers in our community. They will be released under ODC-By 1.0 license.

Our GPT4-Vision prompt is defined as "Renderings show different angles of the same set of 3D objects. Concisely describe 3D object (distinct features, objects, structures, material, color, etc) as a caption" accompanied by six image tokens. On average, the context encompasses approximately 1,867 tokens, while the average number of tokens generated stands at approximately 26.72. Notably, we employed the "GPT-4-1106-vision-preview" model for this study.

As described in Section 3.3, given a 3D object, we generate 28 views using two distinct rendering methods [20, 35]. For each view, we generate 5 captions with BLIP2. Subsequently, we apply the DiffuRank algorithm (Algorithm 1) to evaluate the alignment of the 28 renderings relative to the input 3D object by doing inference ovew 140 captions and the 3D object. Ultimately, we select the best 6 views for further caption generation using GPT4-Vision.

For the ray-tracking render engine, we used Blender render engine 'CYCLES' with samples 16. Additionally, we adopted 'OPTIX' denoiser for the cycle engine. For the real-time render engine, we used Blender render engine 'EEVEE' with 'taa_render_samples' 1.

B.2 Captions: overcome the failure cases in Cap3D

The Cap3D captions we used to compare throughout the whole paper are from their dataset page. Specically, the version described in their paper.

Here, we provide direction comparisons with the failure cases mentioned in their paper "Limitations and Failure Cases". Our captions have obviously eliminated lots of hallucinations, such as 'butterfly' and 'flowers' in Figure 9, and 'dump truck' in Figure 10.

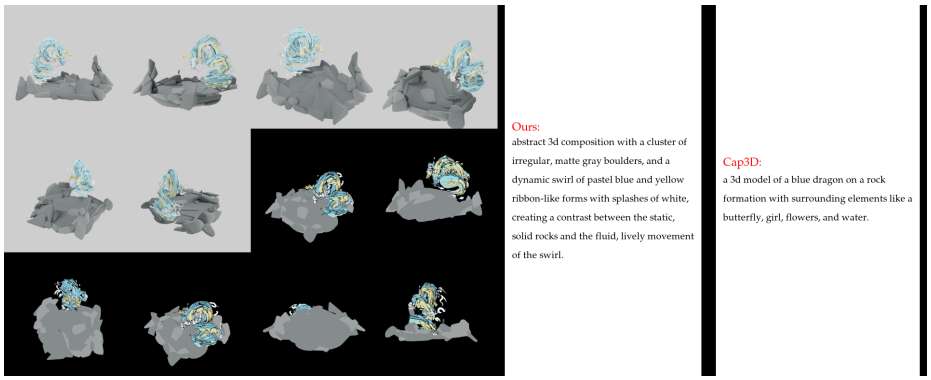

Fig. 9: Comparisons between our captions and Cap3D captions.

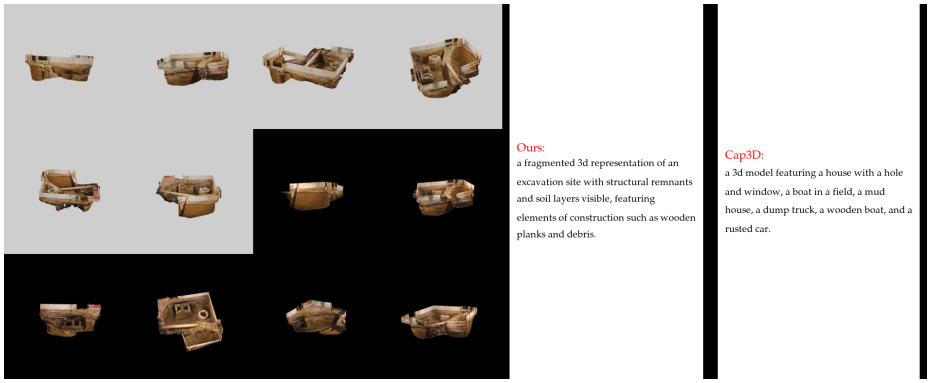

Fig. 10: Comparisons between our captions with Cap3D captions.

B.3 Captions: Ours vs. Cap3D vs. human-authored

We present a variety of qualitative comparisons: those generated by our model, those produced by Cap3D, and captions written by humans, all of which were selected through random sampling. The below qualitative results show the captions generated by our method usually contain more details and less hallucinations.

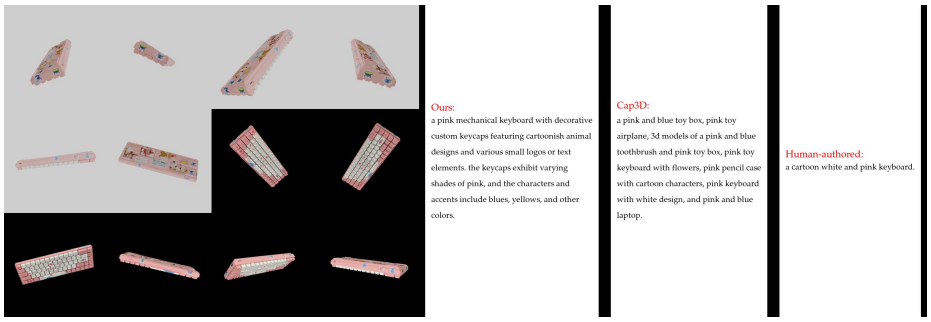

Fig. 11: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

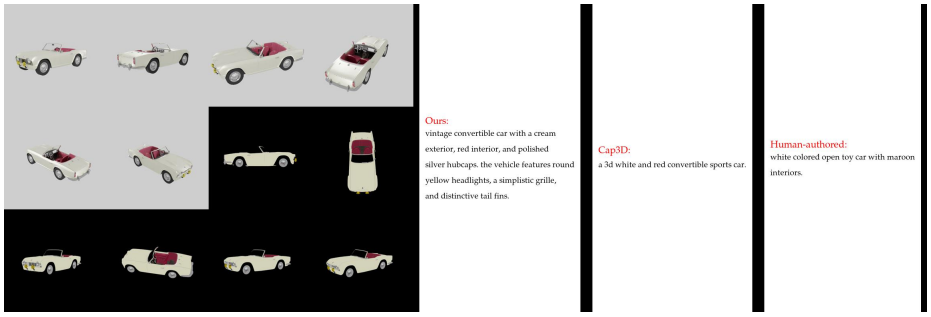

Fig. 12: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

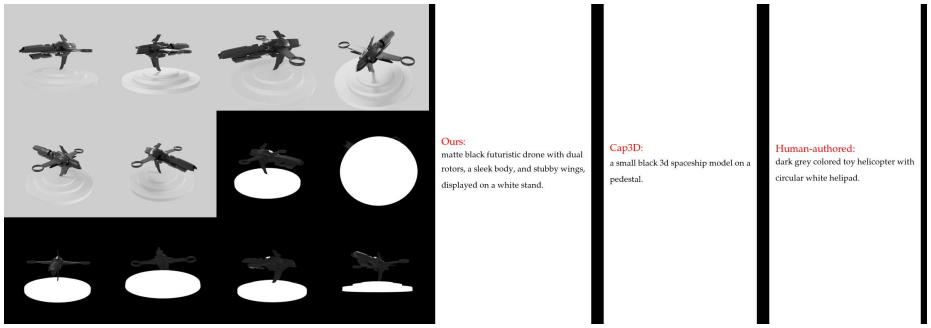

Fig. 13: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

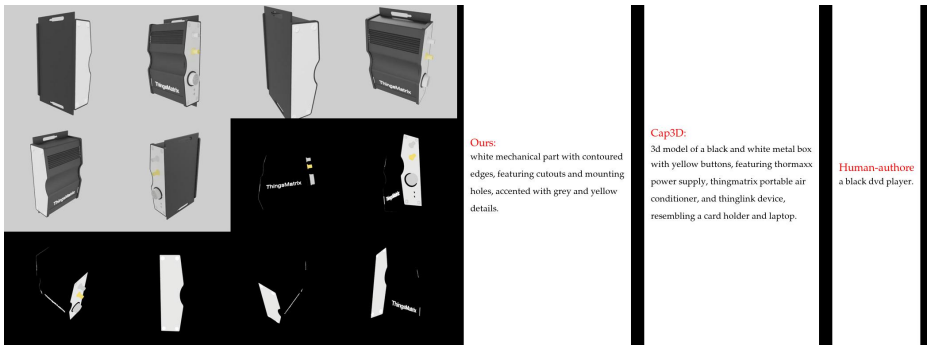

Fig. 14: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

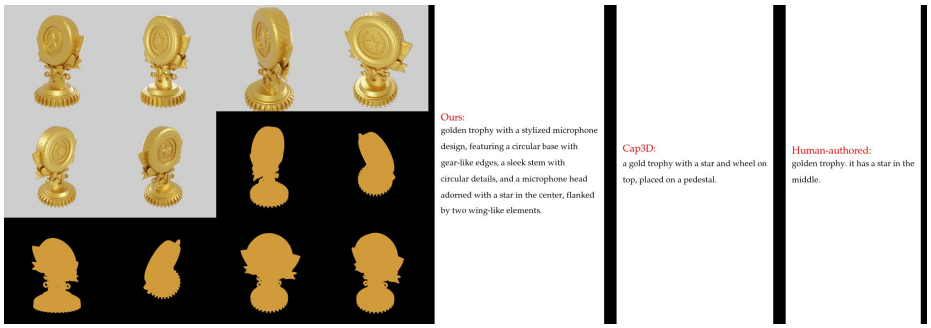

Fig. 15: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

Fig. 16: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

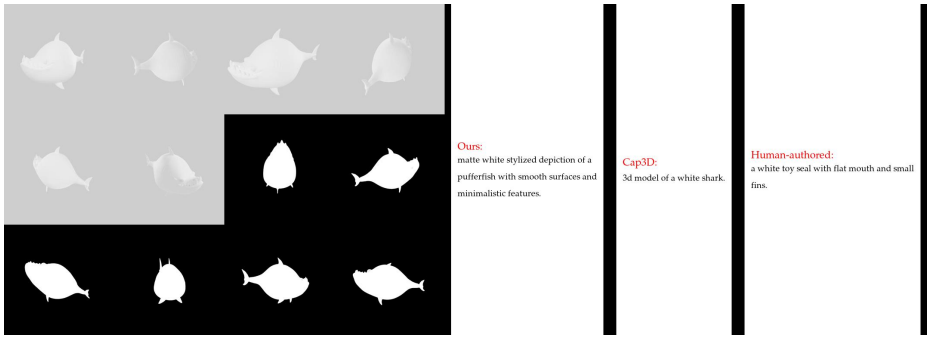

Fig. 17: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

Fig. 18: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

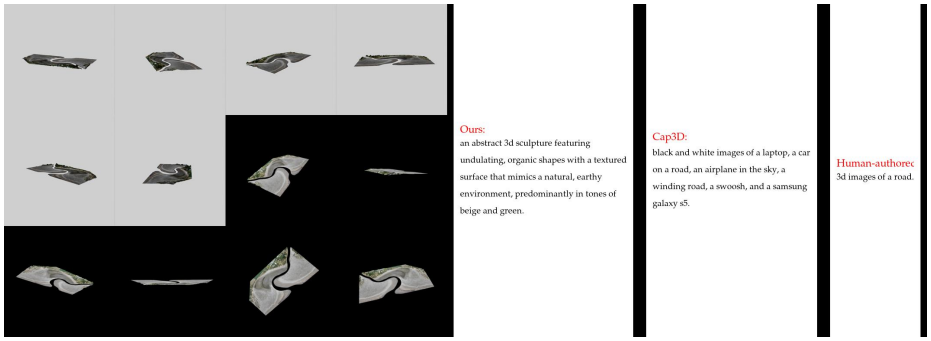

Fig. 19: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

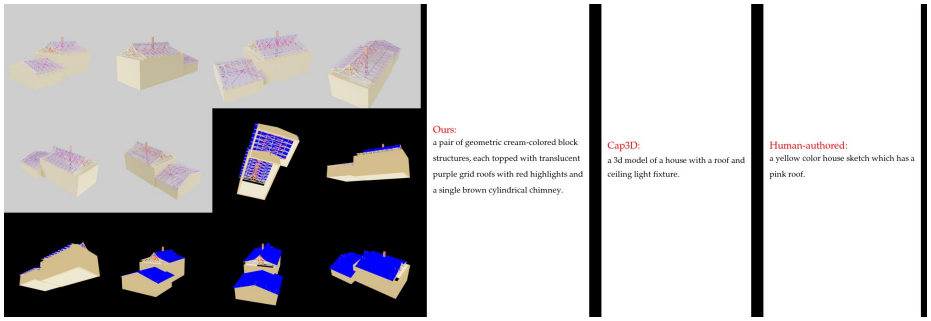

Fig. 20: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

Fig. 21: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

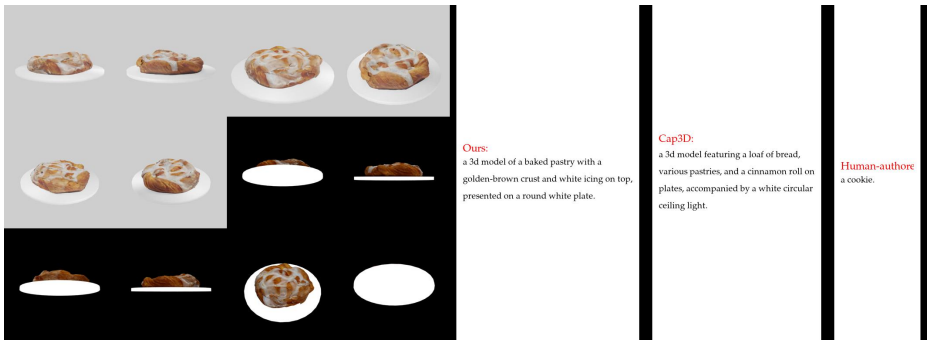

Fig. 22: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

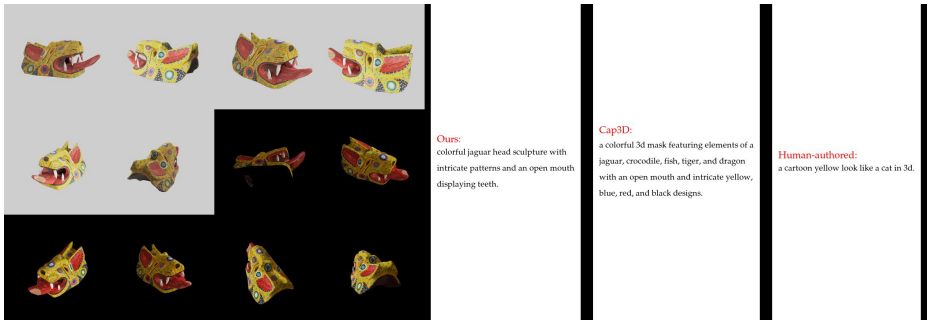

Fig. 23: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

Fig. 24: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

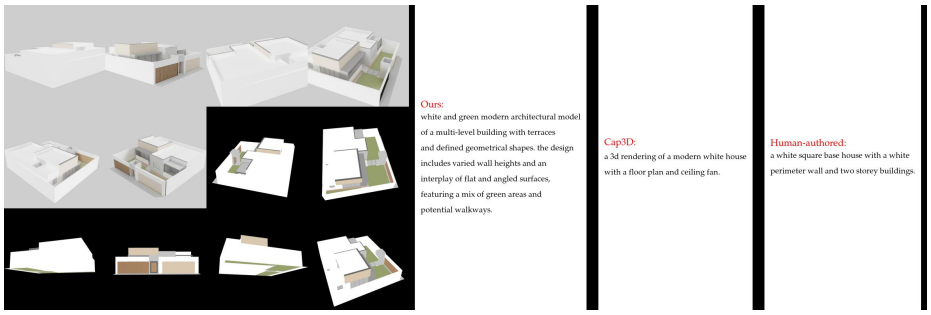

Fig. 25: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

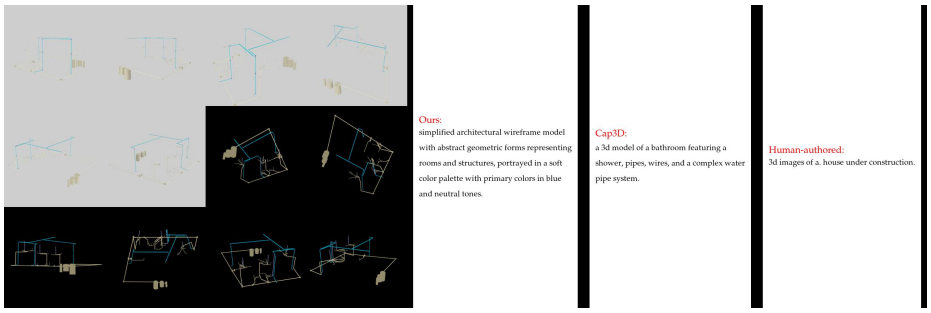

Fig. 26: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

Fig. 27: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

Fig. 28: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

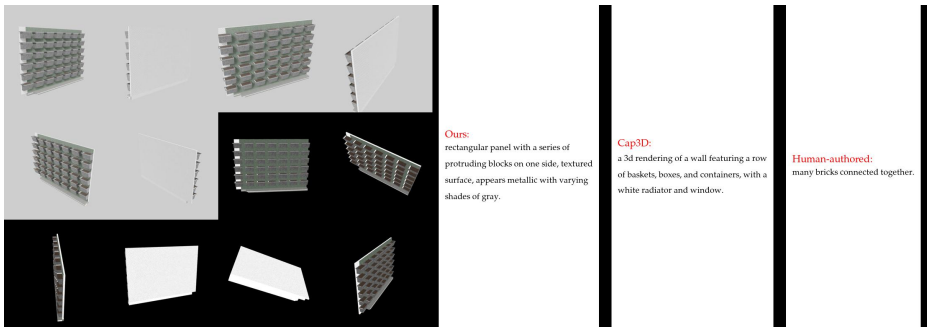

Fig. 29: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

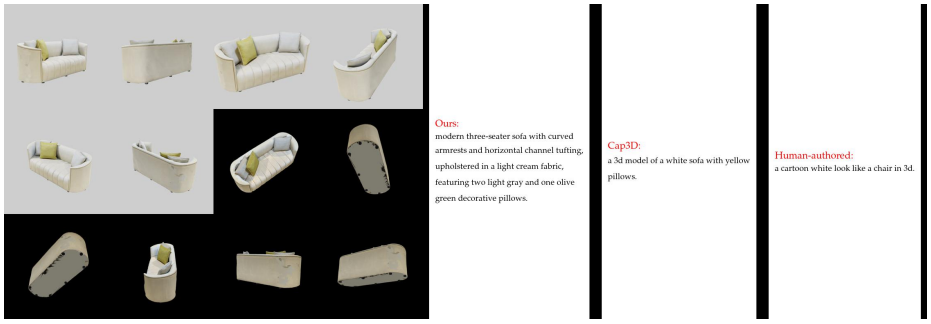

Fig. 30: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

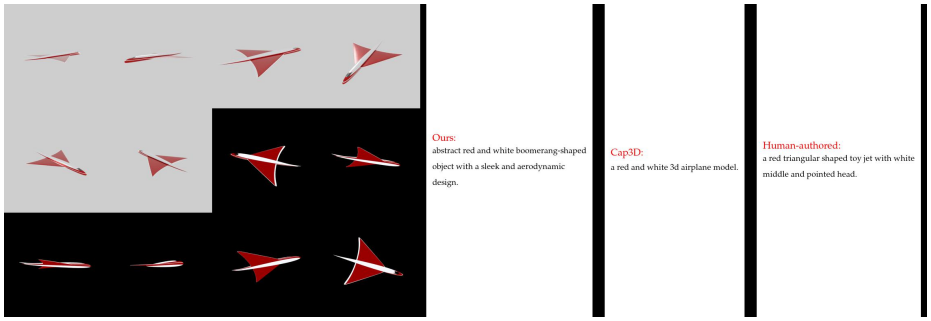

Fig. 31: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

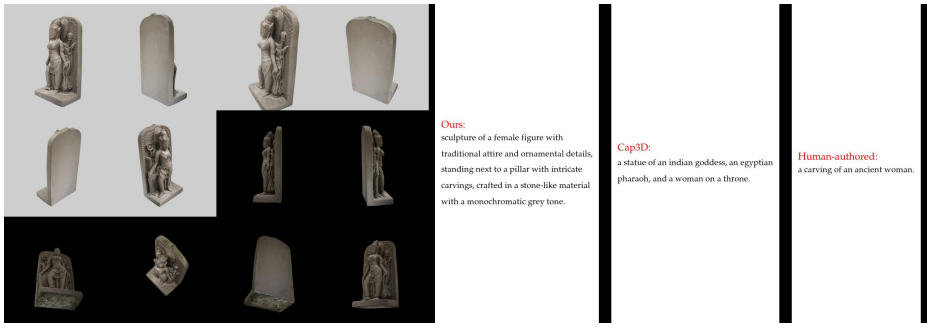

Fig. 32: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

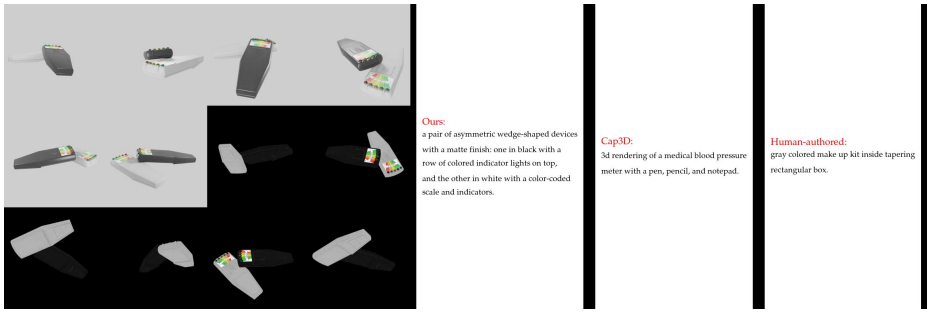

Fig. 33: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

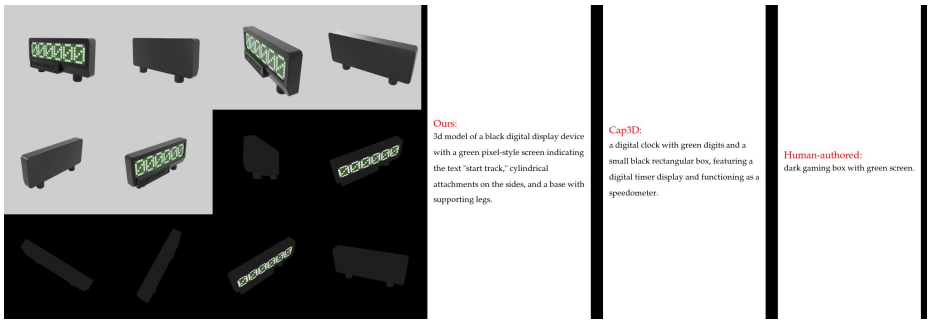

Fig. 34: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

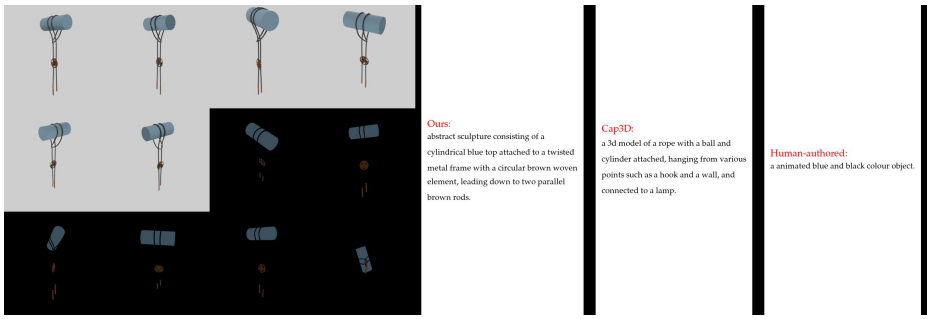

Fig. 35: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

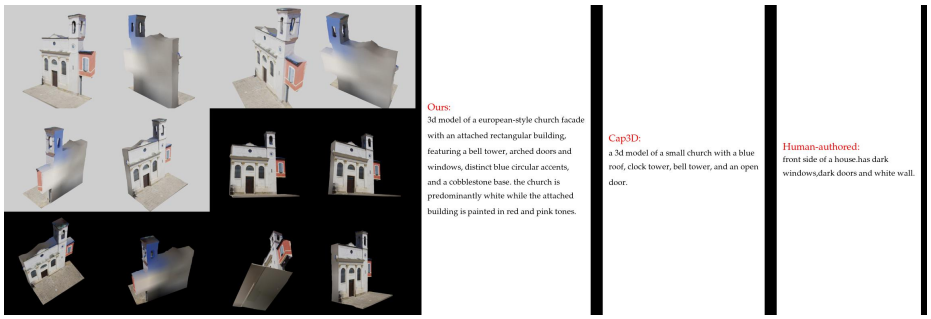

Fig. 36: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

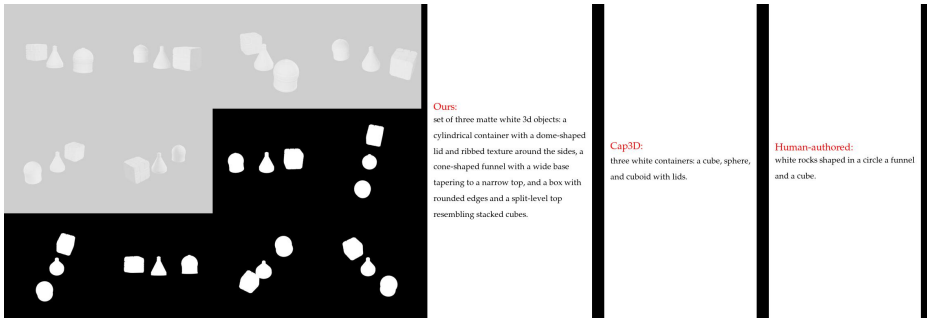

Fig. 37: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

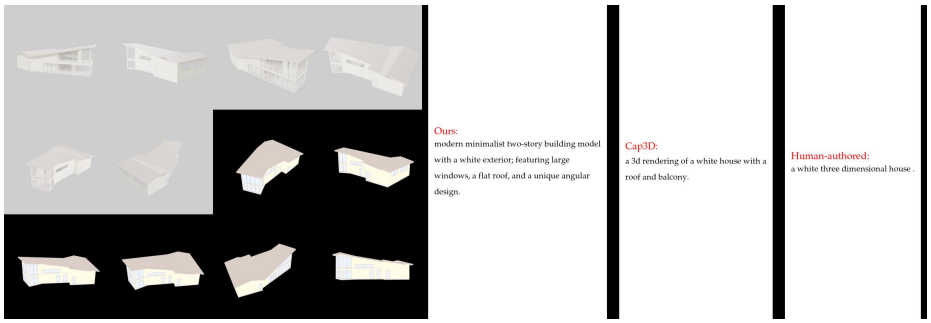

Fig. 38: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

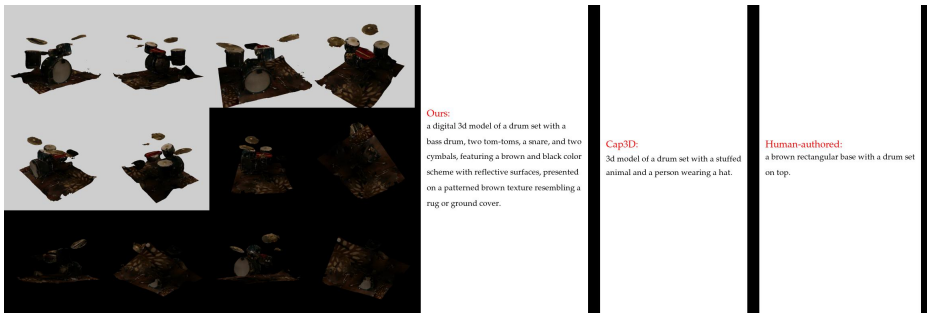

Fig. 39: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

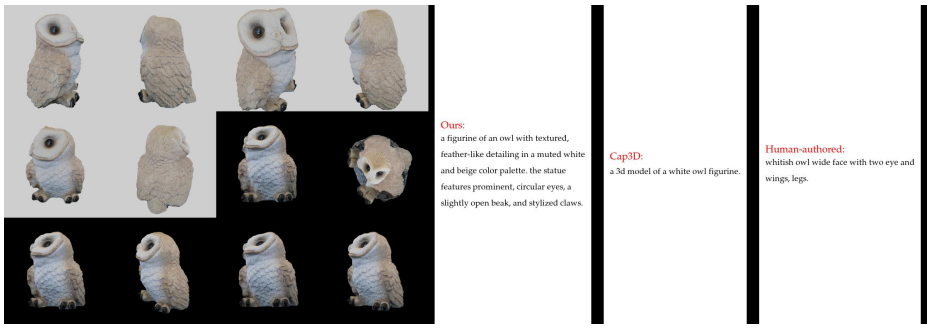

Fig. 40: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

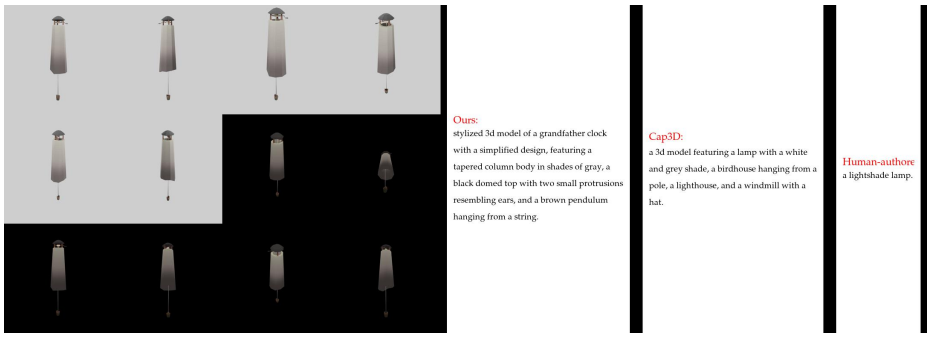

Fig. 41: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

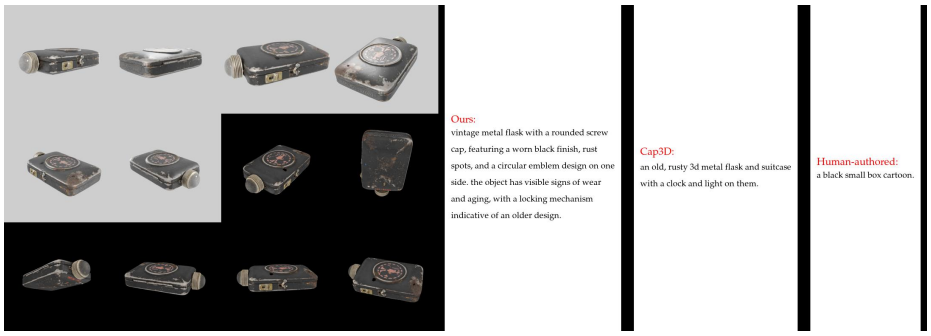

Fig. 42: We compare captions through random sampling, including those generated by our method, by Cap3D, and those authored by humans.

B.4 Captions: Ours vs. ablated variants

We list several qualitative comparisons here to demonstrate the effectiveness of our method compared to (1) **Bottom 6-views**, we employ the 6 renderings identified as having the lowest alignment scores, as determined by our DiffuRank algorithm (refer to Alg. 1); (2) **Allviews 28-views**, which involves utilizing all 28 rendered views as inputs for the GPT4-Vision; and (3) **Horizontal 6-views**, this configuration involves selecting 6 rendered views that position the camera horizontally relative to the object's default orientation, adhering to the same vertical positioning guidelines used by Cap3D. Results generally show the captions generated by our method (i.e., Top-6) contain more accurate, detailed, and less hallucinated information.

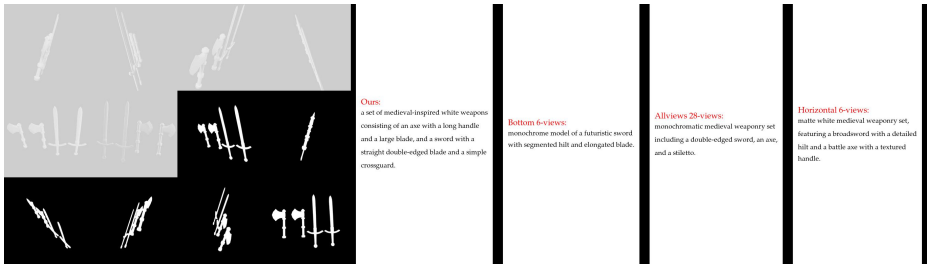

Fig. 43: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

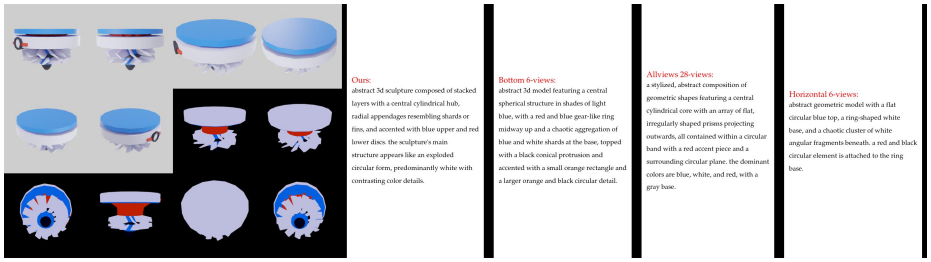

Fig. 44: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

Fig. 45: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

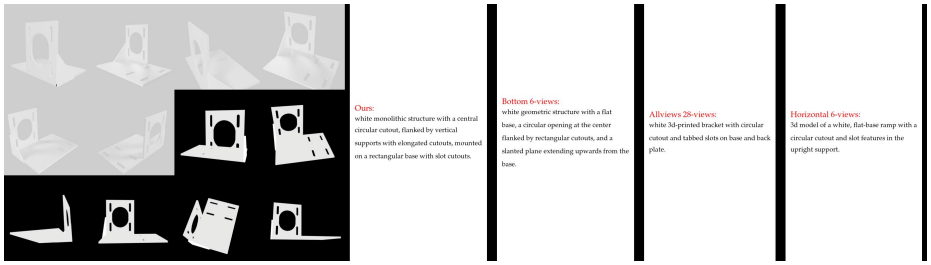

Fig. 46: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

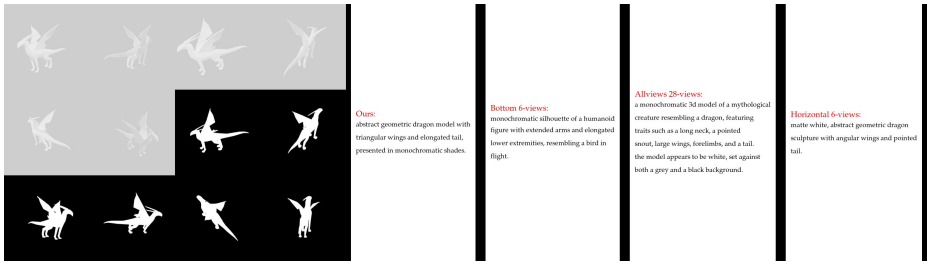

Fig. 47: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

Fig. 48: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

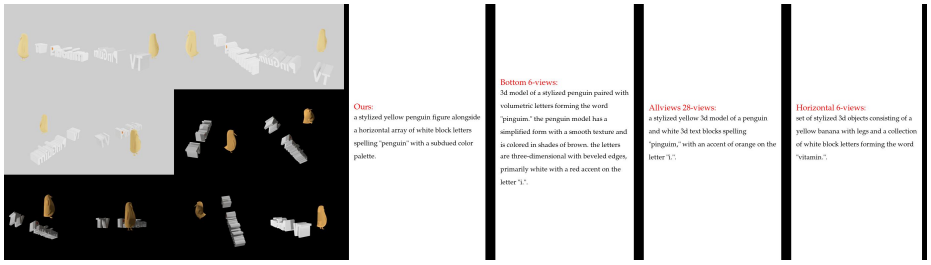

Fig. 49: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

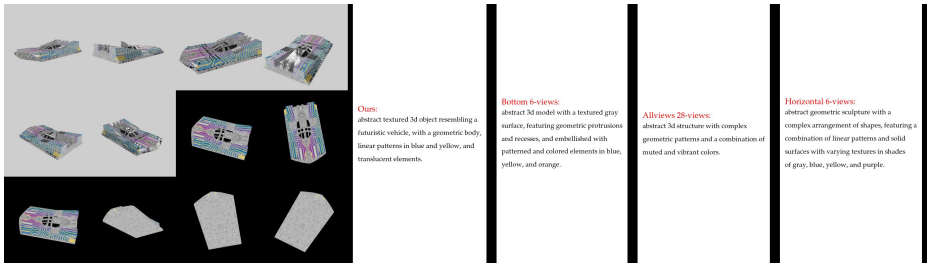

Fig. 50: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

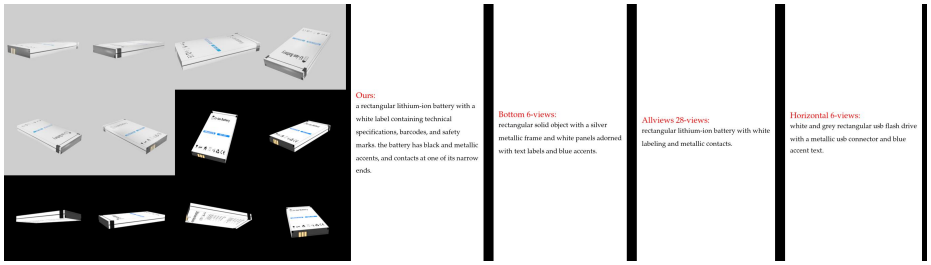

Fig. 51: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

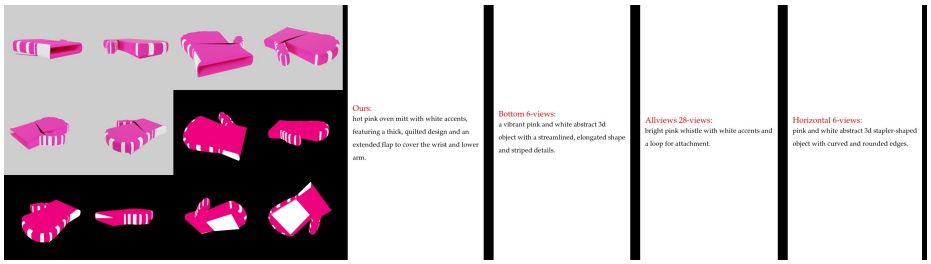

Fig. 52: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

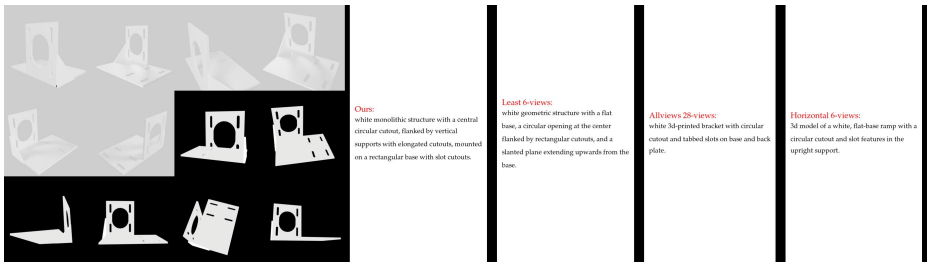

Fig. 53: We evaluate captions by randomly sampling and comparing them across different methods: our approach (Top 6-views), using the bottom 6-views, utilizing all 28-views, and employing horizontal 6-views.

B.5 Diffu: Ours vs. bottom 6-views vs. horizontal 6-views

This section lists several randomly sampled DiffuRank results of Top 6-views with 6 highest alignment scores (our method), bottom 6-views, and horizontal 6-views. According to the results, we can see (1) Top 6-views obviously outperforms Bottom 6-views on Figures 57, 58, 59, 61, 66, 67, 68, 71; (2) Compared to Horizontal 6-views, DiffuRank can adaptly choose angles and types of rendering as shown in Figures 55, 70, 72; (3) in some cases (Figures 54, 63, 65), there are no significant difference.

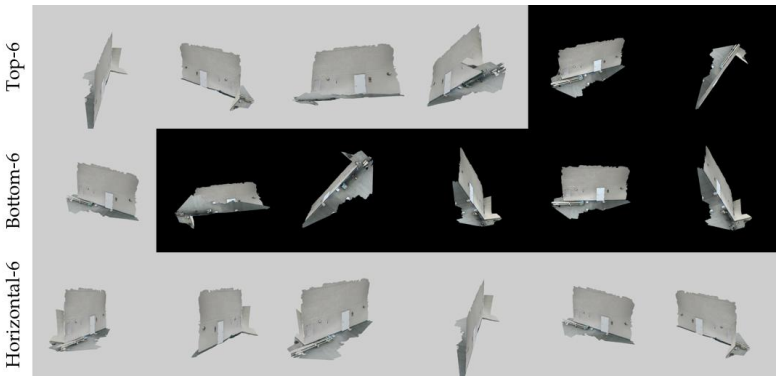

Fig. 54: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

Fig. 55: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

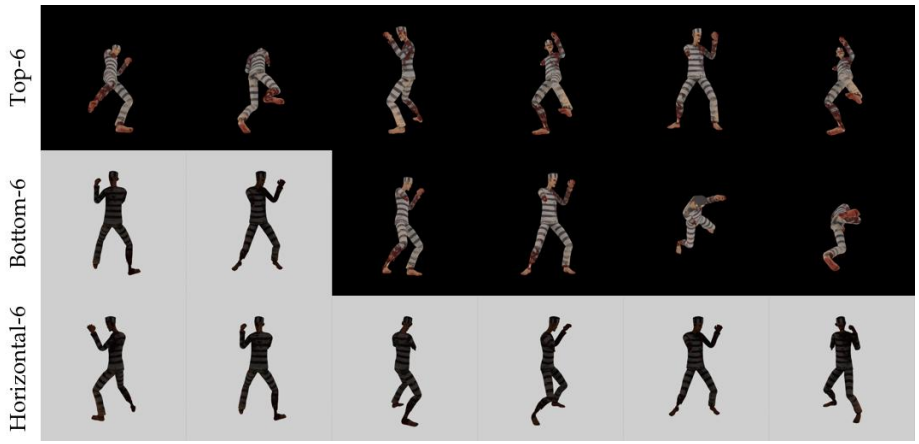

Fig. 56: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

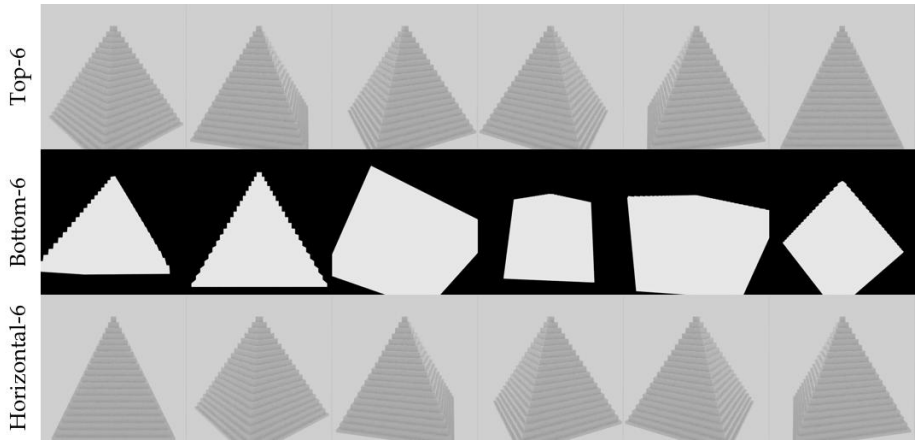

Fig. 57: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

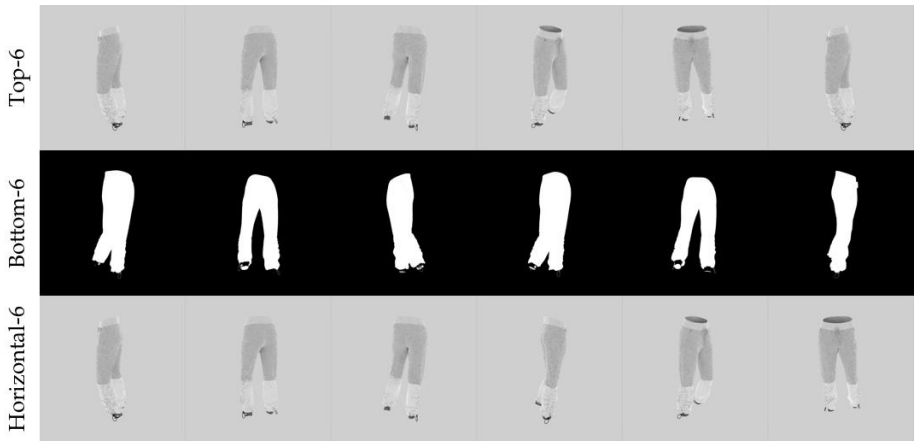

Fig. 58: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

Fig. 59: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

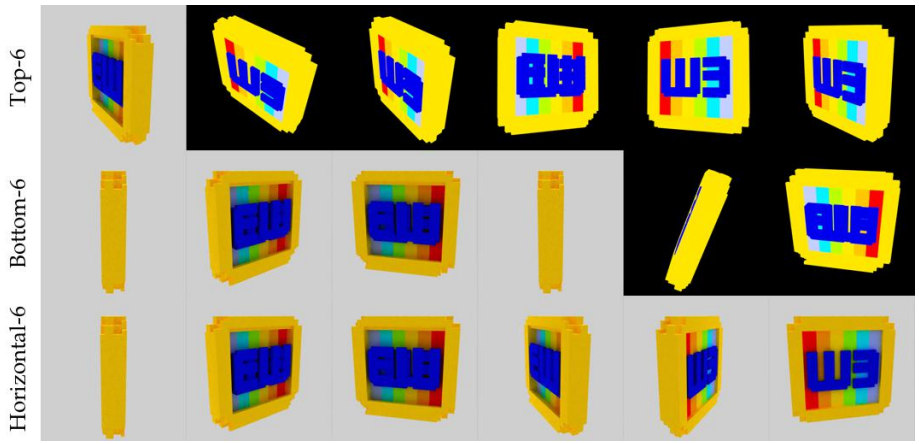

Fig. 60: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

Fig. 61: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

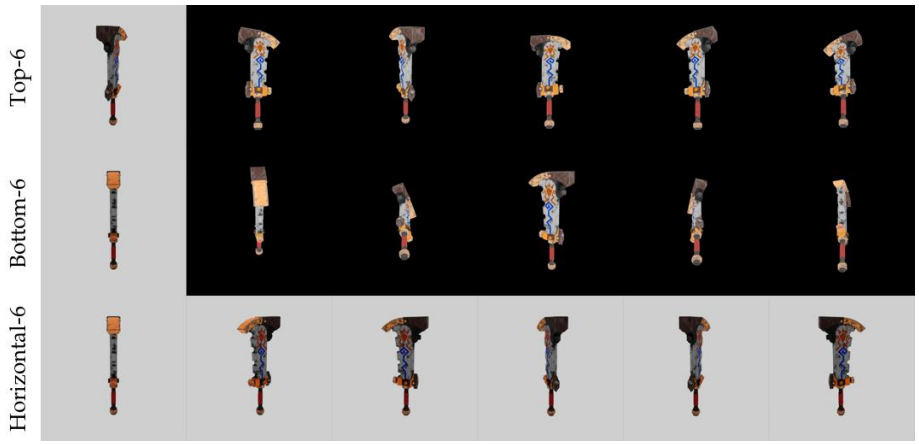

Fig. 62: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

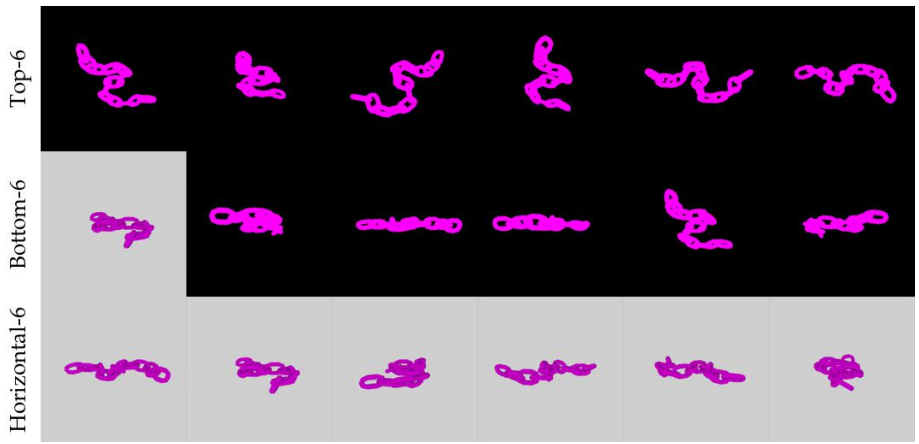

Fig. 63: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

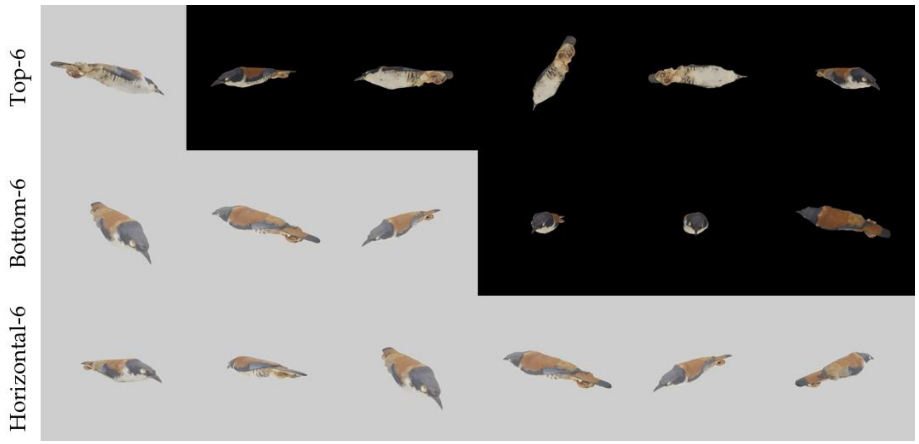

Fig. 64: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

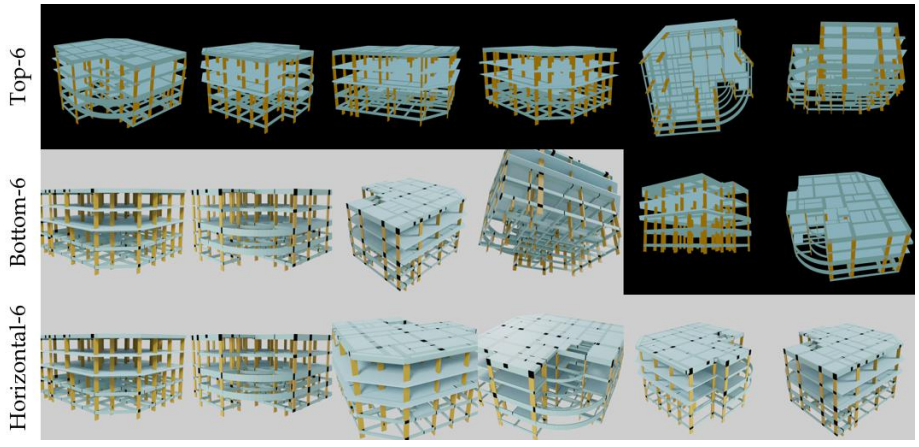

Fig. 65: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

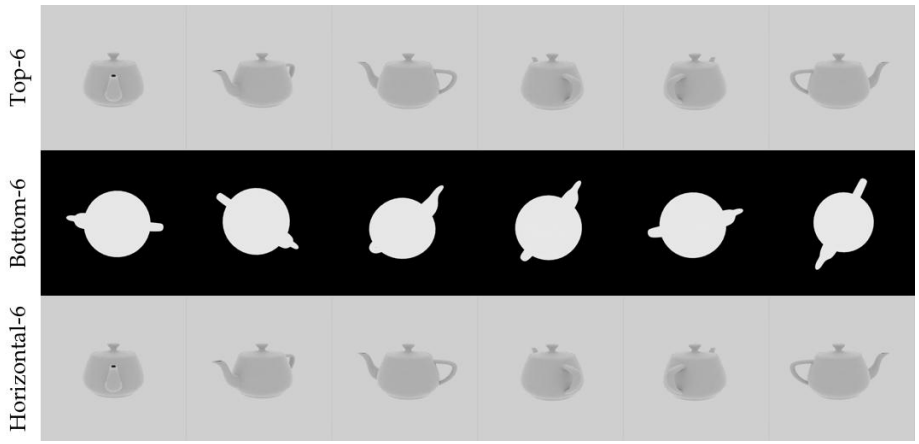

Fig. 66: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

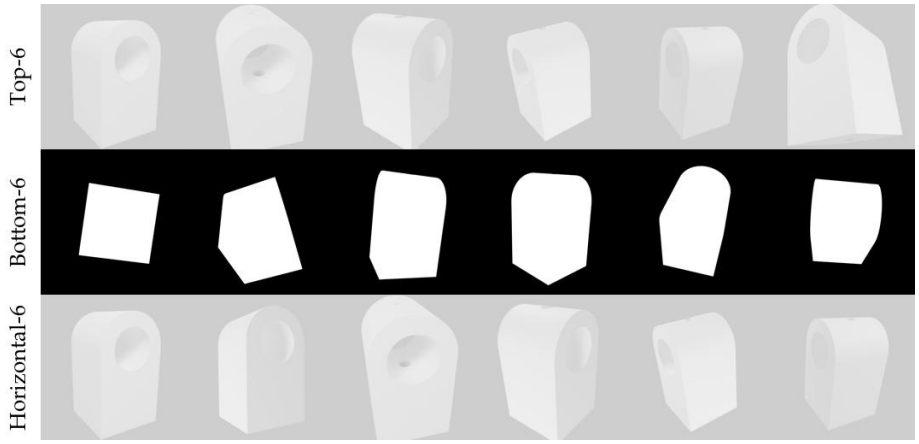

Fig. 67: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

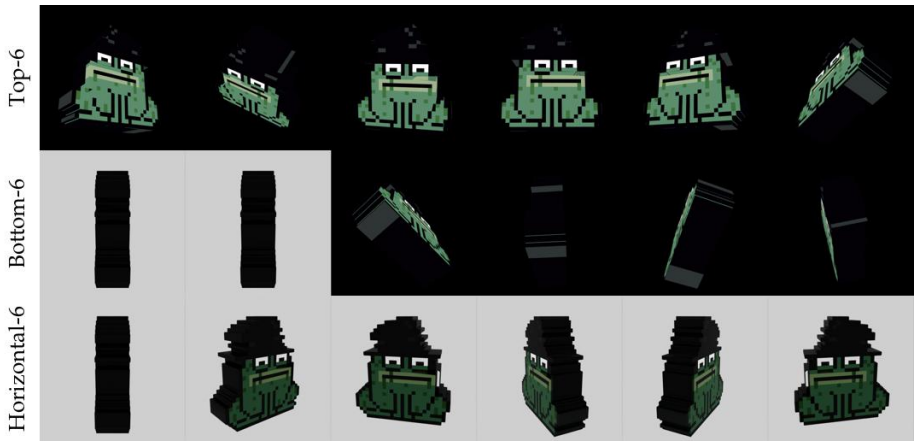

Fig. 68: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

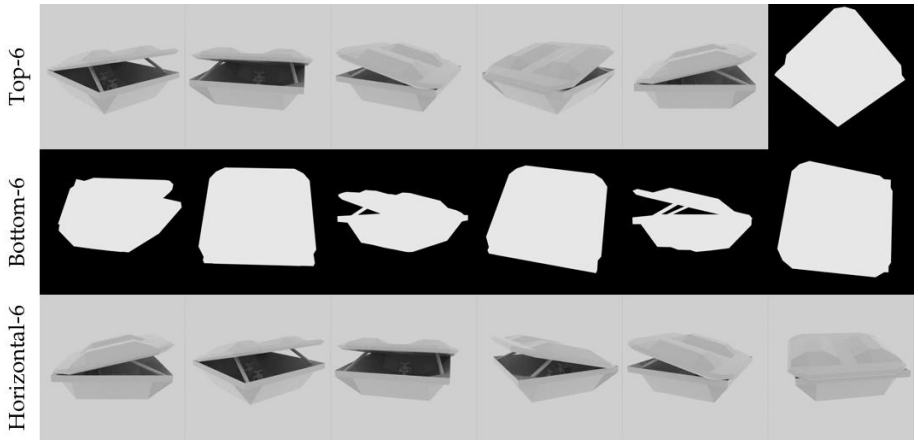

Fig. 69: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

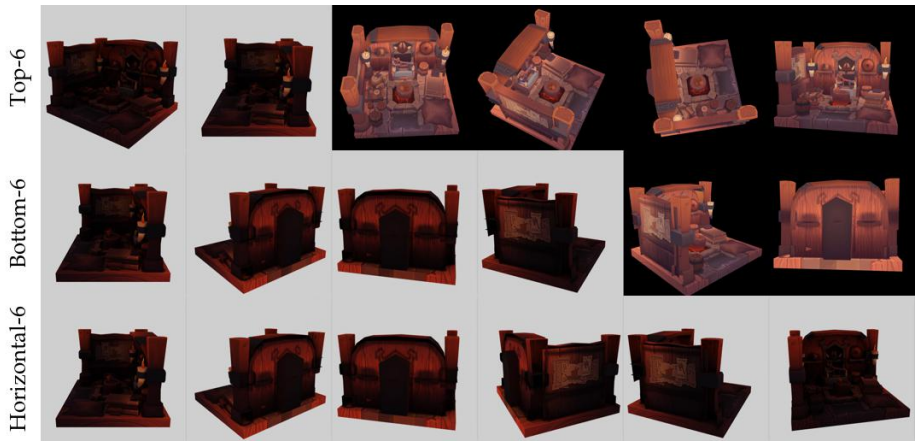

Fig. 70: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

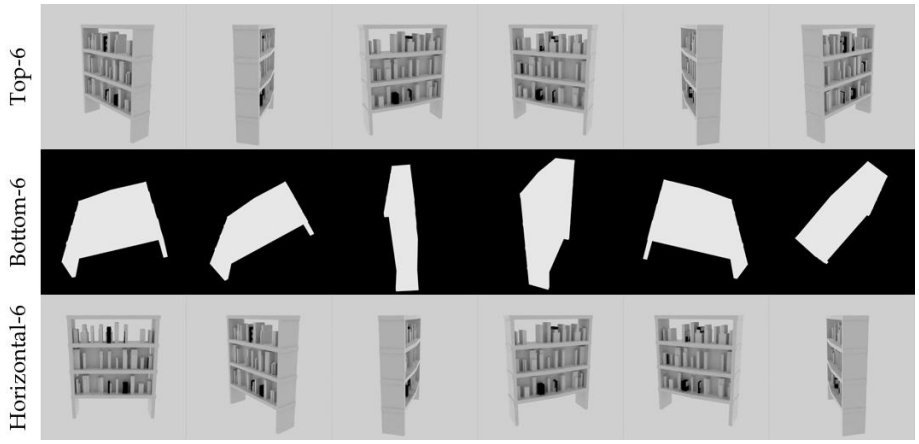

Fig. 71: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

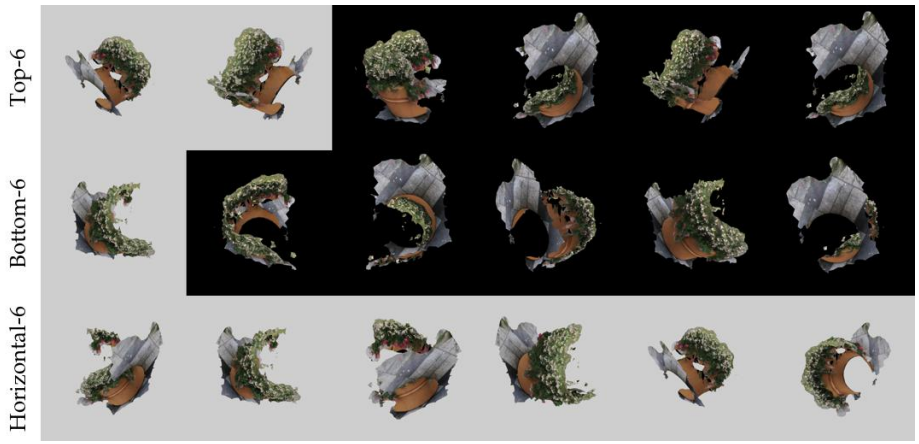

Fig. 72: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

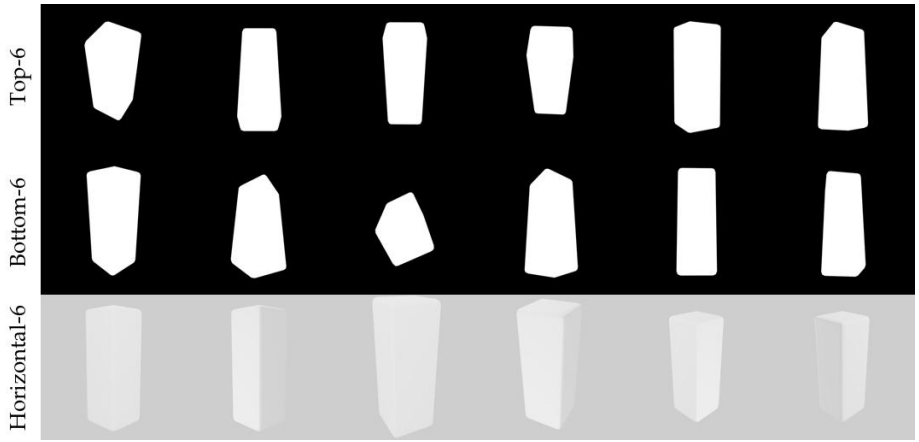

Fig. 73: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

Fig. 74: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

Fig. 75: Randomly sampled DiffuRank comparisons. **Top-row**: Top 6-views selected by DiffuRank; **Middle-row**: Bottom 6-views selected by DiffuRank; **Bottom-row**: Horizontal 6-views.

B.6 Failure cases

We have observed three types of failure cases: (1) DiffuRank fails due to BLIP2 captioning fails or alignment compute not accurate. As shown in Figure 77, where BLIP2 captions contain a lot of “a tree in the dark". Since our DiffuRank needs the initial captioning results to compute alignment scores, BLIP2 captioning fails will cause rendering selection poorly and further cause final caption inaccurate. This could be solved via stronger captioning model, such as GPT4-Vision. Also, as mentioned in Future work (Appendix E), with better captions, we can fine-tune stronger Text-to-3D models, which help to obtain more accurate alignment scores. (2) sometimes, our captioning method fails to capture small object. One example is in Figure 9, where there is a small black person above the rock, while the caption fails to describe it. Also, it may contain hallucinations with small chances (according to our eyeballs over 10k captions) as shown in Figure 76. (3) for some scene renderings, the model failed to capture meaningful characteristics for Figure 78 with caption “Abstract 3D composition with fragmented, textured surfaces in shades of beige, white, and charcoal". However, human may also not distinguish this kind of renderings.

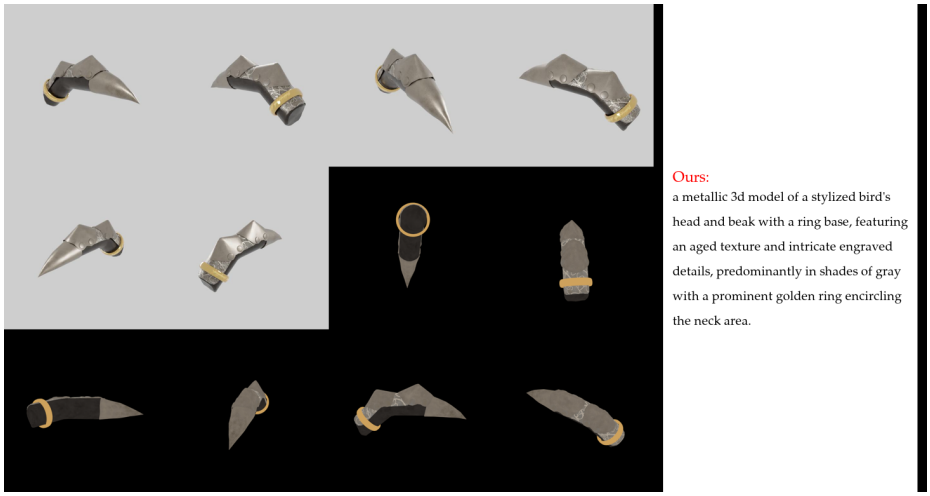

Fig. 76: Failure cases: hallucination.

B.7 Human evaluation details

We utilize the Hive platform for conducting crowdsourced A/B testing. In this process, participants are presented with an image accompanied by two different captions as shown in Figure 79. They are asked to judge which caption is more suitable based on a 5-point scale, where a score of 3 indicates neither caption

Fig. 77: Failure cases: BLIP2 captioning fails or alignment compute not accurate.

is preferred over the other. Scores of 1 and 2 suggest a preference for the left caption, with 1 indicating a strong preference and 2 a moderate preference. The sequence in which the captions are presented (left or right) is varied randomly in each case.

Participants receive guidelines on how to perform this task, including examples that set the standard for quality. We have two distinct types of tasks as shown in Table 1: quality and hallucination. For quality tasks, workers are advised to focus first on the accuracy of their choices, followed by the level of detail provided in terms of type, structure, and appearance. For hallucination tasks, workers are advised to focus on if the caption contain hallucination or false information.

We totally hired 46 workers from Hive without access to their personally identifiable information. They are paid approximately $35 per 1k tasks for our caption evaluation tasks. The entire procedure was carried out in compliance with the ECCV ethics guidelines.

The platform automatically excludes workers who fail to meet the required standards on essential test examples set by us. However, our review revealed that some workers managed to meet the criteria for these essential examples but engaged in deceitful practices for the rest. The prevalent forms of deceit included consistently choosing the same option (always choose left or right) or selecting captions based on their length, either the shortest or the longest. Consequently, we conducted a thorough examination of all workers and excluded those found to be engaging in these deceptive practices, also disregarding their evaluations.

C Text-to-3D: more details & results

In this section, we provide a detailed examination of our Text-to-3D experiments, along with a comprehensive set of qualitative comparisons. It is important to

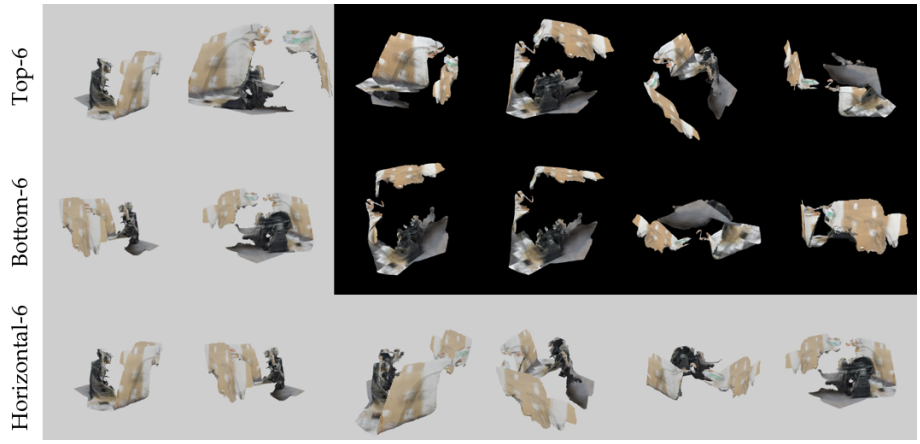

Fig. 78: Failure cases: for some scene renderings, our framework fails to capture meaningful characteristics.

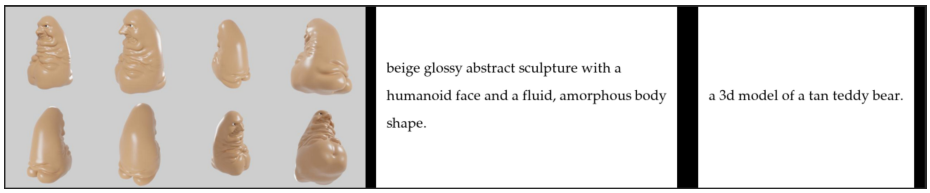

Fig. 79: Example hive case. Caption are from ours and Cap3D.

note that employing captions generated by our method typically enhances the performance of Shap·E pre-trained models, a trend that is clearly supported by the data presented in Table 2. However, when we fine-tune the Shap·E pre-trained model using Cap3D, we observe a decline in performance across all CLIP-based metrics.

C.1 Setting

We adopted the same fine-tune strategy used in Cap3D [35] for fair comparisons. We employed the AdamW optimizer alongside the CosineAnnealingLR scheduler, setting the initial learning rate at $1e-5$ for fine-tuning both the Point·E and Shap·E models. The batch sizes were set to 64 for Shap·E and 256 for Point·E. For training epochs, we set the training epoch which would cost approximately three days. The training was performed on four A40 GPUs.

The evaluation times, measured in seconds per iteration and inclusive of rendering, are as follows:

– For Point·E, the total time is 37 seconds, with 28 seconds dedicated to text-to-3D processing and 9 seconds to rendering.

– Shap·E (stf) requires 16 seconds in total for both text-to-3D processing and rendering.
– Shap·E (NeRF) takes significantly longer, with a total of 193 seconds for both text-to-3D processing and rendering.

C.2 Qualitative comparisons

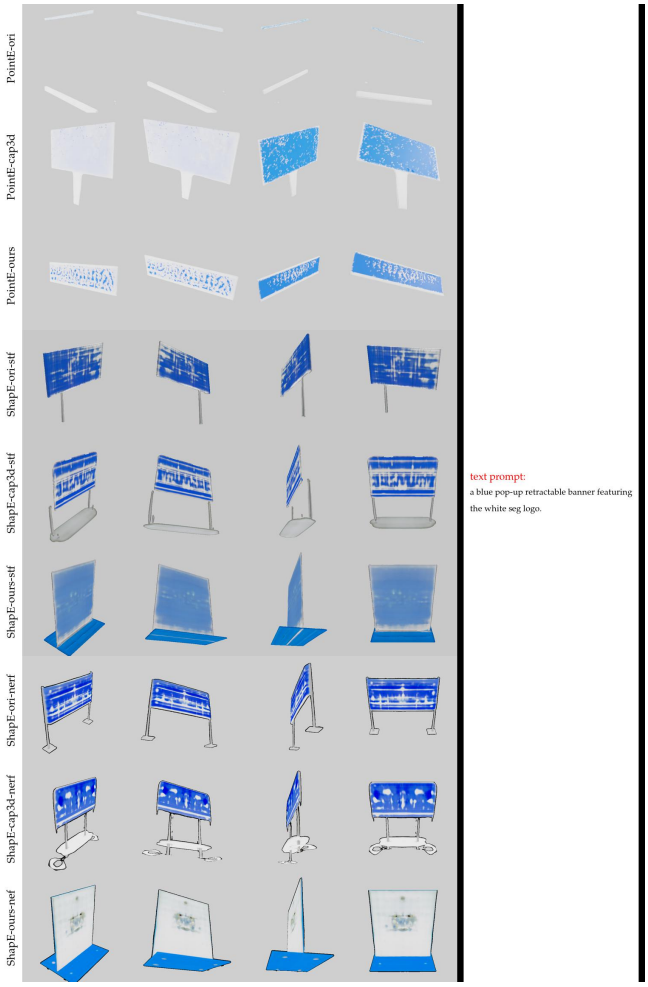

Fig. 80: Randomly sampled Text-to-3D results.

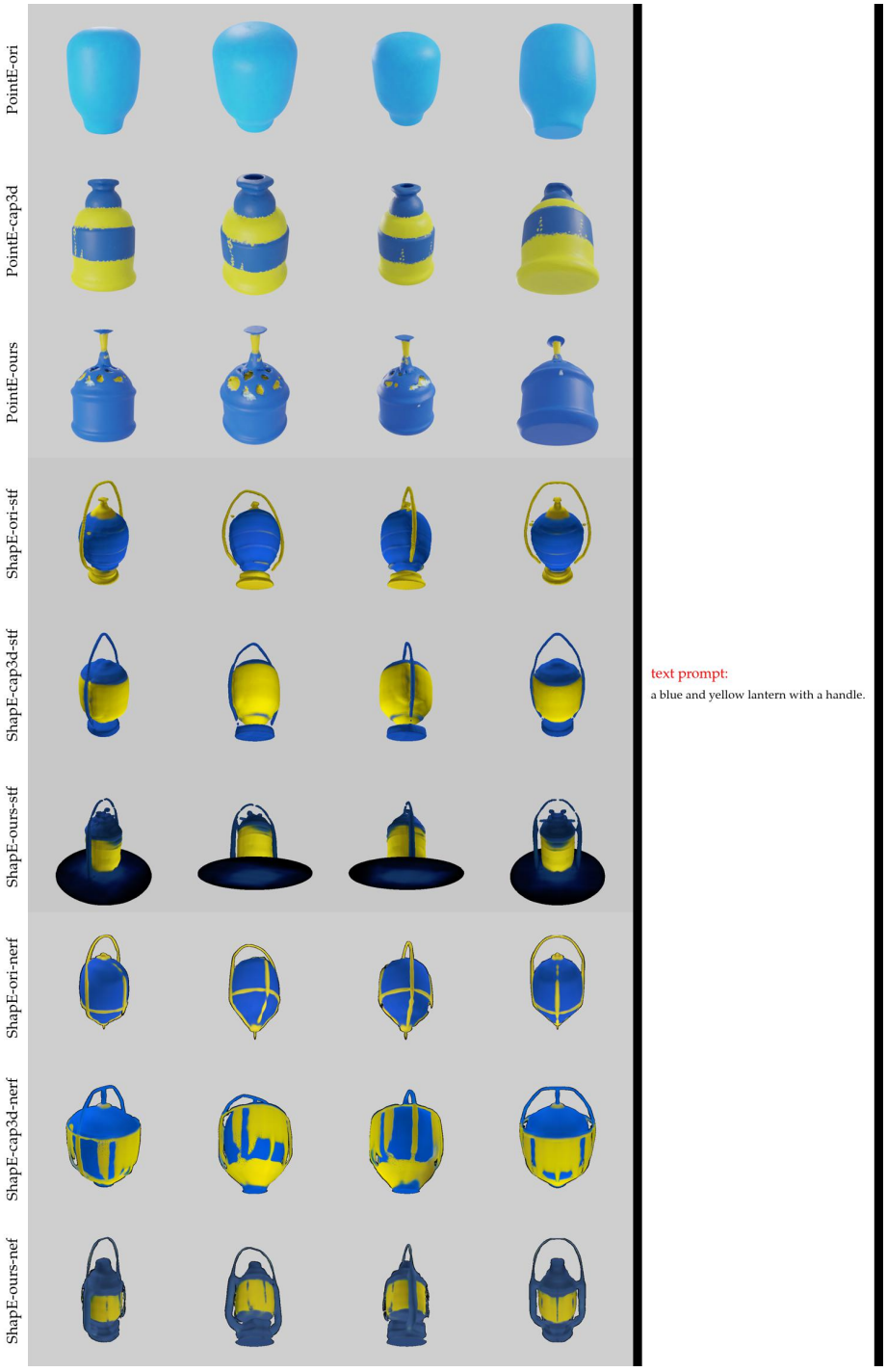

Fig. 81: Randomly sampled Text-to-3D results.

Fig. 82: Randomly sampled Text-to-3D results.

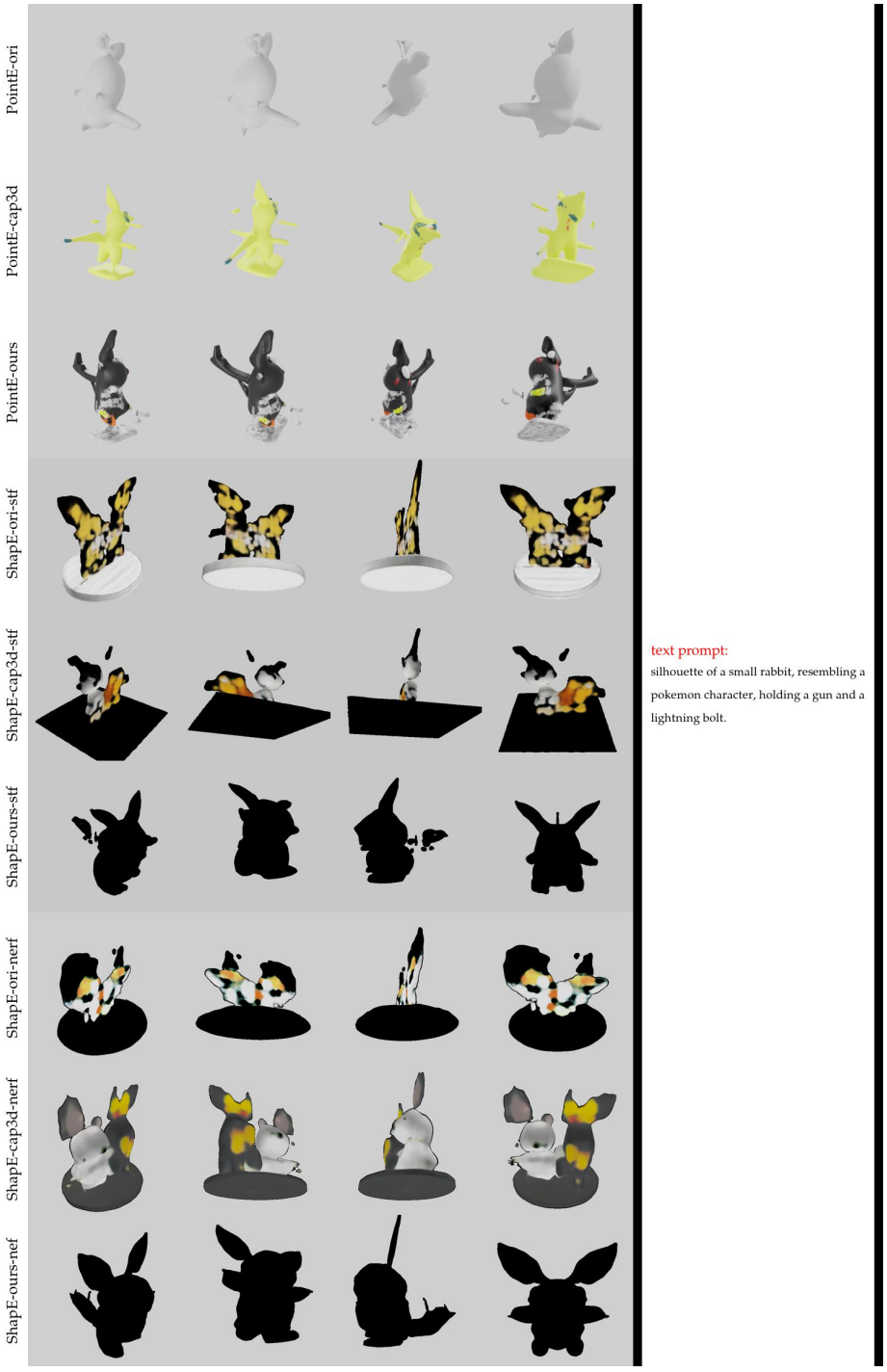

Fig. 83: Randomly sampled Text-to-3D results.

Fig. 84: Randomly sampled Text-to-3D results.

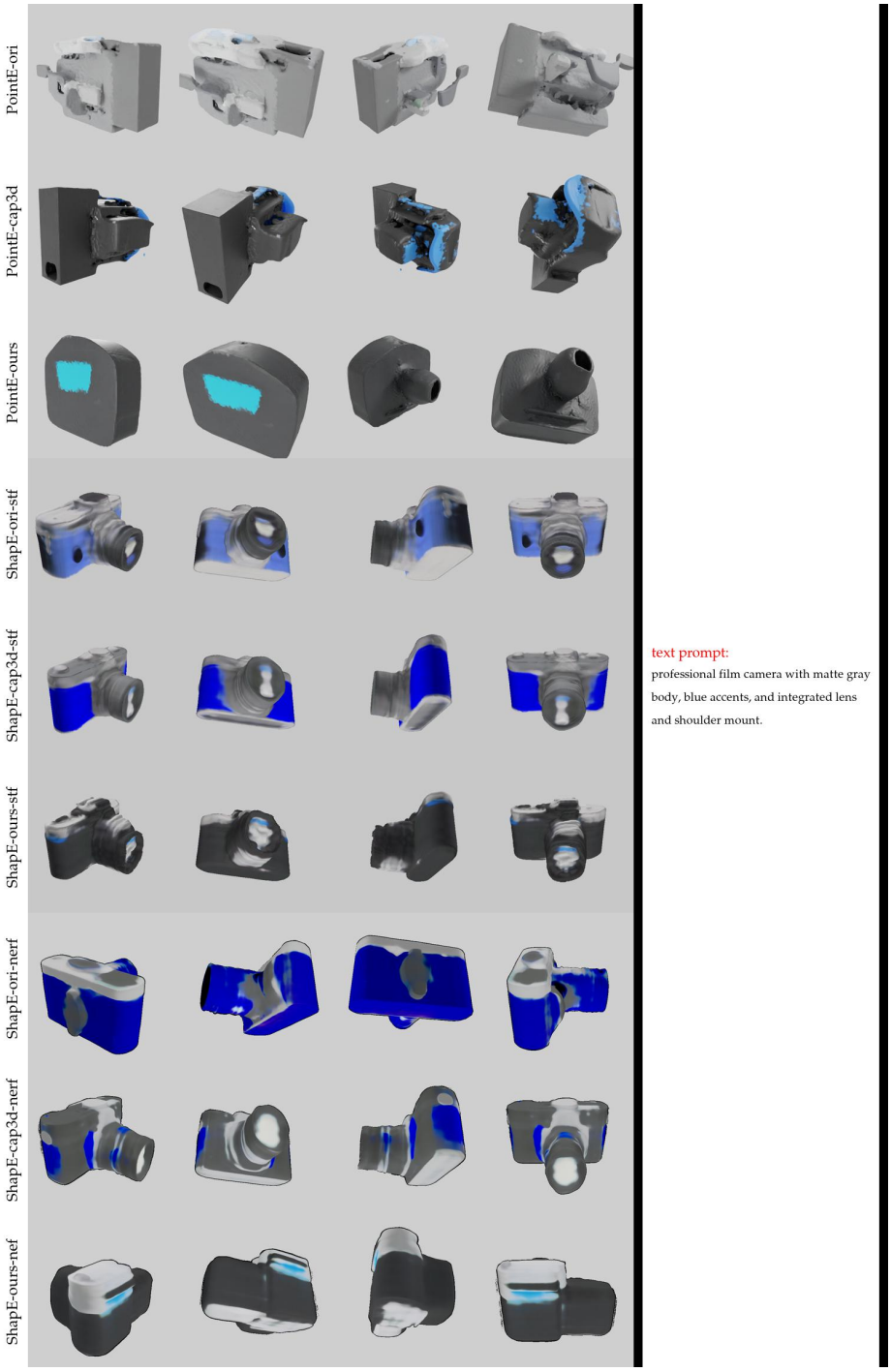

Fig. 85: Randomly sampled Text-to-3D results.

D DiffuRank on VQA

Algorithm 2 demonstrates the DiffuRank approach to the task of 2D Visual Question Answering. Initially, the process involves converting the question and each potential answer/option into a coherent statement. As shown in Figure 8, we convert Question: "Is the school bus driving towards or away from the camera?" and options "(a) Towards the camera (b) Away from the camera" into statements (1) "The school bus is driving towards the camera and statement" and (2) "The school bus is driving away from the camera". Another example shows converting Question: "Is there a shadow on the flower?" and options "(a) Yes (b) No,(a)" into statements (1) "There is a shadow on the flower." and (2) "There is not a shadow on the flower."

This conversion is accomplished through the utilization of GPT-4 in our implementation. Subsequently, we determine the alignment scores by evaluating the correspondence between each generated statement and the provided 2D image. The statement that exhibits the highest alignment score, along with its associated option, is then selected as the definitive answer.

Different from Algorithm 1, our objective here is computed over noise difference, the way adopted in our used stable-diffusion models [49].

Algorithm 2 DiffuRank for modeling the alignments between 2D images and answers for VQA tasks

Require: Given a Visual Question Answering (VQA) task, which consists of images $\mathcal{O}$, a question q, and multiple options o_i, and a pre-trained text-to-2D model Dtext-to-2D

\# 1. Turn question q and multiple options $\{o\}_{i=1,\cdots,M}$ into multiple corresponding statements $\{s\}_{i=1,\cdots,M}$;

\# 2. Compute average alignment scores

for each statement s_i **do**

  for $k \leftarrow 1$ to num_samples **do**

    Sample timestamp $t_k \sim \text{Uniform}(0, 1)$.

    Sample noise $\epsilon_k \sim \mathcal{N}(0, I)$.

    Compute noised input $\mathcal{O}_{t_k} = \sqrt{\bar{\alpha}_{t_k}}\mathcal{O}_0 + \sqrt{1 - \bar{\alpha}_{t_k}}\epsilon_k$.

    Compute loss $\mathcal{L}_{s_i,k} = \|D_{\text{text-to-3D}}(\mathcal{O}_{t_k}|s_i) - \epsilon_k$.

  end for

  Compute average loss for each statement s_i, $Cor(s_i, \mathcal{O}) = -\mathbb{E}_k\mathcal{L}_{s_i,k}$.

end for

return Top-1($\{Cor(s_i, \mathcal{O})\}_{i=1,\cdots,M}$)

E Future Work & Limitations

Future Work: DiffuRank leverages a pre-trained text-to-3D diffusion model for rendering view ranking, enhancing 3D object captioning. Improved captioning enables the refinement of the diffusion model, creating a feedback loop that cyclically utilizes the model for data generation and employs this data to fortify the model further. Besides, due to our limited computational resources and funding, it is not feasible to encompass all Objaverse-XL objects, presenting an opportunity for industrial entities.

Limitations: During our subtitling process, we use DiffuRank to select 6 rendered views out of 28 views. This process requires us to render more views, generate captions, and perform inference using a pre-trained text-to-3D diffusion model to compute alignment scores. All of the steps take calculation and time.

As highlighted in the related work (Section 2), DiffuRank faces challenges with speed, requiring multiple samplings for each option and necessitating forward model processing for all options. Our process for a single 3D object involves 28 rendered views, 5 captions per view, and performing sampling 5 times (num_{sample} in Alg. 1), resulting in a total of 700 inference operations. While parallel processing (large batch size) can mitigate delays, the system's performance is inherently slow. We show a VQA extension in Section 5.3 as it only has two options. But, generally, DiffuRank's design is not optimal for tasks requiring numerous options, such as classification and image-text retrieval.

Our discussion around broader impact is listed in Appendix A. Some of the failure cases and analysis are included in Appendix B.6.