Gen3DEval: Using vLLMs for Automatic Evaluation of Generated 3D Objects

Supplementary Material

8. Training Dataset

8.1. Examples from the Pre-training Dataset

In this section, we present examples from the pre-training dataset utilized to train Gen3DEval. Figures 5 and 6 illustrate the different types of input data generated from a single 3D asset, accompanied by the corresponding Question-Answer prompts.

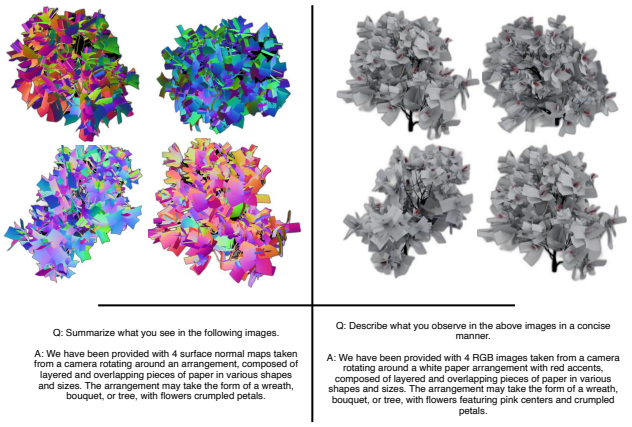

Figure 5. **Pre-training Dataset** We use multiple views of RGB and surface normal maps rendered from a 3D object, accompanied by a Question-Answer prompt that summarizes the object.

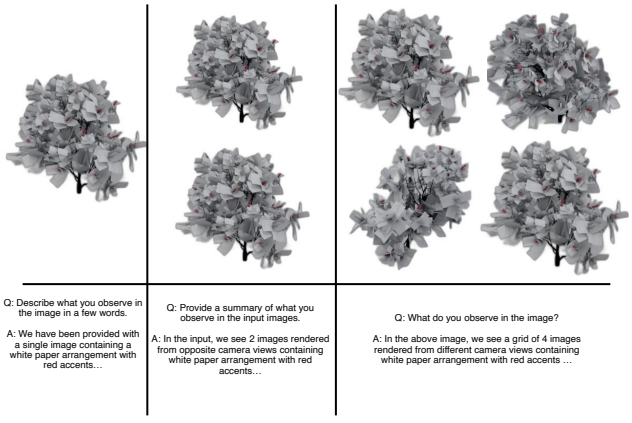

Figure 6. **Pre-training Dataset** We use single and multiple views rendered from a 3D object as well as an image grid composed of the aforementioned multi-view (4) RGB images.

8.2. Examples from the Supervised Fine-tuning Dataset

Figures 7 and 8 provide more examples from the supervised fine-tuning dataset employed in training Gen3DEval. The SFT dataset distribution is displayed in Figure 9.

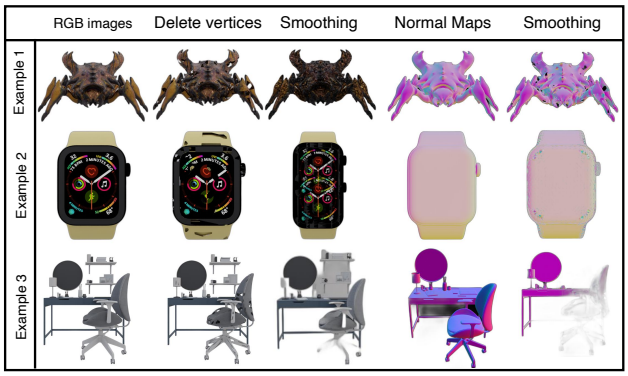

Figure 7. **Deep dive into Supervised Fine-tuning Dataset:** We use single and multiple views of RGB and surface normals rendered from a 3D object generated from a prompt. Further, we take these objects and perturb them to simulate common appearance, surface and text-related artefacts in generative 3D methods. In this figure, we showcase Laplacian smoothing and random deletion of vertices in the original meshes.

Figure 8. **Deep dive into Supervised Fine-tuning Dataset:** Further, we use artist-drawn meshes of 3D objects and perturb them to simulate common appearance, surface and text-related artefacts in generative 3D methods. In this figure, we showcase textual and structure specific perturbations, i.e., by generating objects using NeRFs and Gaussian splatting.

9. Dataset Ablation

Please refer to Figure 10 to see how the performance of Gen3DEval varies with the removal of different subgroups

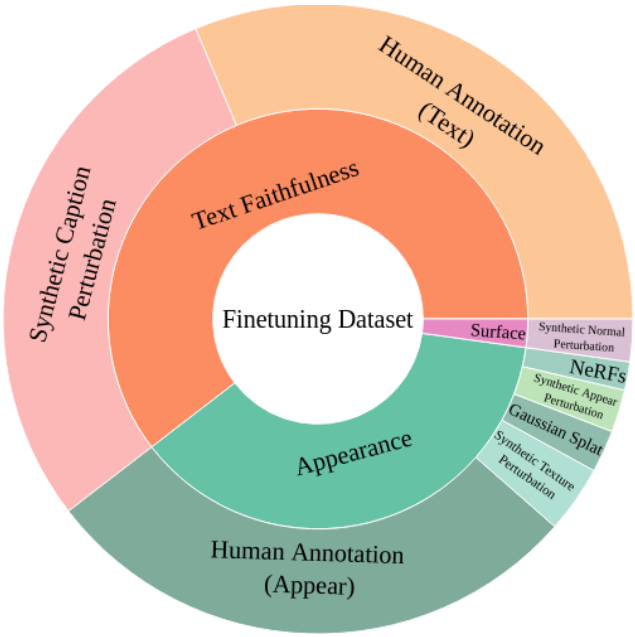

Figure 9. **Data distribution** for the SFT dataset used in training Gen3DEval. It consists of appearance, surface quality and text fidelity comparison data that are synthetically generated from artist-created meshes as well curated from user annotation with outputs from text-to-3D methods.

of the dataset.

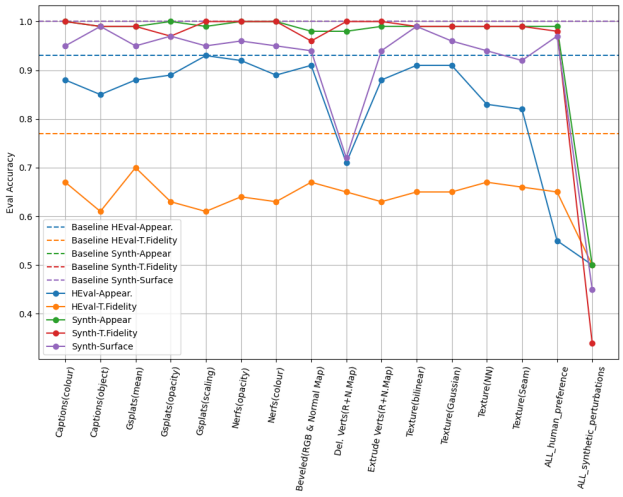

Figure 10. Ablation on training data by REMOVING subsets of data from the final fine-tuning dataset and evaluated on the held-out evaluation datasets. Dotted lines: accuracy when fine-tuned using the entire SFT dataset (same random seed).

10. Limitations

Gen3DEval exhibits erratic performance when there is Janus (subparts repeated in an object like multiple faces), and on out-of-domain surface evaluation. Figure 11 provides an example of this limitation, where the method compares multiple assets generated from the same prompt and ranks them from best to worst.

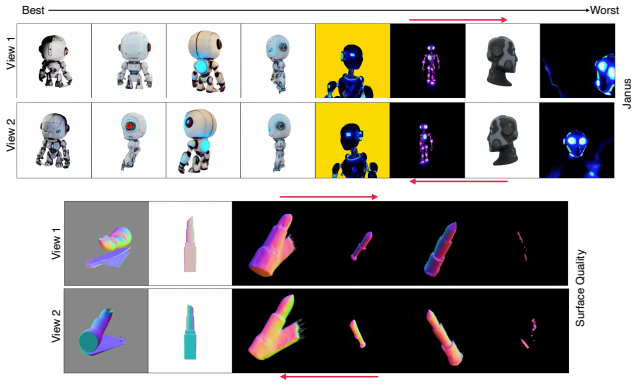

Figure 11. **Limitation of Gen3DEval.** Gen3DEval has limited success detecting janus and out-of-domain surface normal images. The image shows how Gen3DEval ranks the objects. The red arrow points to the expected ranking of the object.

11. Benchmark Analysis

Please refer to Table 3 for comparison of Gen3DEval-Bench with different existing benchmarks of generation prompts. Our aim was to come up with a small and diverse dataset containing an even split in terms of object type (animate like humanoids and animals vs. inanimate such as chairs, tables, football) and composition (single vs composite objects or scenes) to allow for granularity in the evaluation of 3D assets. Moreover, we also wanted to increase the mean and variance for the length of the prompts.

12. Comparison of objects generated from different prompts

Since Gen3DEval disambiguates evaluation on the basis of 1) Appearance 2) Text Fidelity and 3) Surface Quality, we additionally test its performance on a benchmark containing pairs of 3D objects generated by different prompts and annotating on the basis of appearance and filtered to remove any ambiguous samples. Gen3DEval has an accuracy of 0.88 on this benchmark. Qualitative examples are provided in Figure 12 .

	General		Object Type		Composition	
	Num. Prompts	Avg. word length	Animate	Inanimate	Single Obj	Multi-object
T3Bench [17]	300	7.98	36	264	100	200
ChatGPTEval3D [57]	110	11.49	18	92	65	45
DreamFusion [40]	404	6.98	211	192	154	250
Gen3DEval-Bench	80	12.863	40	40	43	37

Table 3. **Comparing Gen3DEval-Bench with existing 3D generation prompt benchmarks**.

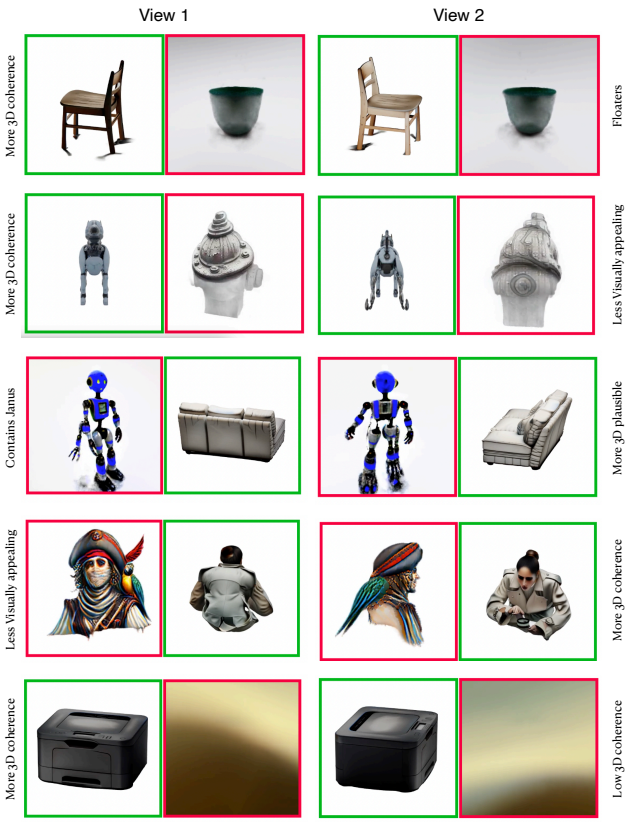

Figure 12. **Qualitative result from Gen3DEval on comparing objects generated from different text prompts on Appearance.** This image displays 5 examples of the preference of Gen3DEval from an annotated evaluation dataset where we conduct pairwise comparison of objects generated from different text prompts on the basis of their appearance only. Green border is for the preferred object and red for the other object. We use 4 views as input but in the image, we display two views side-by-side.

13. Results

13.1. Qualitative ablation study for Gen3DEval's image encoder choices

In our ablation study involving different image encoders, we evaluated the quantitative metrics of using CLIP, DinoV2, and Fit3D [62], as well as combinations of these with CLIP [42]. Our findings indicate that while the Gen3DEval with CLIP consistently performed well across all evaluation datasets, the pairing of DinoV2 [39] and CLIP was not too far behind. On investigating fuurther, we noticed that Gen3DEval with CLIP and DinoV2 gave more weight to 3D coherence and plausibility where Gen3DEval with standalone CLIP leans towards more visually appealing objects. Given that these image embeddings capture distinct object features, we provide qualitative examples generated on Gen3DEval-Bench to compare and contrast Gen3DEval's asset preferences. Figure 13 contrasts the strengths and weaknesses of Gen3DEval with CLIP and with the combination of CLIP and DinoV2 respectively. Overall, both embeddings capture relevant 3D features for comparison as shown in Figure 14.

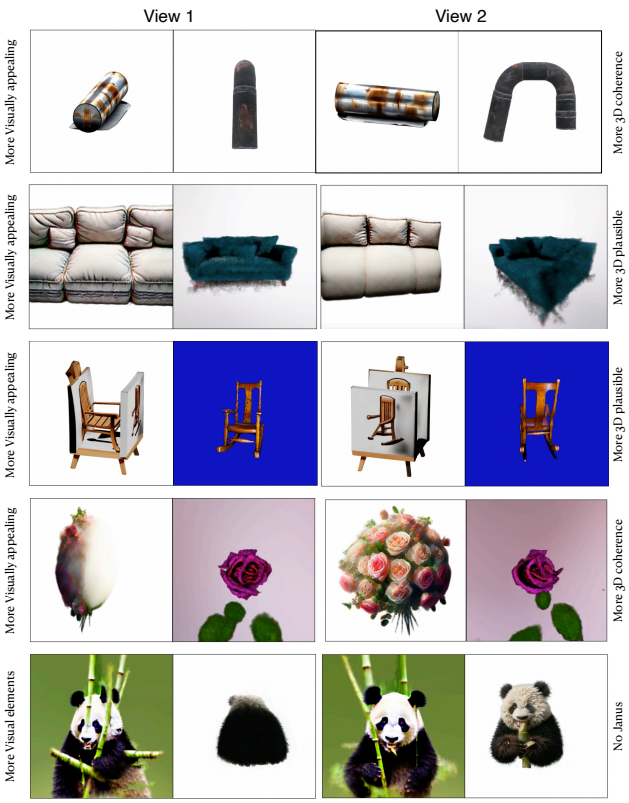

Figure 13. **Qualitative result from ablation study on different input image embeddings on Appearance.** Object Preference - *L.H.S* (Gen3DEval w/ CLIP), *R.H.S* (Gen3DEval w/ CLIP+DinoV2): We demonstrate five examples containing (displaying two views side-by-side to provide some clarity). Observation: CLIP evaluates more favourably on visual/appearance/surface properties whereas CLIP+DinoV2 prefers more on the basis of 3D coherency (lack of janus) and plausibility.

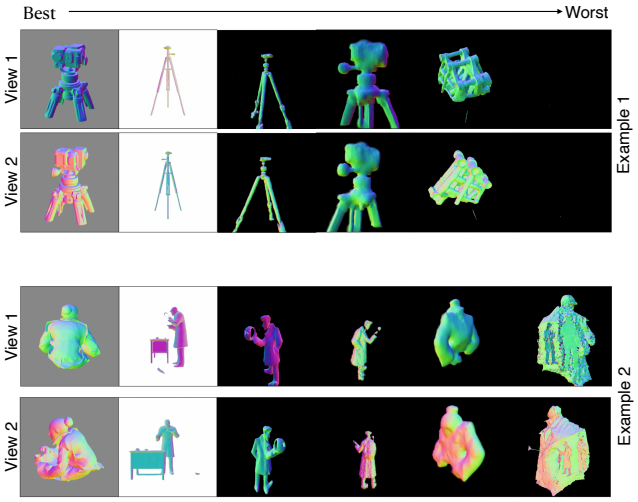

Figure 17. **Comparison of leaderboard methods (Surface Quality).** Qualitative examples from applying Gen3DEval to evaluate 3D generative methods on the surface quality parameter. Left to Right: Best object to worst in pairwise comparison of all assets for the same prompt.

13.2. Qualitative Comparison of Leaderboard Methods

We present qualitative examples of pairwise evaluation and ranking of generative 3D methods from our leaderboard on Gen3DEval-Bench. These examples focus on their performance on appearance (Figure 15), surface quality (Figure 16), and text faithfulness (Figure 17).

Figure 14. **Qualitative result from ablation study on different input image embeddings on Appearance.** This image displays a degree of correlation between the preferences of Gen3DEval when using either standalone CLIP embeddings or CLIP combined with DinoV2 for two examples containing (displaying two views to provide some clarity), since both the encoders select these assets in the same order using two examples where the objects were ranked in a similar manner.

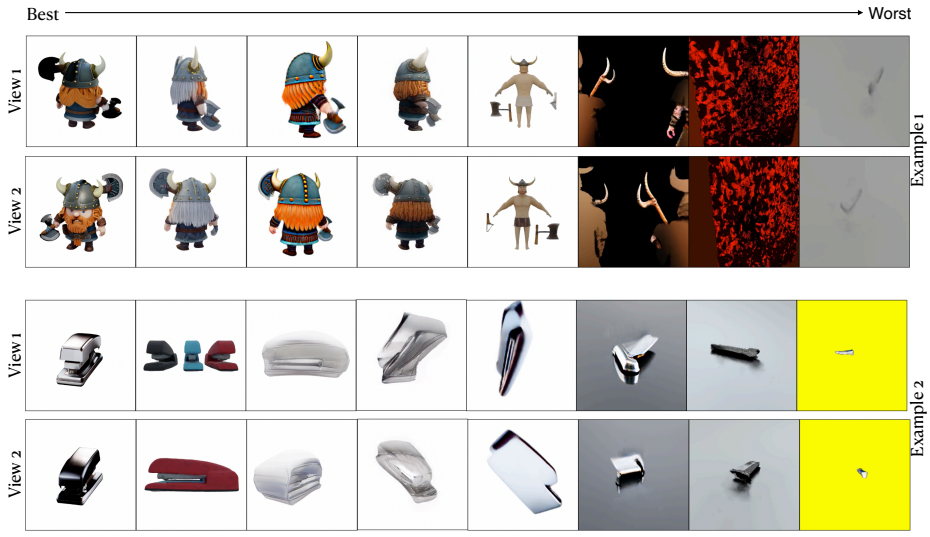

Figure 15. **Comparison of leaderboard methods (Appearance).** Qualitative examples from applying Gen3DEval to evaluate 3D generative methods on appearance quality parameter. Left to Right: Best object to worst in pairwise comparison of all assets for the same prompt.

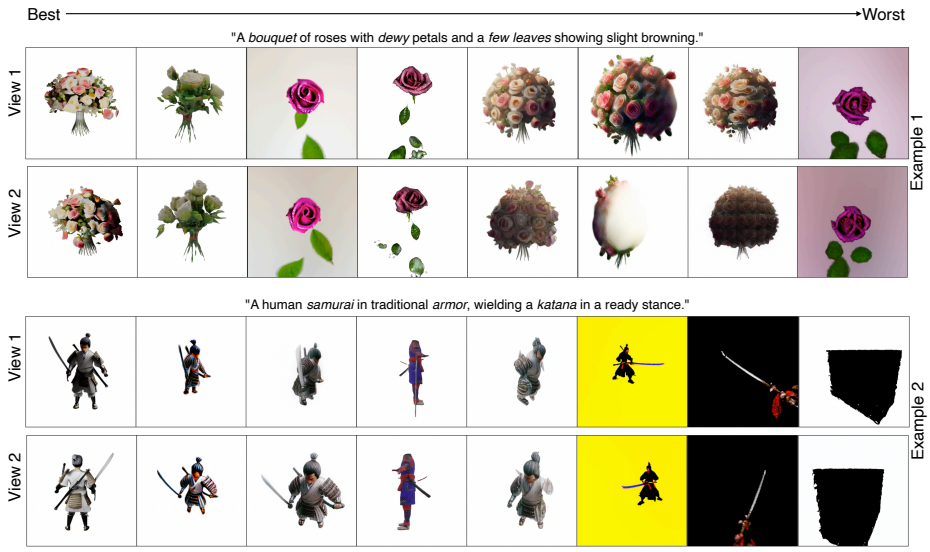

Figure 16. **Comparison of leaderboard methods (Text Fidelity).** Qualitative examples from applying Gen3DEval to evaluate 3D generative methods on the text fidelity parameter. Left to Right: Best object to worst in pairwise comparison of all assets for the same prompt.