Objaverse++: Curated 3D Object Dataset with Quality Annotations

Chendi Lin[1]* Heshan Liu[1] Qunshu Lin[2] Zachary Bright[3]
Shitao Tang[4] Yihui He[1] Minghao Liu[5] Ling Zhu[3] Cindy Le[6]†
[1]Carnegie Mellon University [2]Zhejiang University [3]Exascale Labs [4]Simon Fraser University [5]2077AI [6]Columbia University

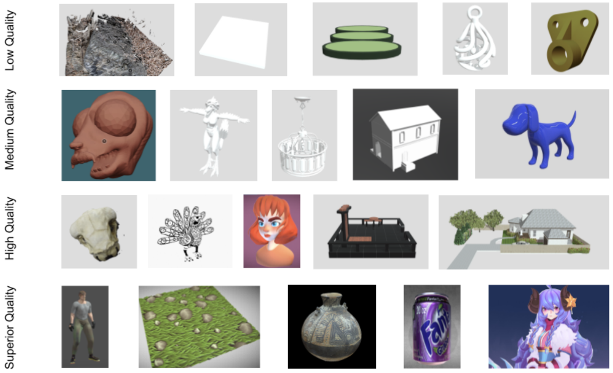


Figure 1. Examples of different quality scores assigned to 3D models.

Abstract

This paper presents Objaverse++, a curated subset of Objaverse enhanced with detailed attribute annotations. To address the prevalence of low-quality models in Objaverse[3], human experts manually annotate 10,000 3D objects with quality and characteristic attributes. Then, we trained a neural network capable of annotating the tags for the rest of the Objaverse dataset. We show that models trained on a quality-focused subset achieve better performance than those trained on the larger Objaverse dataset in image-to-3D generation tasks. In addition, our experiments show that the higher the data quality, the faster the training loss converges. These findings suggest that careful curation and rich annotation can compensate for the raw dataset size, potentially offering a more efficient path to develop 3D generative models. We release our enhanced dataset of approximately 790,000 curated 3D models [1] [2] to facilitate further research in 3D computer vision and aim to extend our annotations to cover the entire Objaverse dataset.

1. Related Work

The generation of 3D models has been an important research topic for years. Early 3D generation research was influenced by 2D text-guided generation, starting with GAN-based approaches like 3D-GAN [13]. After the introduction of Contrastive Language-Image Pre-Training (CLIP) [11] in 2021, several works adapted CLIP-based 2D generation methods to the 3D space [5, 8, 9]. Dreamfields [5] and DreamFusion [10] pioneered a new era of pre-trained text-to-image models for 3D generation through Score Distillation Sampling (SDS), followed by subsequent diffusion-based works [1, 6, 7, 12].

The emergence of large-scale 3D datasets has been crucial for advancing 3D vision and generation tasks. Objaverse [3] represents a significant milestone in this domain, providing over 800,000 annotated 3D objects under Creative Commons licenses. As a comprehensive collection widely used by the research community, Objaverse still has limitations such as a lack of high-quality, texture-rich objects valuable to train texture generation models.

2. Methods

We manually annotated 10,000 3D objects from Objaverse with quality scores and additional binary traits and trained a neural network capable of annotating the rest of the dataset.

Our data labeling follows a systematic multistage annotation and validation process. During the preliminary assessment phase, we consulted artists to determine the essential components of our dataset and randomly sampled 1,000 objects for evaluation. Building on these initial insights, we develop comprehensive labeling rubrics that standardize the evaluation process. Then, our human annotators manually annotated 10,000 objects with quality and aesthetic tags. Using these data, we trained an annotation network that classifies models and outputs the quality and aesthetic tags based on 2D-rendered multiviews.

*Email: chendil@alumni.cmu.edu

†Corresponding author. Email: xl2738@columbia.edu

[1] https://github.com/TCXX/ObjaversePlusPlus

[2] https://huggingface.co/datasets/cindyxl/ObjaversePlusPlus

2.1. Annotation Tags

2.1.1. Quality and Aesthetics Score

We define quality score as a metric to tell how useful a 3D object is for machine learning training. We assume a neural network for 3D-related tasks may want to learn two aspects: the semantic meaning of the geometric shape and the color information from the surface texture. The following are the enumeration values for the quality annotation.

- **Low Quality**: No semantic meaning. Specifically, if the annotators are not able to identify the object, or if the object is corrupted, it will fall under this category.
- **Medium Quality**: The object is identifiable, but missing the basic material texture and color information. Items that are single-colored by nature, for example, a garden statue, are still considered as having a texture.
- **High Quality**: High Quality indicates an acceptable quality with a clear identity of the object. The object is properly textured with some material and color details.
- **Superior Quality**: The object is of excellent quality with high semantic clarity. The object is professionally textured with strong aesthetic harmony. This type of object can be used in specific gaming scenario setups without a sense of violation.

A rubric and various samples, as shown below and in Fig.1 are provided to human annotators to determine the score of a 3D model from Objaverse. Note that the quality score is meant to guide researchers and does not necessarily conclude the artistic value of the 3D objects.

2.1.2. Binary Traits

In addition to the quality score on a linear scale, our Objaverse++ annotates several binary tags for each 3D object.

- **Transparency**: Identifies models with see-through parts, where some areas allow visibility through objects in front of them. Entirely opaque objects do not have this tag. Some 3D generation algorithms rely on 2D multiviews and may not handle transparent parts properly.
- **Scene**: Identifies whether the model represents a scenario or an environment, rather than a standalone object. Since scene generation [2, 14] and object generation differ greatly in algorithms and training data needs, this tag will make an important distinction.
- **Single Color**: Tags models that are unintentionally monochromatic, meaning that they consist of only one color without any shading, texture, or other visual variation. Models with deliberate monochromatic design (e.g., a sculpture), shading, or texture do not receive this tag. This tag filters out 3D objects that are meaningful for learning texture generation.
- **Not a Single Object**: Marks models that consist of multiple separate components rather than a single unified object. The tag refocuses on model learning for the generation and understanding of single objects.
- **Figure**: Indicates if the model represents a character, person, or figure. The tag creates a subset of data for potential training in character generation.

Note that a 3D model could own multiple binary tags.

2.2. Annotation Network

The data distribution of binary tags in the human-annotated training dataset is presented in Fig.9 in Supplement Materials. To scale our approach to cover a larger portion of the Objaverse dataset, we develop and train a 3D model classifier using our manually annotated data.

2.2.1. Classifier Architecture

The 3D object classifier uses a multiview approach, combining convolutional and recurrent neural networks with an attention mechanism, enhanced by object-specific metadata. The training data includes 40 screenshots of each 3D model, captured from different angles, along with metadata from Objaverse. The model architecture is shown in Fig.6 in Supplement Materials and is composed of five main components: feature extraction (pre-trained ResNet50), sequence modeling (RNN), attention layer, metadata integration, and classification heads that predict a specific attribute of the 3D model, such as style, quality score, and binary tags.

Cross Entropy Loss is used for score annotation, and BCEWithLogitsLoss is used for binary tag labeling.

2.2.2. Classifier Validation

The evaluation metrics performed on the test set (1971 samples) are shown in Table 1.

Table 1. Metrics of Annotation Network

Metric	Accuracy	F1 Score	mAP
score*	0.5945	–	–
relaxed score accuracy	0.8221	–	–
is_multi_object	0.8621	0.703	0.6927
is_scene	0.8667	0.731	0.9176
is_figure	0.9448	0.844	0.6815
is_transparent	0.9372	0.835	0.7435
is_single_color	0.9169	0.796	0.6745

Among all the metrics, it is clear that the metrics for the binary tags (is_multiple_object, is_scene, is_figure, is_transparent) demonstrate strong accuracies. These metrics highlight the reliability of the network in identifying different characteristics of the model, which is valuable for various 3D modeling applications. “Score” (marked with * in Table 1) is a relatively weak metric of the annotation network due to subjective nuances, as mentioned above. Here we include a “relaxed accuracy for score” that allows scores 2 and 3 to be interchangeable. The “relaxed score accuracy” improves significantly to 82.21%, showing that the network captures general quality distinctions effectively, though missing subtle differences among models of high quality. The high accuracy of our annotation network shows that our proposed tags are learnable by a carefully designed classifier.

3. Dataset Evaluation

We set up an image-to-3D generation task as a practical and reproducible approach for future studies on 3D dataset curation. OpenLRM [4], an open-source framework designed to generate 3D models from a single image input. Given our computational constraints, we utilized OpenLRM's small model architecture, which is optimized for limited computing resources while still providing effective 3D generation capabilities. Table 3 in the supplement describes the model configuration used in this work[4].

We randomly sampled 100,000 objects from Objaverse to create Training Set A, and its binary quality tag distribution is described in Table 2 in Supplement Materials. Using quality filtering criteria, we then selected approximately 50,000 high-quality objects to form Training Set B. Filtering criteria included selecting models of high or superior quality, excluding monochromatic models, excluding scenes, and excluding models with any transparent part. This setup is well-suited for single-object generation, as the filtering criteria ensure a subset that closely aligns with the input distribution of the generation task.

We trained the same model on Training Set A. This training, starting from scratch, took approximately 9 hours on 8 H100 GPUs. Subsequently, we trained the same model on Training Set B, which required about 6 hours under the same conditions.

To assess the performance of the model, we conducted a user study with 47 participants. Each participant was presented with 10 pairs of generation results: one generated from Training Set A and the other from Training Set B. The generated 3D models are inferenced given the sample inputs of the open-source OpenLRM. The generated results were shown in random order, and participants were asked to choose their preferred result or to select "no preference", without knowing the origin of each generation. To reduce bias, the question order and the option order were randomized. The participants came from diverse backgrounds, including artists, machine learning researchers, game developers, and software engineers.

3.1. Better Generation Quality

The user study shows significant results (see Figure 3) in favor of our generation model. Of the 10 pairs of objects, 8 showed a preference for our model over the baseline. Despite the presence of the no-preference option, participants favor our generation more. Our model receives 83.5% more votes than the baseline.

We have also conducted quantitative research using chamfer distance. The pre-trained OpenLRM-obj-small-1.1 trained using the whole Objaverse is used as ground truth, and we compare the performance of our model (high-quality subset) and the baseline model (randomly sampled 100k subset) as shown in Fig.4.

Figure 2. A comparison of image-to-3D generation results by randomly sampled 100,000 dataset vs. our model. Overall, our model consistently produces results with more refined details, better texturing, and more realistic material properties in surface details, shadows, and material rendering.

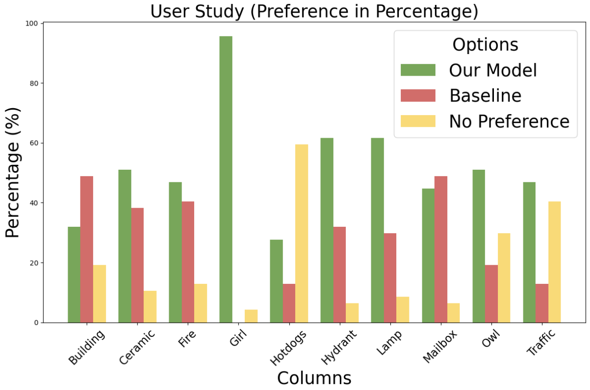


Figure 3. User Study Results. Of the 10 pairs of objects, 8 showed a preference for our model over the baseline.

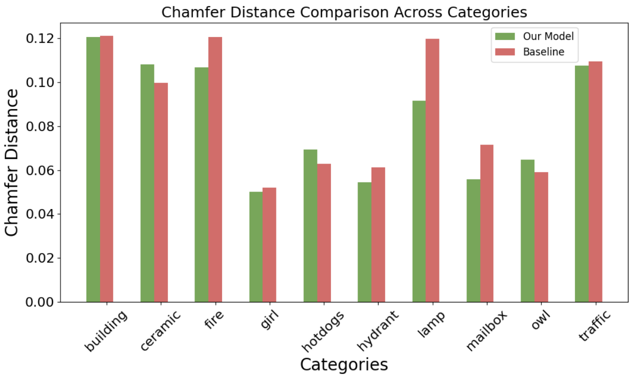


Figure 4. Chamfer distance comparison of the model trained using the randomly-sampled 100,000 dataset vs. our model using high-quality subset.

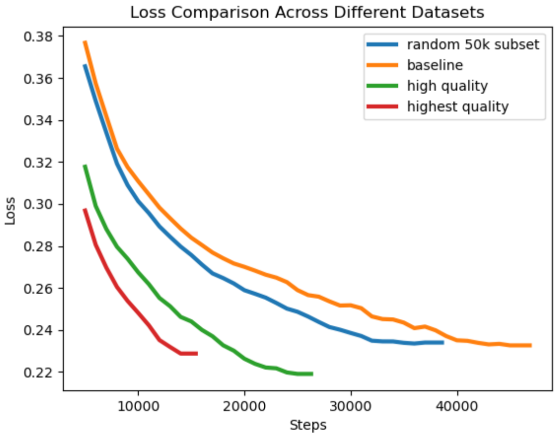


Figure 5. **Baseline**: A randomly sampled subset of 100,000 objects from Objaverse, called Training Set A.
Random 50k Subset: A randomly sampled subset of 50,000 samples from Training Set A, used to isolate the effect of dataset size on convergence.
High-Quality Subset: A quality-filtered dataset containing samples with high and superior quality, excluding monochromatic models, excluding scenes, and excluding models with any transparent part, called Training Set B.
Superior Quality Subset: A highly curated subset containing only samples with superior quality, excluding monochromatic models, excluding scenes, and excluding models with any transparent part.

3.2. Faster Convergence in Training

As shown in Fig.5, our model demonstrates faster convergence on a carefully curated dataset compared to a randomly selected subset of the Objaverse. Training in a refined selection allows the model to require fewer epochs and steps to achieve optimal performance. This improvement is evident in two main aspects:

Quality Over Random Sampling. To ensure that dataset size is not the primary reason for faster convergence, we randomly sampled 50,000 objects from Training Set A to create a subset of comparable size to our "high-quality dataset". As illustrated in Fig.5, the random 50k subset does not result in a significantly faster convergence than the baseline dataset. This implies that simply reducing the dataset size does not guarantee faster convergence. In comparison, our "high-quality dataset", similar in size to the random 50k subset, produces substantially faster convergence, demonstrating that the curated quality of the samples is crucial.

Impact of Annotation Quality. The "superior quality subset", containing only top quality samples, converges faster than the "high quality subset" dataset, which includes both high and superior quality models. This finding supports the idea that our scoring rubrics are efficient, as higher-quality data directly contributes to faster and more efficient model training.

In summary, the findings show that faster convergence is driven by the quality of the curated dataset rather than its size. This underscores the importance of our quality-filtered data and validates that our labeling process improves the quality of the original Objaverse dataset.

4. Conclusion

We used a combination of human labeling and an annotation network to tag models from Objaverse. The tagging results for the 100,000 models sampled from Objaverse will be open-sourced as the initial result, allowing users to create custom training sets tailored to their specific needs. Using quantitative metrics and a user study, we show how high-quality training data can simultaneously enhance effectiveness, efficiency, and performance in 3D generation.

To our knowledge, our work is the first -

1. to provide quality annotations for objects in the Objaverse dataset and manually annotated the largest scale of 10,000 unique 3D objects;
2. to examine the correlations between the quantity of training data and the quality of generation in the research domain of 3D modeling; and
3. to develop standard rubrics of quality scores and other relevant traits for 3D objects.

Future directions. We acknowledge that further experiments comparing Objaverse++ with Objaverse-XL are important. In addition, we will explore comprehensive and quantitative filtering criteria, possibly leveraging more advanced annotation networks to tag additional model attributes, such as structural complexity or aesthetic style.

References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation, 2023. 1

[2] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Scenedreamer: Unbounded 3d scene generation from 2d image collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15562–15576, 2023. 2

[3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1

[4] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. `https://github.com/3DTopia/OpenLRM`, 2023. 3

[5] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields, 2022. 1

[6] Cindy Le, Congrui Hetang, Chendi Lin, Ang Cao, and Yihui He. Euclidreamer: Fast and high-quality texturing for 3d models with stable diffusion depth, 2024. 1

[7] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation, 2023. 1

[8] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes, 2021. 1

[9] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*. ACM, 2022. 1

[10] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[12] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion, 2023. 1

[13] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling, 2017. 1

[14] Xiuyu Yang, Yunze Man, Jun-Kun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation, 2024. 2

Objaverse++: Curated 3D Object Dataset with Quality Annotations

Supplementary Material

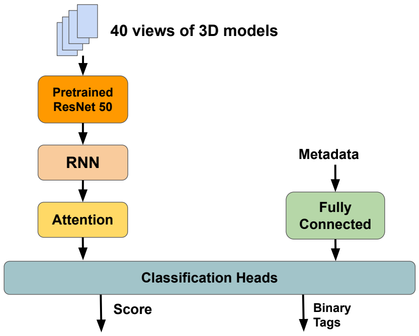


Figure 6. Structure of the annotation network.

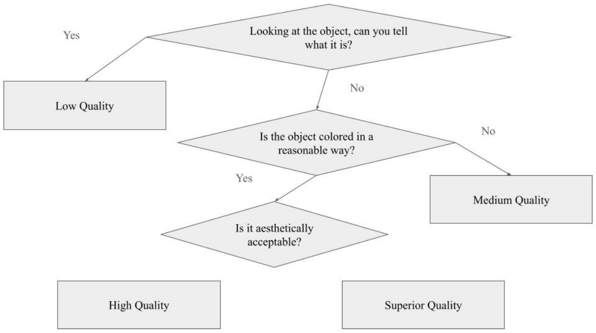

Looking at the object, can you tell what it is?
Yes
Low Quality
No
Is the object colored in a reasonable way?
No
Medium Quality
Yes
Is it aesthetically acceptable?
High Quality
Superior Quality

Figure 7. Decision tree for human annotators to categorize the quality level of a 3D object.

5. Annotation Model Structure

The annotation network structure described in 2.2 is presented in Fig.6.

6. Quality Score Rubrics

Due to the specialty of judging 3D objects, we composed additional training material for our human annotators, including Fig. 7 and the rubrics below. For quality score, here are some criteria to consider when an object has proper semantic meaning and texture.

High-Quality Criteria:

- Basic color scheme present but lacks richness and aesthetic appeal.
- Acceptable geometric shapes - not too rough but not highly detailed.
- Basic textures present - goes beyond flat surfaces but lacks sophistication.
- Visually comfortable and harmonious, but lacks refinement in details (like color rendering and fabric textures).

Superior-Quality Criteria:

- High-quality modeling with rich textures, vibrant colors, and aesthetic value.
- Rich, harmonious color combinations that feel natural or appropriate to the style.
- Geometric proportions that match either real-world references or suit the intended artistic style.
- Detailed surface texturing with effective lighting/shading.
- Aesthetically pleasing or visually impactful.
- Abundant detailed elements such as decorations, patterns, etc.

Binary Tags Examples: The examples of each binary tag are presented in Fig.8.

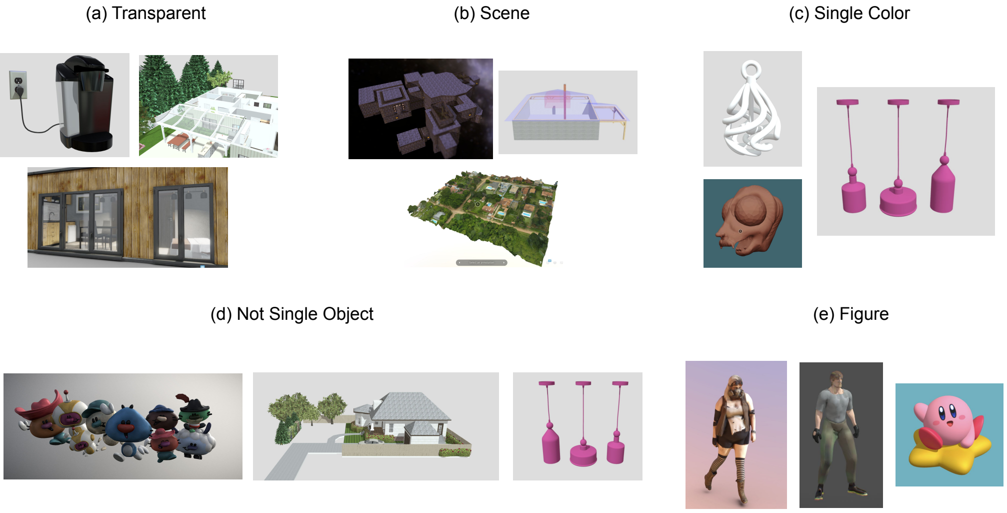
The data distribution for each binary tag in the human-annotated dataset is presented in Fig.9, and the data distribution for each binary tag in the randomly sampled 100,000 objects from Objaverse is described in Table 2.

7. Training Loss

Figure 10 demonstrates the impact of dataset quality on training loss. The high and superior quality subsets show faster and more stable convergence than the baseline, a randomly sampled subset of 100,000 objects from Objaverse, and random 50k subsets. Quality-filtered data reduces noise, accelerates optimization, and enhances learning stability, allowing the model to converge more efficiently. In contrast, the baseline dataset's noisy samples hinder optimization. The superior quality subset achieves the best results among the four datasets, underscoring the importance of high-quality data over dataset size for efficient model training.

8. User Study Results

We include an additional Table 4 for the percentage breakdown of our user study and Fig. 2 as examples of the result comparison included in the user study.



(a) Transparent

(b) Scene

(c) Single Color

(d) Not Single Object

(e) Figure

Figure 8. Examples of different binary tags assigned to 3D models. (a) **Transparency**: Identifies models with see-through parts. (b) **Scene**: Distinguishes scene-like models from standalone objects, enabling differentiation in 3D model generation suited for environments versus single objects. (c) **Single Color**: Tags unintentionally monochromatic models, filtering out non-texture-rich objects in texture generation learning. (d) **Not a Single Object**: Identifies models with multiple separate components, focusing learning on single-object generation tasks. (e) **Figure**: Marks character or figure models, creating a subset for character generation that may benefit from specialized training.

Label	**is_multi_object**	**is_scene**	**is_figure**	**is_transparent**	**is_single_color**
0 (No)	94.98%	59.45%	97.64%	97.67%	81.32%
1 (Yes)	5.02%	40.55%	2.36%	2.33%	18.68%

Table 2. Distribution of Selected Binary Tags in the 100,000 Annotation Dataset.

Type	**Layers**	**Feat. Dim.**	**Attn. Heads**	**Triplane Dim.**	**Input Res.**	**Image Encoder**	**Size**
small	12	512	8	32	224	dinov2_vits14_reg	446M

Table 3. Large language model configuration details.

	Our model	**Baseline**	**No preference**
Building	31.9	**48.9**	19.2
Ceramic	**51.1**	38.3	10.6
Fire	**46.8**	40.4	12.8
Girl	**95.7**	0	4.3
Hotdogs	**27.7**	12.8	59.5
Hydrant	**61.7**	31.9	6.4
Lamp	**61.7**	29.8	8.5
Mailbox	44.7	**48.9**	6.4
Owl	**51.1**	19.1	29.8
Traffic	**46.8**	12.8	40.4

Table 4. User study results in percentage. Of the 10 pairs of objects, 8 preferred our model (in bold) over the baseline.

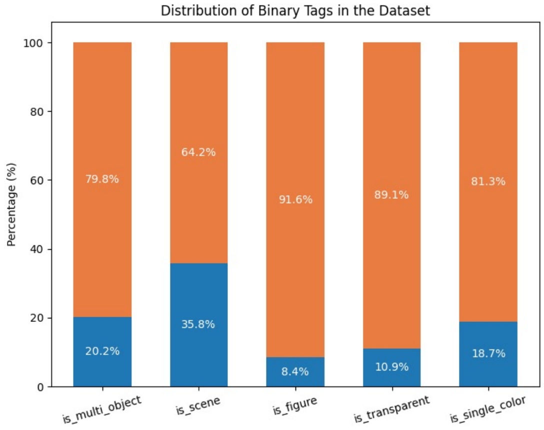


Figure 9. The distribution of binary tags in the human-annotated training dataset.

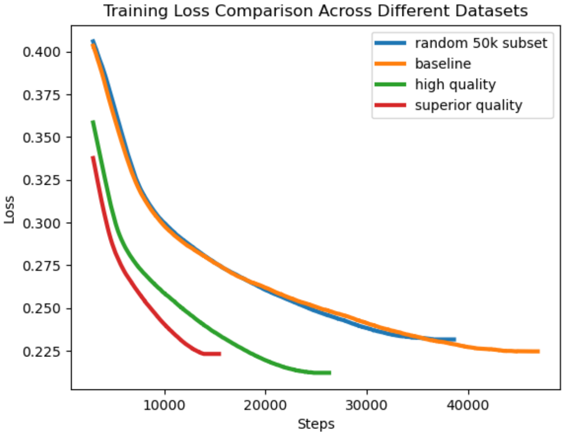


Figure 10. Training loss comparison across different datasets. Similar to the validation loss result 5, the model converges significantly faster on high-quality and superior-quality datasets, and converges roughly at the same speed on a random 50k subset and baseline, which is a randomly sampled subset of 100,000 objects from Objaverse.