# UnCommon Objects in 3D

Xingchen Liu[∞] Piyush Tayal[∞] Jianyuan Wang[∞] Jesus Zarzar[+∞] Tom Monnier[∞]
Konstantinos Tertikas[∞*] Jiali Duan[∞] Antoine Toisoul[∞] Jason Y. Zhang[†]
Natalia Neverova[∞] Andrea Vedaldi[∞] Roman Shapovalov[∞] David Novotny[∞]

*NKUA, Greece [+]KAUST [†]Carnegie Mellon University [∞]Meta AI

uCO3D `https://uco3d.github.io` `https://github.com/facebookresearch/uco3d`

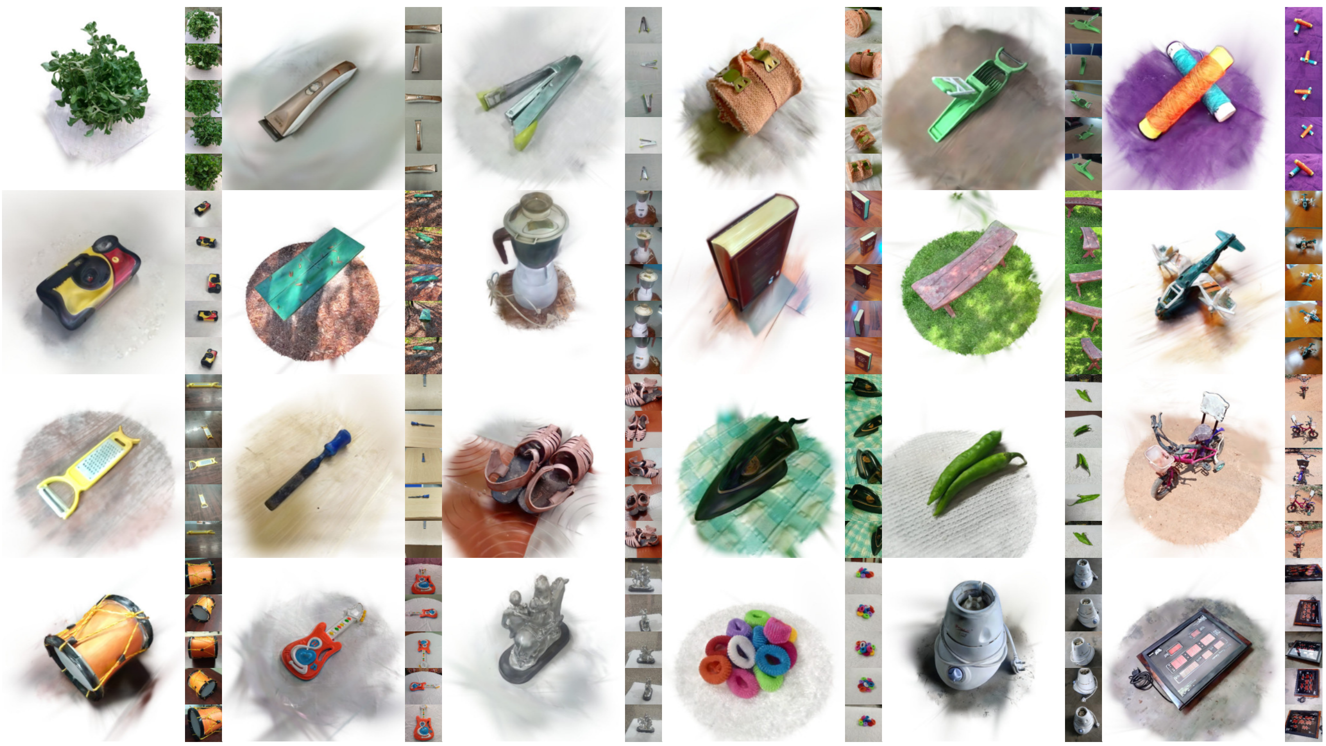

Figure 1. We introduce **UnCommon Objects in 3D (uCO3D)**, a large and diverse dataset of high-quality 360° videos covering over 1,000 object categories. Each video frame is 3D-annotated with accurate SfM cameras, point cloud, and a 3D Gaussian Splatting reconstruction.

## Abstract

*We introduce Uncommon Objects in 3D (uCO3D), a new object-centric dataset for 3D deep learning and 3D generative AI. uCO3D is the largest publicly-available collection of high-resolution videos of objects with 3D annotations that ensures full-360° coverage. uCO3D is significantly more diverse than MVImgNet and CO3Dv2, covering more than 1,000 object categories. It is also of higher quality, due to extensive quality checks of both the collected videos and the 3D annotations. Similar to analogous datasets, uCO3D contains annotations for 3D camera poses, depth maps and sparse point clouds. In addition, each object is equipped with a caption and a 3D Gaussian Splat reconstruction. We train several large 3D models on MVImgNet, CO3Dv2, and uCO3D and obtain superior results using the latter, showing that uCO3D is better for learning applications.*

# 1. Introduction

The primacy of data has been the defining characteristic of the last decade of machine learning, alongside deep learning. The most powerful models in language, speech, and computer vision are large deep networks trained on massive amounts of data, and then further fine-tuned on its high-quality subset. This paradigm is expected to extend to all machine learning applications, including 3D vision.

However, 3D training data is much harder to come by than data for text, audio, and image processing.

Seeking training data for large 3D neural networks [27], many have turned to synthetic datasets like Objaverse [13]. However, synthetic data is a poor substitute for real data in applications like *digital twinning*, which creates 3D models of real-life objects. This is why many photorealistic reconstruction networks [6, 23, 57, 58, 63, 67] are trained using *real object-centric datasets* like CO3D [47], MVImgNet [76], GSO [15], and OmniObject3D [68]. Real data is also crucial for generalization, as demonstrated by DUSt3r [64] for point map prediction and DepthAnything [74] for depth prediction, both of which are trained on numerous real datasets. Even non-curated image datasets like the billion-scale LAION [52] are applicable to 3D vision. For instance, text-to-3D generators [35, 45, 53, 54] build on LAION-trained large text-to-image models [5, 11].

Given the importance of 3D datasets, but also their relative scarcity, in this paper we ask what is the next step for real data in 3D vision. To answer this question, we note that, while the size of a dataset is crucial, in most cases its *quality* is just as important. For example, text-to-3D models [35, 55] are notoriously sensitive to fine-tuning data quality, training only with the best-looking models (*e.g.*, Instant3D [35] uses only about 1% of Objaverse). We conclude that high-quality datasets balancing scale and fidelity are more valuable than simply amassing low-quality data.

Based on this observation, we argue that there is a gap in the real object-centric 3D datasets that are currently available, as none strikes the optimal balance between quality and scale. For example, the 3D object scans in OmniObject3D [68] and GSO [15] provide very accurate geometry and textures, but only count a few thousand objects. Conversely, datasets like CO3D [47] and MvImgNet [76] contain orders-of-magnitude more objects, but lack reliable 3D scans. Instead, they provide many *views* of the objects together with lower-quality 3D cameras and point clouds reconstructed with structure-from-motion (SfM).

In this paper, we address this gap with a new dataset, *Uncommon Objects in 3D* (uCO3D), which better balances data quality and size (Tab. 1). Similar to CO3D, it comprises full-360° crowd-sourced videos capturing objects from all sides, annotated with cameras and point clouds using SfM. Furthermore, uCO3D has much greater data diversity (Fig. 2) than prior alternatives as it contains objects from the 1,070 visual object categories of the LVIS [22] taxonomy, which has long tails. For reference, MVImgNet and CO3Dv2 contain only 238 and 50 categories, respectively. These fine-grained categories are organized in super-categories, also shown in Fig. 2. Furthermore, uCO3D contains 170k scenes, which is more than four times larger than CO3Dv2's 40k. While this is less than MVImgNet's 220k, uCO3D's videos cover each object from all sides, as opposed to MVImgNet's partial object captures.

| | Real | Count | # Classes | Data type | Annotations |
|---|---|---|---|---|---|
| ShapeNet [8] | ✗ | 51k | 55 | 3D meshes | mesh |
| Objaverse [13] | ✗ | 800k | 21k | 3D meshes | mesh |
| Objaverse-XL [12] | ✗ | 10M | 2M | 3D meshes | mesh |
| ABO [10] | ✗ | 8k | 63 | 3D meshes | mesh |
| OmniObject3D [68] | ✓ | 6k | 190 | Videos w/ meshes | cameras, mesh |
| GSO [14] | ✓ | 1k | 17 | Views w/ meshes | cameras, mesh |
| Objectron [2] | ✓ | 15k | 9 | Limited vp. videos | cameras, 3D box |
| MVImgNet [76] | ✓ | 220k | 238 | Limited vp. videos | cameras, pcl |
| CO3D [47] | ✓ | 19k | 50 | 360° videos | cameras, pcl |
| CO3Dv2 [47] | ✓ | 40k | 50 | 360° videos | cameras, pcl |
| **uCO3D** (ours) | ✓ | 170k | 1k | 360° videos | cameras, 3DGS, caption |

Table 1. **Overview of 3D object datasets.** We compare the number of objects / classes, the type of data and associated annotations.

Besides improving size and diversity, uCO3D also raises the quality bar. This was achieved by checking extensively both the collected videos and their 3D annotations. Differently from datasets like CO3Dv2 that still contain a certain portion of low-quality videos, in uCO3D we manually verified that each video provides full 360° turn-table covering all sides of the object. Additionally, 60%+ of the videos have 1080p+ resolution, higher than CO3Dv2. To ensure 3D-annotation quality, we improved both the reconstruction algorithm and the reconstruction validation. For camera reconstruction, we used VGGSfM [62], which is currently the best SfM system available, and is more robust and accurate than COLMAP [50], used in CO3Dv2 and MvImgNet. We also improve on CO3Dv2's active-learning camera quality evaluation by combining it with novel-view synthesis accuracy after reconstructing each scene using 3D Gaussian Splatting (3DGS) [29]. The latter also guarantees that scenes are reconstructible to a high quality, which is important for training of 3D models.

We validate uCO3D's benefits in applications. We train two popular 3D models, LRM [27] and CAT3D [18], using uCO3D and demonstrate improved results compared to training on MVImgNet and CO3Dv2, which makes uCO3D the better data source for real object-centric 3D learning. We also use uCO3D to train a text-to-3D model following Instant3D's [35] two-stage design. The latter requires objects to be rendered from canonical viewpoints, and thus, so far, was limited to synthetic data. By using our 3DGS reconstructions, we 're-shoot' uCO3D's from these viewpoints, which allows to train a more realistic generator.

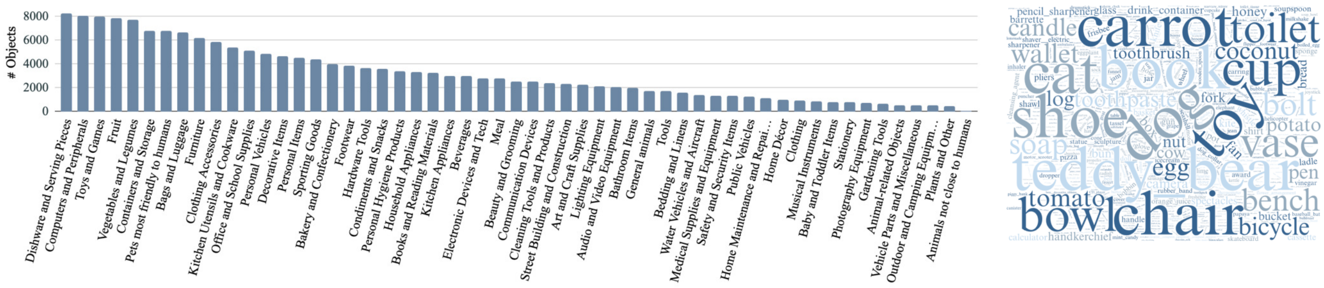


Figure 2. **Statistics of uCO3D.** (Left) We plot the number of objects per super-category. In total, the dataset contains 50 super-categories, each gathering around 20 sub-categories. (Right) We show a word cloud of all 1,070 visual categories represented in the dataset.

## 2. Related Work

**Datasets of synthetic 3D objects.** Historically, object-centric 3D datasets have predominantly been synthetic, composed of artist-generated 3D models. A prominent example is ShapeNet [8], with 51,000 meshes across 55 object categories. The meshes have detailed geometries, but relatively simplistic textures. Datasets such as 3D-FUTURE [17], IKEA [36], Pix3D [56], and ABO [10] are less diverse, concentrating on furniture and other consumer goods. In contrast, ModelNet [69], DeepCAD [66], and ABC [33] provide CAD models with clean geometry but no texture. Objaverse [13] is perhaps the most impactful dataset following ShapeNet. It is significantly larger, comprising 800,000 3D meshes. Objaverse-XL [12] expands this collection to 10M objects. These datasets have been pivotal in enabling the first 3D deep generative models, including text-to-3D [35, 53, 54] and image-to-3D [6, 27, 73] models.

**Real 3D object datasets.** Acquiring real-world 3D data presents many challenges, resulting in only a few real 3D-object datasets. Early datasets like Pascal3D [70] include multiple categories but provide only isolated object views and approximate 3D annotations. Conversely, DTU [28], BlendedMVS [51], GSO [15], OmniObject3D [68], Aria Digital Twin [44], and Digital Twin Catalog [1] provide 3D scans of objects, featuring high-quality 3D geometry and textures, but have only a few thousand objects.

The use of 3D scanners significantly restricts the scale of data acquisition; consequently, other datasets capture multi-view turntable-like videos of objects using consumer cameras. CO3D and CO3Dv2 [47] crowd-sourced 40,000 360° object videos, providing 3D annotations by reconstructing point clouds and cameras using COLMAP SfM [50]. MvImgNet [76] collected even more videos (220,000) across more object categories (238), but their videos capture objects only partially, preventing full reconstruction. Objectron [2] is similar to MvImgNet, but with fewer videos (10,000). A common challenge is that large-scale datasets often rely on SfM for video reconstruction, which can lead to imprecise 3D annotations. uCO3D also employs SfM, but using VGGSfM, which has greater accuracy than COLMAP, and with a more reliable data validation setup. Furthermore, uCO3D is five times larger and significantly more diverse than CO3Dv2, encompassing 20 times more visual categories, and provides caption and 3D Gaussian Splat reconstructions of each object.

**Applications.** In order to assess the quality of uCO3D, we measure how it benefits a number of popular downstream applications. First, we consider feedforward few-view 3D reconstruction models. Among those, LRM [27] is a transformer that maps an input image to a neural radiance field supported by a triplane [7]. LightplaneLRM [6] adds splatting layers and a memory-efficient renderer. Further extensions use different representations like 3D Gaussian Splats [59, 73, 77, 80] and meshes [65, 71].

We also consider text-to-3D generators, which create 3D assets from text, and focus on the two-stages approach of Instant3D [35]. This is based on training a text-to-multi-view diffusion model [11, 48] which generates several 2D views of the object, followed by a 3D reconstruction network that outputs the 3D asset, all in a matter of seconds. The multi-view diffusion is improved in ViewDiff [26], MVDiffusion [60], IM-3D [42], CAT3D [18] and many others. AssetGen [55] further extends Instant3D by modelling material properties instead of baking in the radiance function and adds a texture refiner that outputs relightable PBR textures. As an illustration, we use uCO3D to train a model like CAT3D, which results in better new-view synthesis than the one trained on alternative datasets. We also show that the Gaussian Splat reconstructions provided with uCO3D can supervise, for the first time, an Instant3D-like pipeline using solely real-life data.

## 3. Uncommon Objects in 3D

In this section, we introduce uCO3D, our new dataset of real-life 3D objects. uCO3D comprises 360° turn-table-like videos of objects, crowdsourced and annotated with 3D cameras, point clouds, 3D Gaussians, and textual captions.

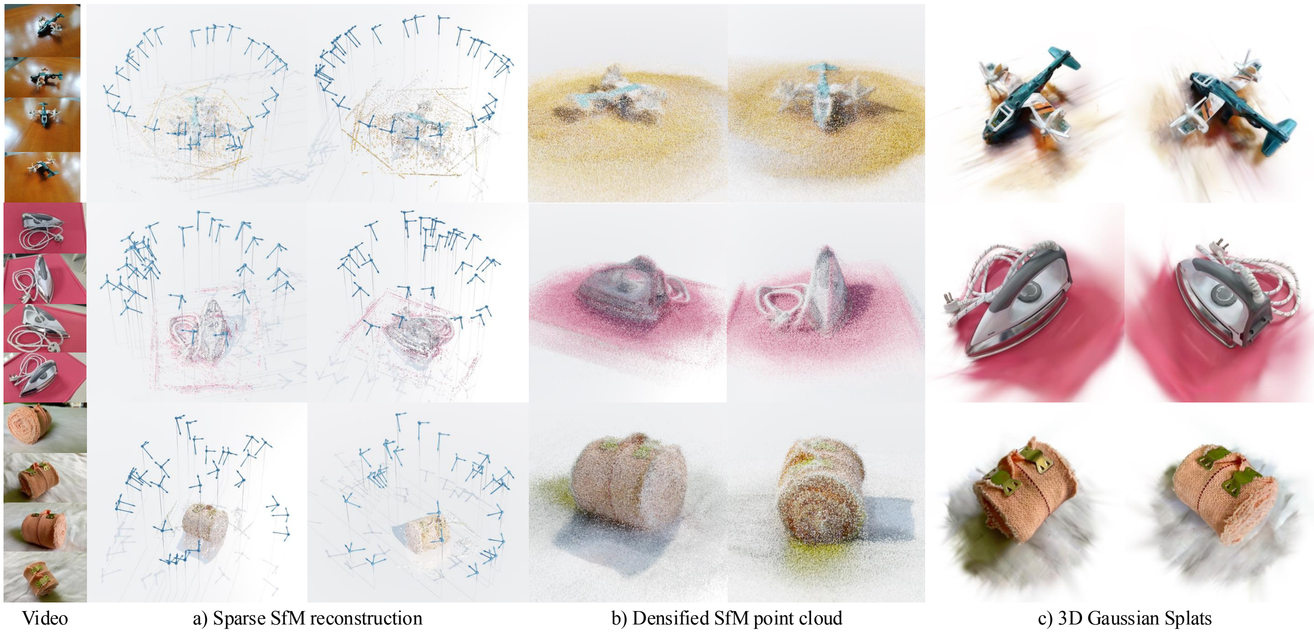


Figure 3. **Data annotation overview.** Each scene in uCO3D is reconstructed in three different ways: a) per-frame cameras with sparse point cloud calculated by VGGSfM [62], b) semi-dense point cloud comprising triangulations of additional denser tracks from VGGSfM's tracker, c) 3D Gaussian Splat [29] reconstruction optimized separately for each scene.

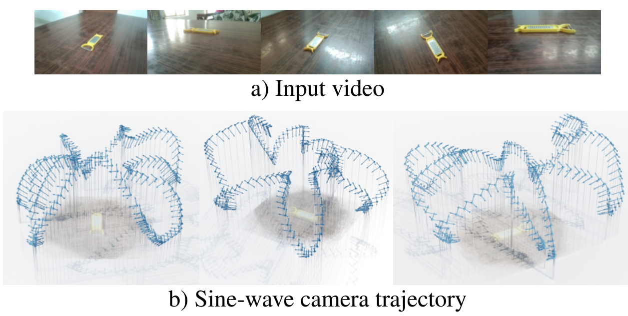


Figure 4. **Data collection example.** For each video, the cameras follow a sine-wave trajectory to ensure good viewpoint coverage.

Compared to older datasets like CO3Dv2 [47], uCO3D comes with many improvements. First, uCO3D is much larger and more diverse than CO3Dv2: it contains more than 1k different categories and more than 170k objects, compared to the 50 and 38k of CO3Dv2. While CO3Dv2's categories are taken from MS COCO [38], the categories in uCO3D are taken from the LVIS [21] taxonomy. Hence, we inherit the LVIS focus on covering the long-tail of the visual-category distribution. To simplify data analysis, we grouped the 1k+ LVIS categories to 50 super-categories, each containing approximately 20 subcategories. Figure 2 shows the number of videos collected per super-category, and the LVIS category distribution.

Second, uCO3D improves quality significantly compared to CO3Dv2, ensuring that videos are of high resolution, cover each object well, and that the 3D annotations are accurate. uCO3D also contains rich textual descriptions of each object, missing in other datasets, and useful to train large generative models. It also comes with additional 3D Gaussian Splat reconstructions of all objects, each rigidly aligned to a canonical object-centric reference, which make it possible to re-render the dataset from a fixed, canonical set of cameras, simulating synthetic data acquisition, which is very useful for training generative models [35, 55].

**Dataset collection.** Videos of objects were captured by workers on Amazon Mechanical Turk. To ensure high video quality, workers were required to submit videos of a sufficient resolution. As a result, more than 60% of videos in uCO3D are of 1080p resolution or higher, compared to 33% in CO3Dv2. Furthermore, to aid the 3D reconstruction, workers followed a sine-wave capture trajectory instead of the plain circular trajectory of CO3Dv2, ensuring varying camera elevations (cf. Fig. 4). Finally, each video was individually manually assessed to make sure that it adheres to these requirements, a process more rigorous than the rough eyeballing used in CO3Dv2 [47].

**Video object segmentation.** We used text-conditioned Segment-Anything (langSAM) [20, 32] to segment the object of interest in each video frame given text-conditioning in form of the object-category name, which had been provided by Turkers at collection time.

To improve frame-to-frame consistency, CO3Dv2 used a simple Viterbi algorithm, which often led to segmentation

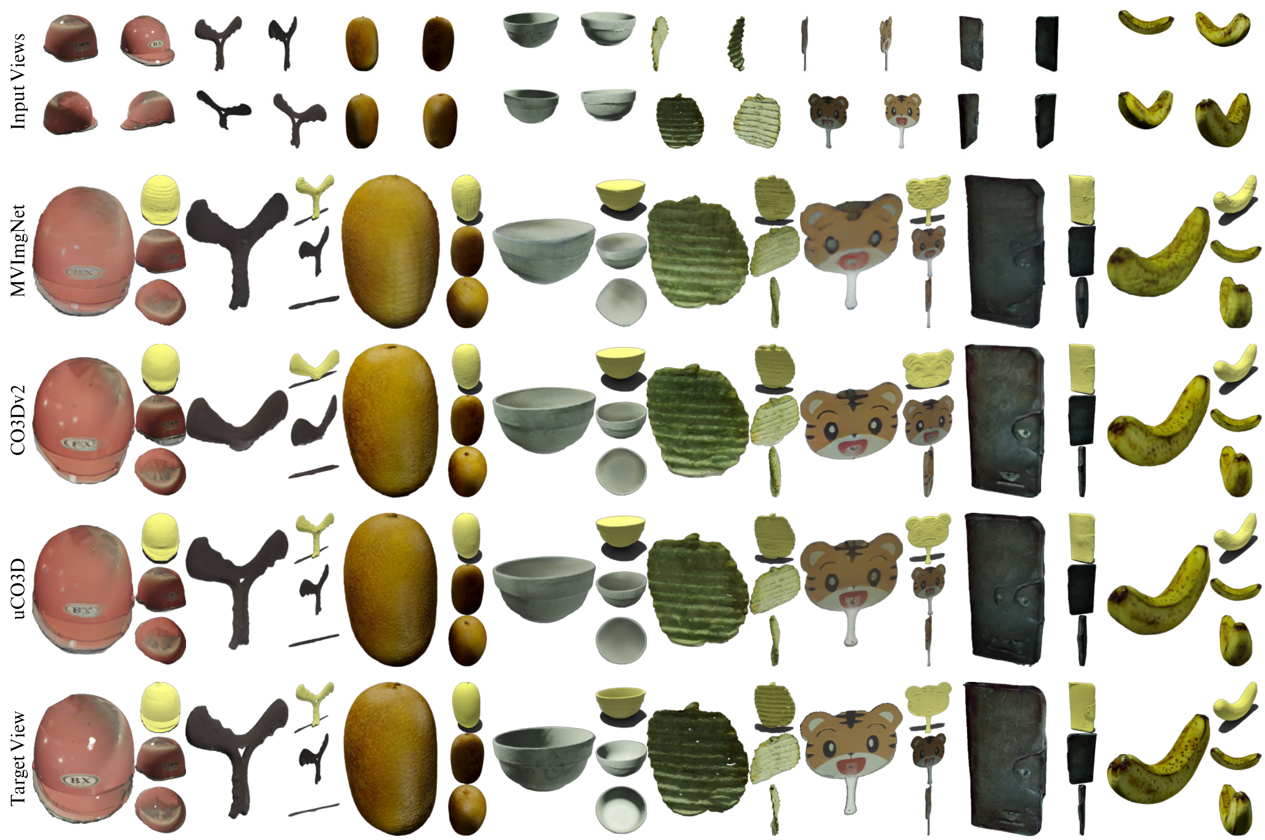


Figure 5. **3D reconstruction comparison**. We show results of LightplaneLRM [6] models trained on MVImgNet, CO3Dv2 and uCO3D.

| | OmniObject3D | | | StanfordORB | | |
|---|---|---|---|---|---|---|
| Train dataset | LPIPS↓ | PSNR↑ | IoU↑ | LPIPS↓ | PSNR↑ | IoU↑ |
| MVImgNet [76] | 0.109 | 23.39 | 0.928 | 0.070 | 24.451 | 0.939 |
| CO3Dv2 [47] | 0.095 | 23.62 | 0.926 | **0.056** | 25.617 | 0.956 |
| uCO3D (ours) | **0.093** | **24.61** | **0.946** | 0.057 | **25.715** | **0.957** |

Table 2. **3D reconstruction benchmark.** We compare LightplaneLRM [6] models trained on CO3Dv2, MVImgNet, and uCO3D. We report novel-view synthesis performances on OmniObject3D [68] and StanfordORB [34].

flickering, impairing the final 3D reconstruction quality. Instead, in uCO3D, we refine the SAM segmentations with state-of-the-art deep video-segmenter based on XMem [9], leading to more temporally-stable object segmentations.

**3D annotation with VGGSfM.** For each video, we use the state-of-the-art VGGSfM [62] Structure from Motion (SfM) system to estimate the parameters of the cameras (intrinsic and extrinsic) for 200 frames sampled uniformly. VGGSfM also outputs a sparse 3D point cloud, and its denser version obtained by triangulating additional 3D points from VGGSfM's tracker. Examples of sparse and densified SfM point clouds are shown in Fig. 3.

**Scene alignment.** While the coordinate system of VGGSfM cameras is defined only up to a rigid transformation, it is crucial for applications like generation and reconstruction to train on a dataset of rigidly aligned objects. We thus align all objects so they have a horizontal ground plane, similar scale, centring, and orientation. Details of the scene alignment procedure are in the supplementary material.

**Gaussian Splat reconstruction.** Sparse and even dense SfM point clouds provide an accurate but still quite sparse 3D reconstruction of the scene's surface. To further densify it, uCO3D provides a 3D Gaussian Splat reconstruction [29] for each scene, fitted using gsplat [75].

**Scene captioning.** uCO3D also provides textual captions for all scenes, useful for generative modelling. Motivated by Cap3D [41], we first caption each view separately using a vision-language model, and then summarise these into a single scene caption using LLAMA3 [16].

## 4. Applications

In this section, we demonstrate uCO3D's merit on three popular 3D learning tasks: feedforward sparse-view 3D reconstruction (Sec. 4.1), new-view synthesis using diffusion

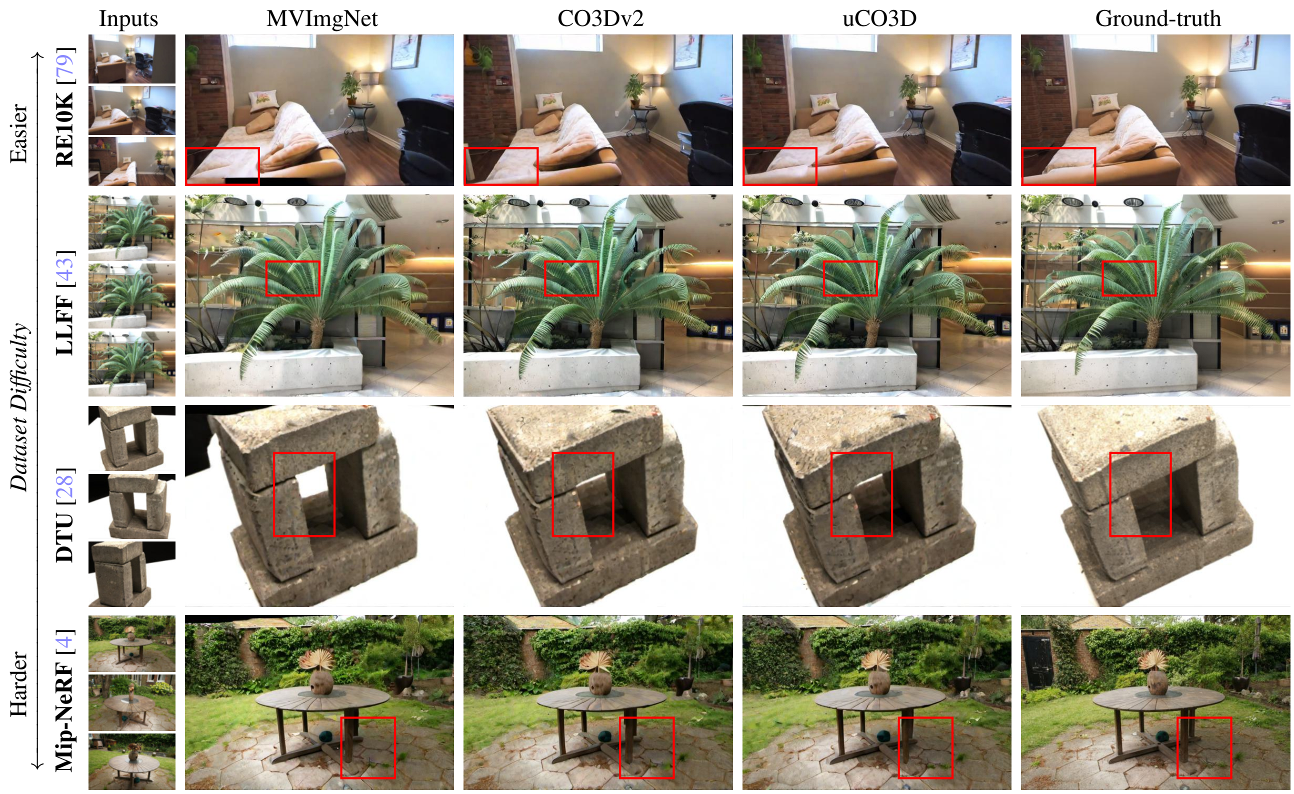


Figure 6. **Novel-view synthesis comparison.** We compare results of CAT3D-like [19] models trained on different datasets (MVImgNet, CO3D, uCO3D) and evaluated on standard NVS datasets (top-to-bottom: RealEstate10K [79], LLFF [43], DTU [28], Mip-NeRF 360 [4]).

| | Easier | | *Dataset Difficulty* | | | | Harder | |
|---|---|---|---|---|---|---|---|---|
| | **Re10K** [79] | | **LLFF** [43] | | **DTU** [28] | | **Mip-NeRF** [4] | |
| Train dataset | LPIPS↓ | PSNR↑ | LPIPS↓ | PSNR↑ | LPIPS↓ | PSNR↑ | LPIPS↓ | PSNR↑ |
| MVImgNet [76] | 0.310 | 18.77 | 0.426 | 14.38 | 0.377 | 12.79 | 0.605 | 12.39 |
| CO3Dv2 [47] | 0.281 | **20.02** | **0.418** | 14.95 | 0.329 | 16.42 | 0.532 | 14.19 |
| uCO3D (ours) | **0.278** | 19.77 | **0.418** | **15.16** | **0.315** | **16.97** | **0.528** | **14.37** |

Table 3. **Novel-view synthesis benchmark**. We evaluate CAT3D-like [19] models trained on MVImgNet, CO3Dv2 or uCO3D. We report NVS performances on RealEstate10K [79], LLFF [43], DTU [28] and Mip-NeRF 360 [4].

(Sec. 4.2), and text-to-3D (Sec. 4.3).

## 4.1. Few-view 3D Object Reconstruction

Traditionally, multi-view 3D-annotated datasets such as CO3D or MVImgNet have been used to supervise few-view 3D reconstructors. In this section, we train LightplaneLRM [6], an evolution of the seminal LRM [27], and show that doing so on uCO3D leads to better performance than training on alternative datasets.

LRM is a large transformer [61] that accepts few input images of an object and predicts a 3D representation of the latter. The transformer, conditioned on the tokens of the observed images via cross attention, converts a set of learnable input tokens to a 3D representation. The 3D representation is a triplane [7] supporting an opacity/radiance implicit shape. LightplaneLRM improves LRM by adding so called "splatting layers" and a memory-efficient renderer.

During training, LightplaneLRM receives four random source frames from a training uCO3D video sequence, and renders the predicted triplane into held-out target views. Learning minimizes the photometric loss between the renders and the corresponding ground-truth targets. Both source and target views are masked using the extracted segmentation masks to make sure that LightplaneLRM only reconstructs the foreground object. Training uses the Adam optimizer and is warm-started following the original LRM training protocol by pre-training the model on a large dataset of synthetic objects similar to Objaverse [13].

**Baselines.** Our main goal is to demonstrate that uCO3D contains higher quality data than existing object-centric datasets. As such, starting from the model pre-trained on the synthetic data, we finetune either on uCO3D, or on two other baseline datasets, namely MVImgNet and CO3Dv2.

**Evaluation protocol.** We evaluate each trained model in a novel-view synthesis setting on two small-scale high-quality object-centric datasets: OmniObject3D [68] and Stanford-ORB [34]. Given four views of a held-out test scene, the model reconstructs the scene which is then rendered to unseen target views. We report the average LPIPS [78] loss and Peak-signal-to-noise ratio (PSNR) between each render and the corresponding ground-truth image. We also report the intersection-over-union (IoU) between the rendered object alpha mask and the target view segmentation mask.

**Results.** Table 2 and Fig. 5 report the quantitative and qualitative results, respectively. The LightplaneLRM trained on uCO3D is better than the other baselines in most metrics on both datasets. The latter confirms that uCO3D is currently the most reliable source of real data for training feedforward few-view 3D reconstructors.

## 4.2. Novel-view synthesis using diffusion

We now consider application of uCO3D to training new-view image diffusion generators. These generators can, given one or a few views of an object and a target camera pose, output new arbitrary views as observed from the target camera, hallucinating missing details based on a statistical prior they learn. They can thus complement and integrate the feed-forward reconstruction models of the previous section, which are deterministic and thus unable to deal with ambiguity well. To this end, we train a diffusion model similar to the recent CAT3D [19], but reimplement it from scratch given lack of source code (see details in the supplementary). We call this model CAT3D-like.

**Evaluation protocol.** As in Sec. 4.1, we compare versions of CAT3D-like trained using uCO3D, MVImgNet, and CO3Dv2 and test them on held-out datasets. A feature of CAT3D is the ability to reconstruct both the principal object in the images as well as the background. We thus benchmark the method using new-view synthesis datasets that do contain background, namely DTU [28], containing structured light scans of various objects, LLFF [43], containing scenes captured from fronto-parallel camera trajectories, RealEstate10k [79], containing real-estate walk-throughs, and Mip-NeRF 360 [3] with complex indoor and outdoor scenes. For evaluation, we take three known views as input and use the model to predict a new view. We report LPIPS and PSNR but not the IoU since CAT3D only generates new RGB views without reconstructing the 3D shape.

**Results.** Table 3 and Fig. 6 contain the results: training CAT3D-like on uCO3D leads to the best performance across all four datasets. Even when compared to MVImgNet, which is slightly larger than uCO3D, the latter improves PSNR by 3–4 points, and reduces the LPIPS error by 5% to 20%.

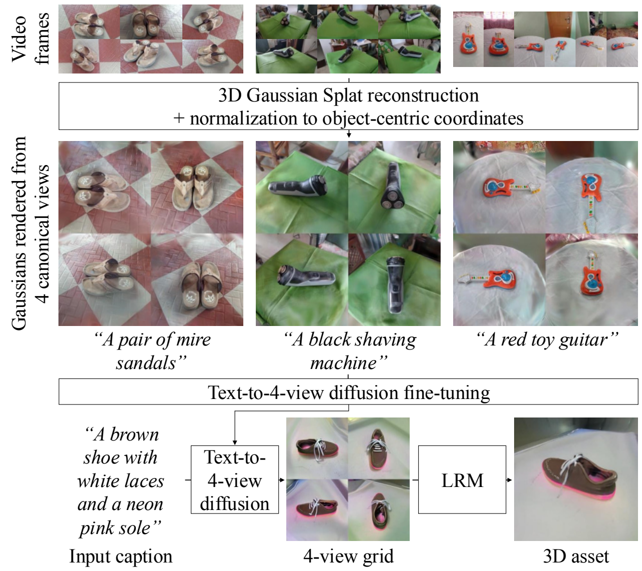


Figure 7. **Supervising Instant3D with 3DGS.** For each training scene, its 3DGS is rendered from 4 canonical views yielding a captioned image dataset for finetuning an image diffuser. Samples from the latter are then reconstructed with LRM.

## 4.3. Photorealistic Text-to-3D

Next, we show that uCO3D enables training photorealistic text-to-3D generators. Methods like CAT3D and others [39, 42, 53] generate several views of the object first, and then fit a 3D model, such as a NeRF or 3DGS, via optimization. This can work well, but it is not particularly robust or fast. An alternative, popularized by Instant3D [35] and follow-ups [54, 72], is to use a feedforward reconstructor in the second step, similar to LightplaneLRM from Sec. 4.1, which is faster and more robust. However, these models require *canonical* views of the objects — for example, Instant3D considers 4 orthogonal viewpoints, covering all 'sides'. The requirement of such training canonical views complicates training on real data, where viewpoints are arbitrary, and explains why such models are usually trained on synthetic data, limiting realism.

**Imaging 3DGS from canonical views.** Our new idea is to 're-shoot' the 3DGS reconstructions provided with uCO3D from canonical viewpoints, making our data compatible with any method requiring canonical views for training (Fig. 7). To do so, we render the normalized reference frames (Sec. 3) into four views for each object, and arrange them in a grid as a target for the text-to-4-view generator. We double check the quality of the renders by calculating their CLIP similarity [24] to the object caption, and discard a sample if this is below 0.3.

We use this data to fine-tune a text-to-4-view image diffusion model using these 4-view image grids and the corre-

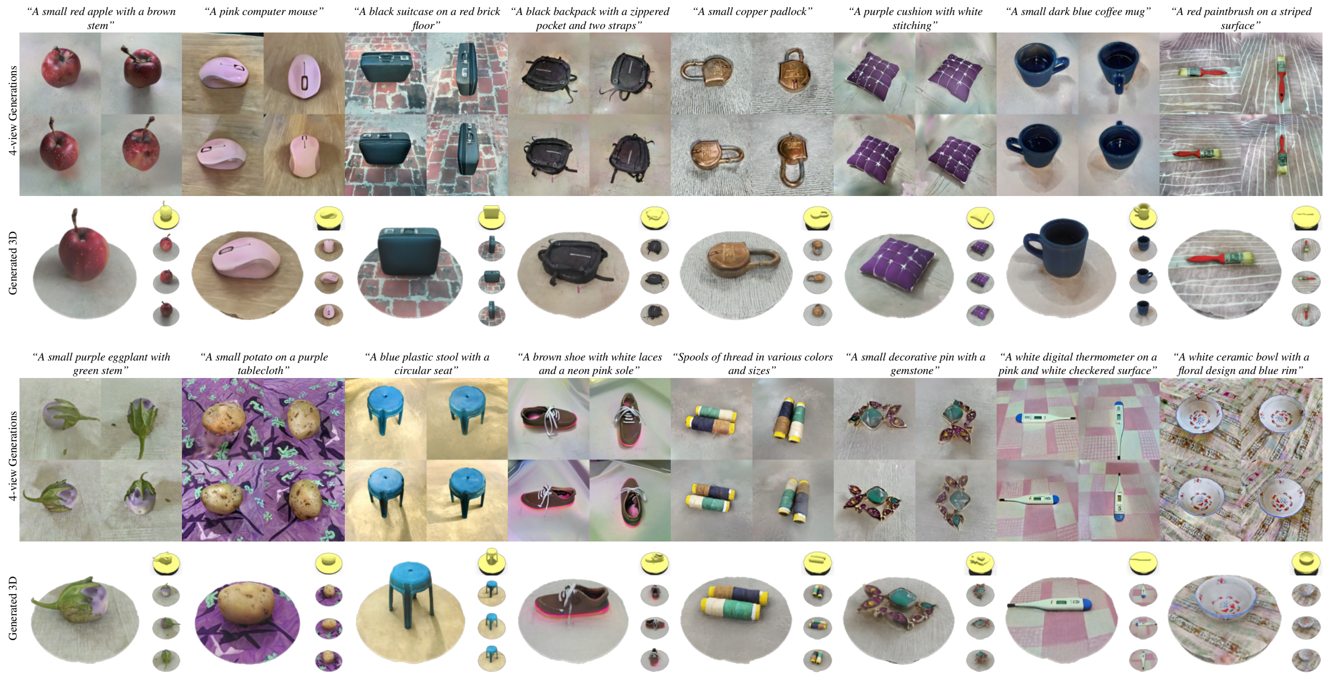


Figure 8. **Qualitative results for text-to-3D** generation displaying the 4-view grids generated by our Instant3D-like model given the input caption, and the 3D asset obtained by reconstructing the latter. The 4-view grid generator was trained using the canonical 4-view renders of uCO3D's 3DGS scene reconstructions.

| Train dataset | **Real** - FID↓ | **Surreal** - FID↓ |
|---|---|---|
| Synthetic | 82.8 | 42.3 |
| uCO3D (ours) | 63.9 | 68.9 |

Table 4. **Text-to-3D evaluation.** We compare Instant3D-like models trained on uCO3D or a dataset of synthetic renders from artist-created meshes. We report FID on two sets of data corresponding to real and surreal objects, see text for details.

sponding scene captions. At inference time, given a caption describing the desired object, we use the model to sample a new 4-view grid and feed the latter, together with the corresponding cameras, to the LightplaneLRM model (Sec. 4.1) for 3D reconstruction.

**Baselines.** We train another 4-view generator on a dataset of synthetic assets similar to Objaverse [13] and use it with the original LightplaneLRM model [6] trained on the same data and thus optimally matched to it.

**Evaluation protocol.** We report metrics evaluating the alignment between the distributions of the generations and the ground-truth objects. Specifically, we report the Frechet Inception Distance (FID) [25] between the renders of the generated 3D shapes and images of ground-truth objects.

The main purpose of this experiment is to show that, by training on the uCO3D dataset, the 3D generations are more realistic. We assess this using two sets of prompts: **Surreal**, containing 100 captions of objects from the synthetic dataset, and **Real**, containing 100 random captions form the held-out evaluation sequences of uCO3D. We report FID between the generated 3D shapes and the images/renders corresponding to the objects of each prompt-set.

**Results.** Table 4 and Fig. 8 contain the quantitative and qualitative results respectively. The table reveals that the uCO3D-trained generator outperforms the synthetic generator when evaluated on real prompts. The latter verifies our hypothesis that a generator trained on uCO3D yields more realistic samples than a model trained on synthetic data.

## 5. Conclusions

We have introduced uCO3D, a new object-centric 3D dataset of real-life objects. uCO3D strikes a balance between size and quality, ensuring the quality of the collected turntable-like videos and of the 3D annotations, while at the same time significantly expanding the scale of the data compared to CO3Dv2 and the diversity compared to CO3Dv2 and MVImgNet. We have shown the benefits of using this dataset compared to alternatives when training models for feedforward few-view 3D reconstruction, multi-view generation, and text-to-3D generation. Equipped with 3D cameras, point clouds, masks, textual captions, and 3DGS reconstructions of objects, uCO3D is a ready-to-use resource for training large generative models and for exploring 3D deep learning.

## References


[1] Aria digital twin catalog. https://www.projectaria.com/datasets/dtc/. Accessed: 2024-12-20. 3

[2] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021. 3

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proc. ICCV*, 2021. 7

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proc. CVPR*, 2022. 6

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv.cs*, abs/2311.15127, 2023. 2

[6] Ang Cao, Justin Johnson, Andrea Vedaldi, and David Novotny. Lightplane: Highly-scalable components for neural 3d fields. *arXiv*, 2024. 2, 3, 5, 6, 8, 1

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proc. CVPR*, 2022. 3, 6

[8] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet an information-rich 3d model repository. *arXiv.cs*, abs/1512.03012, 2015. 3

[9] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 5

[10] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 3

[11] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam S. Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack. *CoRR*, abs/2309.15807, 2023. 2, 3, 1

[12] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. *CoRR*, abs/2307.05663, 2023. 3

[13] 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 *Proc. CVPR*, 2023. 2, 3, 6, 8, 1

[14] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[15] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google Scanned Objects: A high-quality dataset of 3D scanned household items. In *Proc. ICRA*, 2022. 2, 3

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, and Kevin Stone. The Llama 3 herd of models. *arXiv*, 2407.21783, 2024. 5

[17] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 3

[18] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *arXiv.cs*, 2024. 2, 3

[19] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: create anything in 3d with multi-view diffusion models. *arXiv*, 2405.10314, 2024. 6, 7, 1

[20] Paul Guerrero. Lang-sam. https://github.com/paulguerrero/lang-sam, 2024. 4

[21] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[22] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proc. CVPR*, 2019. 2

[23] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[24] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. NeurIPS*, 2017. 8, 1

[26] Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. ViewDiff: 3D-Consistent Image Generation with Text-to-Image Models. *arXiv preprint*, 2024. 3

[27] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *Proc. ICLR*, 2024. 2, 3, 6

[28] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, pages 406–413, 2014. 3, 6, 7

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *Proc. SIGGRAPH*, 42(4), 2023. 2, 4, 5

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proc. ICLR*, 2015. 1

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proc. CVPR*, 2023. 4

[33] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. ABC: A big CAD model dataset for geometric deep learning. In *Proc. CVPR*, 2019. 3

[34] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-orb: a real-world 3d object inverse rendering benchmark. In *NeurIPS*, 2023. 5, 7

[35] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3D: Fast text-to-3D with sparse-view generation and large reconstruction model. *Proc. ICLR*, 2024. 2, 3, 4, 7

[36] Joseph J Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing ikea objects: Fine pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2992–2999, 2013. 3

[37] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv.cs*, abs/2305.08891, 2023. 1

[38] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. ECCV*, 2014. 4

[39] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proc. ICCV*, 2023. 7

[40] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Proc. NeurIPS*, 2022. 1

[41] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *arXiv preprint*, 2023. 5

[42] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. IM-3D: Iterative multiview diffusion and reconstruction for high-quality 3D generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 3, 7

[43] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 6, 7

[44] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Carl Yuheng Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception, 2023. 3

[45] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *Proc. ICLR*, 2023. 2

[46] Christian Reiser, Richard Szeliski, Dor Verbin, Pratul P. Srinivasan, Ben Mildenhall, Andreas Geiger, Jonathan T. Barron, and Peter Hedman. MERF: memory-efficient radiance fields for real-time view synthesis in unbounded scenes. abs/2302.12249, 2023. 1

[47] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common Objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *Proc. ICCV*, 2021. 2, 3, 4, 5, 6

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 3, 1

[49] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *Proc. ICLR*, 2022. 1

[50] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016. 2, 3

[51] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proc. ECCV*, 2016. 3

[52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[53] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *Proc. ICLR*, 2024. 2, 3, 7, 1

[54] Yawar Siddiqui, Filippos Kokkinos, Tom Monnier, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. Meta 3D Asset Gen: Text-to-mesh generation with high-quality geometry, texture, and PBR materials. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3, 7

[55] Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *NeurIPS*, 2024. 2, 3, 4

[56] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[57] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3D generative models from 2D data. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2, 1

[58] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter Image: Ultra-fast single-view 3D reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[59] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view Gaussian model for high-resolution 3D content creation. *arXiv*, 2402.05054, 2024. 3

[60] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. MVDiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv*, 2402.12712, 2024. 3

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017. 6

[62] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *CVPR*, pages 21686–21697, 2024. 2, 4, 5

[63] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proc. CVPR*, 2021. 2

[64] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proc. CVPR*, 2024. 2

[65] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. MeshLRM: large reconstruction model for high-quality mesh. *arXiv*, 2404.12385, 2024. 3

[66] Rundi Wu, Chang Xiao, and Changxi Zheng. Deepcad: A deep generative network for computer-aided design models. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 6772–6782, 2021. 3

[67] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. ReconFusion: 3D Reconstruction with Diffusion Priors. *arXiv preprint*, 2023. 2, 1

[68] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 2, 3, 5, 7

[69] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proc. CVPR*, 2015. 3

[70] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *Proc. WACV*, 2014. 3

[71] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. InstantMesh: efficient 3D mesh generation from a single image with sparse-view large reconstruction models. *arXiv*, 2404.07191, 2024. 3

[72] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 7

[73] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. GRM: Large gaussian reconstruction model for efficient 3D reconstruction and generation. *arXiv*, 2403.14621, 2024. 3

[74] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proc. CVPR*, 2024. 2

[75] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *arXiv preprint arXiv:2409.06765*, 2024. 5

[76] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, pages 9150–9161, 2023. 2, 3, 5, 6

[77] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: large reconstruction model for 3D Gaussian splatting. *arXiv*, 2404.19702, 2024. 3

[78] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, pages 586–595, 2018. 7

[79] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 6, 7

[80] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets Gaussian splatting: Fast and generalizable single-view 3D reconstruction with transformers. *arXiv.cs*, abs/2312.09147, 2023. 3