

CanoVerse: 3D Object Scalable Canonicalization and Dataset for Generation and Pose

Li Jin^{1*}, Yuchen Yang^{1*}, Weikai Chen^{2†}, Yujie Wang^{3†}, Dehao Hao⁴, Tanghui Jia⁵, Yingda Yin², Zeyu Hu², Runze Zhang², Keyang Luo², Li Yuan⁵, Long Quan⁴, Xin Wang², and Xueying Qin^{1†}

¹School of Software, Shandong University ²Lightspeed Studios

³University of North Carolina at Chapel Hill ⁵Peking University

⁴Hong Kong University of Science and Technology

{jinli, yangyc}@mail.sdu.edu.cn {weikaichen, zeyuhu}@global.tencent.com
{adamyin, ryanrzzhang, keyangluo, alexinwang}@tencent.com
yujiew.cn@gmail.com {dhaoab, quan}@cse.ust.hk
jiatanghui@stu.pku.edu.cn yuanli-ece@pku.edu.cn qxy@sdu.edu.cn

Abstract. 3D learning systems implicitly assume that objects occupy a coherent reference frame. Nonetheless, in practice, every asset arrives with an arbitrary global rotation, and models are left to resolve directional ambiguity on their own. This persistent misalignment suppresses pose-consistent generation, and blocks the emergence of stable directional semantics. To address this issue, we construct CanoVerse, a massive canonical 3D dataset of 320K objects over 1,156 categories – an order-of-magnitude increase over prior work. At this scale, directional semantics become statistically learnable: CanoVerse improves 3D generation stability, enables precise cross-modal 3D shape retrieval, and strengthens zero-shot point-cloud orientation estimation even for out-of-distribution data. This is achieved by a new canonicalization framework that reduces alignment from minutes to seconds per object via compact hypothesis generation and lightweight human discrimination, transforming canonicalization from manual curation into a high-throughput data generation pipeline. Project page: <https://github.com/123321456-gif/Canoverse>

Keywords: 3D Canonicalization · Pose Estimation · Large-scale Dataset

1 Introduction

A persistent source of uncontrolled variation in 3D vision is *orientation*. Web assets, 3D scans, and 3D generative models all appear in arbitrary unknown frames. While scale and translation are routinely normalized, orientation is often ignored, yet it systematically degrades representation stability. This is not a minor nuisance: orientation variation could fragment shape retrieval (the same

*: equal contributions, †: corresponding author



Fig. 1: We present CanoVerse, a large-scale canonical dataset standardizing object orientation, size, and position with intra- and inter-category alignment. It contains 1,156 categories and 320k objects, making it the largest canonical dataset to date, an order of magnitude larger than existing ones.

instance becomes dozens of rotated identities), confuse generative models (global pose becomes inconsistent and uncontrollable), and contaminate cross-category evaluation (generalization cannot be measured fairly). Crucially, directional semantics like “front” or “up” cannot be discovered from geometry alone. They only become learnable priors when they appear consistently across *many* canonicalized instances. In other words, canonical orientation must exist at *dataset scale* before 3D models can internalize reliable orientation-consistent 3D semantics.

However, existing canonical 3D datasets remain limited in scale: 32K (COD [11]), 14.8K (Objaverse-OA [18]), and 6K (OmniObject3D [35]) – numbers that are two orders of magnitude smaller than the raw 3D asset universe. The reason is simple: current canonicalization workflows still rely heavily on manual alignment, making it prohibitively expensive to scale up the dataset. Consequently, canonical orientation today exists only in small pockets of data, and is far too sparse to support robust generalization studies or to serve as a foundation for orientation-consistent 3D representation learning.

To address these limitations, we introduce a scalable canonicalization framework that dramatically reduces human cost of alignment: manual annotation time per object drops from minutes to seconds. In particular, we design a tailored pipeline that reformulates alignment as hypothesis generation plus lightweight human discrimination. Unlike prior efforts that primarily rely on geometry heuristics or fully manual alignment, Our framework jointly leverages *geometry* and *semantic* cues, and designs category-specific judgment criteria for both horizontal and vertical dimensions, to generate a small set of canonical pose candidates that include the ground-truth canonical pose. This hybrid strategy preserves the accuracy of human judgment while eliminating the search over $SO(3)$, yielding substantially higher throughput with high coverage and accuracy.

With this engine in place, we canonicalize the Objaverse and Objaverse-XL collection and construct *CanoVerse* – **320K canonicalized objects across 1,156 categories** – which, to our knowledge, is the largest canonical 3D dataset

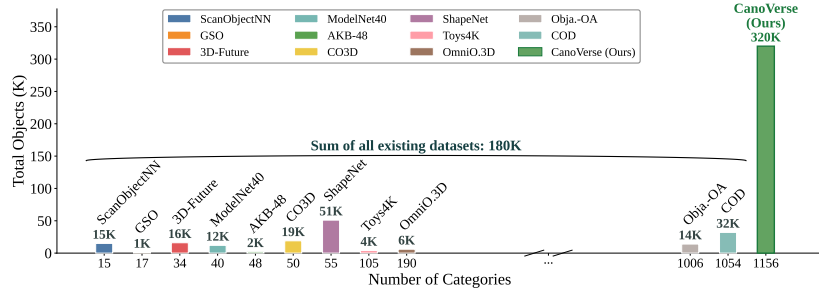


Fig. 2: Scale and category coverage of different canonical datasets. Our dataset boasts the largest number of categories and objects among existing canonical datasets—far surpassing each individual one and even the total of all aforementioned datasets.

to date (See Figure 2). CanoVerse provides sufficient statistical mass to make directional semantics learnable across broad coverage. At this scale, directional semantics become statistically learnable. We observe substantial gains across multiple tasks, including improved point-cloud orientation estimation with stronger generalization to unseen objects and real-world scanned out-of-distribution data, more orientation-stable 3D generation when models are trained on canonical data, and more accurate cross-modal 3D shape retrieval. Finally, CanoVerse exhibits higher canonicalization quality than COD [11] and Objaverse-OA [18] (the largest and latest canonical dataset with a comparable number of covered categories) under quantitative comparison, and a user study confirms that our pipeline reduces annotation time from minutes to seconds per object while maintaining alignment reliability. We summarize our contributions as follows:

- A scalable canonicalization framework. We propose a novel pipeline that fuses geometric and semantic cues, designs multiple canonicalization criteria, generates canonical candidates, and converts 3D object canonical annotation from 3D continuous optimization to 2D selection, cutting per-object manual annotation time from minutes to seconds.
- A large-scale canonical 3D dataset coded CanoVerse with 320K objects across 1,156 categories – an order-of-magnitude increase over prior canonical datasets.
- Substantial gain on downstream tasks. CanoVerse improves 3D generation stability, scales to cross-modal 3D shape retrieval, and, critically, enables zero-shot point cloud orientation estimation, which was previously impractical due to insufficient canonical data.

2 Related Work

3D Canonicalization Datasets. Early 3D canonical datasets such as ShapeNet [1], ModelNet40 [36], and OmniObject3D [35] have made significant contributions

to related research. However, these datasets are limited in scale: for example, ShapeNet contains approximately 51k objects across 55 categories, ModelNet40 includes 12k models spanning 40 categories, and OmniObject3D comprises 6k objects covering 190 categories. Such scales remain insufficient to meet the data demands of large-model training, making it difficult to enable directional semantics to be learnable across broad category coverage. The key reason for the difficulty in large-scale expansion of these datasets lies in their heavy reliance on manual alignment, which incurs high alignment costs. The COD [11] dataset adopts a one-shot method for object alignment, eliminating manual costs but suffering from poor alignment quality. In contrast, the method proposed in this paper can ensure high alignment quality while significantly reducing manual alignment costs: it shortens the alignment time from the minute level to the second level, greatly improving data annotation efficiency and thus enabling the expansion of larger-scale datasets.

Object Orientation Estimation. A feasible approach for object orientation estimation is to leverage 6D object pose estimation methods, which can be broadly divided into two categories. Instance-level pose estimation [12, 20, 26, 32] predicts relative poses based on known 3D CAD models or reference images. However, in practical scenarios, such models or references are often unavailable, limiting the applicability of these methods. To improve generalization, category-level pose estimation methods [3, 8, 14, 30] learn shape priors shared within a category. Additionally, Orient Anything [31] trains an object orientation estimation model using large model-processed data. While more flexible, they require large-scale orientation-aligned 3D datasets, which are costly to construct and thus remain scarce. In contrast, our method enables the creation of large-scale canonicalized datasets covering diverse categories and instances, allowing category-level pose estimation methods to scale beyond the limited category coverage of existing approaches.

3D Object Generation. Recent advances in 3D generation have largely followed the latent VAE-diffusion paradigm [2, 10, 13, 33, 43], inspired by the success of 2D diffusion models [23]. Current frameworks can be broadly grouped into VecSet-based approaches [13, 39, 41, 43], which encode shapes as latent vector sets and decode them via neural implicit functions, and sparse-voxel-based approaches [22, 34, 37], which use structured voxel grids for explicit spatial reasoning. While these systems achieve impressive geometric and visual fidelity, they remain sensitive to inconsistencies in data orientation and geometry. As training assets often lack a consistent canonical frame, generative models tend to produce unstable poses and duplicated symmetric parts, highlighting the need for large-scale, orientation-consistent data supervision. By providing 320K canonicalized objects with consistent global orientation, CanoVerse can directly address this limitation, enabling 3D generators to internalize directional priors and produce pose-stable and coherent results.

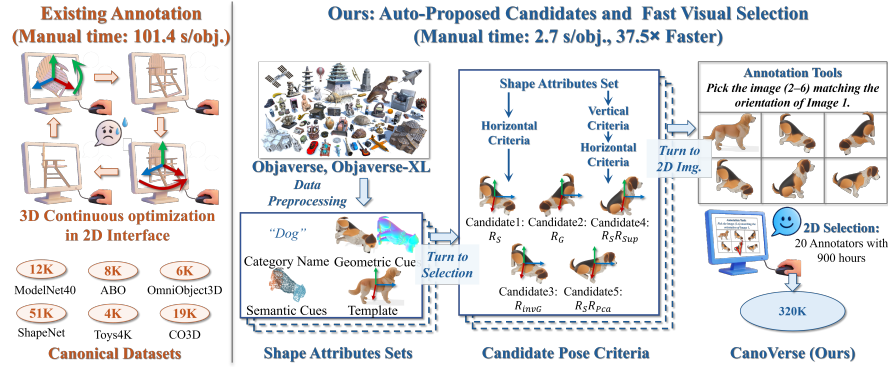


Fig. 3: 3D canonicalization pipeline. Left: prior manual alignment (~ 101.4 s/object) yields small-scale datasets. Right: our candidate generation (~ 3 s) plus one-click selection (~ 2.7 s) gives $\sim 37.5\times$ faster annotation and enables a 320K canonical dataset.

3 Method

We develop a two-stage pipeline for large-scale, category-level canonicalization (Figure 3): **(I) candidate-pose generation** (Section 3.1) and **(II) interactive selection** (Section 3.2). Stage I operates in a *one-shot, category-level* manner: given a single template per category, it automatically produces a compact set of reliable pose candidates per instance. Stage II provides an annotator-friendly GUI for one-click selection among the top-5 candidates produced from Stage I. This design amortizes effort while preserving accuracy at scale.

3.1 Candidate-Pose Generation

Data Preprocessing. We extract geometric and semantic cues, per-category templates, and (for symmetric objects) symmetry axes and angles. For each shape, we sample a colored point cloud and normalize its translation and scale to a unit sphere, then use Uni3D [44] for category assignment and Find3d [19] for semantic part segmentation; the per-category semantic text used in segmentation follows COD [11] and is obtained via large language models.

For each category, we choose a representative template as the canonical frame, with orientation following human convention—for a bicycle, the z-axis is front and the y-axis is up. We harmonize across categories so that related ones share aligned axes; the front of cars and bicycles, for instance, both map to the z-axis. For symmetric objects, following NOCS [30], we define a symmetry axis and rotational symmetry angle per category (Figure 4, left), shared by all instances; defining these once per category keeps annotation tractable. Extended preprocessing details are in the supplementary material.

Candidate Pose Criteria. Humans align objects based on different cues depending on the category. For example, cameras are oriented according to seman-

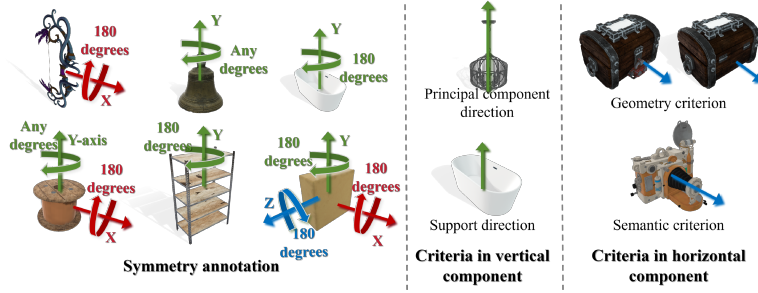


Fig. 4: For symmetric categories, we annotate the symmetry axis and angle (left). Due to diverse object shapes, unified annotation is impractical. We thus design separate standards based on object characteristics and generate candidate poses in the vertical (middle) and horizontal (right) directions.

tic features such as lens direction, while cartons are aligned based on geometric structure. Consequently, there is no universal rule for canonicalizing 3D object orientations across diverse categories. To ensure that the canonical pose is covered within a candidate set, we design multiple alignment criteria capturing both geometric and semantic cues commonly used in human perception. We decouple object orientation into horizontal and vertical components (Fig. 4, middle and right) and define separate alignment criteria for each. Candidate poses are then generated by applying these criteria independently or sequentially to the input object. These criteria form the basis of our candidate pose generation stage.

Criteria in Vertical Component. Vertical canonicalization with respect to gravity generally falls into two categories. The first aligns objects using a supporting surface, while the second relies on principal component analysis (PCA). For example, a cup can be canonicalized by placing it stably on a horizontal plane, whereas swords or chandeliers require alignment via suspension.

For the first category, we compute candidate supporting surfaces that allow the object to remain in static equilibrium. For a rigid body, the center of mass must lie within the supporting polygon [9]. We therefore enumerate faces satisfying this condition and restrict to poses where the supporting face is parallel to the ground. The resulting candidate poses are rendered, and a vision-language model (VLM) selects the one that best matches human expectations for upright orientation. The corresponding rotation is denoted as $\mathbf{R}_{\text{Sup}}$. Implementation details are provided in the supplementary material.

For the second category, we align the object to the category template using Principal Component Analysis (PCA). Specifically, we compute the first two principal axes of both the object and the template from the covariance eigen-decomposition of their centered point clouds. As PCA eigenvectors are sign-ambiguous, the object axes admit four polarity combinations relative to the template axes, yielding four candidates:

$$\mathbf{R}_i = [\mathbf{v}_1^t \ \mathbf{v}_2^t \ \mathbf{v}_1^t \times \mathbf{v}_2^t] [\sigma_{i1} \mathbf{v}_1^o \ \sigma_{i2} \mathbf{v}_2^o \ \sigma_{i1} \mathbf{v}_1^o \times \sigma_{i2} \mathbf{v}_2^o]^\top, \quad i \in \{1, 2, 3, 4\}, \quad (1)$$

where $\mathbf{v}_1^o, \mathbf{v}_2^o$ and $\mathbf{v}_1^t, \mathbf{v}_2^t$ denote the first two principal axes of the object and template point clouds, respectively, and $\sigma_{i1}, \sigma_{i2} \in \{+1, -1\}$ resolve the sign ambiguity of PCA eigenvectors.

We resolve this ambiguity by utilizing semantic cues to determine the correct polarity of each axis. The semantic part distributions of the object and template provide additional constraints for calibrating the object’s axial directions. The optimal rotation is selected by minimizing the semantic alignment error:

$$\mathbf{R}_{\text{Pca}} = \arg \min_{\mathbf{R}_i} \frac{1}{m} \sum_{k=1}^m \text{CD}(\mathcal{S}_k^t, \mathbf{R}_i \mathcal{S}_k^o). \quad (2)$$

m is the semantic part amount, $\mathcal{S}_k^o$ and $\mathcal{S}_k^t$ are the k -th semantic point sets of the object and template, and $\text{CD}(\cdot, \cdot)$ is the Chamfer distance. This selects the polarity combination that best aligns the semantic part layout with the template.

Criteria in Horizontal Component. Similarly, we introduce two criteria for horizontal alignment, leveraging geometric and semantic cues to align object orientations with the category template.

Specifically, the geometry criterion mainly aligns the categories with a similar shape and topology. For instance, Benches share a strip-like geometric form in shape and a support structure composed of legs and a backrest frame in topology. Thereby, we evaluate the geometric similarity by calculating the chamfer distance between the object and the template along the horizontal direction, and obtain the alignment results by selecting the most similar candidate. The geometry criterion is: $\theta^* = \arg \min_{\theta \in [0, 2\pi)} \text{CD}(\mathcal{P}^t, R_z(\theta) \mathcal{P}^o)$, where θ is the horizontal rotation angle around the vertical axis z ; $R_z(\theta)$ is rotation matrix for horizontal rotation; $\mathcal{P}^o, \mathcal{P}^t$ is point sets of the target object and template, respectively; $\text{CD}(\cdot, \cdot)$ is Chamfer Distance, which measures the geometric similarity between two point sets; θ^* is optimal horizontal rotation angle that minimizes the Chamfer Distance. Substituting θ^* into $R_z(\theta)$, we obtain the geometric alignment matrix $\mathbf{R}_G$ and its 180° inverted version $\mathbf{R}_{\text{inv}G}$ for symmetric objects:

$$\mathbf{R}_G = R_z(\theta^*), \quad \mathbf{R}_{\text{inv}G} = R_z(\theta^* + \pi) \quad (3)$$

In addition, in categories with significant shape diversity, the semantics criterion play a dominant role in guiding alignment. For instance, cameras can be aligned by lens’ orientation, a core semantic cue, regardless of whether they are instant cameras with miniaturized lenses or telephoto cameras with a bulky lens. Therefore, inspired by COD [11], we use the joint energy function combined semantic and geometry to fully leverage the dominant semantic guidance while compensating for the lack of fine-grained geometric details in pure semantic alignment. The semantic criterion is defined as follows:

$$E_g(\theta) = \text{CD}(\mathcal{P}^t, R_z(\theta) \mathcal{P}^o), \quad E_s(\theta) = \frac{1}{m} \sum_{k=1}^m \text{CD}(\mathcal{S}_k^t, R_z(\theta) \mathcal{S}_k^o) \quad (4)$$

$$\theta^* = \arg \min_{\theta \in [0, 2\pi)} - \left(e^{-E_s(\theta)} \cdot \sum_{\omega_k \in \Omega} \mathcal{N}(\theta \mid \omega_k, \mathbf{I}) \right)$$

3.2 Interactive Selection

Task Formulation. During annotation, annotators perform a discrete visual selection task. For each instance, annotators are presented with the reference canonical template and renderings under 5 candidate rotations, and are asked to select the best-aligned one.

Annotation Protocol. To ensure consistent understanding, each annotator completes a short calibration before large-scale annotation; we monitor per-category consistency and conduct lightweight rechecking for low-confidence categories. Annotators discard instances that are misclassified or exhibit severe geometric degradation such as substantial mesh loss or incomplete structure. Protocol details, including calibration setup and workforce scale, are given in the supplementary material.

Quality Control and Annotation Statistics. Across 750K annotated samples, 320K are retained after screening; the average time per instance is under 5 seconds. The selection distribution over the five candidates (Fig. 5) shows that all candidate types contribute to final alignment decisions. Our final dataset covers a large number of categories with a long-tailed distribution (Fig. 6) and enriches most categories compared to COD [11].

4 Experiments

We evaluate CanoVerse on three downstream tasks: 3D object orientation estimation (Section 4.1), 3D object generation (Section 4.2), and 3D shape retrieval (Section 4.3). Two additional studies assess the efficiency and quality of our annotation pipeline (Section 4.4).

4.1 3D object Orientation estimation

In 3D object orientation estimation, the objective is to recover the rotation that aligns an arbitrarily orientation object to the canonical space. We evaluate orientation estimation under three settings to validate dataset quality, cross-dataset transferability, and performance scaling with increasing data volume.

In-Distribution Evaluation. We randomly split 100k samples from our curated dataset for training and 10k samples for held-out testing. Based on this split, we retrain multiple state-of-the-art (SOTA) methods, including CleanPose [16], AGPose [17], SecondPose [4], and VI-Net [15], using identical training configurations. Detailed hyperparameters are put in the supplementary material.

Results. As shown in Table 1, all learning-based methods trained on our dataset outperform the traditional PCA method, demonstrating that geometric cues alone cannot ensure robust canonical alignment for large-scale objects with complex shapes. Among network-based approaches, VI-Net trained on our Canoverse dataset achieves the best performance across all evaluation metrics and excellent generalization over diverse object categories. Compared with models trained on the recent dataset Objaverse-OA [18], our method shows clear advantages, verifying its effectiveness.



Fig. 7: Qualitative comparison of the performance between our model and Orient Anything [31] on the OmniObject3D [35] dataset shows that the model trained on our dataset has better generalization ability.

Table 1: In-distribution comparisons for orientation estimation. Best results in **bold**.

Method	Acc@10° ↑	Acc@30° ↑	Abs ↓
PCA	22.98	34.83	91.49
CleanPose(CanoVerse) [16]	35.02	57.36	60.72
AGPose(CanoVerse) [17]	23.80	48.88	70.39
SecondPose(CanoVerse) [4]	49.15	58.39	60.11
VI-Net(Objaverse-OA) [15]	21.77	25.86	97.69
VI-Net(CanoVerse) [15]	55.31	64.14	51.47

Out-of-Distribution Generalization. To evaluate cross-dataset transferability, we test models trained on our dataset on OmniObject3D [35], which is reconstructed from real-world scans, serving as an independent out-of-distribution benchmark. We additionally compare against two mainstream general-purpose 6D pose estimation methods, including GenPose++ [40] and Orient Anything [31]. For a fair comparison, we provide the full object point cloud to GenPose++, while for Orient Anything, we use the croplargeEX2 version for experiments, and we render six orthogonal views and select the highest-confidence prediction as the final output.

Results. As summarized in Table 2, VI-Net retrained on our dataset significantly outperforms the same model trained on NOCS and Objaverse-OA when evaluated on unseen OOD data. Furthermore, compared with other large-scale 6D pose estimation methods, it demonstrates superior cross-distribution generalization. These results highlight the advantages of richer category coverage and large-scale canonical annotations. Figure 7 further confirms that models trained on our dataset produce more consistent and accurate object orientations across in-distribution samples, out-of-distribution objects, and generative outputs.

Scalability Analysis. To quantify data scaling effects, we train separate models using 32k, 100k, 200k, and 310k training samples, respectively, while keeping the evaluation protocol unchanged. We train and evaluate using the VI-Net [15] model and test on the in-distribution dataset.

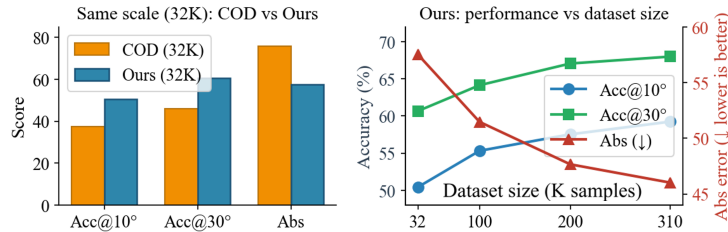


Fig. 8: Scalability Analysis on the in-distribution dataset. Left: COD vs. Ours at 32K samples. Right: Ours performance versus dataset size (32K $\rightarrow$ 310K). For metrics: higher Acc@10° and Acc@30° are better; lower Abs error is better.

Table 2: Out-of-distribution orientation estimation results on the real-world scanned dataset OmniObject3D [35].

Method	Acc@10° $\uparrow$	Acc@30° $\uparrow$	Abs $\downarrow$
GenPose++ [40]	4.33	6.95	105.17
Orient Anything [31]	4.77	7.32	103.11
VI-Net(Nocs) [15]	3.92	6.64	108.51
VI-Net(Objaverse-OA) [15]	6.56	9.60	105.90
VI-Net(CanoVerse) [15]	20.18	24.83	94.60

Results. As shown in Fig. 8, orientation estimation accuracy improves steadily with increasing training data. At the same data scale (32K), the model trained on our dataset outperforms its counterpart trained on COD, highlighting the higher quality of our annotations. This controlled scaling study demonstrates the data efficiency and scalability of our canonical dataset, suggesting that larger-scale 3D datasets can further benefit downstream 3D tasks.

4.2 3D Object Generation

To evaluate the impact of canonical data on 3D object generation, we fine-tune two representative generative backbones—Hunyuan3D 2.1 [29] and Trellis [37], on 100K samples from CanoVerse, limited by available computational resources. Specifically, the task follows a straightforward input-output paradigm: input an image and output a generated 3D asset.

Experiment Settings. To ensure fair comparison, we conduct strictly controlled single-variable experiments, where data canonicalization is the only varying factor. All fine-tuning runs share identical 3D assets, training configurations, and optimization steps. We directly compare models fine-tuned on raw (non-canonicalized) data against those fine-tuned on canonical CanoVerse data. Evaluation is performed on the public Toys4K [25] and GSO [7] datasets.

Metrics. We evaluate geometric and perceptual consistency using Chamfer Distance (CD) [42], LPIPS [42], and CLIP Score [21]. Pose stability is measured by the interquartile range (IQR) of angular errors between generated object poses and canonical poses. CD is computed on 100K mesh points sampled after normal-

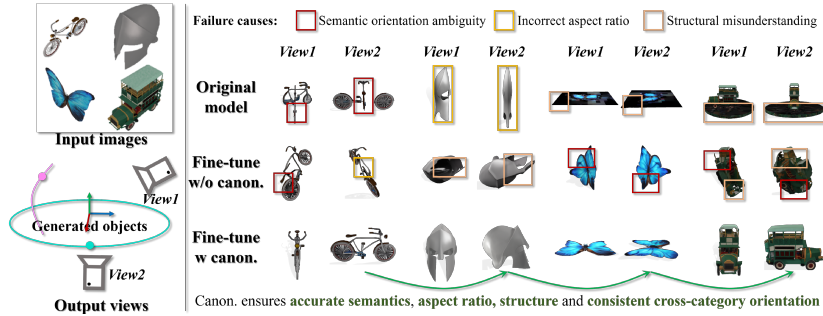


Fig. 9: Qualitative analysis of 3D object generation. Fine-tuning with canonical representation outperforms fine-tuning without it in semantics, aspect ratio, and structure.

Table 3: 3D object generation comparison results. Lower CD, CD_{ICP} , CD_{PCA} , and LPIPS indicate better geometric quality; higher CLIP scores indicate better performance; lower IQR ($^{\circ}$) indicates better pose stability. CD_{ICP} and CD_{PCA} denote Chamfer distance after ICP and PCA alignment, respectively. All variants are fine-tuned for the same total training time.

Method	canon.	Toys4K						GSO					
		CD↓	CD_{ICP} ↓	CD_{PCA} ↓	LPIPS↓	CLIP↑	IQR($^{\circ}$)↓	CD↓	CD_{ICP} ↓	CD_{PCA} ↓	LPIPS↓	CLIP↑	IQR($^{\circ}$)↓
Trellis [37]	w/o	0.227	0.181	0.154	0.436	89.7	48.5	0.272	0.170	0.156	0.468	88.0	74.0
	w	0.095	0.087	0.122	0.254	93.9	7.7	0.072	0.064	0.099	0.244	92.5	8.1
Hun. 2.1 [29]	w/o	0.258	0.163	0.156	0.413	90.8	75.6	0.244	0.174	0.156	0.471	88.1	74.0
	w	0.085	0.084	0.115	0.260	93.7	8.1	0.090	0.089	0.144	0.288	90.6	18.6

ization to $[-1,1]$. For LPIPS and CLIP, meshes are normalized to $[-1,1]$, rendered from six orthogonal views, and averaged across views to evaluate appearance fidelity and orientation alignment. For pose stability, we randomly sample 100 objects, obtain their ground-truth canonical poses from our annotation pipeline, compute angular errors with respect to these poses, and report the IQR.

Results. As illustrated in Fig. 9, models fine-tuned on canonical data generate shapes with more consistent orientations and structurally coherent geometries. Canonical training reduces pose-induced ambiguity during learning, leading to improved part structure and proportion consistency. This also enhances the model’s semantic understanding of object structures, yielding higher-fidelity 3D shapes with fewer structural ambiguities, artifacts, and unreasonable deformations. Quantitative results in Table 3 further confirm that, across both generative backbones, canonical fine-tuning consistently improves orientation alignment, geometric accuracy, and perceptual consistency. These findings highlight the practical benefits of canonicalized data for stable and high-quality 3D generation.

4.3 Cross-modal 3D Shape Retrieval

We evaluate two cross-modal 3D shape retrieval tasks: text-to-3D and image-to-3D. Given a text or image query, the goal is to retrieve the corresponding 3D

instance. We examine whether canonicalization reduces directional ambiguity caused by non-canonical training data and improves cross-modal alignment.

Experimental Setting. We randomly sample 310k canonical instances from CanoVerse for training and 10k for testing. Each sample forms a triplet of a point cloud, rendered images, and text descriptions generated by Qwen2.5-VL [28]. For a strictly controlled comparison, we construct the non-canonical baseline using the same underlying Objaverse assets but apply only scale normalization without canonical alignment, ensuring that canonicalization is the sole varying factor. We fine-tune ULIP [38] and Uni3D [44] under identical configurations. In addition to in-distribution evaluation on our test split, we conduct out-of-distribution testing on OmniObject3D [35]. We report Recall@10 and Recall@30 based on exact instance matching. More details are in the supplementary material.

Table 4: Quantitative Comparison of Retrieval Performance. Higher Recall@10 and Recall@30 scores indicate better retrieval performance.

Method	Canon.	Modality	Canoverse		OmniObject3D	
			Re.@10↑	Re.@30↑	Re.@10↑	Re.@30↑
ULIP [38]	w/o	Text	21.33	35.17	8.19	15.98
Uni3d [44]	w/o	Text	51.29	62.18	34.21	48.38
ULIP [38]	w	Text	25.05	39.43	9.95	19.89
Uni3d [44]	w	Text	53.88	64.31	35.88	49.02
ULIP [38]	w/o	Image	19.03	32.39	8.46	16.55
Uni3d [44]	w/o	Image	79.07	92.29	61.34	77.42
ULIP [38]	w	Image	23.63	39.04	9.66	21.50
Uni3d [44]	w	Image	81.82	92.98	64.86	81.50

Results. As shown in Table 4, models trained on canonical data consistently outperform those trained on non-canonical data across both ULIP [38] and Uni3D [44], for text-to-3D and image-to-3D retrieval on both Canoverse and OmniObject3D [35]. These results suggest that canonicalization reduces directional ambiguity and enables models to better capture intrinsic 3D shape semantics, thereby improving cross-modal alignment and generalization.

Canonical annotation methods fall into two categories: (1) **manual annotation**, which is widely adopted in existing datasets including ShapeNet [1] and OmniObject3D [35]; (2) **self-supervised automatic canonicalization methods**, which remain in the research stage (e.g., CaCa [27], ConDor [24]) with limited real-world deployment. To validate the effectiveness of our proposed method, we conduct comprehensive comparative experiments against both categories.

Experimental Setup. Following COD [11], we conduct all experiments on the DREDS dataset [5], which contains seven object categories. All methods are evaluated on objects that are randomly rotated beforehand.

For comparison with **manual annotation** pipelines, we consider two baselines: Blender-based manual alignment and the COD candidate pose selection pipeline. In COD, four initial random rotations are applied to each object, which are then processed to generate candidate canonical poses, from which annota-

Table 5: Annotation Speed and Precision Comparison. Smaller is better for both speed (s/object) and precision ($^{\circ}$). Best results are in **Bold**, and second best is underlined.

Method	Speed (s/object)			Precision ($^{\circ}$)		
	Low	Medium	High	Low	Medium	High
Blender (Manual)	125.3	101.4	94.5	<u>8.2</u>	4.5	3.7
COD [11]	<u>4.6</u>	<u>3.4</u>	<u>3.0</u>	9.4	8.5	8.1
Ours	4.0	2.7	2.6	6.7	<u>6.3</u>	<u>5.9</u>

Table 6: Comparison of 3D object canonicalization on the DREDS [5] dataset.

Method	Aeroplane		Car		Bowl		Camera		Bottle		Can		Mug		Avg	
	IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC
CaCa [27]	0.785	0.869	1.985	1.881	1.569	1.895	1.294	1.451	2.632	3.080	1.797	2.095	0.661	0.655	1.532	1.704
ConDor [24]	0.823	1.391	0.605	0.782	0.898	1.202	1.035	1.529	1.108	1.261	1.048	1.753	0.126	0.189	0.806	1.158
ShapeMat. [6]	0.659	2.337	2.409	3.596	0.368	0.452	0.377	1.491	1.279	3.791	1.369	2.211	0.399	0.629	0.980	2.072
COD [11]	0.051	0.058	0.103	0.118	0.011	0.012	1.116	1.177	0.031	0.034	0.030	0.037	0.018	0.019	0.194	0.208
Ours	0.129	0.176	0.007	0.012	0.018	0.032	0.033	0.584	0.009	0.012	0.005	0.017	0.022	0.026	0.032	0.123

tors select the best one. We evaluate both annotation efficiency and precision. Ten annotators with no prior pose annotation experience are recruited and given standardized training, and their proficiency is evaluated across multiple annotation rounds. See the supplementary material for more details. For **automatic annotation** methods, we follow the evaluation protocol of COD [11] and adopt two metrics for canonicalization accuracy: Instance-Level Consistency (IC) and Ground Truth Equivariance Consistency (GEC).

Results. Compared with manual annotation methods (Table 5), Blender-based manual canonicalization is intuitive but extremely time-consuming. Our method significantly improves efficiency, requiring only 2.6 seconds per object for high-skill annotators compared to 94.5 seconds for manual annotation (**36 $\times$ speedup**), and achieving an even larger improvement for low-skill annotators (**31 $\times$**). Despite this speedup, our method maintains competitive precision (5.9 $^{\circ}$ for the high-skill group), close to medium-skill manual annotation (4.5 $^{\circ}$), while significantly outperforming the COD baseline. Moreover, training annotators for image selection is substantially easier than learning 3D software operations. Considering that the DREDS dataset itself is manually annotated and object alignment can be inherently ambiguous across different geometries, angular errors within 10 $^{\circ}$ are regarded as acceptable systematic fluctuations.

4.4 Annotation Quality and Efficiency

Against automatic annotation methods (Table 6), our method achieves strong overall performance, improving IC by 83.5% and GEC by 40.9% over the state-of-the-art method, benefiting from the more reliable candidate poses generated by our pipeline. For a small number of categories, our precision is slightly lower than COD due to the difficulty of capturing subtle angular differences when projecting 3D objects to 2D renderings. Nevertheless, the errors remain within the

acceptable fluctuation range, and the resulting annotations are valid canonical data for downstream tasks.

5 Conclusions

We have addressed a persistent bottleneck in 3D vision: the absence of large-scale canonical orientation. We introduced CanoVerse, a 320K/1,156-category canonical 3D dataset, and demonstrated that canonical orientation – when provided at dataset scale – emerges as a useful structural prior for modern 3D models. CanoVerse not only improves pose stability in 3D generation, but also enables zero-shot orientation estimation from point clouds, a capability that was largely inaccessible before. This dataset was made possible by a new canonicalization engine that treats alignment as scalable hypothesis triage instead of exhaustive $SO(3)$ search, reducing human oversight to seconds per object while maintaining high alignment fidelity. We hope CanoVerse will serve as a new foundation for orientation-consistent 3D learning, benchmarking, and generative modeling.

6 Acknowledgements

This work is partially supported by the NSF of China (No.62572276,62172260).

References

1. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015)
2. Chen, R., Zhang, J., Liang, Y., Luo, G., Li, W., Liu, J., Li, X., Long, X., Feng, J., Tan, P.: Dora: Sampling and benchmarking for 3d shape variational auto-encoders. arXiv preprint arXiv:2412.17808 (2024)
3. Chen, X., Dong, Z., Song, J., Geiger, A., Hilliges, O.: Category level object pose estimation via neural analysis-by-synthesis. In: European Conference on Computer Vision. pp. 139–156. Springer (2020)
4. Chen, Y., Di, Y., Zhai, G., Manhardt, F., Zhang, C., Zhang, R., Tombari, F., Navab, N., Busam, B.: Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9959–9969 (2024)
5. Dai, Q., Zhang, J., Li, Q., Wu, T., Dong, H., Liu, Z., Tan, P., Wang, H.: Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects. In: European Conference on Computer Vision (ECCV) (2022)
6. Di, Y., Zhang, C., Wang, C., Zhang, R., Zhai, G., Li, Y., Fu, B., Ji, X., Gao, S.: Shapematcher: Self-supervised joint shape canonicalization segmentation retrieval and deformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21017–21028 (2024)

7. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items (2022), <https://arxiv.org/abs/2204.11918>
8. Fan, Z., Song, Z., Xu, J., Wang, Z., Wu, K., Liu, H., He, J.: Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In: European Conference on Computer Vision. pp. 220–236. Springer (2022)
9. Fu, H., Cohen-Or, D., Dror, G., Sheffer, A.: Upright orientation of man-made objects. *ACM Trans. Graph.* **27**(3), 42 (2008). <https://doi.org/10.1145/1360612.1360641>, <https://doi.org/10.1145/1360612.1360641>
10. Han, X., Xiu, Y., Xu, Z., Lian, Z., Peng, S., Yao, Y., Chen, A., Huang, J., Zhang, B., Xu, L., Xu, F., Zhang, G., Xu, W., Yu, J., Liu, L., Chen, B., Liu, Y., Zhou, X.: Frontiers and prospects of 3D reconstruction and generation. *Journal of Image and Graphics* **31**(6), 2182–2197 (2026). <https://doi.org/10.11834/jig.260070>, <https://doi.org/10.11834/jig.260070>, received: 2026-01-29; Revised: 2026-03-09; Published: 2026-06-16
11. Jin, L., Wang, Y., Chen, W., Dai, Q., Gao, Q., Qin, X., Chen, B.: One-shot 3d object canonicalization based on geometric and semantic consistency. In: CVPR. pp. 16850–16859 (June 2025)
12. Lee, T., Wen, B., Kang, M., Kang, G., Kweon, I.S., Yoon, K.J.: Any6d: Model-free 6d pose estimation of novel objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11633–11643 (2025)
13. Li, W., Liu, J., Chen, R., Liang, Y., Chen, X., Tan, P., Long, X.: Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. arXiv preprint arXiv:2405.14979 (2024)
14. Lin, J., Wei, Z., Ding, C., Jia, K.: Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In: European Conference on Computer Vision. pp. 19–34. Springer (2022)
15. Lin, J., Wei, Z., Zhang, Y., Jia, K.: Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14001–14011 (2023)
16. Lin, X., Peng, Y., Wang, L., Zhong, X., Zhu, M., Yang, J., Feng, Y., Liu, C., Chen, Q.: Cleanpose: Category-level object pose estimation via causal learning and knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
17. Lin, X., Yang, W., Gao, Y., Zhang, T.: Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21040–21049 (2024)
18. Lu, Y., Tian, Y., Jiang, Z., Zhao, Y., Yang, Y., Ouyang, H., Hu, H., Yu, H., Shen, Y., Liao, Y.: Orientation matters: Making 3d generative models orientation-aligned (2025), <https://arxiv.org/abs/2506.08640>
19. Ma, Z., Yue, Y., Gkioxari, G.: Find any part in 3d (2024), <https://arxiv.org/abs/2411.13550>
20. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4561–4570 (2019)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

22. Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4209–4219 (2024)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
24. Sajnani, R., Poulenard, A., Jain, J., Dua, R., Guibas, L.J., Sridhar, S.: Condor: Self-supervised canonicalization of 3d pose for partial shapes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
25. Stojanov, S., Thai, A., Rehg, J.M.: Using shape to categorize: Low-shot learning with an explicit shape bias (2021)
26. Sun, J., Wang, Z., Zhang, S., He, X., Zhao, H., Zhang, G., Zhou, X.: Onepose: One-shot object pose estimation without cad models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6825–6834 (2022)
27. Sun, W., Tagliasacchi, A., Deng, B., Sabour, S., Yazdani, S., Hinton, G., Yi, K.M.: Canonical capsules: Self-supervised capsules in canonical pose. In: Neural Information Processing Systems (2021)
28. Team, Q.: Qwen2.5-vl (January 2025), <https://qwenlm.github.io/blog/qwen2.5-vl/>
29. Team, T.H.: Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material (2025)
30. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2642–2651 (2019)
31. Wang, Z., Zhang, Z., Pang, T., Du, C., Zhao, H., Zhao, Z.: Orient anything: Learning robust object orientation estimation from rendering 3d models. arXiv:2412.18605 (2024)
32. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17868–17879 (2024)
33. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Xu, J., Torr, P., Cao, X., Yao, Y.: Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems* **37**, 121859–121881 (2024)
34. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Yang, Y., Bao, Y., Qian, J., Zhu, S., Cao, X., Torr, P., et al.: Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. arXiv preprint arXiv:2505.17412 (2025)
35. Wu, T., Zhang, J., Fu, X., Wang, Y., Jiawei Ren, L.P., Wu, W., Yang, L., Wang, J., Qian, C., Lin, D., Liu, Z.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
36. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
37. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. arXiv preprint arXiv:2412.01506 (2024)
38. Xue, L., Yu, N., Zhang, S., Panagopoulou, A., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., et al.: Ulip-2: Towards scalable multimodal pre-

- training for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27091–27101 (2024)
39. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)* **42**(4), 1–16 (2023)
 40. Zhang, J., Huang, W., Peng, B., Wu, M., Hu, F., Chen, Z., Zhao, B., Dong, H.: Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In: European Conference on Computer Vision. pp. 199–216. Springer (2024)
 41. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)* **43**(4), 1–20 (2024)
 42. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
 43. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202* (2025)
 44. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. In: International Conference on Learning Representations (ICLR) (2024)