# A APPENDIX

## A.1 EVALUATION DETAILS

**Coverage (COV)**: COV is computed as the percentage of ground-truth models that have at least one closely matching generated model. Formally,

$$COV(S, G) = \frac{|\{\arg\min_{Y \in S} d(X, Y) \mid X \in G\}|}{|S|},$$

where $d(X, Y)$ denote the chamfer distance between two point clouds of $X$ and $Y$, $S$ represents the ground-truth set, and $G$ represents the generated model set.

**Minimum Matching Distance (MMD)**: MMD calculates the average minimum distance between each point in the generated set and the points in the ground-truth set. Formally,

$$MMD = \frac{1}{N} \sum_{i=1}^{N} \min_{j} d(M_i, G_j)$$

where $M_i$ represents the $i$-th model in the ground-truth set, $G_j$ represents the $j$-th generated model, $d$ is a distance metric, and $N$ is the total number of models in the generated set.

**Jensen-Shannon Divergence (JSD)**: JSD measures the similarity between two probability distributions. In the context of point clouds, it can be calculated using the following formula,

$$JSD(P, Q) = \frac{1}{2} D_{KL}(P \,||\, M) + \frac{1}{2} D_{KL}(Q \,||\, M),$$

where $P$ and $Q$ are the probability distributions of the ground-truth and generated data, $M$ is the average of $P$ and $Q$, and $D_{KL}$ is the Kullback-Leibler divergence.

**Novel**: This metric measures the novelty of the generated samples. Let $G$ represent the set of generated samples, and $T$ represent the set of training samples. Formally, Novel is calculated as:

$$Novel = \frac{|G \setminus T|}{|G|} \times 100\%,$$

where $G \setminus T$ represents the set of unique samples in $G$ that do not appear in $T$ and a "unique sample" refers to a generated model that does not match any model in the training set.

**Unique**: This metric measures the uniqueness of the generated samples. Let $G$ represent the set of generated samples, and $U$ represent the set of unique samples that appear only once in $G$. Formally, Unique is calculated as:

$$Unique = \frac{|U|}{|G|} \times 100\%.$$

## A.2 CAD SEQUENCE REPRESENTATION DETAILS

In the sketch-and-extrude model, multiple extruded sketches combine to form a composite design, with extrude parameters detailed in Table 3. The extrusion representation in our model is characterized by six parameters, summing up to ten individual settings. S-Offset refers to the deviation of the sketch plane from the origin, while S-Scaling denotes the scaling factor applied to alter its dimensions. Extrusion signifies the magnitude of extrusion executed on either side of the sketch plane. The Rotation parameter embodies the rotational aspect of the extruded body, and in this work, we employ rotation matrices solely with entries of 0, 1, or -1, aggregating to 26 unique matrices, which can conveniently be represented by a single parameter. Transition indicates the offset of the extruded body relative to the origin. Finally, Boolean Operation specifies the operation type, *i.e.*, intersection, union, or subtraction. The aforementioned parameters fully determine an extrusion. Each sketch is

an amalgamation of various faces, with each face being structured by multiple loops. Diving deeper, each loop is defined by a series of curves, which could manifest as a line, an arc, or a circle, comprising 2, 3, or 4 points, respectively. Each point within these structures is denoted by a geometry token, while end-primitive token is employed to signify the termination of each primitive entity. When a face has multiple loops, the first outlines the external boundary, while the rest defines internal holes.

To achieve a code tree representation, we structure CAD models hierarchically into three main levels, *i.e.*, *loops, profiles, and solids* (Xu et al., 2023). At the foundation of this hierarchy is the loop, which is a basic unit composed of interconnected lines, arcs, or circles, defined by sequences of 2D coordinates. These coordinates are separated by specific tokens, with curves arranged counterclockwise based on their initial coordinates. Moving up the hierarchy, the profile level captures the geometry of the loop, defined by 2D bounding box parameters, organized by the bottom-left corners of these boxes. This representation is sufficient for a 2D CAD sequence. With the addition of an extrusion operation, the hierarchy extends to the solid level, representing the 3D structure created by extruding one or more profiles. This level is defined by 3D bounding box parameters of the extruded profiles, also organized by their bottom-left corners.

Table 3: Parameters for representing extrusion

| Description | S-Offset | S-Scaling | Extrusion | Rotation | Transition | Bool Operation |
|---|---|---|---|---|---|---|
| **Parameters** | x,y | S | $E_{up}$,$E_{down}$ | R | $O_x$, $O_y$, $O_z$ | B |

### A.3 Ablation experiment

Table 4 shows the ablation study conducted on the key components of VQ-CAD. Four different configurations are delineated: (1) without diffusion, replacing it with a transformer; (2) without using VQ-VAE, directly diffusing on the command sequence; (3) employing data augmentation; (4) our final proposed VQ-CAD. This table reports the numerical results across various metrics including COV, MMD, JSD,Novel, and Unique.

In the setting of "Ours w/o diffusion", the quality indicator is high, while the diversity is limited. This may be attributed to the insufficiency of the dataset size, making the Transformer prone to overfitting on such a dataset, thereby compromising the model's generalization capability.

For the "Ours w/o VQ-VAE", the quality of the generated results is inferior, especially for the JSD metric, which is fourfold that of our baseline method. This implies that diffusing on the code tree representation in VQ-VAE significantly enhances the quality of generation.

In the "Ours with data augmentation" configuration, the quality of generation is comparable to our baseline, with better divergence performance. However, it relinquishes some implicit rules, as evident from the random generation results depicted in Figure 7 and Figure 8. In contrast, our method generates more regular results.

In summary, our model manages to attain a good quality of generation while ensuring diversity. A commendable trade-off between quality and diversity is achieved, retaining the implicit rules. This balanced achievement demonstrates the efficacy of our proposed VQ-CAD model, making a persuasive case for its application in conditional generation tasks.

Table 4: Ablation studies.

| **Method** | COV % ↑ | MMD ↓ | JSD ↓ | Novel % ↑ | Unique % ↑ |
|---|---|---|---|---|---|
| Ours w/o diffusion | **88.20** | **1.02** | **0.64** | 78.1 | 98.8 |
| Ours w/o VQ-VAE | 84.74 | 1.20 | 3.17 | 96.8 | 99.8 |
| Ours w/ data augmentation | 87.73 | 1.08 | 0.66 | **99.1** | **99.9** |
| **Ours** | 88.11 | 1.05 | **0.64** | 98.0 | **99.9** |

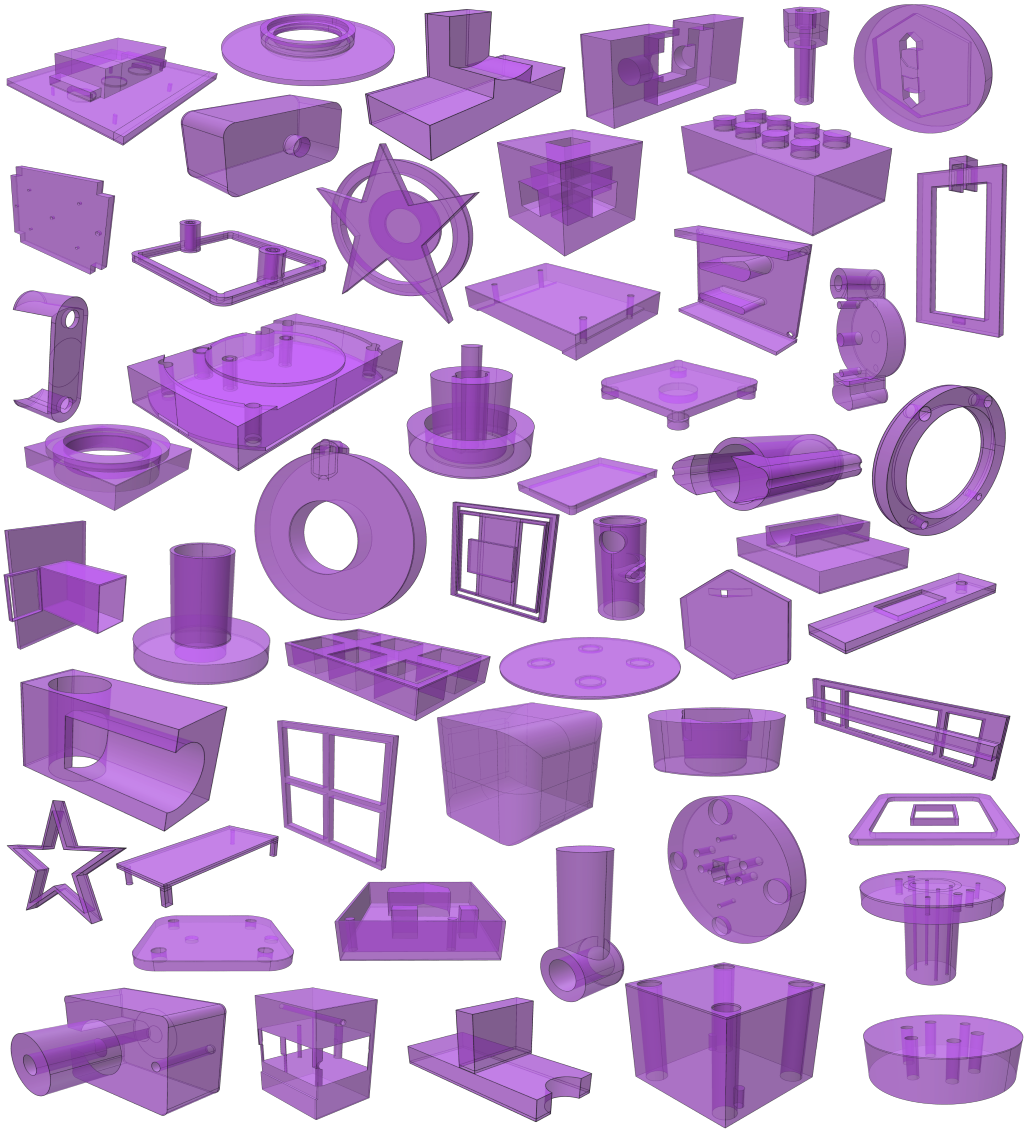

Figure 7: Randomly generated 3D CAD models without data augmentation.

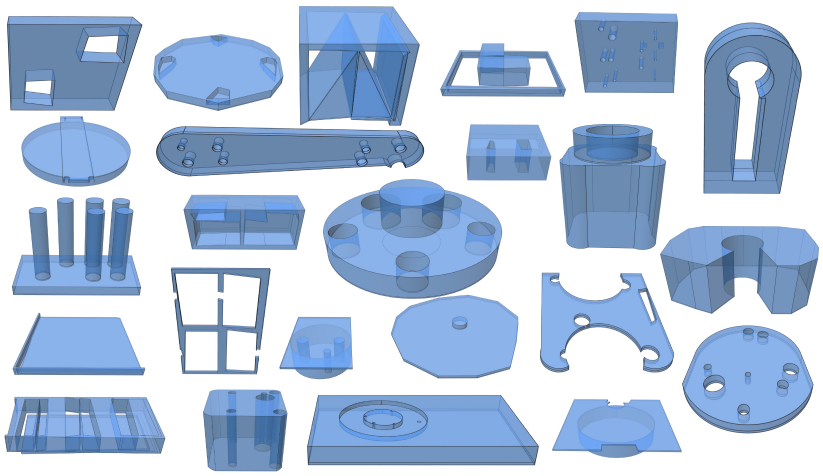

Figure 8: Randomly generated 3D CAD models with data augmentation.

### A.4 More Qualitative Results

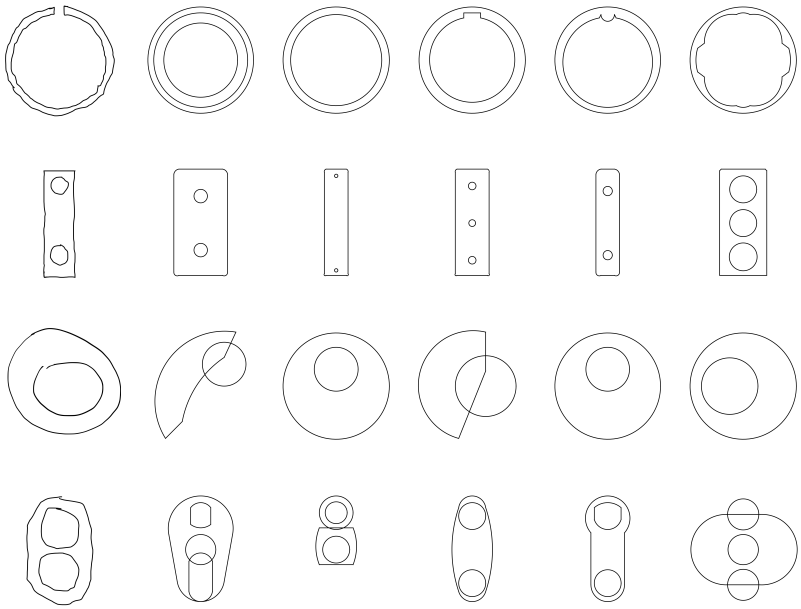

Figure 9: More results of our hand-drawn Sketch-to-CAD generation: First column shows input, following five columns shows various generated results of each sketch.

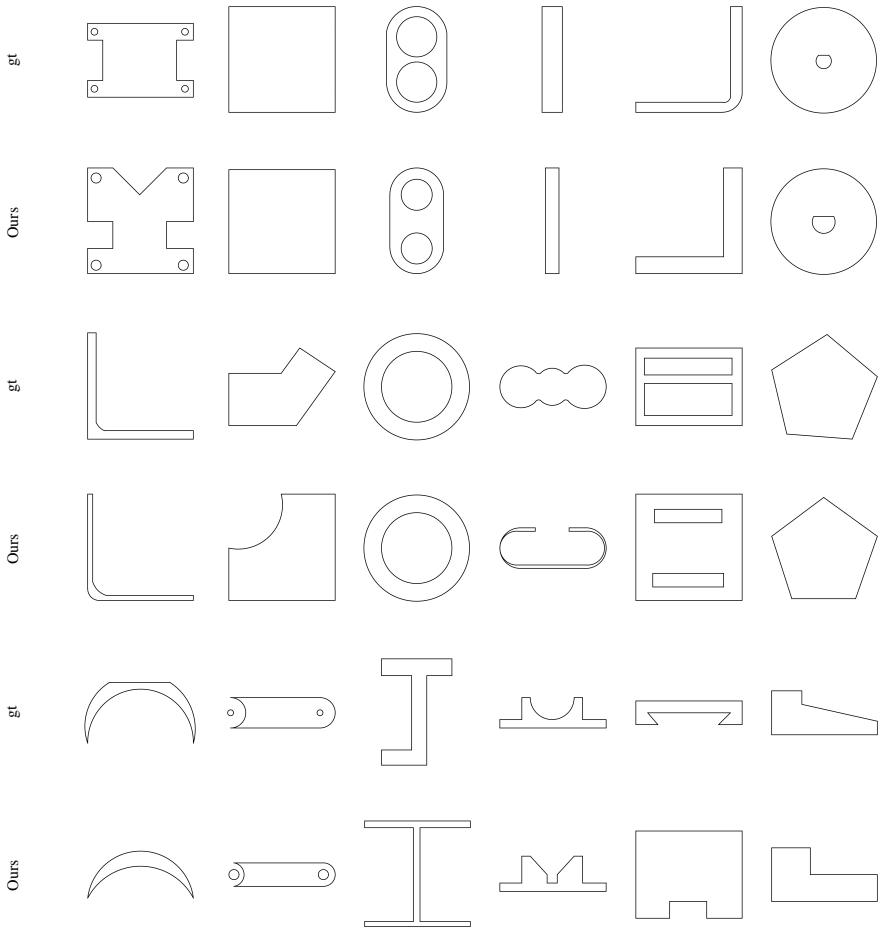


Figure 10: More results of our Picture-to-Command generation.