CMT: A Cascade MAR with Topology Predictor for Multimodal Conditional CAD Generation

A. mmABC dataset

A.1. Construction

Data Augmentation. After removing the multi-body models and similar models from the ABC [4] dataset, less than 60% (approximately 600,000 models) of the data remains. Further analysis reveals that many complex multi-body CAD models are composed of relatively simple basic models, as shown in Fig. 1 (b). Inspired by the principle that complex structures can emerge from simple building blocks, we hypothesize that more intricate models, including multi-body structures, can be generated by first mastering the generative capabilities of simpler, basic models. Therefore, as shown in Fig. 1 (b), we can split the complex multi-body models into multiple basic single models, thereby expanding the dataset and increasing its diversity. Following this augmentation, the dataset is expanded from 600 thousand to over 2.5 million models.
Data Filtering. After augmentation, we first remove models that fail in the process of converting into Mesh required for subsequent annotation. Then, by comparing the similarity points sampled from the remaining models under 6-bit quantization, the identical repeated models are deleted. Finally, about 1.3 million B-Rep models are obtained.
Data Annotation. We first use OpenCASCADE [1] to render 14 fixed Angle multi-view images for each B-Rep model. At the same time, we take points sampling on each CAD surface with its corresponding normal vectors as point clouds input. Finally, to generate a textual description of the model, we utilize the previously rendered multi-view images to generate the caption of the model through Vision-Language Models (VLMs). Specifically, we employ the open sourced VLM InternVL2-40B [3] and randomly input two or more views into it, with randomly selected prompts from the pre-designed prompt templates, so as to generate diverse and high quality text descriptions.

A.2. Prompts

This section lists the prompts we fed to a Vision–Language Model (VLM) in order to obtain descriptions for our CAD models.

We concatenate all rendered views of a single CAD model with one of the prompts below, then append (i) a *lexicon filter* that forbids colour/appearance terms, and (ii) some *reference examples*. The resulting string is provided to the VLM together with the rendered views.
Instruction pool (uniformly sampled)

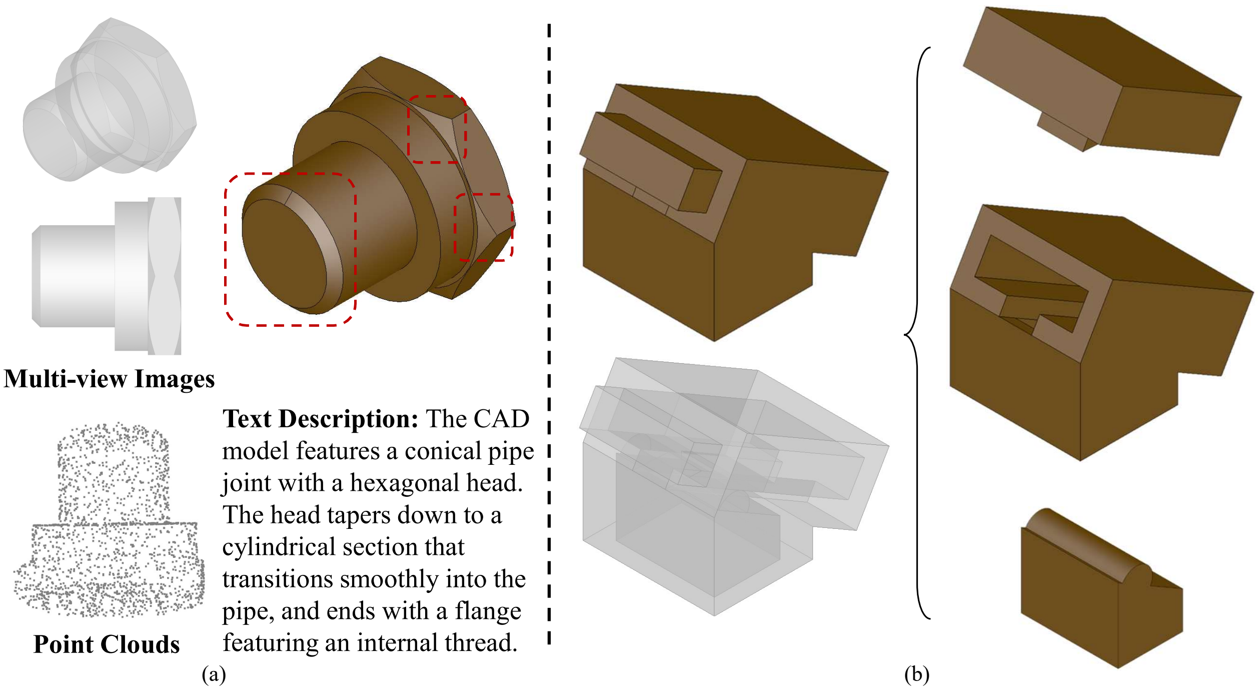
1. *“You are a veteran CAD specialist examining several orthographic and isometric views of a part. Write a clear, engineering-grade caption that details its overall geometry and key topological features so that an entry-level designer could model it from scratch.”*

2. *“Given multiple rendered perspectives of a CAD component, act as a senior mechanical drafter and describe the object’s shapes, profiles, and feature hierarchy in plain language sufficient for a junior colleague to replicate the design.”*

3. *“You are reviewing 3-D CAD snapshots from different angles. Produce a concise natural-language brief that captures dimensional proportions, primary surfaces, and connective topology, enabling an apprentice designer to rebuild the model accurately.”*

4. *“Act as a lead product engineer. From the provided multi-view CAD images, generate an instructional caption that thoroughly explains the model’s geometric primitives, Boolean operations, and mating relationships for a beginner to follow.”*

5. *“Observe the series of CAD visualisations. Draft an informative description outlining the solid’s key volumes, feature order, and interface topology, aimed at guiding a junior CAD user through recreating the part.”*



Text Description: The CAD model features a conical pipe joint with a hexagonal head. The head tapers down to a cylindrical section that transitions smoothly into the pipe, and ends with a flange featuring an internal thread.

Figure 1. Example of data in mmABC. From left to right, it depicts the B-Rep model, multi-view images, a point cloud, and a text description. The CAD model's head tapers from a hexagonal shape down to a cylindrical section, smoothly transitioning into the pipe, and ends with a flange that has an internal thread.

6. *"As an expert CAD educator, inspect these different-view images of a digital part and write a step-by-step geometric summary—mentioning profiles, extrusions, revolutions, and cuts—that a novice can translate into a fresh CAD build."*
7. *"You are a seasoned parametric-modelling engineer. Using the given front, side, top, and isometric CAD views, compose a precise caption highlighting shape symmetries, fillets, holes, and assembly interfaces so a trainee can reconstruct it."*
8. *"From the supplied CAD screenshots, craft a plain-English explanation that enumerates the core bodies, auxiliary features, and their spatial relationships, allowing an inexperienced designer to generate an identical model."*
9. *"Assume the role of a senior design reviewer. Study these multi-angle CAD renders and produce an accurate natural-language overview emphasising geometric intent, feature sequence, and topological connections for reproduction by a junior designer."*
10. *"Here are images of a computer-aided-design (CAD) model from different views. You are a senior CAD engineer who is familiar with the geometric shapes and topological structures of various CAD models. Give an accurate natural-language description about the CAD model to a junior CAD designer who can design it from your simple description. Focus on describing the geometric shape and topological structure of the CAD."*

Common suffix appended to every instruction

> *"Do not use words like – "blue", "shadow", "transparent", "metal", "plastic", "image", "black", "grey", ..., "abstract", "orange", "purple", "golden", "green".*
> *There are a few examples of descriptions for reference:*
> *1. The CAD model features a rectangular plate with four holes along its length.*
> *2. The CAD model consists of a stylised letter 'O' shaped object, resembling a curved, hollow ring.*
> *3. ......*

The above template guides the VLM to output concise yet comprehensive captions that emphasise geometry and topology while suppressing irrelevant visual cues, thereby aligning the generated text with the requirements of novice CAD model reconstruction.

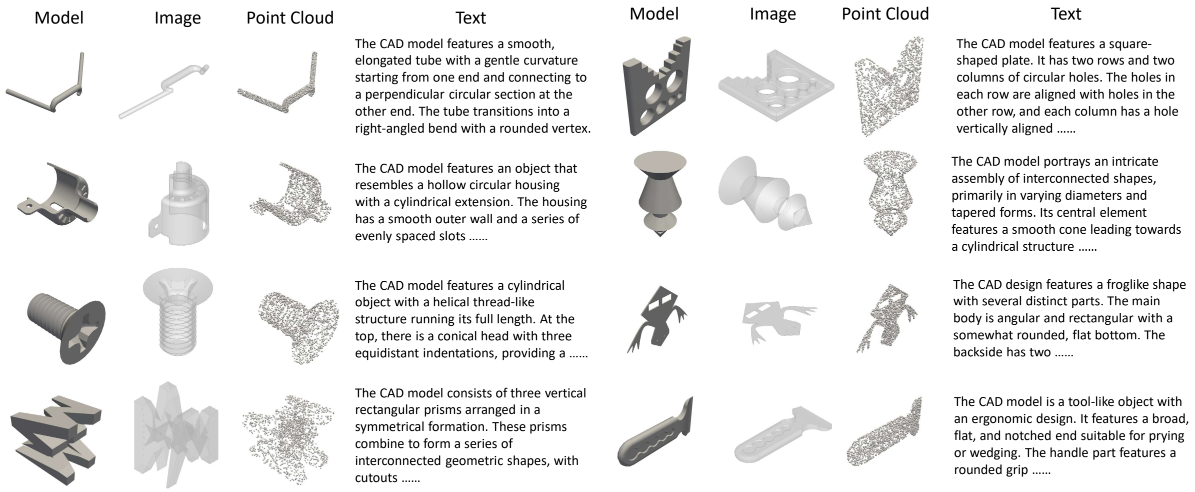

Model
Image
Point Cloud
Text
The CAD model features a smooth, elongated tube with a gentle curvature starting from one end and connecting to a perpendicular circular section at the other end. The tube transitions into a right-angled bend with a rounded vertex.
The CAD model features an object that resembles a hollow circular housing with a cylindrical extension. The housing has a smooth outer wall and a series of evenly spaced slots ……
The CAD model features a cylindrical object with a helical thread-like structure running its full length. At the top, there is a conical head with three equidistant indentations, providing a ……
The CAD model consists of three vertical rectangular prisms arranged in a symmetrical formation. These prisms combine to form a series of interconnected geometric shapes, with cutouts ……
Model
Image
Point Cloud
Text
The CAD model features a square-shaped plate. It has two rows and two columns of circular holes. The holes in each row are aligned with holes in the other row, and each column has a hole vertically aligned ……
The CAD model portrays an intricate assembly of interconnected shapes, primarily in varying diameters and tapered forms. Its central element features a smooth cone leading towards a cylindrical structure ……
The CAD design features a froglike shape with several distinct parts. The main body is angular and rectangular with a somewhat rounded, flat bottom. The backside has two ……
The CAD model is a tool-like object with an ergonomic design. It features a broad, flat, and notched end suitable for prying or wedging. The handle part features a rounded grip ……

Figure 2. Dataset models visualization. In the mmABC dataset, each CAD model includes images of 14 views. Here, we randomly select one view for demonstration purposes.

A.3. Visualization

A visualization of mmABC dataset is shown in Fig. 2, which contains diverse CAD models and corresponding multimodal annotations, including rendered images from multiple perspectives, point clouds sampled from model surfaces, and detailed text descriptions.

B. Experiment Details

B.1. Implementation Details

.

Diffusion [2]: Our denoising process adopts a cosine strategy, with 1,000 steps during training and reduced steps (by default, 100) during inference.

Training Strategy: We first pretrain cascade autoregressive generation network and topology predictor without input condition for 2,100 epochs, with 100 epochs for warming up. After model master the prior distribution of B-Rep through pretraining, we finetune it on our multimodal dataset $mmABC$, aligning condition embedding from the unified multimodal condition encoder with the B-Rep generated by the cascade autoregressive generation network.

B.2. Metrics

We follow the metrics in [6] and [5] for Unconditional CAD Generation and Conditional CAD Generation tasks, respectively. The detailed definitions are as follows:

Unconditional Generation.

- Coverage (COV): The percentage of reference data with at least match after assigning every generated data to its closest neighbor in the reference set based on Chamfer Distance (CD).
- Minimum Matching Distance (MMD): The averaged CD between a reference set data and its nearest neighbor in the generated data.
- Jensen-Shannon Divergence (JSD): Measuring the distribution distance between reference and generated data after converting point clouds into 28^3 discrete voxels.
- Novel: The percentage of data that do not appear in training set.
- Unique: The percentage of data that appears only once in generation.
- Valid: The percentage of B-Rep data that are watertight solids.

Conditional Generation.

- Chamfer: The averaged Chamfer Distance between a reference set data and the generated data.
- F-score: It is calculated by the recall R and precision P of point clouds sampled from the generated data and reference data, the formula is: $F_{score} = \frac{2 \times P \times R}{P+R}$.

Method	Chamfer ↓	F-score ↑	Normal C ↑
InstantMesh	6.91	82.12	50.33
CMT	**3.74**	**88.61**	**64.17**

Table 1. The quantitative results on shadowing image-based generation tasks.

- Normal Consistency: Evaluate whether the generated surface normal vector is consistent with the reference model. Note that the data are normalized to $[-0.5, 0.5]^3$ before evaluation.

B.3. More Results

Nearest neighbor results As shown in Fig. 3, the differences between the generated shapes(Grey) and the top two most similar training shapes(Blue) using CD illustrate the novelty of the shapes.

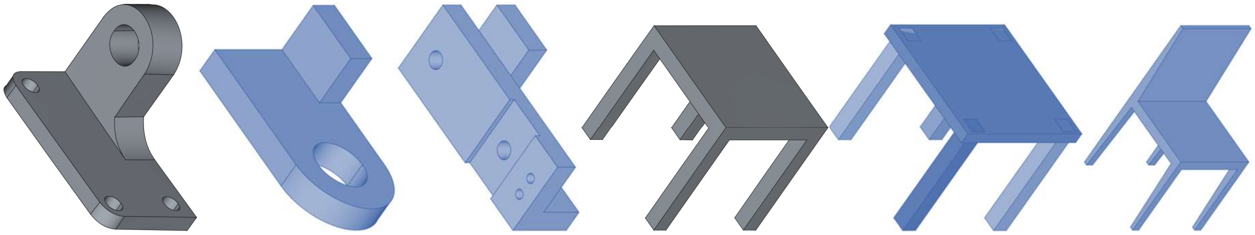


Figure 3. Novelty analysis of the generated CAD models.

Results on shading images With shading images rendered in the same way as InstantMesh in Tab. 1, CMT still outperforms (**-3.17** on Chamfer, **+6.49%** on F-score and **+13.84%** on Normal C).
High-resolution results We also provide high-resolution (3840x2160) images of generated CAD models rendered by *FreeCAD* in Fig. 4.

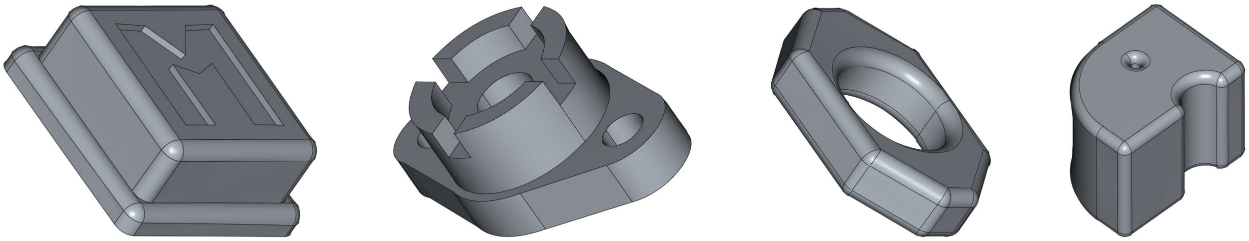


Figure 4. More results generated by CMT.

References

[1] Mladen Banović, Orest Mykhaskiv, Salvatore Auriemma, Andrea Walther, Herve Legrand, and Jens-Dominik Müller. Algorithmic differentiation of the open cascade technology cad kernel and its coupling with an adjoint cfd solver. *Optimization Methods and Software*, 33(4-6):813–828, 2018. 1

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[3] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9601–9611, 2019. 1

[5] Jingwei Xu, Chenyu Wang, Zibo Zhao, Wen Liu, Yi Ma, and Shenghua Gao. Cad-mllm: Unifying multimodality-conditioned cad generation with mllm. *arXiv preprint arXiv:2411.04954*, 2024. 3

[6] Xiang Xu, Joseph Lambourne, Pradeep Jayaraman, Zhengqing Wang, Karl Willis, and Yasutaka Furukawa. Brepgen: A b-rep generative diffusion model with structured latent geometry. *ACM Transactions on Graphics (TOG)*, 43(4):1–14, 2024. 3