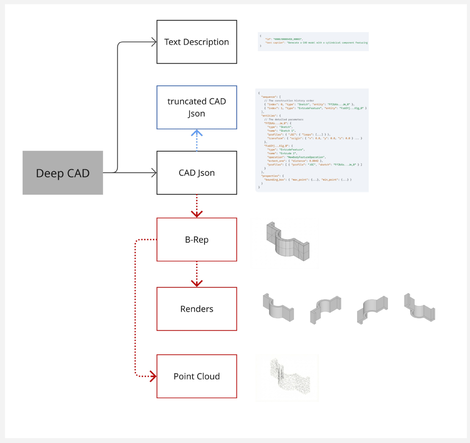
CAD-MLLM: Unifying Multimodality-Conditioned CAD Generation With MLLM

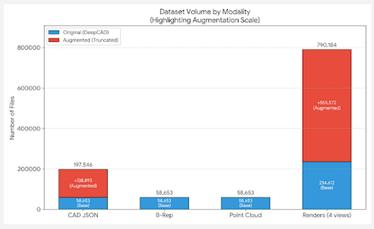
Team 21: Yizhuo Di | David Chen | Karthick Raja | Chia Hui Yan

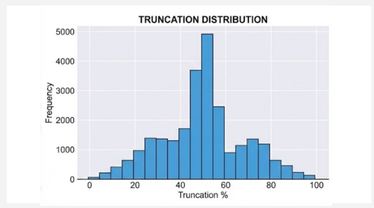
Motivation: Multi-Modal CAD Generation + Autocompletion

We aim to design a unified Computer-Aided Design (CAD) generation system that can easily generate editable CAD models based on the user's inputs in the form of different modalities.

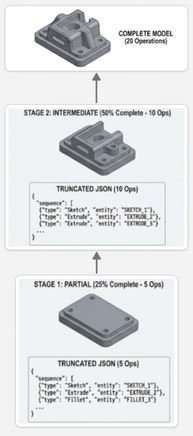
Dataset: DeepCAD subset + autocomplete augmentation

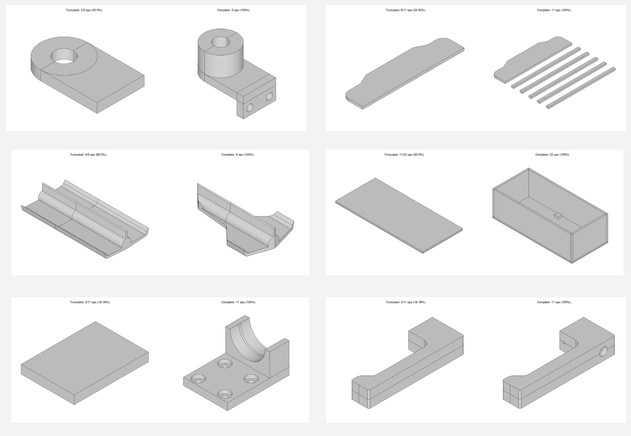






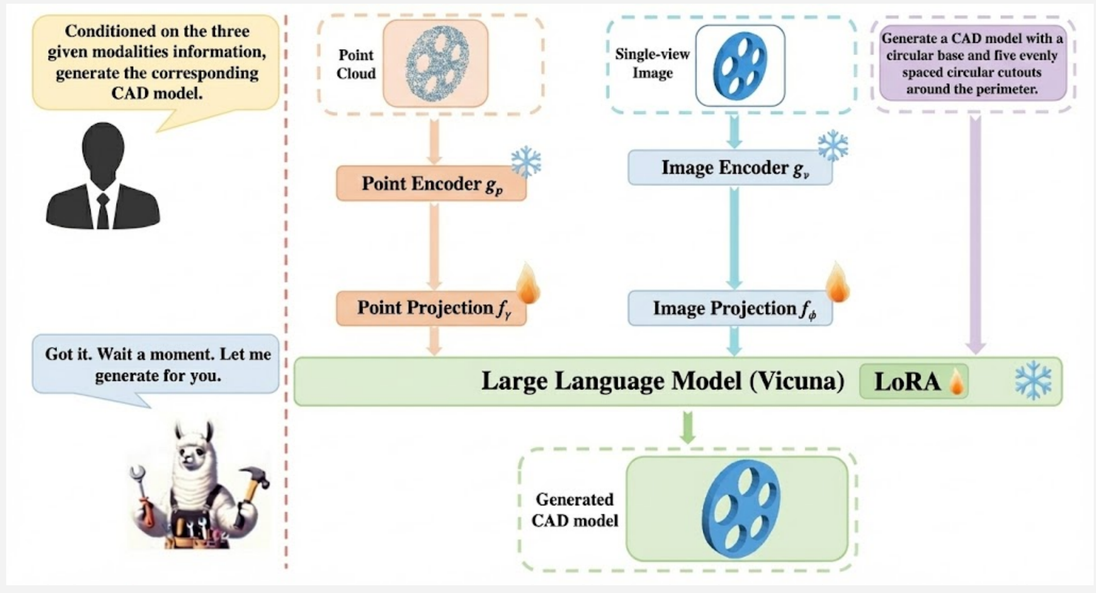
To enable multi-modal autocompletion, we extended a **10% DeepCAD subset (58,653 models)** into a rich synthetic dataset. Using OpenCascade, we generated **STEP files, point clouds, and multi-view renders** for each design—modalities absent in the original dataset. Additionally, we developed an **Intelligent Truncation algorithm** that creates valid partial sequences by recursively tracing entity dependencies. This strategy amplified our data by **3.37×**, resulting in **197,546** training examples across all modalities.





Our algorithm generates valid partial designs by identifying operation boundaries and performing a **dependency trace** to preserve all referenced entities. This ensures every truncated sequence is **geometrically consistent for incremental training**. Consequently, we amplified the original 58,653 models into over 197,000 valid design sequences.

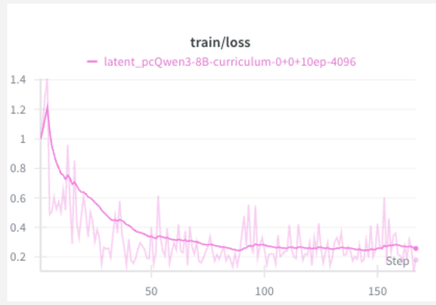
Methodology:



The network could process **three single modalities or any combinations of them**.

Except for the textual descriptions, each modality is first processed through its corresponding **frozen encoder** before being further integrated. Subsequently, they are passed through **a trainable projection layer**, aligning them within a unified language feature space.

The fine-tuned Large Language Models (LLMs), augmented with Low-Rank Adaptation (**LoRA**), then process a combination of the prompt and the projected embeddings, enabling the accurate generation of CAD models.



Training configs:

```
modality_sample_probs={
      "text": 0.2,
      "text+point_cloud": 0.3,
      "text+image": 0.2,
      "text+point_cloud+image": 0.3,
},
max_seq_length=4096
```

Evaluation:

Evaluation Settings:
(Eval set, Checkpoint using, max new tokens generated, input modality)

- **Eval 1,** checkpoint-epoch0-step140-20251128_072200, 10,240, Text Only
- **Eval 2,** checkpoint-epoch0-step140-20251128_072200, 2,048, Text Only
- **Eval 3,** stage3-epoch0-step100-20251128_220651, 4,096, Image+Point Cloud+Text
- **Eval 4,** stage-3-4096-20251129_213538, 4,096, Image+Point Cloud+Text

Ground truth JSON sequence length is less than 2,048 tokens.

Evaluation Metrics

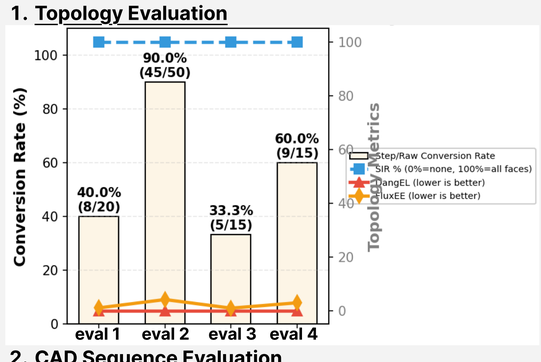
1. Topology Evaluation Metrics：
   - **STEP/raw conversation rate**
   - **DangEL** (Dangling Edge Length): Length of boundary edges
   - **SIR** (Self-Intersection Ratio): Percentage of self-intersecting faces
   - **FluxEE** (Flux Enclosure Error)
2. CAD Sequence Metrics:
   - **Entity Count Acc:** 1.0 = perfect match, 0.0 = completely wrong
   - **Type Seq Acc (**Accuracy of type sequence): 1.0 = perfect sequence match
   - **Type Dist Sim** (Entity Type Distribution Similarity): Jaccard similarity between the distribution of entity types, 1.0 = identical type distribution
3. Visualization: **Visual compare gt and generated STEP**

Summary:

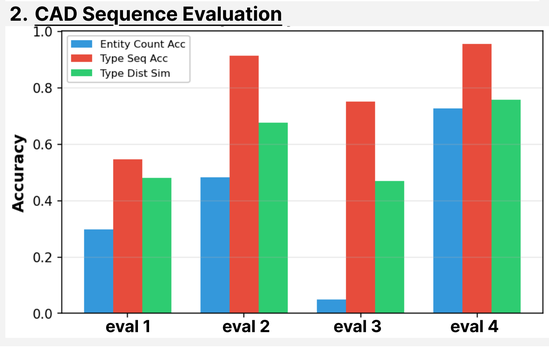
The success rate for STEP file generation (Topology) is higher when using text-only input (Eval 1, 2), potentially due to simpler output generation. Conversely, the accuracy of the CAD sequence significantly improves when providing multi-modal input (image + point cloud + text), strongly indicating that richer input information increases sequential fidelity.

Overall, the pipeline achieves an average 60% successful STEP generation rate and maintains CAD sequence accuracy above 70% across the successful outputs. The current models are constrained to testing on simple shapes. This limitation stems from the small training data size and the short text prompts used in development, which results in the generated model sizes often being inaccurate.

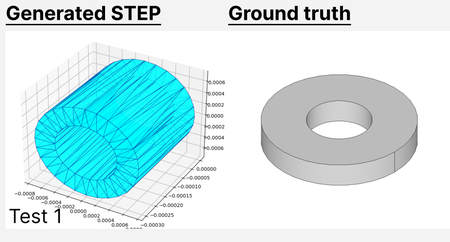
1. Topology Evaluation



2. CAD Sequence Evaluation



Generated STEP **Ground truth**



Generated STEP **Ground truth**