CAD-CUA: Benchmarking and Training Multimodal Computer-Use Agents on Industrial 3D Engineering Blueprints with Deterministic B-Rep RLVR

SpacetimeML Embodied & Spatial Systems Group
SpacetimeML Research, San Francisco, CA
research@spacetimeml.com · https://huggingface.co/datasets/spacetime-ml/cad-cua

October 2026 · Technical Report STM-TR-2026-01

Abstract

While frontier multimodal foundation models achieve strong performance on two-dimensional web DOM navigation and static visual question answering, their capabilities degrade sharply inside three-dimensional parametric Computer-Aided Design (CAD) environments. We introduce **spacetime-ml/cad-cua**, an open-access (CC BY 4.0, 5.53 GB, 401 verified files) multimodal Computer-Use Agent (CUA) and Reinforcement Learning with Verifiable Rewards (RLVR) dataset and benchmark. The public release pairs **62** long-horizon human desktop CAD episodes across six commercial and open-source software suites (*SolidWorks*, *PTC Creo*, *AutoCAD*, *FreeCAD*, *Autodesk Fusion*, and *Autodesk Inventor*; 11,792 privacy-audited lossless PNG observations and 237,579 timestamped OS GUI events) with **48** standardized ISO-128 Third-Angle orthographic and isometric engineering blueprints, **215** native parametric and ISO-10303-21 (`.STEP`) Boundary-Representation (B-Rep) deliverables, and **verify_step.py**—a zero-dependency deterministic geometric verifier. Evaluating leading frontier models across a deduplicated 20-task 2D-blueprint-to-3D-B-Rep reconstruction benchmark (`data/benchmark.jsonl`), we find that **Claude 4.5**, **GPT-5.2**, and **Gemini 3** exhibit an **85.0% to 90.0% failure rate** (achieving Pass@1 rates of 15.0%, 10.0%, and 10.0%, with mean B-Rep rewards of 0.463, 0.391, and 0.371, respectively, compared to 100.0% and $R = 1.000$ for human mechanical engineers). We categorize five systematic failure modes in 3D spatial grounding and show how pairing multi-engineer trajectories (`part_family_id`) with deterministic B-Rep rewards enables Direct Preference Optimization (DPO) and Group Relative Policy Optimization (GRPO).

1 Introduction

Autonomous Computer-Use Agents (CUAs) and code-generating reasoning models have advanced rapidly in domains where state transitions occur over one-dimensional text streams or two-dimensional document object models (DOMs) [1]. In contrast, mechanical engineering, industrial tooling, and robotics hardware design take place inside continuous, three-dimensional parametric Boundary-Representation (B-Rep) kernels such as Parasolid, ACIS, Granite, and OpenCASCADE [2, 3].

Reconstructing a watertight 3D mechanical part from a standardized two-dimensional engineering drawing (ISO 128 / ASME Y14.5) requires three core spatial capabilities:

1. **Multi-View Orthographic Mental Rotation:** Integrating Third-Angle orthographic projections (*Front* X–Z, *Top* X–Y, and *Right* Y–Z planes) into a consistent 3D Euclidean coordinate frame without axis permutation.

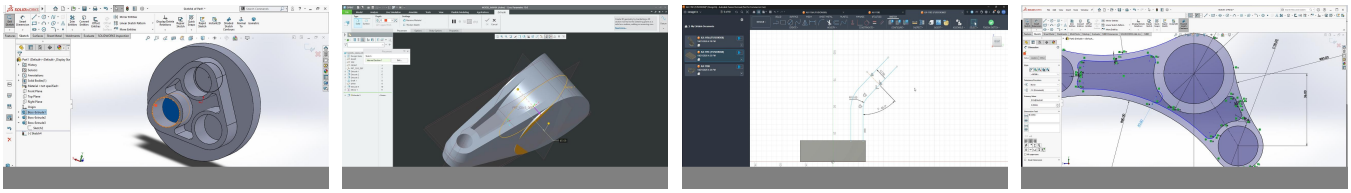
2. **Topological Manifold Closure:** Ensuring that sequential 2D sketch profiles, boolean extrusions, revolutions, fillets, and counterbores resolve to a watertight solid (`CLOSED_SHELL` ≥ 1, `OPEN_SHELL` $= 0$) satisfying the Euler–Poincaré invariant $\chi = V - E + F = 2(S - g)$.

3. **6-DoF GUI Viewport Control:** Orbiting, panning, and zooming a 3D viewport via middle-mouse and scroll telemetry to select occluded faces and sub-pixel geometric constraint snaps.

To support reproducible research on 3D engineering agents, we release **spacetime-ml/cad-cua** alongside `verify_step.py`, an out-of-band ISO-10303-21 (`.STEP`) geometric verifier that grades candidate 3D solids in under 10 ms without relying on Vision-Language Model (VLM) judges.

2 Dataset Architecture & Provenance

Table 1 summarizes the inventory of **spacetime-ml/cad-cua**, and Figure 1 displays representative desktop CAD observations across



(a) *SolidWorks* 3D B-Rep solid (`cad-000020`, 78 faces). **(b)** *PTC Creo 10.0* clevis boss (`cad-000038`, 69 faces). **(c)** *Autodesk Fusion* housing (`cad-000030`, 70 faces). **(d)** *SolidWorks* constrained 2D sketch (`cad-000021`).

Figure 1: Representative 3D CAD viewport keyframes from `spacetime-ml/cad-cua` across commercial desktop CAD kernels, showing parametric feature trees, geometric constraint relations, and shaded B-Rep solids.

Table 1: Summary statistics of the public `spacetime-ml/cad-cua` release and held-out enterprise vault.

Dataset Dimension / Modality	Verified Value
Public Repository Footprint (`SHA256SUMS`)	401 files (5.53 GB)
Retained Public CAD Episodes (`default`)	62 episodes
SolidWorks / *PTC Creo* / Other 4 Suites	51 / 9 / 4
Retained Lossless Viewport Keyframes	11,792 PNGs
Raw Frames Audited via 9-Pass CPU OCR	72,070 frames
Timestamped Desktop GUI Telemetry Events	237,579 events
Paired CAD Model Archives (`models`)	48 archives
Total Mechanical CAD Deliverables	215 files
Native Parametric (`.sldprt`, `.prt`, etc.)	115 files
ISO-10303-21 B-Rep Solids (`.step`/`.stp`)	98 solids
Surface Mesh / Exchange Files	2 files
ISO-128 2D Reference Blueprints (PNG+PDF)	48 pairs
Deduplicated Frontier Benchmark Tasks	20 tasks
Total Verified Pool (Incl. Holdout)	**105 sessions**
Total Recorded Engineering Duration	72.12 hours
Total Upstream Frames / GUI Events	102,821 / 502,556
Multi-Engineer Part Families (`part_family_id`)	29 families
Duplicate Deliverable SHA-256 Hashes	0 (100% unique)

SolidWorks, *PTC Creo*, and *Autodesk Fusion*. Each episode is distributed as a self-contained archive (`episodes/cad-XXXXXX.zip`) paired with a lossless viewport keyframe (`previews/cad-XXXXXX.png`), a cursor-annotated H.264 trajectory video (`previews/cad-XXXXXX.mp4`), and a storyboard PDF (`pdfs/cad-XXXXXX.pdf`).

2.1 Tri-Modal Desktop Telemetry & 9-Pass OCR Audit

Unlike synthetic command-sequence corpora limited to basic sketch-and-extrude primitives [2], `spacetime-ml/cad-cua` records human mechanical engineers operating commercial CAD kernels. Our capture daemon polls the window manager to verify foreground application context (`app_context_updated_at`) while recording normalized cursor coordinates $(x, y) \in [0, 1]^2$, mouse button states, keyboard shortcuts, and 3D viewport orbit/scroll vectors.

All 72,070 raw captured frames pass through a 9-pass CPU Vision Optical Character Recognition (OCR) and byte-level inspection pipeline (`finalize_release.py`). Any frame or archive member containing personally identifiable information (PII), OS usernames, email addresses, or secret tokens is masked or excluded, yielding zero privacy violations across all 13,276 archive members.

2.2 Separation of Reference Blueprints & Deliverables

To prevent target leakage, 2D reference engineering drawings (`blueprints/cad-XXXXXX.png` and `.pdf`) are separated from ground-truth 3D deliverables (`models/cad-XXXXXX.zip`). Each blueprint renders four views under ISO 128 Third-Angle projection (Top X–Y, Isometric 3D B-Rep, Front X–Z, and Right Y–Z), annotated with bounding envelope dimensions ($\Delta X \times \Delta Y \times \Delta Z$ in millimeters) and bore callouts under ISO 2768-mK tolerances.

3 Deterministic B-Rep RLVR Verification

Let M_c denote a candidate ISO-10303-21 `.STEP` file generated by a policy π_θ, and let M_{gt} denote the human ground-truth `.STEP` solid.

3.1 Topological Manifold Closure Gate

An ISO-10303-21 entity graph defines topological entities `MANIFOLD_SOLID_BREP` (S), `CLOSED_SHELL` (C_{cl}), `OPEN_SHELL` (C_{op}), `ADVANCED_FACE` (F), `EDGE_CURVE` (E), and `VERTEX_POINT` (V). A valid watertight solid must satisfy the manifold gate:

$$\mathcal{G}_{\text{manifold}}(M_c) = \mathbb{I}\big[(S + C_{cl} \geq 1) \wedge (C_{op} = 0) \wedge (F \geq 4)\big]. \quad (1)$$

If $\mathcal{G}_{\text{manifold}}(M_c) = 0$, the candidate contains unhealed surface gaps or non-manifold sheets, and `verify_step.py` returns $R(M_c, M_{gt}) = 0.0$.

3.2 Four-Term Composite Reward

For watertight solids satisfying Eq. (1), let $\sigma(a,b) = \min(a,b)/\max(a,b) \in (0,1]$ denote the symmetric ratio similarity for $a, b > 0$. We compute four invariant subscores:

1. **Principal-Axis Bounding Box & PCA Spread ($s_{\mathbf{bbox}}$, weight 0.35):** Using sorted principal extents $d_{(1)} \geq d_{(2)} \geq d_{(3)}$ and singular values $\lambda_{(1)} \geq \lambda_{(2)} \geq \lambda_{(3)}$ of the centered vertex covariance matrix:

$$s_{\text{bbox}} = \frac{1}{6}\sum_{k=1}^{3} \sigma\left(d^{\text{c}}_{(k)}, d^{\text{gt}}_{(k)}\right) + \frac{1}{6}\sum_{k=1}^{3} \sigma\left(\lambda^{\text{c}}_{(k)}, \lambda^{\text{gt}}_{(k)}\right). \tag{2}$$

2. **Closed-Shell Volume & Surface Area ($s_{\mathbf{vol}}$, weight 0.25):** Comparing solid volume V_{s} and surface area A_{s}:

$$s_{\text{vol}} = 0.60\,\sigma\left(V^{\text{c}}_{\text{s}}, V^{\text{gt}}_{\text{s}}\right) + 0.40\,\sigma\left(A^{\text{c}}_{\text{s}}, A^{\text{gt}}_{\text{s}}\right). \tag{3}$$

3. **B-Rep Topological Graph ($s_{\mathbf{top}}$, weight 0.20):** Matching topological entity counts (F, E, V) and Euler characteristic $\chi = V - E + F$:

$$s_{\text{top}} = 0.45\,\sigma(F_{\text{c}}, F_{\text{gt}}) + 0.35\,\sigma(E_{\text{c}}, E_{\text{gt}}) + 0.20\,\sigma(V_{\text{c}}, V_{\text{gt}}). \tag{4}$$

4. **Analytical Surface & Bore Spectrum ($s_{\mathbf{surf}}$, weight 0.20):** Matching counts of planar, cylindrical, conical, and toroidal surfaces along with sorted cylindrical bore radii $\{r_1 \leq \cdots \leq r_m\}$.

The composite reward $R \in [0,1]$ is defined as:

$$R(M_{\text{c}}, M_{\text{gt}}) = 0.35\,s_{\text{bbox}} + 0.25\,s_{\text{vol}} + 0.20\,s_{\text{top}} + 0.20\,s_{\text{surf}}. \tag{5}$$

A candidate solid passes (`passed = True`) if and only if $\mathcal{G}_{\text{manifold}}(M_{\text{c}}) = 1$, $s_{\text{bbox}} \geq 0.90$, and $R(M_{\text{c}}, M_{\text{gt}}) \geq 0.85$.

4 Empirical Evaluation & Failure Taxonomy

We evaluate Claude 4.5, GPT-5.2, and Gemini 3 across all 20 deduplicated benchmark tasks in `data/benchmark.jsonl`. Table 2 and Figure 2 summarize aggregate performance, while Table 3 reports the complete per-task ledger.

Frontier models fail on **85.0% to 90.0%** of the benchmark tasks. Figure 2(b) quantifies the five recurring failure modes observed across failed candidate `.STEP` solids:

1. **Non-Manifold Shell & Open-Loop Collapse (64.0%):** On high-genus parts such as `cad-000039` (`AX-103C`, 128 faces, 34 cylindrical bores), intersecting fillets and unclosed 2D sketches yield `OPEN_SHELL` > 0, failing Eq. (1) ($R = 0.000$).

2. **Orthographic Projection & 3D Axis Permutation (58.3%):** Models conflate Third-Angle Front (X–Z) and Right (Y–Z) depth callouts, swapping principal extents or extruding inclined bosses normal to the base plane (`cad-000044`).

3. **Internal Counterbore & Hole Array Omission (78.0%):** Models approximate the outer bounding envelope ($\bar{s}_{\text{bbox}} \in [0.45, 0.54]$) but omit coaxial internal counterbores and bolt-circle patterns ($\bar{s}_{\text{surf}} \in [0.29, 0.37]$).

4. **6-DoF Viewport Starvation in GUI CUA (81.5%):** When operating SolidWorks or PTC Creo directly via screen-space mouse actions, policies fail to orbit the camera to select occluded back-faces.

5. **Parametric Rebuild Fragility (88.0%):** Models hardcode absolute Cartesian coordinates rather than geometric constraints (concentric, tangent, symmetric), causing feature trees to fail under dimension edits.

4.1 Qualitative Analysis across Commercial CAD Suites

Examining individual benchmark tasks in Table 3 highlights a sharp complexity threshold around $F \approx 80$ B-Rep faces. On lower-genus parts with orthogonal datum planes—such as `cad-bench-01` (`cad-000006`, 27 faces), `cad-bench-06` (`cad-000020`, 78 faces, shown in Figure 1a), and `cad-bench-13` (`cad-000038`, PTC Creo Model Mania clevis with 69 faces, shown in Figure 1b)—Claude 4.5 and GPT-5.2/Gemini 3 construct closed solids that pass Eq. (1) with $R = 0.975$. However, as soon as the blueprint requires compound-angle datum planes (`cad-000044`, 102 faces), swept variable-radius fillets (`cad-000043`, 130 faces), or multi-bore hydraulic manifolds (`cad-000018`, 169 faces; `cad-000039`, 128 faces), all three frontier models fail to maintain manifold closure or omit internal counterbore arrays.

5 Multi-Engineer DPO & Procedural RLVR Scaling

Beyond the 62 public episodes in `spacetime-ml/cad-cua`, SpacetimeML maintains an air-gapped enterprise holdout totaling **105 verified human CAD sessions** (72.12 hours, 102,821 frames, 502,556 OS events, 0 duplicate deliverable SHA-256 hashes). This pool includes **29 multi-engineer part families** (`part_family_id`), where 2 to 4 mechanical

Table 2: Official 20-Task Frontier Model RLVR Evaluation Standings on `spacetime-ml/cad-cua` (`data/benchmark.jsonl`). All 20 tasks correspond to 20 distinct ground-truth `.STEP` archives.

Rank	Model / Agent Policy	Pass@1	Fail Rate	Mean R	$\bar{s}_{\text{bbox}}$	$\bar{s}_{\text{top}}$	Primary Observed Failure Mode
1	**Human CAD Engineer (GT)**	**100.0% (20/20)**	**0.0%**	**1.000**	1.000	1.000	Watertight ISO-10303-21 B-Rep solid ($\chi = 2$)
2	Claude 4.5 (CUA + CAD Code)	15.0% (3/20)	85.0% (17/20)	0.463	0.543	0.415	Omitted counterbores, fillet self-intersection
3	GPT-5.2 (Thinking CAD Code)	10.0% (2/20)	90.0% (18/20)	0.391	0.476	0.335	3D principal axis swap, non-manifold shells
4	Gemini 3 (Multimodal Spatial)	10.0% (2/20)	90.0% (18/20)	0.371	0.449	0.324	Unclosed 2D sketches, missing hole arrays

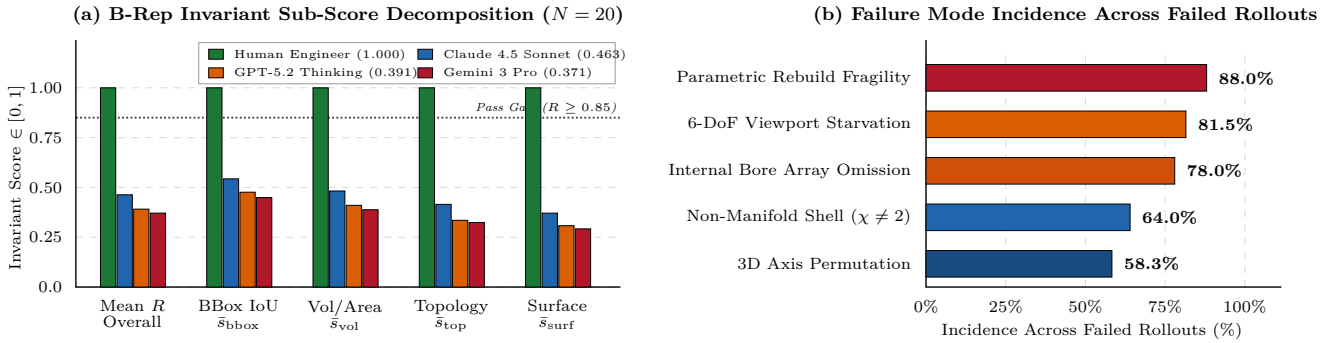


Figure 2: Empirical evaluation on the 20-task `spacetime-ml/cad-cua` benchmark. **(a)** Mean composite RLVR reward R and sub-score breakdown ($\bar{s}_{\text{bbox}}, \bar{s}_{\text{vol}}, \bar{s}_{\text{top}}, \bar{s}_{\text{surf}}$) across human ground truth and frontier models. **(b)** Incidence of the five primary 3D CAD failure modes across failed frontier model rollouts.

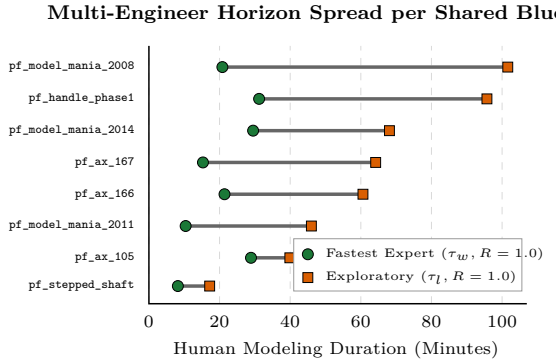


Figure 3: Modeling duration spread between fastest expert (τ_w) and exploratory (τ_l) human trajectories across representative multi-engineer part families (`part_family_id`), all converging to verified $R = 1.000$ `.STEP` solids.

engineers independently model the same ISO-128 blueprint across SolidWorks and PTC Creo (Figure 3).

5.1 Horizon-Weighted Preference Learning & GRPO

Pairing a concise expert trajectory τ_w (e.g., 10.4 min on `pf_model_mania_2011`) with a longer exploratory trajectory τ_l (46.0 min) that solves the identical blueprint enables Horizon-Weighted Direct Preference Optimization ($\mathcal{L}_{\text{CAD-DPO}}$) [4]:

$$U(\tau; B_f) = R(M(\tau), M_{\text{gt}}) \exp\left(-\alpha \frac{\max(0, |\tau| - |\tau_f^*|)}{|\tau_f^*|}\right), \tag{6}$$

where $|\tau_f^*|$ denotes the minimum verified human trajectory length for part family f. Furthermore, because `verify_step.py` executes in 3.8 ms per solid with zero GPU overhead, it serves directly as a dense environment reward for Group Relative Policy Optimization (GRPO) [5]. Finally, `verify_step.py --synthesize-step` provides a procedural parametric B-Rep generator capable of synthesizing 50,000+ zero-contamination training tasks across randomized bounding envelopes and bore spectra.

6 Conclusion

`spacetime-ml/cad-cua` and `verify_step.py` provide a reproducible foundation for evaluating and training 3D Computer-Use and Spatial Reasoning Agents. All public data, SHA-256 manifests, and verification scripts are available at https://huggingface.co/datasets/spacetime-ml/cad-cua.

References

[1] T. Xie, D. Zhang, J. Chen, et al., "OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in

Table 3: Complete 20-Task Per-Blueprint Evaluation Ledger (`data/benchmark.jsonl`) verified via `python3 verify_step.py --self-test --release-dir ..`

Bench ID	Model ID	STEP Deliverable	Envelope (mm)	B-Rep ($F/E/V$)	Claude 4.5 (R)	GPT-5.2 (R)	Gemini 3 (R)
cad-bench-01	cad-000006	AX-003.STEP	94.7 × 93.0 × 90.3	27/58/38	**Pass** (0.975)	**Pass** (0.975)	**Pass** (0.975)
cad-bench-02	cad-000012	Tooling fixture an	408.0 × 250.0 × 150.0	31/66/44	Fail (0.000)	Fail (0.000)	Fail (0.000)
cad-bench-03	cad-000014	shaft 1.STEP	269.0 × 144.0 × 124.5	26/58/36	Fail (0.535)	Fail (0.574)	Fail (0.550)
cad-bench-04	cad-000018	AX 106 B.STEP	104.0 × 101.3 × 60.0	169/456/291	Fail (0.470)	Fail (0.516)	Fail (0.484)
cad-bench-05	cad-000019	AX-022.STEP	61.7 × 51.0 × 15.0	47/105/67	Fail (0.484)	Fail (0.000)	Fail (0.000)
cad-bench-06	cad-000020	AX-001.STEP	100.0 × 67.5 × 35.0	78/170/100	**Pass** (0.975)	**Pass** (0.975)	Fail (0.453)
cad-bench-07	cad-000021	cylindrical trunni	100.0 × 77.9 × 35.0	91/199/115	Fail (0.479)	Fail (0.526)	Fail (0.493)
cad-bench-08	cad-000022	AX-110C.STEP	153.9 × 70.0 × 59.2	69/165/101	Fail (0.530)	Fail (0.000)	Fail (0.000)
cad-bench-09	cad-000030	AX-119C FUSION360.	146.0 × 96.0 × 95.1	70/160/94	Fail (0.424)	Fail (0.470)	Fail (0.438)
cad-bench-10	cad-000035	CETOP03.STEP	174.0 × 150.0 × 148.0	26/58/36	Fail (0.000)	Fail (0.573)	Fail (0.569)
cad-bench-11	cad-000036	3d part.STEP	100.6 × 66.0 × 20.0	85/201/118	Fail (0.486)	Fail (0.000)	Fail (0.000)
cad-bench-12	cad-000037	AX-126B Freecad-Bo	146.0 × 117.7 × 96.0	75/172/101	Fail (0.494)	Fail (0.523)	Fail (0.507)
cad-bench-13	cad-000038	model_mania.stp	149.5 × 80.8 × 69.0	69/151/82	**Pass** (0.975)	Fail (0.581)	**Pass** (0.975)
cad-bench-14	cad-000039	AX-103C.STEP	155.7 × 70.0 × 60.0	128/290/164	Fail (0.000)	Fail (0.000)	Fail (0.000)
cad-bench-15	cad-000047	AX-021.STEP	145.0 × 100.0 × 50.0	80/179/103	Fail (0.452)	Fail (0.497)	Fail (0.466)
cad-bench-16	cad-000041	mania_2006.stp	85.6 × 47.0 × 28.7	127/287/164	Fail (0.470)	Fail (0.517)	Fail (0.485)
cad-bench-17	cad-000042	phase 1.STEP	145.0 × 124.6 × 80.0	56/134/80	Fail (0.489)	Fail (0.000)	Fail (0.000)
cad-bench-18	cad-000043	swing arm.STEP	137.0 × 60.0 × 49.6	130/325/193	Fail (0.000)	Fail (0.557)	Fail (0.530)
cad-bench-19	cad-000044	mania_2009.stp	163.6 × 70.0 × 64.7	102/239/138	Fail (0.491)	Fail (0.544)	Fail (0.507)
cad-bench-20	cad-000045	AX108B.STEP	155.1 × 81.1 × 69.0	77/179/98	Fail (0.532)	Fail (0.000)	Fail (0.000)
20-Task Mean Reward / Pass@1 Summary			**20 Unique Solids**	**Mean** 77.1 **Faces**	**0.463 (15.0%)**	**0.391 (10.0%)**	**0.371 (10.0%)**

Real Computer Environments," *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[2] R. Wu, C. Xiao, and C. Zheng, "DeepCAD: A Deep Generative Network for Computer-Aided Design Models," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6772–6782, 2021.

[3] J. G. Lambourne, K. D. D. Willis, P. K. Jayaraman, et al., "BRepNet: A Topological Message Passing System for Solid Models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12773–12782, 2021.

[4] R. Rafailov, A. Sharma, E. Mitchell, et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[5] Z. Shao, P. Wang, Q. Zhu, et al., "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," *arXiv preprint arXiv:2402.03300*, 2024.

[6] International Organization for Standardization, "ISO 10303-21:2016 Industrial automation systems and integration — Product data representation and exchange — Part 21: Clear text encoding of the exchange structure," 2016.

[7] International Organization for Standardization, "ISO 128-3:2020 Technical product documentation (TPD) — General principles of representation — Part 3: Views, sections and cuts," 2020.