CALIPER: DIRECT POINT–TEXT ALIGNMENT FOR DIMENSION-BEARING CAD RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieving CAD models from a geometry-only gallery with natural-language text differs from open-vocabulary 3D retrieval in what the query contains. Engineers describe a part by its dimensions, so retrieval must match stated numbers against geometry. The prevailing design, which anchors a point encoder to a frozen CLIP text tower, discards this information on both sides, truncating expert captions before their dimensions and normalizing away the scale those dimensions refer to, and it freezes the tower that would have to learn them. This paper presents CALIPER, a point–text bi-encoder for geometry-only CAD galleries that keeps the whole caption and each shape's corpus-frame scale and adapts both towers to paired CAD data. Ablations show that each preserved quantity and each adapted tower is load-bearing. Because the standard Text2CAD split leaks a majority of its evaluation shapes into training through near-duplicates, results are reported on a component-disjoint split, where CALIPER reaches R@10 0.955 under cluster credit and 0.848 by exact uid. Counterfactual probes establish aggregate-level grounding of the numbers in a query, and on human-written CADPrompt queries over held-out geometry CALIPER retrieves dimension-bearing prompts seven times as often as the strongest zero-shot baseline. Code is available at `https://anonymous.4open.science/r/CALIPER-5D13`.

1 INTRODUCTION

Retrieving CAD models from a geometry-only gallery using natural-language text is a fundamental problem in fields such as engineering search, underlying parts-catalog lookup, deduplication before a new design is committed, and retrieval-augmented CAD generation. Formally, given a gallery of shapes $\mathcal{G} = \{g_i\}_{i=1}^{N}$, each a point cloud with no parametric construction history, and a text query q, the task is to find encoders $f : \mathcal{G} \to \mathbb{R}^d$ and $h : \mathcal{Q} \to \mathbb{R}^d$ such that retrieval reduces to

$$i^\star = \arg\max_i \ \mathrm{sim}\big(f(g_i), h(q)\big), \tag{1}$$

where $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity. What separates CAD from general open-vocabulary 3D retrieval is what q contains. Engineers describe a part by its dimensions, for instance a hollow cylinder with an outer diameter of 0.7027 units and a wall thickness of 0.0493 units, where the sizes are stated in the coordinate frame of the corpus (§4). Human-written prompts confirm this habit (Alrashedy et al., 2025), so $i^\star$ is picked out by matching dimensions rather than a category label.

The prevailing design in 3D text retrieval aligns a point encoder to a fixed text space. ULIP, OpenShape, and Uni3D use a pretrained CLIP text tower for this purpose (Xue et al., 2023; 2024; Liu et al., 2023; Zhou et al., 2024; Radford et al., 2021). They optimize only f with

$$\min_f \ \mathbb{E}_{(g,q)} \left[-\log \frac{\exp\big(\mathrm{sim}(f(g), h_{\mathrm{CLIP}}(q))/\tau\big)}{\sum_{g'} \exp\big(\mathrm{sim}(f(g'), h_{\mathrm{CLIP}}(q))/\tau\big)} \right], \qquad h_{\mathrm{CLIP}} \text{ frozen.} \tag{2}$$

This is a sensible default for everyday objects, whose captions emphasize categories and visible attributes already represented in CLIP's text space.

Gallery captioning takes a different path. It produces a description for each gallery shape and retrieves text-to-text. Captions can come from rendered views (Luo et al., 2023; Liu et al., 2023; Xue et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

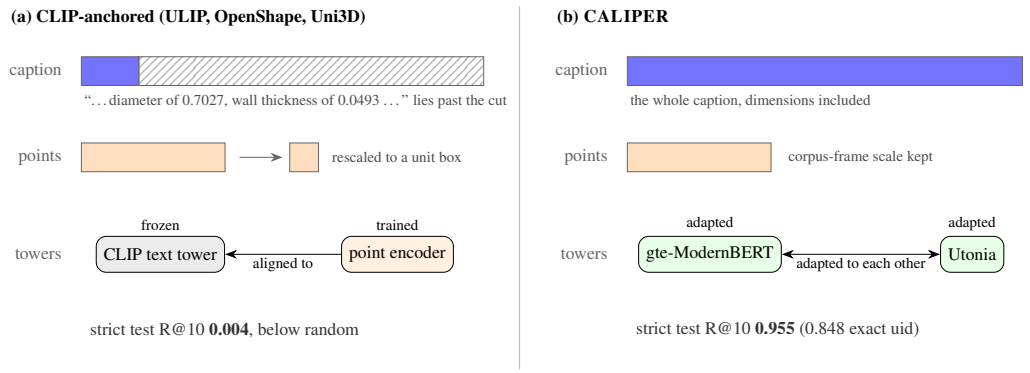


Figure 1: **Two design choices, and what the anchored pipeline does instead.** Rows one and two concern the information a CAD query depends on. The anchored pipeline cuts the caption before its dimensions and rescales the points to a unit box, so the numbers a query states have nothing to match against, whereas CALIPER keeps the whole caption and the corpus-frame scale. Row three concerns who learns. The anchored pipeline freezes the text tower and aligns only the point encoder to it, whereas CALIPER adapts both towers to each other. On the leak-free strict split the anchored baselines fall below random and CALIPER reaches R@10 0.955.

2024) or directly from point clouds (Xu et al., 2024b; Qi et al., 2024). This requires a captioner that can recover dimensions from geometry, which these systems are not designed to do. Even an oracle-caption control trails direct alignment (Appendix D).

Figure 1 shows the information the anchored pipeline loses for CAD queries. In the evaluated anchored models, some of the metric information needed for CAD retrieval is lost before alignment begins. Long Text2CAD expert captions (Khan et al., 2024) are often truncated before their identifying dimensions appear, while per-object unit-box normalization removes size in the corpus coordinate frame. The frozen text tower also cannot adapt to metric details that do remain in a query. The point encoder can therefore align only to distinctions already present in the text embeddings, making closely related parts difficult to separate.

Earlier work has also trained both towers on paired shape–text data, as Text2Shape, TriCoLo, and Parts2Words did from scratch with appearance captions (Chen et al., 2018; Ruan et al., 2024; Tang et al., 2023). In this work, we propose CALIPER, a point–text bi-encoder for geometry-only CAD galleries that takes this route with pretrained towers and dimension-bearing text. It retains dimension-bearing text and each shape's size in the corpus coordinate frame, then adapts both towers to paired CAD data. A gte-ModernBERT text encoder accepts up to 2,048 tokens, while a self-supervised Utonia point encoder processes points in cartesian space without per-object rescaling. Both towers are adapted with a contrastive loss in a shared embedding space, starting from encoders that have not been pretrained on paired text–3D data.

On the strict test split, which prevents near-duplicate overlap with training (§4), CALIPER reaches R@10 of 0.955 with cluster credit and 0.848 for the exact uid. Counterfactual probes show that retrieval responds to the numbers in a query and distinguishes bounding-box-matched shapes with different internal dimensions. These results support aggregate numeric grounding, while binding numbers to named dimensions remains undetected. On human-written CADPrompt queries over geometry held out from training, CALIPER retrieves dimension-rich prompts seven times as often as the strongest zero-shot baseline.

This work makes four contributions.

- CALIPER, a directly aligned point–text bi-encoder for geometry-only galleries that retains long, dimension-bearing descriptions and corpus-frame scale while adapting both towers.
- Controlled tests that separate what the caption keeps, what the points keep, and whether the towers adapt. Identical-text comparisons show that a CLIP-family text tower performs comparably when input and adaptation are matched.

- An audit of near-duplicate geometry in DeepCAD-family data, quantification of leakage in the official split, and a component-disjoint split with a two-track evaluation protocol.
- A retrieval evaluation on human-written CADPrompt queries, conducted on geometry held out from training, with paired-bootstrap comparisons across baselines.

2 Related Work

Text-to-shape retrieval. ULIP/ULIP-2 (Xue et al., 2023; 2024), OpenShape (Liu et al., 2023), and Uni3D (Zhou et al., 2024) align 3D representations to a pretrained CLIP text space. Text2Shape (Chen et al., 2018), TriCoLo (Ruan et al., 2024), and Parts2Words (Tang et al., 2023) learn shared shape–text embeddings from paired descriptions. SCA3D (Ren et al., 2025) adds compositional augmentation, while RoMa (Feng et al., 2025) addresses noisy point–text correspondence. Their evaluations largely use descriptions of object category, appearance, and parts, leaving retrieval with numeric CAD descriptions less explored.

Point-cloud encoders. The geometry backbones used in text-to-shape retrieval range from Point-BERT and sparse convolutions in OpenShape (Liu et al., 2023) to a ViT initialized from 2D in Uni3D (Zhou et al., 2024). Point Transformer V3 (Wu et al., 2024) provides a backbone for self-supervised point representation learning. Sonata (Wu et al., 2025) and Concerto (Zhang et al., 2025) develop this line within individual domains, while Utonia (Zhang et al., 2026) trains across indoor, outdoor, remote-sensing, object, and video-derived point clouds. Utonia serves as the point-tower initialization here, and its CAD-specific input handling is described in §3.

Text-tower adaptation. LiT (Zhai et al., 2022) and LIFT (Yang et al., 2025) study opposite choices of which tower to keep fixed during image–text alignment. Long-CLIP (Zhang et al., 2024) and TULIP (Najdenkoska et al., 2024) extend the text context of CLIP-family encoders. LLM2CLIP (Huang et al., 2024) and ProCLIP (Hu et al., 2025) align visual representations with LLM-based text embedders. For CAD retrieval, the question is how much of the gap comes from the text available to the encoder and how much comes from the encoder itself. The identical-text controls of §5.3 examine that distinction.

CAD captions and retrieval. Text2CAD (Khan et al., 2024) introduced captions at several levels of detail for DeepCAD sequences. CADmium (Govindarajan et al., 2026) provides new descriptions of those models, while CAD-MLLM (Xu et al., 2024a) introduced Omni-CAD with text, images, points, and command sequences. These works provide paired CAD annotations, whereas our gallery contains only geometry at retrieval time. The closest task comparison is IMMA (Pan et al., 2026). It retrieves CAD models from short abstract captions on the official DeepCAD split and encodes each gallery item's parametric command sequence alongside its point cloud. Appendix F gives a reference comparison under IMMA's protocol.

3 Method

CALIPER follows two commitments that answer the losses described in the introduction. The first is to preserve the information a CAD query depends on, keeping the whole caption on the text side and each shape's corpus-frame scale on the point side. The second is to let both towers learn that information, adapting each to the other rather than anchoring one to a space learned from another modality. The architecture, objective, and schedule below instantiate these commitments with off-the-shelf encoders.

3.1 Architecture

CALIPER maps a caption q and a point cloud g to embeddings $h(q)$ and $f(g)$ in a shared 512-dimensional space. The text tower is gte-ModernBERT-base, a 149M-parameter encoder with a 2,048-token context. Its 768-dimensional CLS vector supplies the text representation.

The point tower is Utonia (Zhang et al., 2026), a 137M-parameter PTv3 encoder. Its input is 8,192 xyz points centered on the shape's bounding box without per-object scale normalization.

Utonia's pretraining treats color and normals as optional, allowing xyz-only input. Centering removes translation without changing lengths. A fixed coordinate multiplier shared across shapes sets voxel granularity, leaving corpus-frame lengths comparable across the gallery.

Utonia produces a hierarchy of point tokens. An 8-head attention pool reduces its coarsest-level tokens, with width 576 and a median count of 90, to one vector. The text vector passes through a $768 \to 1024 \to 512$ projection MLP, and the pooled point vector passes through a $576 \to 1024 \to 512$ projection MLP. Their outputs are L2-normalized and compared by cosine similarity. The projection heads contain about 3M trainable parameters in total. At retrieval time, the gallery supplies point clouds without captions or parametric construction histories.

3.2 Cluster-aware alignment

A training batch contains B point–caption pairs (g_i, q_i). Let c_i be the tight near-duplicate cluster of g_i, and let $\mathcal{P}_i = \{j \mid c_j = c_i\}$ contain its positive indices within the batch. The clusters are constructed in §4. We define $s_{ij} = \mathrm{sim}(h(q_i), f(g_j))/\tau$, where τ is learned in log space. The text-to-point and point-to-text losses are

$$\mathcal{L} = \frac{1}{2B} \sum_i \left[\log \sum_k e^{s_{ik}} + \log \sum_k e^{s_{ki}} - \frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \left(s_{ij} + s_{ji} \right) \right]. \tag{3}$$

Both directions average over every available cluster positive. Near-duplicate shapes in the batch therefore contribute as positives rather than negatives.

3.3 Training schedule and caption sources

Training starts with Text2CAD L3 expert captions (Khan et al., 2024). First, both encoders are frozen and the projection heads are trained for 30 epochs with batch size 1024 and learning rate 10^{-3}. Next, the point tower is unfrozen progressively, from its final encoder block through the penultimate block to the full tower. This stage uses batch size 64 and point-tower learning rates up to 10^{-4}. Live encoding replaces the point-token cache as the point tower becomes trainable.

After point-tower adaptation, CADmium captions (Govindarajan et al., 2026) are added to the L3 pairs. This L3+CADmium training produces CALIPER. The CALIPER-mix3 variant also uses short Omni-CAD captions (Xu et al., 2024a), with 71,345 pairs retained after collision filtering (Appendix G). In the final stage, the adapted point tower is frozen as a token cache while the top $k \in \{8, 16, 22\}$ text layers are progressively unfrozen at learning rate 10^{-5}. The last setting adapts the full text tower. CALIPER-mix3 is used for human-written and short queries, since a model trained on a single caption style collapses on the others, whereas the three-source mix gives up about one point in-domain and more than quadruples recall on short, dimension-free queries (Table G.3).

4 Data Audit and Evaluation Protocol

4.1 Corpus and split construction

Text2CAD v1.1 (Khan et al., 2024) provides 175,556 point-cloud uids from DeepCAD (Wu et al., 2021) and ABC (Koch et al., 2019). DeepCAD geometry is normalized to a largest extent of 0.75, and we verified that caption dimensions refer to this shared coordinate frame. Absolute scale in this paper therefore means scale within the corpus frame rather than physical units. The L3 captions average 285 words and have a median length of 355 tokens under the gte-ModernBERT tokenizer. Fewer than 0.1% exceed the 2,048-token context of our text tower.

The published Text2CAD split contains substantial geometric overlap between training and evaluation. We measure this overlap by linking shapes i and j when their Chamfer distance falls below $1.5 \max(\mathrm{floor}_i, \mathrm{floor}_j)$, where each floor estimates the shape's subsampling noise. Under this criterion, 57.6% of shapes have a verified near-duplicate. In the published split, 61.5% of validation shapes and 57.9% of test shapes have a near-duplicate in training. Retrieval on that split can therefore benefit from recognizing geometry already seen during training. Threshold sensitivity is reported in Appendix G.

Table 1: **Text2CAD retrieval** (single-query, cluster credit unless noted). The strict track is primary. The official split leaks a majority of its test shapes into training (§4), and the difference between the two column groups is the inflation. Random R@10 is given under both credit rules. 95% bootstrap CIs have half-width at most 0.011, and full R@1/R@5/R@10 metrics are in Tables C.1–C.2.

	Strict test (7,760)			Official test (8,021)		
Model	R@1	R@10	med	R@1	R@10	med
Reference						
random, cluster credit	0.002	0.017	—	0.000	0.005	—
random, exact uid	0.000	0.001	—	0.000	0.001	—
bbox-dims shortcut	0.146	0.239	—	—	—	—
Zero-shot CLIP-anchored encoders						
Uni3D-g	0.001	0.004	2901	0.000	0.004	3179
OpenShape-G14	0.001	0.004	2805	0.000	0.004	3115
Ours						
CALIPER	**0.738**	**0.955**	1	**0.800**	**0.982**	1
same, exact uid	—	0.848	—	—	0.948	—

We consolidate the links into 88,645 tight clusters and 76,560 single-linkage components. Tight clusters restrict training positives and retrieval credit to closely matching shapes. Single-linkage components collect every shape connected by a chain of near-duplicate links and define split boundaries so that no linked pair crosses between training and evaluation. The *strict track* splits components 90/5/5 into 160,036 training, 7,760 validation, and 7,760 test shapes and supplies the primary results. The *official track* retrains the complete model on the published Text2CAD split of 158,630 training, 8,905 validation, and 8,021 test shapes for comparison with prior work.

4.2 Retrieval evaluation

We report R@1, R@5, R@10, and median rank. Under *cluster credit*, retrieving any member of the target's tight cluster counts as a hit. Under *exact-uid credit*, only the target uid counts. The median number of credited shapes per query is one. A duplicate-free gallery reproduces cluster-credit recall within half a point, with further checks in Appendix G. Cluster credit also cannot conceal dimension differences that the model itself resolves. Credit clusters deviate by a median of 0.03% in extent and only 5.4% contain a member deviating by more than 10%, whereas a 10% uniform change in the queried numbers already costs the model 25 points (§5.4). Because the two credit rules have different chance rates, random recall is stated under both for the main galleries.

Confidence intervals use the 95% bootstrap, and comparisons between models use paired bootstrap. Repeated evaluation varies by approximately ± 0.07 points. All test-set contacts were pre-registered and selected on validation only (Appendix A).

5 Experiments

The experiments report retrieval on both tracks, ablate the two design choices of §3, probe what the aligned model grounds, and test human-written queries on held-out geometry and transfer outside the corpus. Unless noted, controlled comparisons use the strict validation split.

5.1 Text2CAD retrieval

On the strict test, CALIPER places the described shape or a verified duplicate in the top ten for 95.5% of expert queries and the exact shape for 84.8%, against 0.4% for the zero-shot CLIP-anchored encoders and 23.9% for the bounding-box shortcut (Table 1). Test agrees with validation to within 0.2 points. Trained baselines that keep a CLIP-family text tower and share every other component of the recipe reach 0.472, 0.847, and 0.909 on strict validation with 77-, 248-, and 512-token towers (Table 3). CADmium-style queries, a larger text tower, and full R@1/R@5/R@10 breakdowns are in Appendices C and G.

Table 2: **Cost of each intervention** (strict val, cluster R@10, change in points relative to the row's reference condition). Full cells are in Tables G.1, G.2, and E.1.

Axis	Intervention	Δ
Dimensions in the caption	delete all dimensions from captions	−20.2
	truncate the same backbone to CLIP's 77-token window	−20.6
	coarser L2 captions instead of L3	−5.5
	raw ModernBERT instead of contrastive initialization	−1.2
Scale of the geometry	unit-extent normalization, dimension-bearing queries	−4.0
	unit-extent normalization, dimension-free queries	+4.6
Adaptation of the towers	unfreeze the point tower (over frozen towers)	+21.6
	unfreeze the text tower (over a frozen-text control)	+25.7
	CLIP-family text tower on identical text, after adaptation	+5

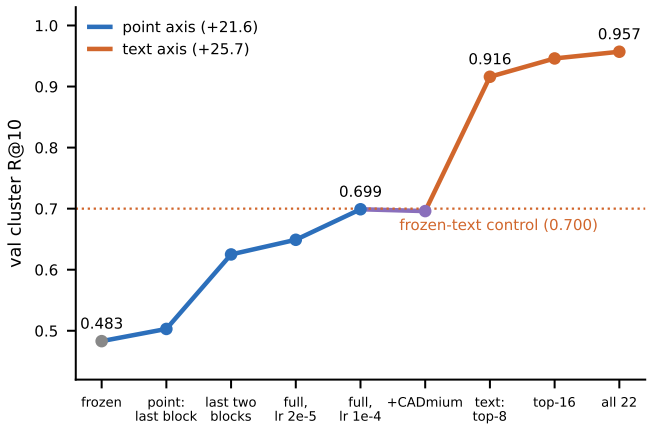


Figure 2: **Trajectory of the staged adaptation** (strict val, cluster R@10, per-stage numbers in Table G.2). Point-tower unfreezing (+21.6pt), then text-tower unfreezing (+25.7pt over a frozen-text control, dotted).

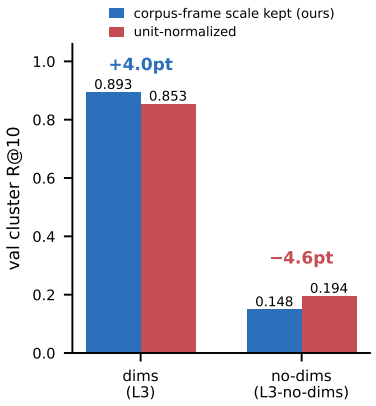


Figure 3: **Effect of scale preservation by query type.**

The official track reaches 0.982. The audit of §4 predicts that its advantage should shrink as the model learns to read the query rather than recognize seen geometry, and under matched schedules it does, from 12.9 points after heads-only training to 5.8 after the point unfreeze and 1.3 after text fine-tuning. The strict track is therefore the evidence and the official track a reference.

5.2 Ablation of the preserved information

The first design choice rests on the claim that dimension-bearing retrieval depends on the dimensions in the caption and on the scale of the geometry they refer to. Table 2 removes each in turn with the rest of the recipe fixed.

On the text side, deleting every dimension costs 20.2 points, nearly four times the bounding-box shortcut ceiling, and truncating the *same* frozen backbone to CLIP's 77-token window costs 20.6, almost exactly the cost of deletion, as expected if the window ends before the dimensions appear. Coarser L2 captions cost 5.5 points and a raw ModernBERT initialization 1.2, so the loss follows the information in the caption rather than the pretraining of the tower. The zero-shot baselines are the limiting case. Their rows sit at or below cluster-credit random because long captions passed through a frozen, truncating tower collapse to near-identical embeddings, and under full adaptation the gap between a tower that reads part of the caption and one that reads all of it persists with none of it attributable to the tower (Appendix E).

On the point side, unit-extent normalization, the Uni3D and OpenShape convention, drops the point stage by 4.6 points, and the sign of the effect depends on the query as predicted. Preserving scale

Table 3: **Identical-text twins and window extension** (strict val, cluster R@10, full staged recipe). Each gte row reads the *same* CLIP-BPE-truncated string as the CLIP-family row above it. The tower accounts for about five points at either window, whereas the text the window discards accounts for the rest of the gap to the full caption, and stretching the CLIP-family tower to 512 tokens closes most of that gap. Frozen-tier cells and the context-length sweep are in Appendix E.

Text tower	reads	R@10	gap to full caption
CLIP-L/14	77 CLIP tokens	0.472	-48.5
gte-ModernBERT	identical 77-token text	0.423	-53.4
Long-CLIP-L	248 CLIP tokens	0.847	-11.0
gte-ModernBERT	identical 248-token text	0.798	-15.9
Long-CLIP-L, stretched	512 CLIP tokens	0.909	-4.8
gte-ModernBERT (CALIPER)	full caption, 2,048 tokens	**0.957**	—

Table 4: **Dimension-perturbation probe** (1,000 strict-val queries, every decimal literal multiplied by f). Rank degrades with the size of the perturbation and the top retrieval's volume follows the queried size, whereas a matched typo control on non-numeric words leaves median rank at 1 and R@10 at 0.900.

f	0.5	0.75	0.9	1.0	1.1	1.25	1.5	2.0
median rank	45	23	3	1	3	13	50	42
R@10	0.318	0.399	0.706	0.951	0.694	0.466	0.315	0.320
top-1 volume ratio	0.31	0.85	0.94	1.00	1.00	1.14	1.09	1.53

helps dimension-bearing queries by 4.0 points and hurts dimension-free ones by 4.6 (Figure 3), since retained size variance is signal when the caption states dimensions and noise when it does not. The effect is the smallest in Table 2 but the only one whose sign flips with the query type, and it excludes the unit-box convention for CAD.

5.3 Ablation of tower adaptation

The second design choice rests on the claim that both towers must adapt, so freezing either should cost accuracy the other cannot recover, and that once both adapt the identity of the text tower should matter little.

Along the trajectory, unfreezing the point tower adds 21.6 points over the frozen-tower ceiling, with the penultimate encoder stage contributing most, and unfreezing the text tower adds 25.7 points over a same-path control that keeps it frozen, monotone in the number of unfrozen layers. Neither gain is recoverable from the other axis, and seed stability and a joint-training control are in Appendix G.

With both towers adapted, a CLIP-family tower and the long-context tower fed the identical truncated string land within five points of each other, with the CLIP tower slightly ahead (Table 3), and the tower gap of the frozen tier disappears under adaptation (Appendix E). Stretching the Long-CLIP tower to 512 tokens by positional interpolation reaches 0.909, within 4.8 points of the full caption, so the gap closes as the window opens. The CLIP text tower lacks adaptation and input, not representation. The point side offers a one-directional check. Frozen Uni3D-g features paired with the same text tower and heads outperform frozen Utonia tokens (0.517 against 0.471), so Uni3D's collapse does not trace to its geometry encoder, and re-randomizing Utonia before the point stage costs only six points (0.634 against 0.695), so the recipe does not depend on its pretrained features either.

5.4 Probing dimension grounding

The ablations show that the numbers in a caption carry most of the retrievable signal. Three probes test what the model does with them at increasing granularity, a uniform rescaling of every number, bounding-box-matched hard pairs, and single-value edits. Grounding here means that retrieval responds to the numeric values with the description held fixed, not that dimensions alone can retrieve, since dimension-only template queries fall below the bounding-box shortcut (Appendix D).

Table 5: **Internal-dimension discrimination on bounding-box-matched hard pairs** (strict val, 2,000 pairs, forced choice). Pairs agree within 2% on every bounding-box axis, belong to different tight clusters, and are confusable (cosine > 0.85) for the unit-normalized control encoder of §5.2, so the mining model is independent of the evaluated one.

Query	forced-choice accuracy	95% CI
full L3 caption	**0.713**	[0.693, 0.733]
caption with numbers stripped	0.561	[0.538, 0.582]
chance	0.500	—

Table 6: **Human-written queries (CADPrompt).** R@10 on a 7,942-shape gallery, n=182 prompts (exact-uid random 0.0013). "dims" marks prompts with annotated dimensions. All CALIPER deltas against every baseline are significant under paired bootstrap, the smallest being Δ=+0.137 [+0.077, +0.198]. The compressed condition LLM-rewrites dims prompts to fit CLIP's window with numbers preserved. Every CALIPER row comes from a chain retrained with the 190 CADPrompt geometries and their 3,241 near-duplicates removed from train and validation. Seen-geometry rows are in Table C.4.

Model	plain	dims	dims, compressed to CLIP window
Uni3D-g (zero-shot)	0.049 [0.022, 0.082]	0.027 [0.005, 0.055]	0.049 [0.022, 0.082]
OpenShape-G14 (zero-shot)	0.033 [0.011, 0.060]	0.016 [0.000, 0.038]	0.038 [0.011, 0.066]
CLIP-twin B/32, frozen	0.011 [0.000, 0.027]	0.016 [0.000, 0.038]	—
CLIP-twin L/14, fully adapted	0.049 [0.022, 0.082]	0.198 [0.143, 0.258]	—
gte frozen, heads only	0.044 [0.016, 0.077]	0.071 [0.038, 0.110]	—
CALIPER-mix3 (held-out geometry)	**0.187** [0.132, 0.247]	**0.357** [0.291, 0.429]	—

Under uniform rescaling, multiplying every decimal literal in 1,000 captions by f degrades the true shape's median rank near-monotonically in $|\log f|$, a 10% perturbation drops R@10 from 0.951 to about 0.70, and the top retrieval's volume tracks f, whereas a matched typo control on non-numeric words leaves retrieval unchanged (Table 4). Retrieval responds to the numbers, but a uniform edit cannot separate internal dimensions from overall size.

Table 5 makes that separation with bounding-box-matched pairs from different clusters, where the model must pick the described member. With the full caption it does so 71.3% of the time, with the numbers stripped 56.1%, against chance of 50%, so the numbers contribute 15.2 points that cannot come from extent. In full-gallery retrieval the hard negative outranks the ground truth for 28.9% of these queries (ground-truth median rank 12), so forced choice measures pairwise discrimination rather than retrieval accuracy.

Single-value edits bound the granularity from above. Doubling one randomly chosen decimal leaves retrieval unchanged (0.957 to 0.958), two-significant-digit rounding costs 0.4 points, and one-significant-digit rounding, roughly a 30% distortion, costs 24. Exchanging the values attached to "length of" and "width of" in the 290 captions that state both leaves rank unchanged (0.959 to 0.962). Retrieval is therefore insensitive to a change in any single stated value, and assignment of numbers to named dimensions is not detected, though a swapped caption can also describe a rotated variant of the same geometry, so the swap probe bounds the claim rather than refuting role binding. The model grounds the numeric profile of a caption as a whole rather than its individual dimensions, and the claim made here is aggregate-level.

5.5 Human-written queries and out-of-corpus transfer

Every result above uses machine captions of in-corpus geometry. The evaluations here test whether the design carries over to queries written by a person about targets never seen in training, and to geometry outside the corpus altogether.

Table 6 reports, to our knowledge, the first retrieval evaluation on CADPrompt's human prompts. Every CALIPER row comes from a chain retrained with the CADPrompt geometries and their near-duplicates removed, under a pre-registered single contact, so every target is unseen, and the held-out chain matches the original on validation (0.947 against 0.946, Appendix C). The model is

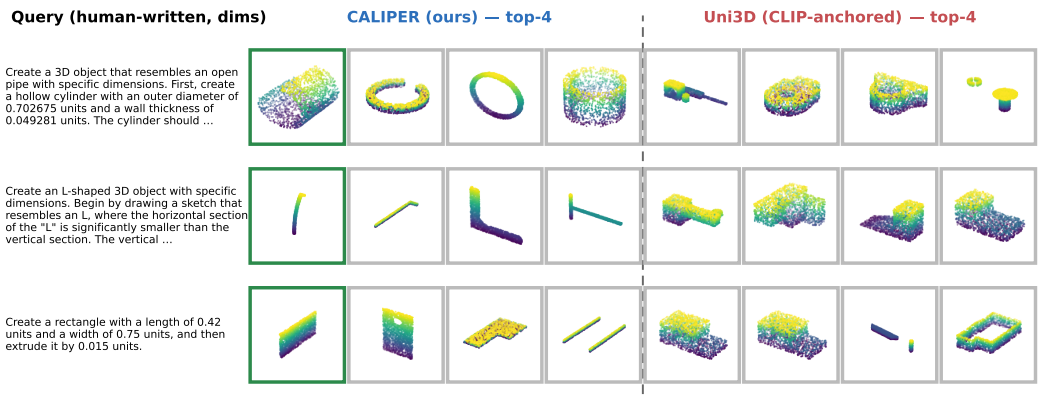


Figure 4: **Retrieval on human-written CADPrompt queries** (dims variant, all query geometries unseen in training). Green frames mark the ground-truth duplicate cluster.

CALIPER-mix3, since the prompts are short and often dimension-free (§3.3). CALIPER exceeds every baseline in every cell, and on dims prompts its median rank is 52 against 930 for Uni3D (Figure 4). Two cells connect back to the ablations. Compressing dims prompts into CLIP's window with the numbers preserved barely moves the zero-shot encoders, so a frozen tower cannot use the numbers even when they fit, and the fully adapted CLIP twin improves substantially on dims prompts yet trails CALIPER by sixteen points, mirroring its in-domain ceiling.

Against prior CAD retrieval, under IMMA's setting of short abstract captions on the official split, an abstract-caption-tuned CALIPER reaches R@10 0.379 and median rank 20 against IMMA's reported 0.361 and 25, without the parametric command sequence IMMA additionally encodes. IMMA has no public code, so this is a reference point only (Appendix F).

The human-query numbers remain far from deployment grade, and two controls bound the gap to machine captions on the same shapes (roughly 0.36 against 0.96). LLM paraphrasing of the prompts moves R@10 by a non-significant point, whereas substituting captions that carry ground-truth-derived information fixes about two thirds of the failures (Appendix D), so the gap is one of information in the query and the practical lever is query elaboration.

Fusion360 Gallery supplies 8,307 human-designed, ABC-disjoint shapes, the only fully out-of-corpus geometry in the evaluation. Because the set was consulted during development, the reported numbers come from a pre-registered sealed probe with frozen checkpoints. On never-consulted VLM captions of the dataset's own renders, CALIPER-mix3 reaches R@10 0.129 against a random 0.0012, 108 times random, and passes the pre-declared criterion. A second sealed set of dimension-only template queries fell below its bar. The same templates fail in-corpus (§5.4), so template information poverty rather than transfer failure is the dominant reading, and out-of-corpus dimension grounding remains untested rather than refuted. The dev-consulted trajectory and the non-CAD boundary are in Appendix G.

6 Limitations and Conclusion

CALIPER retrieves CAD geometry from detailed text by preserving corpus-frame point coordinates and adapting both towers. On the component-disjoint Text2CAD test, it reaches cluster-credit R@10 of 0.955, and on human-written dimension-bearing queries over held-out geometry, CALIPER-mix3 reaches 0.357. Matched controls show that CLIP-family text towers can perform comparably when given the same text and adaptation, making caption coverage and training central to the observed gains. Numeric probes support aggregate dimension grounding within the corpus, but do not establish that individual values bind to named dimensions. Out-of-corpus dimension grounding remains untested, and the human-query evaluation contains only 182 prompts. These limits leave open whether the same gains hold for broader CAD collections and more precise dimensional queries.

AI USE STATEMENT

We used generative AI tools for code implementation, experiment orchestration and monitoring, literature search, and manuscript drafting and editing. Research direction, experimental design, pre-registration commitments, and all final claims were set and reviewed by the authors; AI-generated code was verified against held-out sanity checks, and every reported number traces to archived run artifacts (Appendix B). We take responsibility for the final content of this work.

ETHICS STATEMENT

This work uses only publicly released CAD datasets and human-written prompt sets, involves no human subjects or personal data, and releases weights under the non-commercial terms inherited from its pretraining data. We disclose dataset licensing status and our full test-contact protocol (Appendix A).

REPRODUCIBILITY STATEMENT

Code, deduplicated splits, audit artifacts, and the full test-contact log are available at `https://anonymous.4open.science/r/CALIPER-5D13` and as a frozen snapshot in the supplementary material.

REFERENCES

Kamel Alrashedy, Pradyumna Tambwekar, Zulfiqar Zaidi, Megan Langwasser, Wei Xu, and Matthew Gombolay. Generating CAD code with vision-language models for 3D designs. In *ICLR*, 2025.

Kevin Chen, Christopher B. Choy, Manolis Savva, Angel X. Chang, Thomas Funkhouser, and Silvio Savarese. Text2Shape: Generating shapes from natural language by learning joint embeddings. In *ACCV*, 2018.

Yanglin Feng, Yang Qin, Dezhong Peng, Hongyuan Zhu, Xi Peng, and Peng Hu. PointCloud-text matching: Benchmark datasets and a baseline. *IEEE Transactions on Multimedia*, 2025. arXiv:2403.19386.

Prashanth Govindarajan, Davide Baldelli, Jay Pathak, Quentin Fournier, and Sarath Chandar. CADmium: Fine-tuning code language models for text-driven sequential CAD design. *Transactions on Machine Learning Research*, 2026.

Xiaoxing Hu, Kaicheng Yang, Ziyang Gong, Qi Ming, Zonghao Guo, Yuan Tian, Xiang An, Ziyong Feng, and Xue Yang. ProCLIP: Progressive vision-language alignment via LLM-based embedder. *arXiv preprint arXiv:2510.18795*, 2025.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Chunyu Wang, Xiyang Dai, Dongdong Chen, Chong Luo, and Lili Qiu. LLM2CLIP: Powerful language model unlocks richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024.

Mohammad Sadil Khan, Sankalp Sinha, Talha Uddin Sheikh, Didier Stricker, Sk Aziz Ali, and Muhammad Zeshan Afzal. Text2CAD: Generating sequential CAD designs from beginner-to-expert level text prompts. In *NeurIPS*, 2024.

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 *CVPR*, 2019.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. In *NeurIPS*, 2023.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M. Asano, Nanne van Noord, Marcel Worring, and Cees G. M. Snoek. TULIP: Token-length upgraded CLIP. *arXiv preprint arXiv:2410.10034*, 2024.

Honghu Pan, Zibo Du, Daxiang Liu, Chengliang Liu, and Xiaoling Luo. Text-to-CAD retrieval: a strong baseline. *arXiv preprint arXiv:2605.05572*, 2026.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. ShapeLLM: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision (ECCV)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Junlong Ren, Hao Wu, Hui Xiong, and Hao Wang. SCA3D: Enhancing cross-modal 3D retrieval via 3D shape and caption paired data augmentation. In *ICRA*, 2025. arXiv:2502.19128.

Yue Ruan, Han-Hung Lee, Yiming Zhang, Ke Zhang, and Angel X. Chang. TriCoLo: Trimodal contrastive loss for text to shape retrieval. In *WACV*, 2024.

Chuan Tang, Xi Yang, Bojian Wu, Zhizhong Han, and Yi Chang. Parts2Words: Learning joint embedding of point clouds and texts by bidirectional matching between parts and words. In *CVPR*, 2023.

Rundi Wu, Chang Xiao, and Changxi Zheng. DeepCAD: A deep generative network for computer-aided design models. In *ICCV*, 2021.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *CVPR*, 2025.

Jingwei Xu, Zibo Wang, Zhiwen Zhao, Wen Liu, Yi Ma, and Shenghua Gao. CAD-MLLM: Unifying multimodality-conditioned CAD generation with MLLM. *arXiv preprint arXiv:2411.04954*, 2024a.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. PointLLM: Empowering large language models to understand point clouds. In *European Conference on Computer Vision (ECCV)*, 2024b.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In *CVPR*, 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP-2: Towards scalable multimodal pre-training for 3D understanding. In *CVPR*, 2024.

Jingfeng Yang, Ziyang Wu, Yue Zhao, and Yi Ma. Language-image alignment with fixed text encoders, 2025. URL `https://arxiv.org/abs/2506.04209`.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *ECCV*, 2024.

Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. In *NeurIPS*, 2025.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward one encoder for all point clouds. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.03283.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring unified 3D representation at scale. In *ICLR*, 2024.

A TEST-CONTACT PROTOCOL DISCLOSURE

Strict track. The strict test set (7,760) was contacted exactly twice. (1) 2026-07-07: the point-stage final model, one shot, gated by a validation dry-run anchor reproduction and an adversarial review; embeddings and per-query ranks were preserved, enabling derived metrics without re-contact. (2) 2026-07-09: pre-registered in the lab log *before* execution (candidate frozen, metric suite fixed, single pass, declared final); the tuned model was selected purely on validation. The held-out CADPrompt chain of §5.5 was never evaluated on the strict test set, and the 8 CADPrompt queries whose target geometry lies in that split are excluded from every reported number (see CADPrompt accounting below).

Official track. The official test set (8,021) is a public benchmark on which we impose the same discipline: every contact evaluates a model selected purely on validation. Contacts were: (a) the first chain's finals (gte 6-epoch; mix3 6-epoch); (b) the large-tower final (6-epoch); (c) the two converged 15-epoch models (base, large), under a dedicated pre-registration that declared the headline swap rule in advance; (d) the mix4 (+L0 source) variant (Table C.2); (e) the L0-specialist used in the IMMA comparison (single contact, pre-registered; Appendix F). No training or selection decision anywhere in the paper used an official-test result. Each preserved eval JSON carries a `preregistration` field naming its item above (a propagated strict-track label was corrected on 2026-07-30; the prior value is retained in a `metadata_correction` field; metric payloads untouched).

Fusion360 sealed probe. One pre-registered contact (two frozen candidates $\times$ two query sets), run in a single pass with both outcomes reported (§5.5).

CADPrompt accounting. The 200 prompts comprise 190 with point clouds and 10 without. Of the 190, 174 belong to the strict training split, 8 to validation, and 8 to test. The held-out chain removes all 190 geometries and their 3,241 near-duplicate uids from train and validation before every training stage and was contacted once, on 2026-09-18, scoring all 190 pre-registered queries in a single pass (the pre-registration commit precedes checkpoint creation, and the preserved evaluation record has n=190). The n=182 figure of Table 6 drops the 8 queries whose geometry belongs to the strict test split from the preserved per-query ranks, for protocol consistency with the seen-geometry model of Table C.4, and is a derivation rather than a re-contact. On these 182 queries the held-out chain reaches dims R@10 0.357 (R@1 0.170, median rank 51.5, reported as 52 in §5.5) and plain R@10 0.187 (median rank 127).

B REPRODUCIBILITY

All experiments ran on a single A100-80GB. A heads-only cell takes about 71 minutes; a point-tower full-unfreeze epoch takes 2,100–2,800 seconds over 7 epochs; text fine-tuning takes 2.4–4.2 hours; a full official-track retrain takes roughly 19 hours. Caching Utonia tokens takes 1.9 hours and each text source about 10 minutes. Configs are YAML with CLI overrides; all runs save resolved configs in checkpoints; splits and audit artifacts (dedup report, leakage report, caption statistics) ship with the code. Evaluation noise is quantified by 3-draw re-evaluation ($\pm$0.07 points).

C ADDITIONAL TABLES

This appendix collects full R@1/R@5/R@10 and median-rank tables for the CALIPER test conditions on both tracks (Tables C.1–C.2), the shape$\rightarrow$text direction (Table C.3), the CADPrompt results of the seen-geometry model (Table C.4), the duplicate-free-gallery protocol (Table C.5), and the per-stage validation numbers of the held-out CADPrompt chain next to the original chain (Table C.6). All values derive from the preserved evaluation artifacts of the contacts enumerated in Appendix A; no additional test contact was made. The near-agreement between the duplicate-free protocol (e.g. 0.952 strict-track) and cluster credit on the full gallery (0.955) confirms that cluster credit measures retrieval quality rather than duplicate luck.

Table C.1: **Strict-track test (7,760), full metrics**, text→shape, from the pre-registered test contact of Appendix A.

		exact uid				cluster credit			
Model	queries	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR
CALIPER (L3+CADmium)	L3	0.522	0.773	0.848	1	0.738	0.921	0.955	1
	CADmium	0.559	0.808	0.875	1	0.757	0.930	0.963	1

Table C.2: **Official-track test (8,021), full metrics**, text→shape. Every contact is enumerated in Appendix A. The 15-epoch rows are the converged schedule used in Table 1.

		exact uid				cluster credit			
Model	queries	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR
gte, 6 epochs	L3	0.620	0.886	0.940	1	0.765	0.951	0.973	1
	CADmium	0.623	0.890	0.945	1	0.760	0.953	0.978	1
mix3, 6 epochs	L3	0.606	0.879	0.938	1	0.747	0.943	0.971	1
	CADmium	0.610	0.883	0.943	1	0.748	0.948	0.977	1
mix4 (+L0), 6 epochs	L3	0.588	0.871	0.933	1	0.728	0.935	0.967	1
	CADmium	0.600	0.877	0.938	1	0.736	0.943	0.974	1
large, 6 epochs	L3	0.679	0.909	0.954	1	0.829	0.972	0.987	1
	CADmium	0.669	0.901	0.951	1	0.817	0.965	0.985	1
gte, 15 epochs (Table 1)	L3	0.654	0.900	0.948	1	0.800	0.964	0.982	1
	CADmium	0.651	0.899	0.948	1	0.791	0.963	0.983	1
large, 15 epochs	L3	0.712	0.919	0.958	1	0.865	0.981	0.992	1
	CADmium	0.692	0.913	0.957	1	0.842	0.976	0.990	1

D COUNTERFACTUAL DIMENSION PROBES AND PIPELINE CONTROL

All probes in this section are inference-only on the strict validation split; no test set was contacted and no training was performed. The hard-pair discrimination table of §5.4 is summarized there; mining details follow.

Hard-pair mining. Candidate pairs are drawn from the strict validation split. A pair qualifies when the two shapes agree within 2% on every bounding-box axis, belong to different tight clusters, and are placed at cosine above 0.85 by the unit-extent-normalized control encoder of §5.2, which is trained on the official split and is therefore independent of the evaluated model. The first two criteria yield 26,279 candidate pairs, and 2,000 pairs are drawn from the hard pool that also satisfies the cosine criterion, uniformly at random without replacement under a fixed seed. The query is the L3 caption of one member, and the model scores correctly when it ranks that member above the other. In the numbers-stripped condition every numeric literal, whether decimal, integer, or spelled out in words, is replaced by the placeholder token 'N', so that counts are removed along with dimensions and the shape words are untouched. The two conditions therefore differ only in the numbers.

Dimension-perturbation probe. For 1,000 validation queries, every decimal literal in the L3 caption is multiplied by a factor f (integers, degree-suffixed values, and Euler-angle lists are left unchanged; integers encode counts), and retrieval is rerun against the full 7,760-shape validation gallery with the final strict-track model. Table 4 reports the result. The true shape's median rank degrades near-monotonically and symmetrically in $|\log f|$ (Spearman 0.916, with a small tail reversal at frame saturation). What the model retrieves instead tracks the queried size: the top non-GT retrieval's bounding-box volume follows f (Spearman 0.976) until the corpus frame saturates upward (DeepCAD pins the maximum extent at 0.75, so shapes larger than the frame do not exist). Because sketch coordinates and scale factors are perturbed together, the volume row is a monotone readout, not a calibrated one. A matched typo control that applies the same number of distinct-letter swaps to non-numeric words leaves retrieval essentially unchanged (median rank 1, R@10 0.900), so the sensitivity is specific to the numbers. In the unit-normalized embedding space of the scale-ablated encoder (trained on the official split, so a third of the measured shapes were seen by the instrument, as was true for its random-pair baseline), the shapes retrieved under perturbed queries remain much

Table C.3: **Shape→text direction** (L3 captions), derived from the preserved embeddings without re-contact.

	exact uid				cluster credit			
Model	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR
CALIPER, strict track	0.497	0.758	0.838	2	0.736	0.924	0.959	1
gte, 6 epochs, official track	0.587	0.878	0.937	1	0.743	0.951	0.975	1
mix3, 6 epochs, official track	0.572	0.872	0.936	1	0.726	0.945	0.974	1

Table C.4: **CADPrompt, seen-geometry model,** n=190 (all usable prompts, including the 8 whose geometry lies in the strict test split). The CALIPER-mix3 row is the original chain, whose training set contains 174 of the 190 target geometries. Table 6 in the main text reports the held-out chain at n=182 instead. 7,942-shape gallery, cluster credit.

	plain				dims			
Model	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR
Uni3D-g (zero-shot)	0.000	0.037	0.047	714	0.000	0.011	0.026	930
OpenShape-G14 (zero-shot)	0.000	0.016	0.032	482	0.000	0.011	0.016	894
CLIP-twin B/32, frozen	0.000	0.005	0.011	1780	0.000	0.005	0.016	1746
CLIP-twin L/14, fully adapted	0.000	0.032	0.053	380	0.063	0.137	0.195	128
gte frozen, heads only	0.005	0.026	0.042	747	0.005	0.032	0.068	428
CALIPER-mix3 (seen geometry)	0.026	0.074	0.132	128	0.174	0.289	0.358	32

closer to the described geometry than chance (median cosine 0.46–0.59 against a random-pair baseline of 0.13), so retrieval follows the numbers while staying on shape-relevant candidates.

Template information poverty dominates the sealed failure. The sealed Fusion360 probe's dimension-only templates (bounding-box extents plus extrusion count) score R@10 0.074 for the two-source model and 0.039 for mix3 out-of-corpus, against a pre-declared bar of 0.12. Applying the same template family, with ground-truth extrusion counts, to in-corpus validation shapes yields only 0.110 and 0.071 respectively, an order of magnitude below the 0.95 that full captions reach on the same shapes and below the bbox-shortcut bar of 0.249. Dimensions without shape words therefore do not suffice even where grounding is proven, which is the dominant reading of the sealed failure. The in-corpus scores do sit 1.5–1.8× above their sealed counterparts, so a residual transfer gap may remain on top; out-of-corpus dimension grounding is accordingly untested-to-weak rather than refuted.

Residual-gap controls on CADPrompt (§5.5). An LLM rewrite ensemble of the human prompts, which changes style but adds no information, moves R@10 by a non-significant point. Swapping in CADmium captions, which carry ground-truth-derived information, fixes about two thirds of the failures. Because CADmium is also a training style, this second control cannot separate added information from style match, and it is read only as consistent with an information-content account of the residual gap.

Caption-then-retrieve pipeline control. A CLIP-free alternative to alignment training is to caption the gallery and retrieve text-to-text. An oracle version indexes the gallery with CADmium captions, which derive from the parametric JSON that the geometry-only setting does not provide, and retrieves L3 queries with frozen gte embeddings. It reaches cluster R@10 0.202 (exact uid 0.106, median rank 287) against the aligned encoder's 0.957 on the same gallery, so the untrained pipeline collapses even with oracle-information captions. Training a retriever helps but does not close the gap: the same 3M heads trained text-to-text on frozen-gte embeddings (L3 queries against CADmium gallery captions, identical loss and schedule) reach 0.803, still 15 points below direct alignment. Both cells index captions that require the parametric JSON for every gallery shape, which the geometry-only setting rules out; no public geometry-to-text captioner produces dimension-bearing CAD text.

Table C.5: **Duplicate-free gallery protocol.** The gallery is reduced to one representative per tight cluster, so that exact-uid and cluster credit coincide. Derived from the preserved test embeddings without re-contact, with L3 queries from all test shapes.

Model	queries	gallery	R@1	R@5	R@10	medR
CALIPER, strict track	7,760	5,386	0.691	0.912	0.952	1
mix3, 6 epochs, official track	8,021	6,335	0.718	0.936	0.967	1
gte, 15 epochs, official track	8,021	6,335	0.775	0.959	0.978	1

Table C.6: **Held-out CADPrompt chain against the original chain** (strict val, cluster R@10). The 190 CADPrompt geometries and their 3,241 near-duplicate uids are removed from train and validation before every stage of the held-out chain. The point-stage value is the best epoch under 3-draw re-evaluation, 0.6915 ± 0.0003, and the largest deviation across stages is 1.0 point.

Stage	held-out chain	original chain	Δ
heads only	0.517	0.507	$+1.0$
+ point tower	0.691	0.696	-0.5
+ text tower (CALIPER-mix3)	0.947	0.946	$+0.1$

E TEXT-TOWER CONTROLS: WINDOW VERSUS TOWER

This section collects the comparisons that separate what the text tower reads from what the text tower is. All cells are strict validation, cluster R@10.

Collapse under truncation. Figure E.1 shows the embedding-level cause of the zero-shot baselines' below-random rows. Short abstract captions of 300 shapes retain structure under Uni3D's CLIP text tower (mean off-diagonal cosine 0.66); the detailed captions of the same shapes truncate to shared boilerplate and collapse (0.88), so recall is bounded by the few clusters in one fixed top-10. The same harness scores these models at 25 to 38 times random on short CADPrompt queries, so the failure is specific to caption length.

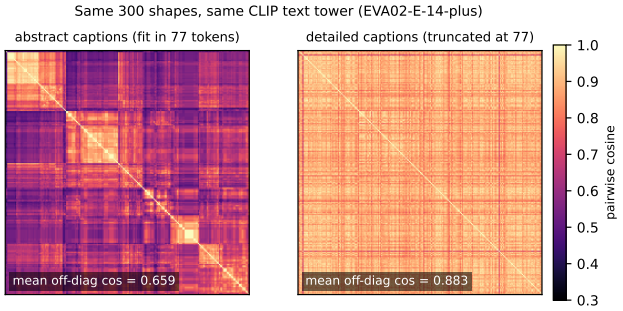


Figure E.1: **Embedding collapse under truncation.** Pairwise cosine between captions of the same 300 shapes under Uni3D's CLIP text tower, shared row order. Short abstract captions fit the window and retain structure (left, mean off-diagonal 0.66); detailed captions truncate to shared boilerplate and collapse (right, 0.88), producing the below-random rows of Table 1.

Matched-backbone context-length sweep. Truncating the *same* frozen gte backbone at increasing windows, under the identical training cell as Table G.1, isolates the window with everything else fixed. Cluster R@10 is 0.265 at 77 tokens, 0.366 at 128, 0.460 at 512, 0.471 at 1,024, and 0.471 at 2,048. The 77-token regime alone costs 20.6 points, roughly 512 tokens recover most of the ceiling given the 355-token median caption, and the 2,048 choice is a costless default that also covers the 1,238-token 95th percentile. Long-CLIP's 248-token window still truncates 91% of expert captions.

Identical-text twins. Table E.1 feeds a CLIP-family tower and the long-context tower the identical truncated string under the identical schedule. At 77 tokens the fully adapted CLIP-L/14 twin reaches

Table E.1: **Window versus tower under the identical staged recipe** (strict val, cluster R@10). Rows marked "identical text" read the *same* CLIP-BPE-truncated string as the CLIP-family row above them. Given the same text, a directly aligned CLIP tower matches or slightly exceeds a directly aligned long-context tower at both windows, so the residual gap to CALIPER is the text the window discards, not the tower. All rows start from the same heads-only stage on the caption mix, so the first column differs slightly from the L3-only cells of Table G.1. "—" marks cells not run.

Text tower	reads	heads only	+ point	+ text
CLIP-L/14	77 CLIP tok	0.199	0.286	0.472
gte-ModernBERT	identical 77-tok text	—	—	0.423
Long-CLIP-L	248 CLIP tok	0.273	0.441	0.847
gte-ModernBERT	identical 248-tok text	—	—	0.798
Long-CLIP-L, stretched to 512	512 CLIP tok	0.294	0.454	0.909
gte-ModernBERT (CALIPER)	2,048 gte tok	0.507	0.695	**0.957**

0.472 and the gte twin on the same string 0.423; at 248 tokens Long-CLIP reaches 0.847 and the gte twin 0.798. Given the same text, the CLIP-family tower is five points ahead at both windows. What CLIP-anchored encoders lack is adaptation and input, not representation.

Frozen-tier tower gaps and their disappearance. In the frozen tier, swapping gte for CLIP at the same 77 tokens costs a further nine points beyond the window effect (Table G.1), and the order between towers reverses relative to the adapted tier: frozen gte on the identical 248-token string reaches 0.284 against frozen Long-CLIP's 0.253. The appearance-first bias attributed to CLIP's text space is therefore measurable only while the tower is locked, and full adaptation erases it (Table E.1).

Under full adaptation the residual is the window. The Long-CLIP twin at 248 tokens reaches 0.847, eleven points short of CALIPER, and since the gte twin on the identical 248-token text scores lower still, none of those eleven points is the tower. They are the 91% of captions that 248 tokens truncate. Stretching the Long-CLIP positional embeddings a second time, to 512 tokens, with the first 20 positions kept fixed and the remainder interpolated as in the original Long-CLIP procedure, and running the identical staged recipe, gives 0.294 after the heads-only stage, 0.454 after the point stage, and 0.909 after text adaptation, 4.8 points short of the full caption. 512 CLIP tokens still truncate 50.1% of validation captions (the median caption is 514 CLIP tokens), discarding a median 35% of each truncated caption, which is the content the remaining 4.8 points correspond to. The gap therefore closes as the window opens, and content discarded from every training caption is not something adaptation can recover.

F Reference comparison against IMMA

Under IMMA's own setting (official test, short abstract L0 captions, and IMMA's strict metric, which is exact-uid credit here), an abstract-caption-tuned CALIPER selected on validation with a single pre-registered test contact reaches R@10 0.379 with median rank 20, where IMMA (Pan et al., 2026) reports 0.361 and 25, using one fewer input modality (IMMA additionally encodes the parametric command sequence of each gallery item). IMMA retains an R@1 edge that traces to caption non-uniqueness only the sequence input can disambiguate. IMMA has no public code, so this is a reference point under its self-reported protocol, not a controlled comparison. Under this L0 short-caption condition our default models are weaker than a system trained for it; style coverage is a training-data choice, not a free property of the recipe.

G Deferred analyses

Frozen-tier ablations. The loss grid spans about one point, which is comparable to seed noise, so no ranking claim is made across single-seed cells and InfoNCE with cluster positives is kept for its stability. On the text side, swapping to CLIP at the same 77 tokens costs a further nine points beyond the window effect, moving from B/32 to L/14 recovers barely one, and frozen Long-CLIP at 248

Table G.1: **Frozen-tier ablations** (val, cluster R@10; 30 epochs, batch 1024, seed 0 unless noted). Loss and false-negative handling are second-order; text-side choices are first-order. † marks 3-seed cells; the default is chosen for stability.

Loss × fn-mode	R@10	Axis	R@10 (Δ)
InfoNCE, none	0.476	attn→mean pool	0.457 (−1.4)
InfoNCE, mask	0.473	gte→raw ModernBERT	0.459 (−1.2)
InfoNCE, positive†	**0.483** ±0.0002	gte→CLIP-B/32 (77 tok)	**0.177** (−29.4)
SigLIP, none	0.480	gte→CLIP-L/14 (77 tok)	0.191 (−28.0)
SigLIP, mask†	0.471 ±0.009	gte→gte@77 (same backbone)	**0.265** (−20.6)
Debiased, none	0.479	gte→Long-CLIP-L (248 tok)	0.253 (−21.8)
		L3→L2 captions	0.416 (−5.5)
		L3→L3 without dimensions	**0.269** (−20.2)

tokens clears the bbox-shortcut bar only marginally. These tower gaps exist only in the frozen tier (Appendix E).

Scaling the text tower. Swapping the 149M gte tower for a 2.6×-larger ModernBERT-large and running the identical official-track chain lifts every stage, from 0.636 to 0.660 after the heads-only stage and from 0.754 to 0.824 after the point stage, and the final official test to 0.992 (R@1 0.865; Table 1). Both towers ride the same staged schedule, so the recipe transfers unchanged across encoder sizes.

Table G.2: **Additive trajectory** (strict val 7,760, cluster R@10; 3-draw re-eval ± shown where measured). The two large axes are point-tower unfreezing (+21.6) and text-tower unfreezing (+25.7 vs. control). Visualized as Figure 2.

Stage	R@10	Δ
Frozen towers + heads (Stage 1)	0.483	—
+ point last block (enc4) unfreeze	0.503 ±0.001	+2.0
+ enc4+enc3	0.625 ±0.000	+12.2
+ full tower, lr $2{\cdot}10^{-5}$	0.649 ±0.001	+2.5
+ lr 10^{-4}	0.699 ±0.000	+4.9
+ CADmium caption mix	0.696 ±0.001	−0.3*
+ text top-8 unfreeze	0.916	+22.0
+ text top-16	0.946	+3.1
+ text all-22 (CALIPER)	**0.957**	+1.1
frozen-text control (same path)	0.700	text axis total: +25.7

*mix trades −0.3 in-domain for robustness gains (Table G.3).

Seed stability, joint training, and schedule length. Three full-chain seeds land at val R@10 0.9548 ± 0.0018; test numbers are seed-0 only by protocol. Training both towers jointly from the same heads-only initialization reaches the same val R@10 of 0.978 on the official track as the staged chain, with R@1 slightly lower (0.771 against 0.787), so staging is kept only because a frozen cache halves peak memory and lets one adapted point cache serve every text-side experiment. A converged 15-epoch text schedule lifts the official-split final by about 0.9 points over the 6-epoch schedule (val 0.978 for gte and 0.989 for the large tower); the official rows of Table 1 use it.

Caption-style robustness (§3.3). Style specialization is brittle in both directions. A model trained only on L3 collapses on CADmium queries, and a CADmium-only model collapses on L3 while losing to the mix even on its own style. The three-source mix gives up about one point in-domain and more than quadruples recall on dimension-free short queries.

Cluster-credit dimension tolerance (§4). Over the 14,169 multi-member tight clusters, the per-axis relative deviation of bounding-box extent between members has median 0.03%, 90th percentile 5.0%, 95th percentile 10.9%, and 99th percentile 27.5%; 10.1% of clusters contain a member deviating by

Table G.3: **Caption-style robustness** (common 6,360-shape gallery, cluster R@10 (median rank)). Single-source training collapses cross-style; the 3-source mix is a second Pareto point.

Model	L3	CADmium	Omni-CAD (short)
CALIPER (L3+CADmium)	**0.957** (1)	0.967 (1)	0.121 (157)
CALIPER-mix3 (3-source)	0.946 (1)	0.960 (1)	**0.537** (9)
CADmium-only control (frozen tier)	0.078	0.647	—
frozen, L3+CADmium mix	0.507	0.657	0.103
frozen, L3-only	0.476	0.198	0.053

more than 5% and 5.4% by more than 10%. A 10% uniform change in the queried numbers already costs the model 25 points (Table 4), so the model's resolution is finer than the tolerance of nearly all credit clusters, and the exact-uid metrics co-reported throughout remove the remaining 5.4% from consideration.

Fusion360 transfer details (§5.5). Meshes are surface-sampled to 8,192 points and rescaled to the same 0.75 max-extent frame as our training data, so "absolute scale" is preserved in the DeepCAD-normalized sense rather than physical units; random R@10 is 0.0012 on the 8,307-shape gallery. The *dev-consulted* zero-shot trajectory on CADmium-style captions tracks the design decisions: frozen towers reach 0.118, the point unfreeze lifts this to 0.172, the caption mix to 0.532, and text fine-tuning ends at strict R@10 0.761 with median rank 2. Because this set was consulted, 0.761 is not used as evidence in the main text. The sealed contact used n=1,000 queries against the full gallery; the passing VLM-caption result is R@10 0.129 (108× random, median rank 138). On non-CAD benchmarks the boundary is sharp: 7.9% top-1 on ModelNet40 (Uni3D-g reaches 88.2) and R@1 of 0.2% on Text2Shape, where 72% and 58% of captions mention color and material respectively, both invisible to an xyz-only encoder.

Absolute-scale ablation details (§5.2). Unit-extent normalization drops the point stage from 0.754 to 0.708 on official-track validation. Split by query type, preserving scale helps dimension-bearing queries by 4.0 points and hurts dimension-free ones by 4.6 (Figure 3). The granularity-rescale constant itself is worth 2.0 points; preservation, not the particular constant, is what is load-bearing.

Point-tower control details (§5.2). Frozen Uni3D-g features with the same text tower and heads reach 0.517 against 0.471 for frozen Utonia tokens. Re-randomizing the Utonia tower and rerunning the identical point stage reaches 0.634 against the pretrained 0.695, far above the fully adapted CLIP twin's 0.286 at the same stage.

Near-duplicate threshold sensitivity. The duplicate test links shapes whose chamfer distance falls below $1.5\times$ the per-shape sampling noise floor. Re-running the pipeline on the validation split at margins 1.2, 1.5, and 2.0 shows this choice sits in a natural gap of the pair-distance distribution. No candidate pair falls below $1.29\times$ the floor, since independent subsampling separates even exact copies by slightly more than the same-shape floor, so a 1.2 margin degenerates to the strict metric; the near-duplicate mode occupies 1.3–1.5 times the floor (17.5% of candidate pairs) while the non-duplicate mass starts near 2.4 (the 25th and 50th percentiles of candidate-pair distances sit at 2.5 and 4.3 times the floor). Widening the margin from 1.5 to 2.0 moves cluster-credit R@10 by only 0.2 points (0.957 to 0.959) and the within-val duplicate fraction from 39.2% to 41.5%. Strict metrics are threshold-free and co-reported throughout, and the duplicate-free-gallery protocol (Table C.5) reproduces the credit numbers independently.

Omni-CAD collision filtering (§3.3). Omni-CAD one-line captions are filtered in two passes. Exact duplicate caption strings are removed first (54% of pairs), then captions whose string maps to more than one tight cluster are dropped as collisions (10.3%), since such a caption cannot identify its shape even in principle. The remaining 71,345 pairs enter the mix.

Failure analysis. Failures (rank $\geq$10) are 4.3% of strict-val queries and concentrate in long procedural captions (7.9% above 500 words against 0.3% below 150) and many-extrusion shapes.

Truncating failed captions worsens rank (median 22 to 103), so the content is needed, while re-querying the same shapes with CADmium captions fixes 64.9% of failures (median rank 4). This is consistent with phrasing rather than shape difficulty being the bottleneck, although CADmium is also a training style, so the control does not isolate phrasing alone. Single-CLS encoding of rambling procedural text is the identified bottleneck; multi-vector text pooling is the natural next architectural move.

Full limitations enumeration. (1) Final-chain seed variance is quantified on val (3 full-chain seeds, std 0.18pt) but test contact remains single-seed by protocol. (2) External baselines on the main galleries use raw L3 queries (100% truncated); the compressed-query fairness condition is reported on CADPrompt only (§5.5). (3) Fusion360 was consulted during development; the pre-registered sealed contact covers new query text but the gallery shapes were part of the consulted metric. (4) The encoder is xyz-only and CAD-specialist, so color and material text are invisible and non-CAD transfer is intentionally sacrificed. (5) Compositional understanding (relation flips, count changes) is not yet probed, and role-level binding of numbers to named dimensions is undetected under a probe that is not confound-free (§5.4). (6) Released weights inherit non-commercial terms from Utonia (CC BY-NC 4.0) and Text2CAD captions (CC BY-NC-SA 4.0); CADPrompt carries no explicit license (we evaluate only and redistribute nothing). (7) All three training caption sources annotate DeepCAD geometry; the only truly out-of-corpus evaluation is Fusion360 Gallery, including its sealed probe. (8) Under IMMA's L0 short-caption condition our default models are weaker than a system trained for it (Appendix F). (9) Held-out human-query evaluation is single-contact and rests on 182 prompts with wide intervals.