LEARNING 3D ATTACHMENT RANKINGS FROM GEOMETRIC PROXIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Can geometric supervision teach a model which component an assembled part attaches to? We study parent ranking from a reference-pose mesh with supplied component assignments. C2LT learns from estimated patch-connection labels computed from ShapeNet geometry, without annotated parent relations, and scores pairs of surface patches to rank candidate parents. We evaluate the trained model without fine-tuning on 4,069 queries from 931 PartNet-Mobility objects. The full model reaches 44.74% parent accuracy, compared with 38.49% for a model trained under the same protocol but without cross-patch attention. Its advantage remains when every candidate receives the same number of patch-pair scores. However, when sampling guarantees at least one patch per component and allocates the remaining patches by area, the full model reaches 65.37%, whereas choosing the candidate with the largest whole-component surface area in the supplied mesh reaches 90.17%. Cross-patch attention therefore improves the learned predictor, but this improvement does not surpass a simple geometric rule on the evaluated task.

1 INTRODUCTION

Given the geometry of an assembled object, can we infer how its parts are connected? For a cabinet door, the task is to identify its *attachment parent*, here the cabinet body, not merely a nearby handle: the component to which the door connects in the assembly. The object is shown in its reference pose, before any part is moved. We are also given which mesh faces belong to the door, body, and other *physical components*; discovering these components is not the task. The output is a ranking of possible parents for each queried component. We focus on this structural question rather than surface reconstruction.

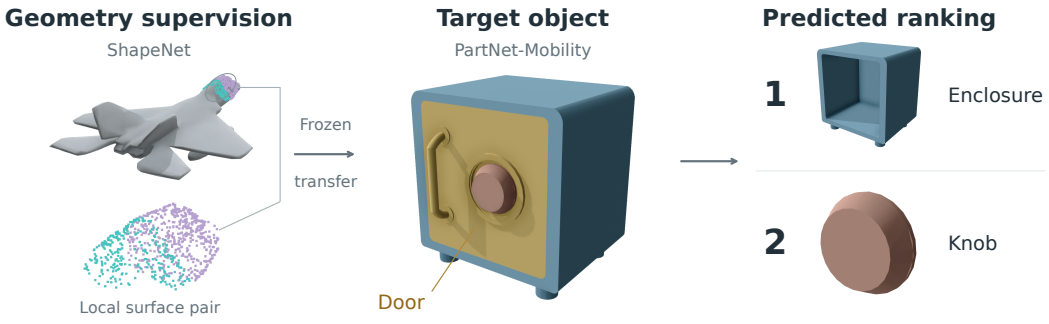


Figure 1: *Geometry-derived training and parent evaluation.* ShapeNet patches supply estimated connection labels without parent annotations (left). Given a PartNet-Mobility safe and a queried door (middle), the model ranks enclosure above knob (right). Meshes are inputs; supplied components define candidates, and parent annotations determine correctness.

The ShapeNet CAD meshes used for training do not provide attachment-parent annotations. We instead construct approximate connection targets from surface proximity, boundaries, orientations, and occupancy. These *geometric proxies* are estimated labels, not ground truth. Since they need not identify the actual parent, we evaluate the trained predictor on separately annotated PartNet-Mobility

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.

objects without fine-tuning (Xiang et al., 2020). ShapeNet validation selects checkpoints; PartNet-Mobility parent annotations define evaluation queries and correctness, but enter neither training nor model selection.

C2LT divides the surface into small neighborhoods, or *patches*, and learns a shape feature for each. The sequence of patch features is fed into a Transformer, where patches can exchange information across the object. A connection head scores ordered patch pairs. The supplied face-to-component assignments are then used to aggregate these scores into one score for each candidate parent of a queried physical component. We ask how well this geometry-trained predictor ranks separately annotated parents, whether information exchange between patches improves that ranking, and whether the learned predictor improves upon simple rules using the supplied mesh.

On 4,069 queries from 931 objects, the full object-context model reaches 44.74% parent accuracy versus 38.49% for a trained control without cross-patch attention. A 4.10-point gain remains when each candidate receives the same number of patch-pair scores. When sampling guarantees at least one patch per component and allocates the remaining patches by area, the full model reaches 65.37%, while choosing the candidate with the largest whole-component surface area in the supplied mesh reaches 90.17%. Near-uniform patch allocation adds correct predictions on newly covered queries but loses nearly as many on previously covered ones.

Contributions. First, we introduce C2LT, a geometry-supervised predictor that learns local surface features and contextual connection scores, then aggregates them into attachment-parent rankings for meshes with supplied component assignments. Second, a trained control with the same architecture and source data but without cross-patch attention shows that object-wide information exchange improves parent ranking; the gain remains when candidates receive equal numbers of patch-pair scores. Third, we evaluate against separately annotated parents and a simple rule using the supplied mesh. Changing which components receive patches shows why this within-model gain does not establish an advantage over choosing the largest component by surface area.

2 Related Work

Proxy supervision and transfer. Data programming and Snorkel (Ratner et al., 2016; 2017) learn from heuristic labels; domain randomization (Tobin et al., 2017) studies synthetic-to-real transfer. We test geometry-derived training against independently annotated attachment parents.

Geometry and structure representations. Discrete latents (van den Oord et al., 2017; Mentzer et al., 2023), mesh representations (Nash et al., 2020; Weng et al., 2025), and compositional shape models (Mo et al., 2019a; Paschalidou et al., 2021; Gao et al., 2025) address related geometry and structure problems. Particulate (Li et al., 2026a) predicts parts, kinematic structure and motion constraints from a static mesh using articulated-asset supervision. Instruct-Particulate (Li et al., 2026b) can condition on a kinematic specification containing connectivity; its connectivity-conditioned setting has a different input regime. Its configuration that predicts a specification automatically would require a separate system-level comparison. We rank parents in an existing assembly with supplied components, using PartNet-Mobility relations (Mo et al., 2019b; Xiang et al., 2020) only for evaluation.

Learning-based assembly. JoinABLe (Willis et al., 2022) predicts CAD joint axes and searches pairwise poses; AutoMate (Jones et al., 2021) predicts CAD mates from B-Rep topology. Both use assembly supervision and output joint/mate information, unlike our proxy-trained parent ranking on assembled meshes; accuracies are not directly comparable.

Aggregation and geometric priors. Assembly constraints (Coros et al., 2013) motivate direct geometric comparisons. Patch-to-component maxima are a fixed multiple-instance aggregation rule; learned permutation-invariant pooling has a separate literature (Ilse et al., 2018). Shortcut learning cautions against mistaking a successful cue for transferable structure (Geirhos et al., 2020), but component size is a legitimate task-input prior on this cohort. We test both matched patch-level controls and the stronger task-level rule. The latter asks whether learned scores add value beyond geometry already present in the input. Numeric component IDs are analyzed separately as a dataset shortcut.

3 Task and Information Sources

The input mesh M is an existing object in its reference placement. A *surface patch* (a *chart* in the implementation) is a sampled neighborhood of that mesh. A *geometric group* collects connected faces, splits large regions by face-centroid clustering, and merges small fragments. Its group ID tells the Transformer which patches share a geometry-derived region; it is not a semantic part label. A *physical component* is a supplied rigid link, such as a door or cabinet body. Patches and groups need not coincide with physical components.

ShapeNet meshes provide the geometry from which we compute training labels; ShapeNet validation selects checkpoints. At evaluation, PartNet-Mobility supplies the reference mesh, face-to-component assignments and annotated child–parent queries. The assignment groups predicted patch-pair scores into component scores only after the network has scored the pairs. Each patch receives the component ID of its sampled anchor face, although its other sampled points may cross a component boundary. Table 1 separates model inputs, learned outputs, and evaluation-only labels.

Table 1: *Information for training, prediction, and evaluation.* Supplied component membership aggregates model scores; annotated parents determine correctness and do not train or select the model.

Quantity	Origin	Use
Points, normals, local frames	Reference-pose input mesh	Patch encoding and pair geometry
Geometric-group IDs	Fixed mesh preprocessing	Context cue, not a component label
Shape features, pair scores	Learned on ShapeNet	Frozen target prediction
Patch-to-component map	Supplied face membership	Aggregate scores; allocate covered-bank anchors
Geometric proxy measurements	ShapeNet meshes; target mesh for direct rule	Training labels; separate direct baseline
Annotated adjacency and parents	PartNet-Mobility annotations	Query eligibility and correctness only

4 Learning Attachment Relations from Geometry

Figure 2 provides an overview over the steps to compute the parent ranking. Section 4.1 describes local patch features and their processing by a Transformer. Section 4.2 describes the intermediate pair predictions and how their labels are derived for training. Section 4.3 turns pair scores into component rankings; Section 4.4 gives the learning objectives and optimization. The figure distinguishes geometric group IDs used by object context from the supplied component map used for parent ranking.

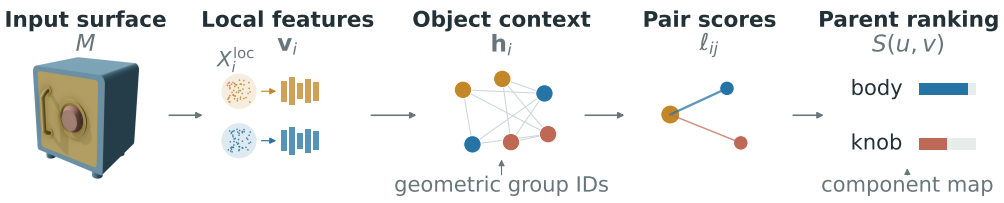


Figure 2: *Prediction path on the example from Figure 1.* The Grouped Transformer updates local features before the head scores ordered patch pairs. The supplied component map aggregates those scores into parent rankings for physical components. Geometric group IDs guide context but do not define the components. Nodes, links, and bars are schematic.

4.1 Local Features and Object Context / Transformer-Processed Features

Patch i contains points and normals sampled from a local surface neighborhood. Its geometry-derived center $\mathbf{a}_i \in \mathbb{R}^3$ and rotation $\mathbf{R}_i \in SO(3)$ define a local frame; $SO(3)$ is the set of three-dimensional rotations. Extent $s_i > 0$ is the larger of the nominal radius and the farthest sample distance from

$\mathbf{a}_i$. Centering and rotating yields X_i^{loc}. Coordinates retain their scale because they are not divided by s_i. The local frame is an orientation convention, rather than a guarantee of rotation invariance. A PointNeXt-style encoder (Qian et al., 2022) and finite scalar quantization (Mentzer et al., 2023) produce $\mathbf{v}_i \in \mathbb{R}^{263}$ from two embedded shape codes and continuous residual features.

These are assembled into a sequence and are processed by a Transformer. It employs four layers, four heads per layer and 256 channels. The pre-context feature $\mathbf{z}_i^{\text{pre}}$ and context input are

$$\mathbf{v}_i = F_\theta(X_i^{\text{loc}}), \quad \mathbf{z}_i^{\text{pre}} = W_{\text{in}}\mathbf{v}_i + \mathbf{b}_{\text{in}}, \quad \mathbf{h}_i^{(0)} = \mathbf{z}_i^{\text{pre}} + E_g(g_i) + P_a(\mathbf{a}_i) + P_s(s_i). \tag{1}$$

F_θ is the local extractor; $W_{\text{in}}, \mathbf{b}_{\text{in}}$ project its output; g_i is the geometric-group ID embedded by E_g; P_a, P_s project position and scale. Four Transformer blocks and output LayerNorm map $\mathbf{h}_i^{(0)}$ to $\mathbf{h}_i$. Attention spans patches with relative-position, scale and same-group biases.

The *Grouped* model, our full-context model, uses the Transformer output feature $\mathbf{h}_i$ for pair scoring. The *Spatial* baseline retains cross-patch attention but omits the group embedding in Eq. (1) and the same-group attention bias. The *No cross-patch attention* control retains Grouped's Transformer blocks and input cues but masks attention between distinct patches during context and head training and evaluation. Thus it is the no-attention comparison for information exchange, not a model without Transformer layers. The *Local* baseline sends $\mathbf{z}_i^{\text{pre}}$ to the same head before Transformer processing. Appendix Sections B.5 and B.6 give the full equations.

4.2 Intermediate Components for Pair Scoring

This section describes the intermediate prediction fields used for pair scoring in Section 4.3. *Outputs:* Given an ordered pair i, j of local patches, the predicted fields are $(\hat{C}_{ij}, \widehat{\Delta\boldsymbol{\rho}}_{ij}, \hat{y}_{ij}^{\text{coll}}, \hat{y}_{ij}^{\text{inv}})$. The three scalar logits predict compatibility, a collision-proxy score, and an invalidity-proxy score, respectively. $\widehat{\Delta\boldsymbol{\rho}}_{ij}$ predicts an auxiliary relative-pose correction: its translation target aligns matched boundary regions, while its rotation target is the identity, not a physical joint pose. At inference time, these are computed by a two-layer MLP H_ψ as

$$(\hat{C}_{ij}, \widehat{\Delta\boldsymbol{\rho}}_{ij}, \hat{y}_{ij}^{\text{coll}}, \hat{y}_{ij}^{\text{inv}}) = H_\psi(\mathbf{h}_i, \mathbf{h}_j, \boldsymbol{\rho}_{ij}, r_{ij}, e_{ij}). \tag{2}$$

Inputs: It uses the patch features $\mathbf{h}_i, \mathbf{h}_j$, relative translation and rotation $\boldsymbol{\rho}_{ij}$ between their local frames, log-relative scale r_{ij}, and $e_{ij} = \mathbf{1}[g_i \neq g_j]$, which is one when the patches belong to different geometric groups. At training time, ShapeNet mesh preprocessing computes four measurements in $[0, 1]$ for each selected patch pair: bidirectional surface overlap O_{ij}, boundary proximity B_{ij}, normal agreement N_{ij}, and occupancy agreement U_{ij}. These define the soft training labels, not annotated attachment parents:

$$\begin{aligned} C_{ij} &= [0.35O_{ij} + 0.25B_{ij} + 0.2N_{ij} + 0.2U_{ij}]_0^1, \\ y_{ij}^{\text{coll}} &= [0.45O_{ij}(1 - U_{ij}) + 0.3O_{ij}(1 - N_{ij}) + 0.25(1 - B_{ij})]_0^1, \\ y_{ij}^{\text{inv}} &= [0.55(1 - C_{ij}) + 0.45y_{ij}^{\text{coll}}]_0^1. \end{aligned} \tag{3}$$

Here $[z]_0^1 = \min(1, \max(0, z))$. Each row combines bounded geometric measurements with coefficients summing to one, so the resulting target stays in $[0, 1]$: overlap has the largest individual coefficient in compatibility, overlap-dependent terms dominate collision, and incompatibility has a slightly larger coefficient than collision in invalidity. These are fixed heuristic preprocessing choices, not calibrated physical probabilities; we have no documented study establishing their particular values as optimal for parent ranking. The invalidity *target* is derived from the other two targets, but its predicted logit is a separate head output. At inference time, we predict these proxy quantities from patch features and pair geometry rather than recomputing their geometric targets, which would require additional boundary sampling and occupancy queries from the target mesh. Appendix Section B.3 defines the measurements in detail.

4.3 Pair Scores and Component Ranking

The pair score combines compatibility and the other outputs from Section 4.2. Larger scores denote higher compatibility.

$$\ell_{ij}^{\text{eval}} = \log[\sigma(\hat{C}_{ij})]_\epsilon^1 + .5\log[\sigma(-\hat{y}_{ij}^{\text{coll}})]_\epsilon^1 + .75\log[\sigma(-\hat{y}_{ij}^{\text{inv}})]_\epsilon^1, \tag{4}$$

where $[p]_\epsilon^1 = \min(1, \max(\epsilon, p))$ is a clamp function, $\sigma(z) = (1 + e^{-z})^{-1}$ and $\epsilon = 10^{-6}$. Clipping is used in evaluation; training omits it. The $.5/.75$ values downweight the two penalty terms relative to compatibility's unit coefficient; they are not likelihood parameters. Section 5.2 also evaluates compatibility alone, without these weights. For child component u and candidate parent $v \neq u$,

$$S(u, v) = \max_{\substack{(i,j)\in\mathcal{E}_{\text{eval}}:\\ m(i)=u,\ m(j)=v}} \ell_{ij}^{\text{eval}}, \qquad \max \varnothing = -\infty. \tag{5}$$

where $m(i)$ denotes the supplied physical-component index of patch i, and $\mathcal{E}_{\text{eval}}$ contains the directed non-self patch pairs scored at evaluation. An attachment may be supported by a small surface region, motivating a maximum over patch-pair scores. Candidates with more pairs also get more opportunities for a high score; Section 5.3 tests this sensitivity. An absent pair scores $-\infty$. For a queried child u, parents are ordered by decreasing $S(u, v)$, with the smallest component ID breaking ties. The ordered pair inputs do not require $S(u, v) = S(v, u)$; queries are independent and no globally valid tree is enforced. This score is not a calibrated parent probability.

4.4 Learning Objectives and Optimization

Table 2 separates the three ShapeNet training stages.

Table 2: *Three-stage ShapeNet training.* All modules are frozen for PartNet-Mobility parent evaluation.

Stage	Updated / frozen modules	Training signal
Local geometry	Train local tokenizer, field projection, and decoder	Occupancy, normals, and geometry regularizers
Object context	Train context projection, Transformer, and field decoder; freeze local tokenizer	Geometry reconstruction
Patch relations	Train connection head; freeze learned representation	Geometry-derived proxy targets and source ranking

Geometry pretraining. During geometry pretraining, a decoder learns to reconstruct each patch's occupancy and surface normals first from its local feature and then from its Transformer-contextualized feature; quantization and surface-consistency losses are additional regularizers used during these geometry-learning stages. Appendix Sections B.7 and B.4 give losses and schedules; no parent annotation is used.

Relation learning. The final relation objective, shared by the matched arms, is

$$\mathcal{L}_C = .5\mathcal{L}_{\text{comp}} + .02\mathcal{L}_{\text{pose}} + .2\mathcal{L}_{\text{coll}} + .5\mathcal{L}_{\text{inv}} + .05\mathcal{L}_{\text{sep}} + \mathcal{L}_{\text{rank}}. \tag{6}$$

$\mathcal{L}_{\text{comp}}$, $\mathcal{L}_{\text{coll}}$ and $\mathcal{L}_{\text{inv}}$ are binary cross-entropies (BCE) between predicted logits and the soft targets in Eq. (3); masks exclude unavailable targets. $\mathcal{L}_{\text{pose}}$ uses smooth-L_1 error on the geometry-derived correction $\Delta\boldsymbol{\rho}_{ij}$ where its target is available. $\mathcal{L}_{\text{sep}}$ imposes a .2 margin: positive patch pairs have $C_{ij} \geq .55$ and both $y_{ij}^{\text{coll}}, y_{ij}^{\text{inv}} < .5$; negative pairs have $C_{ij} < .35$ or either collision/invalidity *target* at least .5. These categories use training labels, not predicted logits.

During training on ShapeNet, patches are put into a fixed geometry-derived order, and for each patch the ranking loss compares its candidate connections only to patches appearing earlier in that order; this ordering constructs a training task, not a predicted assembly sequence or a limit on inference-time parents. For each training patch c, $\mathcal{A}_c^+$ contains connections to previously ordered patches from a different geometric group that satisfy the proxy-positive criteria above. $\mathcal{A}_c^-$ contains every other eligible earlier connection, including same-group alternatives; these ranking negatives need not meet the separation-loss negative thresholds. Using the unclipped training score ℓ_e^{train} from Eq. (4), the hard-ranking term used in Eq. (6) is

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{T}|} \sum_{c\in\mathcal{T}} \text{softplus}\left(.35 + \max_{e\in\mathcal{A}_c^-} \ell_e^{\text{train}} - \max_{e\in\mathcal{A}_c^+} \ell_e^{\text{train}} \right), \quad \mathcal{T} = \{c : \mathcal{A}_c^+ \neq \varnothing, \mathcal{A}_c^- \neq \varnothing\}. \tag{7}$$

Empty $\mathcal{T}$ gives zero loss. The fixed ShapeNet training order begins with geometric groups from largest to smallest number of patches. Appendix Eq. (33) specifies edge directions and tie rules. The coefficients in Eq. (6) weight training losses; the $.5/.75$ values in Eq. (4) weight predicted logits in the score.

Optimization. We use two relation-training settings. The historical pipeline starts with a pretrained representation, fits the connection MLP for compatibility, and then continues the same head with the ranking objective. For the controlled experiments in Section 5.2, the shared local tokenizer is followed by 40 epochs of context training and one 30-epoch connection-head fit with Eq. (6); the representation and decoder are frozen during head fitting. The score in Eq. (4) is a fixed combination of that head's outputs, not a separately trained scoring head. Geometry reconstruction losses are off during relation training, and ShapeNet validation selects checkpoints before frozen PartNet evaluation.

5 Experiments

We test the contribution of cross-patch attention, compare the learned rankings with direct geometry and simple rules, and examine how these comparisons change when patches are redistributed at a fixed budget.

5.1 Training Data, Transfer Evaluation, and Protocol

Training data. We compute geometric training targets from ShapeNet meshes (Chang et al., 2015), without attachment-parent annotations. The earlier geometry-reconstruction schedule samples 40,000 processed, single-component CAD assets, whereas the attention comparison uses fixed sets of 21,309 training objects and 1,190 validation objects for context and relation learning. These training meshes differ from the articulated assemblies used for evaluation. All checkpoints are selected on ShapeNet validation data; Appendix Section A.1 details the data used in each training phase.

Transfer evaluation data. We evaluate the frozen models on PartNet-Mobility (Xiang et al., 2020), using reference-pose meshes, supplied component assignments, and annotated parent relations. An eligible query has an annotated directed parent, an adjacent component, and a non-adjacent alternative. Screening 2,347 assets yields 931 objects across 42 categories and 4,069 queried children. The original geometry-only sampler draws 64 patches per object, each containing 512 point–normal samples, and the model scores all 4,032 ordered pairs of distinct patches. This sampling does not guarantee a patch on every component. Supplied component assignments define the candidate parents and aggregate pair scores after prediction; they are excluded from the neural features (Appendix Section A.2).

Outcome and uncertainty. For query q, let p_q^* denote the annotated parent and $\hat{p}_q$ the highest-scoring candidate with a finite score, or $\perp$ if none exists. We measure the fraction of correctly ranked parents over the same $Q = 4{,}069$ queries for every method:

$$\mathrm{ParentAcc} = \frac{1}{Q}\sum_{q=1}^{Q} \mathbf{1}[\hat{p}_q = p_q^*]. \tag{8}$$

A query whose annotated parent is absent from the candidates is counted as incorrect. We report accuracies in percent and differences in percentage points. ParentAcc omits the additional coverage and joint checks used by the earlier checked Parent@1 metric.

We estimate paired 95% confidence intervals from 10,000 bootstrap samples of whole objects, keeping queries from the same object together. These intervals are conditional on the fitted models and fixed sampling results; they exclude uncertainty from retraining or selecting another data split. The PartNet-Mobility evaluation set has been examined repeatedly during analysis, so the follow-up comparisons are exploratory. Exact-ID screening found no overlap with four training indices or eight additional early-tokenizer IDs, although geometric near-duplicates have not been audited (Appendix Sections A.1 and A.2).

5.2 Does Cross-Patch Attention Help?

We compare the full model, *Grouped*, with a trained *No cross-patch attention* control. The two conditions share the local tokenizer, ShapeNet data, group/position/scale cues, Transformer depth, paired head initialization, training budget, and checkpoint-selection rule. For each of three paired seeds, we train the context network for 40 epochs and the relation head for 30 epochs, freezing the

representation during head training. The control retains the Transformer blocks but masks attention between distinct patches throughout training and evaluation. Its relation head still reads both patches, so the comparison specifically tests information exchange through attention (Appendix Section C.1).

With the complete score, which combines predicted compatibility with the two auxiliary outputs, Grouped reaches 44.74% accuracy versus 38.49% for the control: a gain of 6.25 percentage points (95% CI $[4.43, 8.15]$). The difference is positive for all three paired seeds. Using only predicted compatibility from the same checkpoints, with component maxima recomputed for that score, retains a gain of 3.24 points ($[1.70, 4.84]$). The gain thus persists without the auxiliary predictions in the inference score. Appendix Sections C.1 and A.4 report the group-cue, initialization, and fixed-descriptor comparisons. The last of these matches parameter count but also changes architecture and training history, so it does not isolate geometry pretraining.

5.3 Does the Gain Survive Equal Pair Opportunities?

A candidate parent's score is the maximum over its patch-pair scores. Candidates with more pairs therefore have more opportunities to receive a high score. To control this number, we set K_q to the smallest positive pair count among the candidates for query q and sample K_q pairs without replacement for each candidate. We repeat this 32 times, using the same sampled pairs across methods and choosing them without parent labels. Model parameters and candidate sets remain fixed; queries without a finite-scoring candidate remain incorrect.

Although subsampling lowers both models' accuracies, Grouped retains a 4.10-point advantage ($[3.05, 5.24]$) with the complete score (Table 3). The compatibility-only comparison also remains positive.

Table 3: *Cross-patch attention with all or equally many patch pairs.* ParentAcc (%) on all 4,069 queries, averaged over three paired models selected on ShapeNet validation. Equal-K_q results additionally average 32 shared subsamples. Complete combines the three predicted fields; Compatibility uses only the compatibility prediction. Differences are Grouped minus No cross-patch attention, in percentage points, with paired object-bootstrap 95% confidence intervals.

Score / patch pairs	No cross-patch attention	Grouped	Difference (95% CI)
Complete / all	38.49	44.74	$+6.25$ $[+4.43, +8.15]$
Complete / equal K_q	24.69	28.79	$+4.10$ $[+3.05, +5.24]$
Compatibility / all	39.13	42.36	$+3.24$ $[+1.70, +4.84]$
Compatibility / equal K_q	23.65	26.21	$+2.56$ $[+1.68, +3.50]$

Equalizing pair counts also removes local geometric evidence. This is a sampling diagnostic, rather than a proposed inference improvement or a complete correction for sampling bias. Appendix Section D.1 analyzes high-score and proxy-positive-pair retention, reports per-seed results, and details the 2.15-point reduction in the complete-score attention gain.

5.4 Why Not Compute Geometry Directly?

We next compare learning from geometry with using it directly. *Dense Support* measures how often one patch's sampled points lie near the other, averaging directions. It uses the network's point samples, patch pairs, candidates and maximum aggregation, with no extra input. It reaches 47.43% versus Grouped's 44.74% on original patches, but 25.94% versus 28.79% under equal-K_q subsampling: their ordering depends on pair availability.

On 48 fixed objects, median scoring takes 34 ms for Grouped's GPU forward versus 626 ms for single-threaded CPU Dense Support; both need patch preprocessing (224 ms). This mixed-device diagnostic is not an end-to-end speed comparison, and Dense was not tested as an added network feature (Appendix Section A.5).

The *direct full-proxy* rule recomputes training-target geometry from the same reference mesh, including boundary and occupancy measurements, and reaches 46.72%. All these measurements are obtainable at inference. On eight objects, reconstruction takes a median 3.12 s including geometry

assembly and preprocessing, not a matched model comparison. Geometric NN reaches 42.91%. Once the mesh is assembled, whole-component area reduction takes 0.36 ms median and is more accurate here; the evidence does not justify replacing direct geometry with the learned ranker when parent accuracy alone is the objective. Appendix Sections A.5 and A.4 give the cost protocol and rule definitions.

5.5 Can Learned Scores Improve Simple Rules?

Choosing the candidate with the most sampled patches reaches 58.54% accuracy on the original patch set. Patch count reflects the sampler's area bias. Choosing the candidate with the largest surface area in the supplied mesh reaches 62.25%; this rule measures the whole component rather than its sampled patches. Both rules use information already available in the task input.

To test whether the learned scores can improve the count rule, we use its prediction by default and switch to Grouped only when the two disagree and the softmax value of Grouped's top candidate exceeds a threshold. The softmax is computed over finite component scores, after patch-pair aggregation. At $\tau = 0.5$, this gate reduces accuracy to 56.40%: across the three fitted models, it corrects 129 count-rule errors but replaces 390 correct count predictions with incorrect ones. These totals count query–seed decisions. None of the thresholds in the reported grid improves upon count on the original patch set.

A more restricted hybrid uses Grouped only to break ties among candidates with the largest patch count. This requires no additional fitting and reaches 58.71%, an increase of 0.17 percentage points (95% CI $[0.03, 0.38]$). Its effect is confined to the 111 queries with count ties; all other predictions are unchanged (Table 4).

Table 4: *Learned scores and simple rules on the original patch set.* ParentAcc (%) on all 4,069 queries. Grouped and both hybrids use the same three frozen models; accuracies are computed for each model and then averaged. The gate applies the fixed diagnostic threshold $\tau = 0.5$ to uncalibrated softmax scores. Surface area is measured from the supplied component meshes.

Prediction rule	ParentAcc ↑
Grouped only	44.74
Most parent patches	58.54
Count + confidence gate ($\tau = 0.5$)	56.40
Count + model tie-break	58.71
Largest component surface area	**62.25**

The tie-break resolves some count ties, although the surface-area rule remains more accurate. A further count-restricted ensemble averages normalized scores from the three models and ranks only candidates meeting a patch-count cutoff. It reaches 59.01%, with a 95% interval of $[-0.22, +1.22]$ for its difference from count. The cutoff was explored after examining the target data, so this result does not establish an independently selected improvement. Appendix Section D.4 details the fusion rules and their selection conditions.

For comparison, choosing the largest numeric component ID reaches 62.60%, but falls to 20.38% when IDs are permuted without using parent labels. Among the 2,824 queries whose annotated parent is sampled, that parent's ID exceeds the child's in 2,617 cases (92.7%). The IDs are supplied PartNet-Mobility metadata with a strong label correlation, not geometry or a network prediction; this is a dataset-order shortcut, not a comparable geometry-only baseline (Appendix Section D.2).

5.6 What Changes When Every Component Is Sampled?

Under the original sampler, the queried child has a sampled patch in 2,967 of the 4,069 queries, and both child and annotated parent have sampled patches in 2,824. In particular, 1,102 queried children receive no patch anchor. We therefore resample the same 931 meshes with the same budget of 64 patches per object, using the supplied component assignments to place one anchor on every component. The remaining anchors are allocated either in proportion to component area (*Area-covered*) or as evenly as possible across components (*Uniform-covered*). Each scoring method uses the same new patches and candidates within a sampling condition; queries, labels, and frozen

checkpoints remain unchanged. This comparison changes the inputs on the existing evaluation set (Appendix Section D.3).

Table 5: *Sampling changes ranking.* Ungated ParentAcc (%) on the same 4,069 queries with 64 patches per object. Covered banks represent every gold pair; the original bank can omit components. Direct-geometry results are in Appendix Table 17.

Method	Original	Area-covered	Uniform-covered
No cross-patch attention	38.49	60.40	38.06
Grouped	44.74	65.37	44.86
Most parent patches	58.54	88.62	11.06
Largest component surface area	62.25	**90.17**	**90.17**

Both covered conditions include every annotated child–parent pair. With area-proportional allocation, Grouped reaches 65.37% accuracy, compared with 90.17% for the surface-area rule. Under near-uniform allocation, they reach 44.86% and 90.17%, respectively. The area values are unchanged: the rule improves from 62.25% because all candidate components are now available. Grouped remains above No cross-patch attention in both conditions, yet its near-uniform accuracy is almost unchanged from the original sampling.

The near-uniform result illustrates why. On queries whose child–parent pair was already represented, we first evaluate the new scores on the original candidates, then admit the additional candidates. The two steps contribute -9.69 and -0.86 percentage points, respectively. Correct predictions on newly covered queries contribute $+10.67$ points, almost offsetting those losses. All contributions use the full 4,069-query denominator and sum to $+0.11$ points before rounding. This accounting follows a specified sequence of evaluations; it does not separately identify causal effects of local sampling, context, or aggregation. Appendix Figure 4 and Section D.3 give the full decomposition. That section also reports a separate selector that defaults to surface area and switches to the model prediction only when all three fitted models agree.

5.7 Sensitivity and Scope

Our primary ParentAcc gives each query equal weight. With equal weight per object, the direct full-proxy rule reaches 74.83% versus Grouped's 69.97% under area-proportional allocation. Under near-uniform allocation, the corresponding accuracies are 61.45% and 57.20%. Giving each category equal weight reverses the ordering of these two methods. The surface-area rule remains highest under either weighting (Appendix Table 19). These summaries show how unequal query counts affect the comparison.

Covered sampling also changes patch purity, and a separate permutation of geometric-group IDs changes the frozen model's predictions. The sampling comparisons therefore do not isolate coverage from changes in patch content or contextual inputs (Appendix Section D.3). Finally, Appendix Section C.2 distinguishes the effect of training with the auxiliary invalidity target from the effect of including its predicted penalty in the inference score.

6 Conclusion

Grouped's cross-patch attention improves ranking over the matched No cross-patch attention control, but a task-available component-area rule remains stronger. Near-uniform coverage trades new-query gains for old-query losses; learned value beyond strong geometric priors remains unestablished. Proxy-supervised patch relations therefore transfer enough information to expose an architectural effect, but not to beat the strongest input-derived cue on this benchmark. Candidate coverage, unequal pair opportunities, and component-ID ordering each change apparent performance and should accompany parent accuracy in comparisons. Establishing a task-level advantage requires independent data with these cues controlled and matched access to geometry and computation.

AI USE STATEMENT

Generative AI tools assisted with drafting and polishing manuscript text, discovering related literature, proposing diagnostic checks, writing and debugging analysis code, interpreting numerical results, and preparing figures. They did not generate a synthetic dataset or establish a mathematical proof for this submission. AI-assisted numerical claims were checked against saved outcomes and audit code. The authors take responsibility for the final text, analyses, claims and artifacts.

REPRODUCIBILITY STATEMENT

Sections 4 and 5 define the method, evaluation and controls; the appendix gives formulas, splits, compute and implementation details. The anonymous audit contains split IDs, frozen query-level outcomes, checkpoint identifiers, numerical evaluators and object-level bootstrap code. It can recalculate reported numbers but cannot rerun preprocessing, training or model inference: the compact upload omits third-party meshes, model weights and the full implementation. The target cohort was inspected repeatedly, and the intervals condition on its fixed objects and fitted models rather than an independent validation set or new training runs. Implementation and geometry evaluators are intended for release after de-anonymization, subject to third-party licenses.

ETHICS STATEMENT

Third-party 3D assets remain subject to their research terms; the anonymous audit redistributes no meshes. Predicted connections and finite checks do not certify physical safety or justify autonomous use. Appendix Section F.2 covers licensing and responsible use.

REFERENCES

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Stelian Coros, Bernhard Thomaszewski, Gioacchino Noris, Shinjiro Sueda, Moira Forberg, Robert W Sumner, Wojciech Matusik, and Bernd Bickel. Computational design of mechanical characters. *ACM Transactions on Graphics (TOG)*, 32(4):83:1–83:12, 2013. doi: 10.1145/2461912.2461953.

Daoyi Gao, Yawar Siddiqui, Lei Li, and Angela Dai. Meshart: Generating articulated meshes with structure-guided transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 618–627, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Benjamin Jones, Dalton Hildreth, Duowen Chen, Ilya Baran, Vladimir G. Kim, and Adriana Schulz. AutoMate: A dataset and learning approach for automatic mating of CAD assemblies. *ACM Transactions on Graphics*, 40(6):227:1–227:18, 2021. doi: 10.1145/3478513.3480562. URL https://arxiv.org/abs/2105.12238.

Ruining Li, Yuxin Yao, Chuanxia Zheng, Christian Rupprecht, Joan Lasenby, Shangzhe Wu, and Andrea Vedaldi. Particulate: Feed-forward 3d object articulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27708–27718, 2026a.

Ruining Li, Yuxin Yao, Matt Zhou, Chuanxia Zheng, Christian Rupprecht, Joan Lasenby, Shangzhe Wu, and Andrea Vedaldi. Instruct-particulate: Scaling feed-forward 3d object articulation with kinematic control. *arXiv preprint arXiv:2606.14699*, 2026b.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019a.

Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 909–918, 2019b.

Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International conference on machine learning*, pp. 7220–7229. PMLR, 2020.

Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3204–3215, 2021.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: Revisiting PointNet++ with improved training and scaling strategies. *Advances in neural information processing systems*, 35:23192–23204, 2022.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Alexander J. Ratner, Christopher M. De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/6709e8d64a5f47269ed5cea9f625f7ab-Abstract.html.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *arXiv preprint arXiv:1703.06907*, 2017.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Haohan Weng, Zibo Zhao, Biwen Lei, Xianghui Yang, Jian Liu, Zeqiang Lai, Zhuo Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, et al. Scaling mesh generation via compressive tokenization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11093–11103, 2025.

Karl D. D. Willis, Pradeep Kumar Jayaraman, Hang Chu, Yunsheng Tian, Yifei Li, Daniele Grandi, Aditya Sanghi, Linh Tran, Joseph G. Lambourne, Armando Solar-Lezama, and Wojciech Matusik. JoinABLe: Learning bottom-up assembly of parametric CAD joints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15849–15860, 2022. URL https://arxiv.org/abs/2111.12772.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11097–11107, June 2020.

APPENDIX

Evidence guide. Sections A–B specify the task and model; Section C reports matched controls, Section D the sampling and candidate interventions, Section E sensitivity to task inputs and aggregation, and Section F the released numerical checks and asset scope. The three main claims are supported respectively by Sections C–D, D, and C–E. The appendix and archived experiment outputs also use *Full-context* for the main-text *Grouped* model and *self-only* for the *No cross-patch attention* control; these are names for the same respective checkpoint families, not additional methods.

A Task, Data, and Evaluation

A.1 Source and Target Populations

ShapeNet (Chang et al., 2015) supplies geometry, geometry-derived targets and source validation; it supplies no attachment-parent annotations. The historical local-geometry pool has 44,473/2,535/2,510 train/validation/test entries, and its reconstruction schedule samples 40,000 training objects. The matched context and relation fits instead use fixed 21,309-training and 1,190-validation object archives. Their three context/head seeds share a source-selected tokenizer initialization. These phase-specific populations must not be treated as one common end-to-end training split.

The source-ID screen pooled four relation/local-field indices with 33,444 distinct model IDs. An earlier tokenizer smoke index contains eight further IDs outside that union; none match the retained target IDs. This is an exact-ID check, not a geometric-near-duplicate audit. All checkpoints for the source-only ranking experiment were selected using ShapeNet validation.

PartNet-Mobility (Xiang et al., 2020) supplies assembled reference-pose meshes, component membership and released relations. Starting from 2,347 local assets, the frozen eligibility screen excludes 572 source-lineage overlaps, 710 without a nontrivial hinge/slider query, 108 with unmapped URDF links and 26 above the 64-link cap. The remaining 931 objects span 42 categories and two preprocessing cohorts of 512 and 419 assets. Each object is centered by its bounding box and divided by its box diagonal. The original geometry-only sampler draws 64 patches with 512 point–normal samples each, without a per-component guarantee, and scores all 4,032 directed non-self patch pairs. A patch is assigned to the component of its anchor source face. Supplied membership is used after prediction for aggregation; released relations do not enter neural features, fitting or checkpoint choice.

A.2 Queries, Coverage and Statistical Unit

A query requires a released directed parent, an adjacent component and a non-adjacent alternative. The fixed bank contains 4,069 queried children; every method is scored on that same denominator and an absent finite gold candidate is wrong. There are 2,967 queries with a sampled child, 3,870 with a sampled parent component and 2,824 with a sampled gold child–parent pair. The main ParentAcc in Eq. (8) has no coverage or kinematic gate. Earlier checked Parent@1 additionally required represented alternatives and a released-joint certificate; it is a different historical endpoint. The 4,069 queries are not independent observations: paired intervals resample the 931 whole objects and retain every query, seed and readout belonging to each sampled object. Object and category macros change the weighting unit and are reported separately rather than replacing the declared query mean.

A.3 Reconstructed Target Proxy for Diagnostics

The original target inference archives contain zero-valued proxy placeholders, only one occupancy query per patch and no complete boundary supports. Those placeholders are not proxy ground truth. For target-side diagnostics only, we reconstruct evaluator samples from the same reference mesh under each cohort's recorded seed, checking the stored anchors, groups and component membership. Each patch obtains 64 geometric-group boundary samples and 256 occupancy queries: 128 near-surface samples with Gaussian scale $.12s_i$ and 128 uniform samples in $[-1.25s_i, 1.25s_i]^3$. Object-hash seeds fix these draws. The positive-x ray-parity routine assigns inside/outside labels; a 64-query per-asset check compares the accelerated and uncropped implementations. Non-watertight assemblies still make this a geometric proxy, not physical ground truth.

The compatibility, collision, and invalidity targets use Eq. (3); an edge is proxy-positive when $C_{ij} \geq .55$, $y_{ij}^{\text{coll}} < .5$ and $y_{ij}^{\text{inv}} < .5$. For child u_q and candidate set $\mathcal{V}_q$, the proxy-positive parent set is

$$\mathcal{P}_q = \{v \in \mathcal{V}_q : \exists (i,j) \in \mathcal{E}_{\text{patch}}, m(i) = u_q,\ m(j) = v,\ C_{ij} \geq .55,\ y_{ij}^{\text{coll}} < .5, y_{ij}^{\text{inv}} < .5\}. \quad (9)$$

It may contain several candidates, unlike the unique released parent. Of the 4,069 released parents, 528 satisfy the patch-pair proxy-positive test (12.98% of all queries; 18.70% of the 2,824 covered gold pairs). Predictions and candidate lists are held fixed when comparing these label definitions. Reconstructed boundary and occupancy samples are never passed to the frozen learned predictor.

A.4 Geometric and Learned Controls

Geometric NN and Dense Support score the same patch-pair bank as the learned ranker. With patch centers $\mathbf{a}_i$, group IDs g_i, object-space surface samples X_i and $\tau_{ij} = .12 \min(s_i, s_j)$, their scores are

$$\ell_{ij}^{\mathrm{NN}} = -\|\mathbf{a}_i - \mathbf{a}_j\|_2 + .05\mathbf{1}[g_i = g_j], \tag{10}$$

$$\ell_{ij}^{\mathrm{dense}} = \frac{1}{2}\left[\frac{1}{|X_i|}\sum_{x\in X_i}\mathbf{1}[d(x, X_j) \le \tau_{ij}] + \frac{1}{|X_j|}\sum_{y\in X_j}\mathbf{1}[d(y, X_i) \le \tau_{ij}]\right], \tag{11}$$
$$d(x, X) = \min_{y\in X}\|x - y\|_2.$$

The NN group bonus makes it more than center distance alone. Dense Support is a direct bidirectional near-surface fraction. Both use the same component maximum and ID tie rule as the learned patch scorer.

The direct full-proxy rule recomputes the three source-formula measurements from target geometry rather than using network predictions:

$$\ell_{ij}^{\mathrm{proxy}} = \log[C_{ij}]_\epsilon^1 + .5\log[1 - y_{ij}^{\mathrm{coll}}]_\epsilon^1 + .75\log[1 - y_{ij}^{\mathrm{inv}}]_\epsilon^1, \quad \epsilon = 10^{-6}. \tag{12}$$

It then uses the same candidate list, pair maximum and 4,069-query ParentAcc. It reaches 46.72%; compatibility alone reaches 46.79%. Unlike the learned patch scorer, it additionally computes reference-mesh boundary and occupancy queries. It is a task-level geometric diagnostic, not an equal-derived-feature or equal-compute control. The parameter-count-matched learned control instead uses a fixed 263-dimensional point–normal/pose descriptor, an independent residual MLP, width-256 four-layer context Transformer and connection head, trained from random initialization on the same source relation archive. Its total parameter count matches C2LT's 3,988,604, but architecture, geometry pretraining and optimization history do not match. The closest architecture attribution is therefore the trained self-only attention control in Section C.1.1.

A.5 Stagewise Inference-Cost Diagnostic

We selected 48 of the 931 original-bank objects by a fixed hash of object ID, without reading their parent outcomes. On one RTX 5090, with one CPU thread for Dense Support, median measured stages per object were: reference-mesh assembly 85 ms, whole-component face-area reduction 0.36 ms, 64-patch preprocessing 224 ms, stored patch loading and GPU transfer 6.5 ms, Grouped forward 34 ms, and Dense pairwise proximity 626 ms. Component-score aggregation was under 1 ms for either scorer. The Grouped seed-0 checkpoint was loaded once before timing; its load took 259 ms. A separate pilot on the first eight selected objects rebuilt the direct full proxy in a fresh output directory: all eight succeeded, with a median 3.12 s per object (range 2.32–3.65 s). This pilot includes reference-mesh reconstruction, patch preprocessing, boundary sampling, occupancy queries, and proxy scoring, whereas the model-forward stage does not; its sample is too small for a general speed comparison. The stagewise protocol and per-object durations are released in the separate anonymous numerical audit. OS caches were not cleared, model loading is excluded from per-object stages, and model and Dense calculations ran on different devices. These measurements describe this implementation's cost components, not a hardware-matched or cold-I/O end-to-end speed comparison.

B Model and Training Specification

B.1 Modules Used by the Ranking Path

The model first encodes local point–normal patches, then contextualizes their features, predicts ordered pair fields, and aggregates them by supplied component membership. The field decoder is trained during geometry learning but does not run in target ranking.

Table 6: *Ranking-path architecture and training dependency.* The final model has 3,988,604 parameters in total; frozen modules are still part of inference unless marked training-only. The local decoder and support correction are geometry-learning dependencies, not parent-ranking outputs.

Module	Input / width	Role
Local patch encoder	Point–normal samples; four PointNeXt-style blocks, width 64, $k = 24$	Pools a 192-dimensional local feature
Dual finite scalar quantizers	6 geometry and 4 boundary slots, seven levels each	Project each stream to a 128-dimensional feature, without a full learned lookup table
Residuals and projection	Two 128-D streams, 6-D pose and 1-D scale residual	Concatenate 263 channels and project to context width 256
Context network	Four layers, four heads, width 256, feed-forward width 1024	Object-wide attention with position, scale and geometric-group cues
Connection head	Pair features; two-layer width-256 MLP	Compatibility, collision/invalidity proxy logits and auxiliary 7-D pose correction
Local field decoder	Four-layer width-256 MLP plus support correction	Geometry pretraining only; occupancy and normal supervision

B.2 Information and Terminology

Surface patches, geometric groups, physical components and candidate scores are different objects. The group labels are preprocessing cues, not semantic part labels; a predicted score is not an observed physical joint.

Table 7: *Terminology crosswalk.* A stored quantity, a prediction, and an evaluation label are different objects.

Term	**Meaning in this paper**
Chart / surface patch	Local point–normal neighborhood, not necessarily a semantic part.
Anchor / local frame	Patch center and orthonormal coordinate axes computed from geometry.
Support samples	Points sampled from an input surface; not an inferred load-bearing relation.
Geometric group / partition	Deterministic mesh grouping used as context; g_i is its ID.
Physical component / link	Supplied rigid asset component. It can contain several geometric groups.
Token	Discrete local code; network features also include continuous residuals.
Contextual feature	A patch feature after attention to other patches in the same object.
Connection head	Predicts ordered patch-pair logits; no seam curve is extracted.
Compatibility / collision / invalidity scores	Predicted source-rule targets, not physical validity or calibrated parent probabilities.
Candidate parent	A supplied component other than the queried child, ranked by its patch-pair score.

B.3 Exact Source-Target Construction

Source candidate graph. For each patch, source preprocessing retains up to eight nearest same-group and four nearby cross-group neighbors. Cross-group search first considers anchor distances within $2.5 \max(s_i, s_j)$, falling back to nearest cross-group patches when none fall inside. A cross-group edge is retained only if overlap is at least .02, boundary proximity at least .12, or occupancy agreement at least .55. PartNet transfer instead scores all directed non-self patch pairs; graph rules are shared by scorers within each protocol, not across datasets.

Soft target construction. Let O_{ij} denote symmetric surface overlap, D_{ij}^{bnd} symmetric boundary distance, N_{ij} normal agreement and U_{ij} occupancy agreement. The three soft targets are defined in Eq. (3). With $\epsilon = 10^{-6}$, boundary proximity is

$$B_{ij} = \exp\left(-\frac{D_{ij}^{\mathrm{bnd}}}{.15 \max(\min(s_i, s_j), \epsilon)}\right). \tag{13}$$

These deterministic targets are not measured collision frequencies, force constraints, or independent human validity judgments. The positive-edge test $C_{ij} \geq .55$, $y_{ij}^{\mathrm{coll}} < .5$, $y_{ij}^{\mathrm{inv}} < .5$ has a redundant

last threshold: the first two imply it, though separately predicted logits need not obey the target identity. Boundary samples come from geometric-group mesh boundaries; occupancy queries have binary mesh inside/outside labels.

The following definitions expand the measurements in Eq. (3). All points in this section are in the same object coordinate system. X_i is patch i's surface support, Z_i its sampled geometric-boundary support, and Y_i its occupancy-query set. The unit normal at $x \in X_i$ is $\mathbf{n}_i(x)$ and the occupancy label at $y \in Y_i$ is $o_i(y) \in \{0, 1\}$. Let $\pi_X(x)$ be a nearest neighbor of x in the nonempty set X and $d(x, X) = \|x - \pi_X(x)\|_2$. The pair's distance thresholds are $\tau_{ij} = .12 \min(s_i, s_j)$ and $\tau_{ij}^{\text{occ}} = .18 \min(s_i, s_j)$.

Support overlap O_{ij} is exactly the bidirectional fraction in Eq. (11). Boundary distance is

$$D_{ij}^{\text{bnd}} = \frac{1}{2}\left[\frac{1}{|Z_i|}\sum_{x \in Z_i} d(x, Z_j) + \frac{1}{|Z_j|}\sum_{y \in Z_j} d(y, Z_i)\right]. \tag{14}$$

If either boundary set is empty, its distance is treated as infinite and its exponential score B_{ij} as zero. Define matched subsets

$$X_{i\to j} = \{x \in X_i : d(x, X_j) \le \tau_{ij}\}, \qquad Y_{i\to j} = \{y \in Y_i : d(y, Y_j) \le \tau_{ij}^{\text{occ}}\}. \tag{15}$$

For a nonempty subset, its directional agreement is

$$\begin{aligned} n_{i\to j} &= \frac{1}{|X_{i\to j}|}\sum_{x \in X_{i\to j}} \text{clip}_{[0,1]}(\mathbf{n}_i(x)^\top \mathbf{n}_j(\pi_{X_j}(x))), \\ u_{i\to j} &= \frac{1}{|Y_{i\to j}|}\sum_{y \in Y_{i\to j}} \left(1 - |o_i(y) - o_j(\pi_{Y_j}(y))|\right). \end{aligned} \tag{16}$$

N_{ij} averages the available $n_{i\to j}, n_{j\to i}$; U_{ij} averages the available $u_{i\to j}, u_{j\to i}$. If neither directional subset is nonempty, the corresponding score is zero. The collision and invalidity targets then follow Eq. (3). They reuse these geometric measurements and do not add an independent source of physical supervision. N_{ij} rewards aligned, rather than opposed, normals; opposing contact faces can therefore score poorly. U_{ij} counts matching outside–outside samples as agreement as well as inside–inside samples. Neither quantity is a calibrated physical attachment or collision probability. A separate contact-orientation and occupancy-class sanity evaluation would be needed to determine how much each proxy reflects actual contact rather than sampling composition.

Auxiliary pose-correction target. Let $Z_{i\to j} = \{z \in Z_i : d(z, Z_j) \le \tau_{ij}\}$. The translational target is the average matched boundary displacement expressed in patch i's frame, clipped coordinatewise:

$$\Delta\mathbf{t}_{ij} = \text{clip}_{[-a,a]}\left(\mathbf{R}_i^\top \frac{1}{|Z_{i\to j}|}\sum_{z \in Z_{i\to j}} (\pi_{Z_j}(z) - z)\right), \qquad a = .25 \max(\min(s_i, s_j), 10^{-6}). \tag{17}$$

If a boundary set or the matched subset is empty, $\Delta\mathbf{t}_{ij} = \mathbf{0}$. The seven-coordinate target concatenates $\Delta\mathbf{t}_{ij}$ with the identity quaternion $(0, 0, 0, 1)$. Thus this auxiliary training target does not supervise arbitrary rotational reassembly or an observed parent relation.

B.4 Geometry-Pretraining Objectives

These objectives describe geometry learning, not additional attachment supervision. Let a_{it} be the decoder occupancy logit at query t of patch i, o_{it} the corresponding occupancy target, and $\hat{\mathbf{n}}_{it}, \mathbf{n}_{it}$ normalized predicted and target normals. $M_{it} \in \{0, 1\}$ marks valid normal targets. In the local phase, query regions distinguish near-surface positive/negative and interior positive/negative samples. With region mask $M^{(r)}$, weight w_r, and BCE from Eq. (30),

$$\mathcal{L}_{\text{occ}} = \frac{\sum_r w_r \dfrac{\sum_{i,t} M_{it}^{(r)} \,\text{BCE}(a_{it}, o_{it})}{\max(1, \sum_{i,t} M_{it}^{(r)})}}{\sum_r w_r}, \qquad \mathcal{L}_{\text{norm}} = \frac{\sum_{i,t} M_{it}(1 - \hat{\mathbf{n}}_{it}^\top \mathbf{n}_{it})}{\max(1, \sum_{i,t} M_{it})}. \tag{18}$$

The context phase uses ordinary occupancy BCE over valid object patches plus normal loss. The decoder receives contextual features instead of local projected features; the tokenizer is fixed.

Quantization and code statistics. For a quantizer stream $b \in \{\text{geo}, \text{bnd}\}$, let $\mathbf{u}_i^b$ be its normalized, pre-scaled encoder input and $\mathbf{z}_i^b$ its quantized output mapped back to encoder width. Its commitment term is the mean squared difference from a stopped-gradient output:

$$\mathcal{L}_{\text{quant}} = \sum_b \text{mean}_i \, \|\mathbf{u}_i^b - \text{sg}(\mathbf{z}_i^b)\|_2^2 / 192. \tag{19}$$

sg stops gradients through its argument. The logged usage term is the mean, over the two streams, of $\sum_{k:p_k>0} p_k \log(p_k/(1/K))$, where p_k is the empirical mini-batch frequency of discrete code k and K is that stream's codebook size. In this implementation it is computed from integer code counts and has no gradient to the encoder. Although included in the configured scalar objective, it must not be interpreted as a differentiable anti-collapse mechanism. Token perplexity is $\exp(-\sum_k p_k \log p_k)$; utilization is the observed nonzero-code fraction.

Overfill and difficult negative queries. Overfill discourages predicting excessive occupied mass:

$$\mathcal{L}_{\text{overfill}} = \text{mean}_i \max\left(0, \frac{\sum_t \sigma(a_{it})}{\max(1, \sum_t o_{it})} - \rho\right). \tag{20}$$

ρ is the allowed ratio, set to 1.10 in the support-correction continuation. Its difficult-negative term selects interior negative queries with logit above zero and averages $T \, \text{softplus}((a_{it} - \mu)/T)$ on that set, with temperature $T = .25$ and margin $\mu = 0$; an empty set gives zero. This penalizes confidently filled empty space.

Surface-logit targets. For a normal-labeled near-surface query $\mathbf{q}$, let $\mathbf{x}^*$ be its nearest support point and $\mathbf{n}_q$ its supplied query normal. The normalized signed offset is $d_q = (\mathbf{q} - \mathbf{x}^*)^\top \mathbf{n}_q / \max(s_i, 10^{-6})$. The support continuation clips d_q to $[-.22, .22]$ and uses the target logit $a_q^* = \text{clip}_{[-3.5,3.5]}(-d_q/.08)$. The surface-logit term is mean smooth-L_1 error between the decoder's base occupancy logit and a_q^*. This encourages the local field to change sign near the surface.

Single-surface and context-line consistency. For a sampled surface point $\mathbf{x}$ with normal $\mathbf{n}$, evaluate occupancy logits a_h along $\mathbf{x} + t_h \mathbf{n}$, where the offsets t_h are evenly spaced over a symmetric interval. The interval radius is a configured multiple of median nonzero nearest-support spacing. Let h_0 be the central index, so $t_{h_0} = 0$. The center and sign terms are

$$L_{\text{center}} = \text{mean} \, \text{smooth}_{L_1}(a_{h_0}, 0), \qquad L_{\text{sign}} = \text{mean}_{h \neq h_0} \, \text{BCE}(a_h, \mathbf{1}[t_h < 0]). \tag{21}$$

The margin term averages $\max(0, \delta - a_h)$ inside and $\max(0, \delta + a_h)$ outside, then adds the two means; $\delta > 0$ is a logit margin. The local single-surface term penalizes increasing logits separately on each side of the center. The context-line term uses $L_{\text{mono}} = \text{mean}_h \max(0, a_{h+1} - a_h)$ across all consecutive samples. Each is a weighted sum of center, sign, monotonicity, and margin terms. The support continuation uses 128 support locations, nine offsets, radius multiplier 1.20, margin .35, and weights $4/1/.25/.20$. The context continuation uses 32 locations, seven offsets, multiplier 3.0, margin .45, and weights $1/.75/.5/.5$. These are local field regularizers, not global closure constraints.

Phase-specific composition. The local objective adds occupancy, weighted normals, commitment, the scalar usage statistic, overfill, difficult negatives, surface-logit error, and the local single-surface term. The base/stronger local phases use normal and commitment weights $.25/.08$, usage weight .08, overfill .20, and difficult-negative .02. The surface-logit weight increases from .08 to .14 and single-surface weight from .035 to .08 in the stronger continuation. The support-correction continuation sets normal, commitment, and usage weights to zero, retains overfill/difficult-negative/surface-logit/single-surface weights $.20/.02/.14/.08$, and changes only the support branch. Its region weights are $1.45/1.10/1.05/1.45$ in the four-region order above. The context objective is occupancy plus .1 times normal loss, adding .2 times context-line consistency in its continuation. These geometry objectives produce the shared source-selected initialization for the matched study; they are inactive during relation-head training (Table 8).

Historical naming. Older appendix formulas group geometry terms into $\mathcal{L}_A$ and $\mathcal{L}_B$ for local and contextual training, and write group ID as p_i rather than g_i. All λ symbols in those summaries denote scalar configuration weights, not additional trainable parameters. Section 4.4 defines the training objectives; Appendix Section B.7 expands the masks and candidate directions. These explicit definitions govern the transfer study.

B.5 Local Encoding and Pair-Feature Details

Coordinates. An object has N patches. Patch i contains m_i points $\mathbf{x}_{ir} \in \mathbb{R}^3$ and unit normals $\mathbf{n}_{ir}$, with sample index $r = 1, \dots, m_i$. Its anchor, rotation and support extent are $\mathbf{a}_i, \mathbf{R}_i, s_i$ as defined in the main text. After object normalization,

$$\widetilde{\mathbf{x}}_{ir} = \mathbf{R}_i^\top (\mathbf{x}_{ir} - \mathbf{a}_i), \quad \widetilde{\mathbf{n}}_{ir} = \mathbf{R}_i^\top \mathbf{n}_{ir}, \quad \mathbf{x}_{ir} = \mathbf{R}_i \widetilde{\mathbf{x}}_{ir} + \mathbf{a}_i. \tag{22}$$

The local z-axis follows the anchor normal. The first tangent-axis reference is global $(0, 1, 0)$, replaced by $(1, 0, 0)$ when its absolute dot product with z exceeds .95. The reference is projected into the tangent plane. If its norm is below 10^{-6}, the leading eigenvector of the centered point covariance replaces it and is projected again. The remaining axes are normalized cross products, $y = z \times x$ and then $x = y \times z$. The implementation uses the first maximum eigenvalue returned by `argmax`, without an additional eigenvector-sign or repeated-eigenvalue convention; degenerate patches may therefore depend on numerical eigensolver choices. Coordinates are centered and rotated, without division by s_i. Scale is retained separately for context, pair features and geometric thresholds. This is an orientation convention, not exact rotation or scale invariance.

Quantized and continuous features. A PointNeXt-style encoder f_θ maps the local point–normal set to $\mathbf{f}_i \in \mathbb{R}^{192}$:

$$\mathbf{f}_i = f_\theta(\{(\widetilde{\mathbf{x}}_{ir}, \widetilde{\mathbf{n}}_{ir})\}_{r=1}^{m_i}), \quad t_i^{\mathrm{geo}} = Q_{\mathrm{geo}}(\mathbf{f}_i), \quad t_i^{\mathrm{bnd}} = Q_{\mathrm{bnd}}(\mathbf{f}_i). \tag{23}$$

Q_{geo} and Q_{bnd} each apply LayerNorm and multiply by four, project the 192-dimensional encoder output to respectively six and four scalars, apply a fixed randomly initialized orthogonal rotation, and quantize each scalar to one of seven levels. The seven output levels are symmetric about zero; the floor-rounding tie convention does not make the map strictly odd at bin boundaries. A rotated scalar z is mapped by

$$q_7(z) = \tfrac{1}{3} \operatorname{floor}_{\mathrm{STE}}\big(3(\tanh z + 1) + \tfrac{1}{2}\big) - 1 \in \{-1, -\tfrac{2}{3}, -\tfrac{1}{3}, 0, \tfrac{1}{3}, \tfrac{2}{3}, 1\}. \tag{24}$$

The forward floor is discrete; its straight-through backward derivative is one before the $\tanh$ bound, whose derivative saturates for large $|z|$. No hard clamp is configured. The rotation is a non-trainable buffer; the input/output projections are learned. The implementation also adds an MSE commitment term to geometry pretraining, an implementation choice rather than a property implied by the FSQ citation. Six geometry slots and four boundary slots give 7^6 and 7^4 possible indices. After inverse rotation, the quantized scalars are projected back to 192 dimensions and then linearly to a 128-dimensional feature, $\mathbf{e}_i^{\mathrm{geo}}$ or $\mathbf{e}_i^{\mathrm{bnd}}$; there is no independent 128-dimensional learned lookup vector for every possible index. These stream names do not imply semantic labels or independently recovered boundaries. Linear heads on $\mathbf{f}_i$ give a six-dimensional pose residual $\mathbf{r}_i^{\mathrm{pose}}$ and scalar scale residual r_i^{scale}, concatenated as

$$\mathbf{v}_i = [\mathbf{e}_i^{\mathrm{geo}}; \mathbf{e}_i^{\mathrm{bnd}}; \mathbf{r}_i^{\mathrm{pose}}; r_i^{\mathrm{scale}}] \in \mathbb{R}^{263}. \tag{25}$$

Semicolons denote concatenation. These residual features differ from both supplied frames and the pairwise pose prediction. The local-field phase projects $\mathbf{v}_i$ to width 256 for decoding; the context phase uses its own input projection. F_θ in the main text abbreviates this complete feature computation, including its learned projections and embeddings.

Pair geometry. For candidate (i, j), with $\epsilon = 10^{-6}$,

$$\boldsymbol{\rho}_{ij} = [\mathbf{R}_i^\top (\mathbf{a}_j - \mathbf{a}_i); \operatorname{quat}(\mathbf{R}_i^\top \mathbf{R}_j)] \in \mathbb{R}^7, \qquad r_{ij} = \log \frac{s_j + \epsilon}{s_i + \epsilon}. \tag{26}$$

quat gives a unit quaternion in (x, y, z, w) order. Conversion uses the standard positive-trace branch when available, otherwise the branch of the largest diagonal entry (with the implementation's fixed branch order for ties), then normalizes with denominator at least 10^{-8}. It does not impose a separate $w \geq 0$ rule, so q and $-q$ are not explicitly identified. Translation is in patch i's frame without division by scale. The binary edge type $e_{ij} = \mathbf{1}[g_i \neq g_j]$ has a learned 16-dimensional embedding. The head's auxiliary output $\widehat{\Delta\boldsymbol{\rho}}_{ij} \in \mathbb{R}^7$ is trained against the geometric correction target in Section B.3.

B.6 Context Attention Details

The initial patch feature and embedding maps are defined in Eq. (1). For one attention head, the pair bias is

$$b_{ij} = B_g(\mathbf{1}[g_i = g_j]) + B_a(\boldsymbol{\eta}_{ij}),$$
$$\boldsymbol{\eta}_{ij} = \left[\frac{\mathbf{a}_i - \mathbf{a}_j}{\max(s_i, s_j, \epsilon)}; \log\frac{s_i + \epsilon}{s_j + \epsilon}\right] \in \mathbb{R}^4. \tag{27}$$

Here $\mathbf{1}[\cdot]$ is an indicator, B_g supplies a learned same/different-group bias, B_a is a two-layer geometric-bias network, and $\epsilon = 10^{-6}$ prevents division by zero. Both biases have one output per attention head. At a layer with input $\mathbf{h}_i$, let $\bar{\mathbf{h}}_i = \mathrm{LayerNorm}(\mathbf{h}_i)$ and

$$\mathbf{q}_i = W_Q\bar{\mathbf{h}}_i, \quad \mathbf{k}_j = W_K\bar{\mathbf{h}}_j, \quad \mathbf{w}_j = W_V\bar{\mathbf{h}}_j,$$
$$a_{ij} = \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j/\sqrt{d} + b_{ij})}{\sum_{k=1}^{N}\exp(\mathbf{q}_i^\top \mathbf{k}_k/\sqrt{d} + b_{ik})}. \tag{28}$$

W_Q, W_K, W_V are the head's query, key, and value projections; $d = 64$ is its width, and a_{ij} is the normalized attention weight. The head aggregates $\sum_j a_{ij}\mathbf{w}_j$. Multi-head output projection, residual connections, and a width-1024 feed-forward block complete each layer; padding is masked. The final normalized feature is $\mathbf{h}_i \in \mathbb{R}^{256}$. Attention spans the object's patches; it is not restricted to the candidate-pair graph. Group biases are soft preferences, not hard masks forbidding cross-group interaction.

For the trained self-only control, every attention layer instead uses

$$a_{ij}^{\mathrm{self}} = \mathrm{softmax}_j\left(\mathbf{q}_i^\top \mathbf{k}_j/\sqrt{d} + b_{ij} + M_{ij}\right), \qquad M_{ij} = \begin{cases} 0 & i = j, \\ -\infty & i \neq j. \end{cases} \tag{29}$$

The implementation overwrites off-diagonal entries of the additive attention-bias tensor with -10^4 before PyTorch adds that tensor to the attention logits. In a component-level numerical check on the training device, the bias and attention weights were float32 under fp32, fp16, and bf16 inputs; off-diagonal weights were zero and diagonal weights were one. This retains the input cues, value/output projections, residual path, normalization, feed-forward blocks and depth, but removes cross-patch attention in both context and relation training as well as evaluation. The connection head still consumes a pair of resulting patch features; the control therefore isolates attention-mediated exchange, not all two-patch computation, and the effective attention capacity is different.

B.7 Training Losses and Primary Configuration

Geometry learning. A local field decoder predicts whether a query location lies inside the surface and its normal where a normal target is available. A field is a function evaluated at query coordinates, not a pre-existing mesh output. If a is an occupancy logit and $o \in [0, 1]$ its target, define

$$\mathrm{BCE}(a, o) = -o\log\sigma(a) - (1 - o)\log\sigma(-a). \tag{30}$$

Occupancy learning averages this binary cross-entropy over valid queries, with the configured query-region balancing. The normal loss averages $1 - \hat{\mathbf{n}}^\top\mathbf{n}$ on normal-labeled samples, where both vectors are unit length. Quantization, overfill, and surface-consistency terms regularize local geometry training; the configured objective also includes a code-usage statistic computed from discrete counts. It has no encoder gradient, but it is included in the logged validation total and can affect source checkpoint selection when its weight is nonzero. Definitions and phase-specific weights are given in Appendix Section B.4. They do not use attachment-parent annotations.

Connection learning. The main-text objective in Eq. (6) combines the following masked terms. Let $\langle\cdot\rangle_{\mathcal{E}}$ mean the arithmetic mean over real, non-padded source edges:

$$\mathcal{L}_{\mathrm{comp}} = \langle\mathrm{BCE}(\hat{C}_{ij}, C_{ij})\rangle_{\mathcal{E}},$$
$$\mathcal{L}_{\mathrm{coll}} = \langle\mathrm{BCE}(\hat{y}_{ij}^{\mathrm{coll}}, y_{ij}^{\mathrm{coll}})\rangle_{\mathcal{E}}, \tag{31}$$
$$\mathcal{L}_{\mathrm{inv}} = \langle\mathrm{BCE}(\hat{y}_{ij}^{\mathrm{inv}}, y_{ij}^{\mathrm{inv}})\rangle_{\mathcal{E}}.$$

Missing-target masks additionally exclude unavailable collision or pose labels. Compatibility uses a *soft* target C_{ij} even though the loss is binary cross-entropy. $\mathcal{L}_{\text{pose}}$ is coordinate-averaged smooth-L_1 loss against the geometric pose-correction target on labeled edges with $C_{ij} \geq .5$. Smooth-L_1 is $\frac{1}{2}z^2$ for residual magnitude $|z| < 1$ and $|z| - .5$ otherwise.

For the separation term, positive edges satisfy $C_{ij} \geq .55$, $y_{ij}^{\text{coll}} < .5$, and $y_{ij}^{\text{inv}} < .5$. Negative edges have $C_{ij} < .35$ or either collision/invalidity target at least .5. Writing these mini-batch sets as $\mathcal{P}$ and $\mathcal{N}$,

$$\mathcal{L}_{\text{sep}} = \max\left(0, .2 - \min_{(i,j)\in\mathcal{P}} \hat{C}_{ij} + \max_{(i,j)\in\mathcal{N}} \hat{C}_{ij}\right). \tag{32}$$

It is zero if either set is empty. This term uses compatibility logits, not the combined ranking score. In the matched fits, $\mathcal{P}$ and $\mathcal{N}$ are pooled across edges of the four objects in one forward microbatch. Gradients from 64 such microbatches are accumulated before an optimizer step; the extrema are not recomputed over a joint 256-object pool. The historical 32-by-8 schedule has the same per-forward loss semantics.

The ranking term uses the fixed geometry-derived patch serialization, an ordering of patches stored with each source object. Groups are ordered by decreasing patch count, then centroid coordinates (x, y, z), then integer group ID. Within a group, the first patch is nearest its group centroid (smallest patch index on a tie); subsequent patches follow a nearest-neighbor walk, again breaking ties by index. The first patch of the first group is the root. A stored serialization parent, used to define the source order but not as a released attachment label, is selected among earlier patches using stored edges in either direction: the forward edge is preferred on equal scores, a reverse-only edge receives a .01 penalty, same-group edges receive a .05 bonus, and the first earlier patch wins remaining ties. With no stored edge, the nearest earlier patch is used. At a non-root patch c, ranking compares candidate edges connecting c to previously visited patches, admitting either stored edge direction. $\mathcal{A}_c^+$ contains cross-group edges satisfying the positive criteria above and $\mathcal{A}_c^-$ contains the other eligible edges. Only tasks with both sets nonempty enter $\mathcal{T}$:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{T}|}\sum_{c\in\mathcal{T}} \text{softplus}\left(.35 + \max_{(i,j)\in\mathcal{A}_c^-} \ell_{ij}^{\text{train}} - \max_{(i,j)\in\mathcal{A}_c^+} \ell_{ij}^{\text{train}}\right). \tag{33}$$

Here $\text{softplus}(z) = \log(1 + e^z)$ and ℓ^{train} is the unclipped log-sigmoid score defined alongside Eq. (4). An empty $\mathcal{T}$ gives zero loss. This encourages the best positive to outrank the strongest negative, not every positive to outrank every negative. Source training thus differs from the directed component-parent test both in labels and candidate organization; the separation negatives $\mathcal{N}$ and ranking negatives $\mathcal{A}_c^-$ need not coincide. The denominator $|\mathcal{T}|$ is the number of eligible child-ranking tasks in that forward microbatch, across its objects; it is not the number of edges, objects or accumulated microbatches.

Primary matched configuration. The final network has 3,988,604 total parameters, but only the stated modules update in each phase. The three paired Full-context and self-only fits start from one source-selected local tokenizer initialization and the same fixed 21,309/1,190 source relation indices. Their context checkpoints initialize paired connection-head fits. Table 8 is transcribed from the actual Full-context and self-only context/relation run configurations, not from the earlier 20-epoch continuation or the auxiliary-loss study.

The context stage optimizes geometry reconstruction; the relation stage turns off reconstruction losses while still recomputing contextual features. Validation and target inference use evaluation mode. No target parent label selects a checkpoint. The field decoder is a training dependency but is not used in target parent ranking. The shared local tokenizer has a historical source-only initialization lineage; the matched experiment repeats context and head fitting, not tokenizer training or a fresh data split.

C Complete Matched Comparisons

C.1 Matched Feature Paths and Seed Results

All six arms use three paired seeds, the same source relation indices, 4,069 target queries and source-validation checkpoint selection. The primary Full-context–self-only comparison holds tokenizer,

Table 8: *Primary matched Full-context/self-only training.* Both conditions use AdamW, weight decay .01, mixed precision, three paired seeds and source-validation minimum total loss for checkpoint selection. The self-only mask applies in both stages; all other input cues, widths, depths, source indices and budgets match. Microbatch $\times$ accumulation gives 256 objects per optimizer step but does not enlarge the separation-loss extrema pool.

Stage	Initialization and updated modules	Optimization and selection
Local initialization	Shared ShapeNet-selected tokenizer and local geometry features; not refitted across the three paired seeds	Occupancy/normal and geometry objectives above; source-selected checkpoint
Context geometry	Shared local initialization; update context projection, four-layer Transformer and field decoder; tokenizer frozen	40 epochs, learning rate 10^{-4}, microbatch 32, accumulation 8, dropout .1; minimum source validation loss
Parent-relation head	Selected condition-specific context; paired head initialization; freeze tokenizer, context and decoder	30 epochs, learning rate 5×10^{-5}, microbatch 4, accumulation 64, dropout .1 in live context features; Eq. (6); minimum source validation loss

group/position/scale cues, Transformer width and depth, paired head initialization and epoch budgets fixed. Self-only masks cross-patch attention in context training, relation-head training and evaluation; its pair head still reads both patches (Eq. (29)). Local bypasses context blocks and additive context cues, so it is a broader feature-path comparison.

Table 9: *Complete matched-arm outcomes.* Ungated ParentAcc (%) over three fits on the same 4,069 queries. Complete uses the configured score; Compat. independently reaggregates compatibility alone. The first four arms train only the head after their context fit; FT also updates the representation.

Arm	Complete	Compat.
Frozen representation; head trained		
Local	34.33	34.59
Self-only	38.49	39.13
Full-context	44.74	42.36
Spatial	45.60	42.67
Representation and head fine-tuned		
Pretrained FT	42.67	46.60
Random FT	45.65	47.15

Table 10: *Per-seed ParentAcc (%).* C is the configured complete score and K compatibility-only. These are paired context/head fits rather than the earlier 20-epoch head-continuation family.

	Seed 0		Seed 1		Seed 2	
Arm	C	K	C	K	C	K
Local	34.85	34.78	34.11	33.89	34.04	35.09
Spatial	48.10	42.71	44.85	41.83	43.84	43.47
Full-context	47.78	39.64	42.37	43.40	44.09	44.04
Self-only	38.93	39.00	38.85	39.94	37.70	38.44
Pretrained FT	43.30	46.96	41.75	47.24	42.96	45.59
Random FT	45.51	48.96	44.48	44.70	46.94	47.80

C.1.1 Trained Self-Only Control

The source-selected Full-context–self-only gain is $+6.25$ points [95% paired object-bootstrap CI $+4.43, +8.15$] with the complete score and $+3.24$ [$+1.70, +4.84$] with compatibility alone. Every paired seed has a positive point difference (Table 10). These intervals condition on the three fitted seeds and resample whole objects; they do not measure retraining a tokenizer or sampling another target set. The attention mask is an architecture intervention within this network, not a test of every possible two-patch computation.

Full-context–Local is $+10.41$ $[+7.84, +12.93]$ and $+7.77$ $[+5.66, +9.89]$ points under complete and compatibility scoring. Full-context–Spatial is -0.85 $[-2.80, +1.05]$ and -0.31 $[-2.54, +1.76]$, leaving the separate benefit of group cues unresolved. Pretrained–Random FT is -2.97 $[-4.50, -1.47]$ and -0.56 $[-1.95, +0.82]$; pretraining is not consistently better in these FT conditions. The source-fixed ambiguity high-minus-low interaction for Full-context–Local is -2.17 $[-10.75, +5.91]$, so a high-bin gain does not establish concentration in geometrically ambiguous cases.

C.2 Training Targets Versus Score Readouts

This secondary intervention separates what supervises the frozen-context connection head from which of its outputs are read at inference. Four conditions cross collision supervision a and invalidity supervision b:

$$\mathcal{L}_{a,b} = .5\mathcal{L}_{\text{comp}} + .2a\mathcal{L}_{\text{coll}} + .5b\mathcal{L}_{\text{inv}}, \qquad a, b \in \{0, 1\}. \tag{34}$$

Pose, separation, hard ranking and reconstruction losses are zero in *every* arm. The same three frozen Full-context representations, paired head initializations, source indices and batch order are used across the four objectives. Context dropout is disabled for all these fits. Each head receives 30 epochs of AdamW at 5×10^{-5}, weight decay .01, microbatch four and accumulation 64. Minimum source-validation compatibility BCE selects every checkpoint; all twelve selected checkpoints are from epoch 30. No target parent label enters training or selection.

The primary contrast holds the compatibility-only readout fixed across objectives. If $A_{a,b}(q)$ is query q correctness averaged over the three fits, the prespecified invalidity effect is

$$\Delta_{\text{inv}} = \frac{100}{2Q} \sum_{q=1}^{Q} [(A_{0,1}(q) - A_{0,0}(q)) + (A_{1,1}(q) - A_{1,0}(q))], \quad Q = 4{,}069. \tag{35}$$

Whole-object bootstrap intervals use 10,000 paired draws conditional on the fitted heads. The same checkpoint is then read with each supervised field combination, independently recomputing patch-to-component maxima; an untrained field is never inserted into a purported complete score.

Head training objective	Seed 0	Seed 1	Seed 2	Mean	Checked
Compatibility only	43.23	40.77	36.20	40.07	39.09
Plus collision	42.22	41.12	40.85	41.39	40.39
Plus invalidity	41.29	43.13	40.18	41.53	40.50
All three fields	46.23	39.05	41.09	42.12	41.11

Table 11: *Fixed-feature training-loss intervention.* ParentAcc (%) on the same 4,069 PartNet-Mobility queries, using compatibility-only scores for all 12 source-selected fits. "Checked" applies the original coverage and kinematic gates and is secondary. Means average the three fitted seeds; these arms differ from the historical C2LT objective.

The primary invalidity main effect is $+1.10$ points with a pointwise 95% paired object-bootstrap interval $[+0.37, +1.82]$. Without collision its simple effect is $+1.47$ $[+0.43, +2.48]$; with collision it is $+0.73$ $[-0.09, +1.58]$. The secondary collision main effect is $+0.96$ $[+0.31, +1.64]$, while the factorial interaction is -0.74 $[-1.92, +0.46]$. Final-epoch and source-selected checkpoints give identical query outcomes and contrasts; in every run the selected checkpoint is at epoch 30. The invalidity main effect by seed is $+1.03$, $+0.15$, $+2.11$ points, and the collision main effect is $+1.97$, -1.87, $+2.78$ points. Thus the object-bootstrap intervals do not establish a positive effect across random initializations. This target set was inspected before the study, so these results are exploratory rather than untouched confirmation. Here $A_{a,b}(q)$ in Eq. (35) is the correctness indicator averaged over the three fitted seeds; multiplying its contrasts by 100 gives percentage points.

Head training	Compat.	+ collision	+ invalidity	Complete
Compatibility only	40.07	–	–	–
Plus collision	41.39	40.97	–	–
Plus invalidity	41.53	–	41.43	–
All three fields	42.12	41.53	41.93	41.65

Table 12: *Training objective versus inference readout on the same 12 fits.* Percent ParentAcc, averaged over three paired seeds. A dash means the corresponding predicted field was not supervised and is therefore not read. Every cell uses the same 4,069 queries and separately recomputes the patch-to-component maximum. Rows change training; columns change only the readout of the same checkpoint.

Within the plus-invalidity row, adding the predicted invalidity penalty to compatibility changes ParentAcc by -0.11 points $[-0.29, +0.07]$; this interval includes zero. Within the all-three row, the configured complete readout changes it by -0.48 $[-0.77, -0.18]$, while compatibility plus invalidity changes it by -0.19 $[-0.41, +0.04]$. These are same-checkpoint readout contrasts, unlike the factorial training effects above. In these fits, extra supervision can benefit the compatibility output on average without making a penalized readout preferable. This fixed-feature study is a different checkpoint family from the primary Full-context–self-only comparison.

D Sampling and Candidate Analyses

D.1 Frozen Self-Only Equal-Pair Replay and Coverage

We crossed the trained self-only comparison with the prior equal-pair-budget intervention without fitting any new target model. The three self-only heads were selected by ShapeNet validation, frozen, and re-inferred on the unchanged 931-object target bank. For query q, let $n_q(v)$ be the number of ordered patch pairs whose anchor-face component IDs are child u_q and candidate v. For candidates with $n_q(v) > 0$, set $K_q = \min_v n_q(v)$. Each of 32 fixed, label-blind repetitions samples K_q distinct pairs for every such candidate. The pair indices and candidate set are exactly those in the earlier full-context and Dense Support replay; only the self-only scores are newly inferred. Each readout independently takes its sampled maximum and ranks candidates by the same component-ID tie rule. A query with $K_q = 0$ remains incorrect.

Table 13 reports every seed and both readouts; the main text compares the complete-score means with Dense Support on the same pair draws.

Table 13: *Per-seed effect of equal pair opportunities.* ParentAcc (%) for full-context (FC) and self-only (SO) on the same 4,069 queries. Equal columns average 32 fixed label-blind pair draws per query; all rows use source-selected checkpoints.

Readout	Seed	FC/full	SO/full	FC/equal	SO/equal
Complete	0	47.78	38.93	29.71	25.35
	1	42.37	38.85	28.32	24.97
	2	44.09	37.70	28.35	23.75
Compatibility	0	39.64	39.00	23.62	23.32
	1	43.40	39.94	26.87	24.17
	2	44.04	38.44	28.13	23.45

Finite-bank sampling effect. For a candidate's n fixed pair scores, draw $1 \le K \le n$ without replacement. Let $n(t)$ count scores at most t, M_K be the sampled maximum, and b count geometric-proxy-positive pairs. With $\binom{n(t)}{K} = 0$ when $n(t) < K$,

$$\Pr(M_K \le t) = \frac{\binom{n(t)}{K}}{\binom{n}{K}}, \qquad \Pr(\text{retain a witness}) = 1 - \frac{\binom{n-b}{K}}{\binom{n}{K}}. \tag{36}$$

Equation (36) counts two effects on a fixed candidate bank without independent-score assumptions. Its first numerator is zero when $n(t) < K$; a specified pair is retained with probability K/n. On the fixed target bank, the median K/n is .1145 for the 2,824 covered gold-candidate bags and 1.0

for 16,099 alternative bags; 1,358 gold bags have $K/n \leq .1$. Equal budgeting therefore removes many more pair opportunities from gold candidates in this bank. A specified high-scoring pair is not necessarily a contact witness. We define a *geometric-proxy witness* using the fixed evaluator-side threshold $C \geq .55$, $y^{\text{coll}} < .5$ and $y^{\text{inv}} < .5$, independently of the network's scores. If b of a candidate's n pairs meet this rule, its retention probability follows Eq. (36). Among the 2,824 covered gold bags, 528 have at least one such witness; their mean retention probability is .535. Across all 4,069 queries, including missing and witness-free gold bags, the mean is .0694. Thus equal budgeting changes both extreme-score opportunities and retention of this localized proxy evidence. The proxy is measured from the reference mesh for diagnosis only; it is not physical contact or a parent label. Deidentified counts are checked in `anonymous_evaluation_audit/c18_metrics/`.

The seed-averaged FC–SO differences are $+6.25$ points $[+4.43, +8.15]$ with all pairs and $+4.10$ $[+3.05, +5.24]$ with equal pair budgets under complete scoring. Compatibility gives $+3.24$ $[+1.70, +4.84]$ and $+2.56$ $[+1.68, +3.50]$, respectively. Intervals are paired 10,000-draw object bootstraps conditional on fitted seeds and fixed repetitions. The complete-score equal-minus-full interaction is -2.15 $[-3.53, -0.76]$ points; for compatibility it is -0.67 $[-1.93, +0.54]$. The full-bank self-only correctness matches the previous replay query by query. Anonymous per-query, per-seed, per-repeat binary outcomes and a standard-library verifier are supplied in `c17_equal_pair/`. The audit verifies point estimates and matching conditions; it does not recompute neural inference or bootstrap intervals.

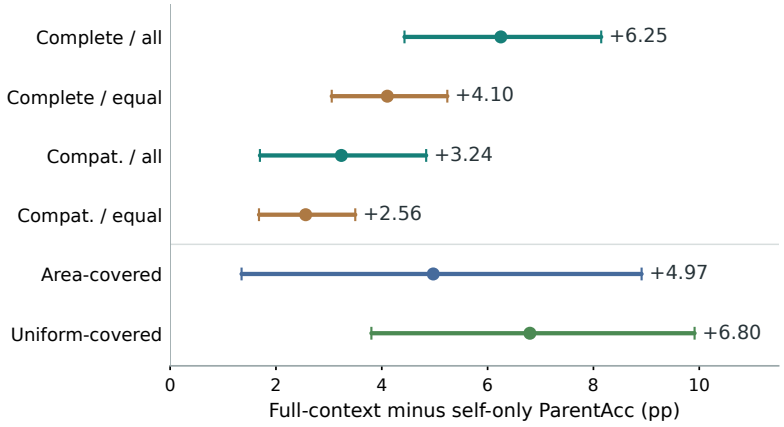


Figure 3: *Full context-gain readout across score budgets and patch banks.* The upper rows use the original bank with all pairs or 32 fixed equal-K draws; the lower rows use the complete score on the area- and uniform-covered banks of Section D.3. Each point is Full-context minus matched self-only ParentAcc on the same 4,069 queries; whiskers are 95% paired object-bootstrap intervals conditional on fitted seeds and draws.

Coverage audit. On the fixed bank, 2,966/4,069 queries have at least one represented candidate. The gold child–parent pair has a finite score in 2,824/4,069. Of the 1,245 failures, 1,102 lack a sampled child anchor, one has a sampled child but no other sampled component, and 142 have candidate pairs but no sampled gold parent. In 1,046 of the child-missing cases the gold parent does have a sampled anchor. Geometry-only patch sampling is not stratified by physical component; the sampled source-face ID supplies an anchor assignment, not patch purity. The coverage decomposition was checked against the fixed target archive and candidate-count record. The separate component-covered resampling experiment in Section D.3 changes the candidate and patch banks while retaining the same objects, labels and total patch budget.

D.2 Identifier and Pair-Opportunity Controls

Component-ID provenance. The target audit reads each physical component's integer `id` and `parent` from the asset's `mobility_v2.json`. It maps URDF visual links to these official node IDs through their part lists, refusing ambiguous or missing mappings, and assigns the resulting ID to each sampled patch through face membership. The preprocessing does not sort or renumber these

physical IDs. In the evaluated `geometry_only` bank, patch anchors are allocated by geometric-group face area, without consulting physical-link IDs or guaranteeing a patch per link; component patch counts are computed afterward from sampled-face membership. This count is not a direct measurement of component volume or surface area. A separate `link_stratified` sampler exists in the codebase but was not used for this bank. In the represented-parent population, the released parent has a larger numeric ID than the child in 2,617 of 2,824 queries, but this association does not establish why the source annotations use that order.

Frozen bijective relabeling. For each of 931 objects, 32 deterministic permutations π of *all* its physical-component node IDs were drawn without using parent labels. This full domain includes 77 node IDs absent from the union of query and candidate IDs across the bank. Patch membership $m(i)$, queried child u, released parent v^* and candidate IDs were jointly relabeled as $\pi(m(i))$, $\pi(u)$, $\pi(v^*)$ and $\pi(v)$. Geometry, geometric-group IDs, model parameters and predicted patch-pair scores were held fixed. Because physical membership is not a neural input, the component maximum obeys

$$S_\pi(\pi(u), \pi(v)) = \max_{m(i)=u,\, m(j)=v} \ell_{ij} = S(u, v). \tag{37}$$

We replayed the stored candidate maxima for every matched fit and checked the original ranking against its recorded evaluation. This is a frozen-score relabeling and exact aggregation-equivariance check, not a newly trained or newly inferred model. With a unique maximum, mapping the selected ID back must recover the original parent choice; only the smallest-ID tie rule can break equivariance.

Table 14: *Frozen ID relabeling of the original five matched arms (15 fits).* This replay predates the separate self-only control. Accuracy is ungated ParentAcc (%) on all 4,069 queries; relabeled accuracy averages 32 bijections per object. Ties count original queries with more than one exact top score. Changes count mapped-back selection changes across $4{,}069 \times 32$ trials per fit. Every change is on an original top tie.

Fit	Original	Relabeled	Top ties	Changes
Grouped 0	47.78	47.78	0	0
Grouped 1	42.37	42.37	0	0
Grouped 2	44.09	44.09	0	0
Local 0	34.85	34.79	65	1,699
Local 1	34.11	34.26	32	725
Local 2	34.04	33.96	73	1,894
Spatial 0	48.10	48.10	0	0
Spatial 1	44.85	44.85	0	0
Spatial 2	43.84	43.84	0	0
Pretrained FT 0	43.30	43.30	0	0
Pretrained FT 1	41.75	41.75	0	0
Pretrained FT 2	42.96	42.96	0	0
Random FT 0	45.51	45.51	0	0
Random FT 1	44.48	44.48	0	0
Random FT 2	46.94	46.94	0	0

The largest-ID direct rule falls from 62.60% to 20.38% averaged over relabelings, while the sampled-patch-count rule remains 58.65% versus 58.54% originally. Its small change comes from ties broken by the newly assigned IDs. Identifier ordering is thus a strong predictor in this bank, not a neural input shortcut. Relabeling cannot remove the geometric and sampling associations that may underlie the patch-count rule.

Equal candidate pair-score budget. For a fixed query component u, candidate v owns $n_{uv} = |\{(i, j) : m(i) = u, m(j) = v\}|$ directed patch pairs. We set the query-specific budget $K_u = \min_v n_{uv}$ over the represented candidates. For each of 32 deterministic, label-blind repeats, we sample exactly K_u distinct pair indices without replacement from every candidate and take the maximum of their *existing* scores. The sampled indices are identical across fitted scorers and readouts for that query and repeat. The original unbalanced maximum, full candidate list and all 4,069 queries remain the reference. We average per-repeat accuracy, not scores or predictions across repeats. This intervention fixes the number of opportunities to attain a large maximum, while also reducing local-surface coverage; it does not equalize component geometry or retrain the model.

Table 15: *Equal pair-score opportunities within each query.* Percent ungated ParentAcc on all 4,069 queries. Full uses every directed patch pair; balanced averages 32 label-blind subsamples with K_u pairs for every candidate of query u. The same sampled pair indices are used across scorers. Compatibility scores are reaggregated separately; Dense has one geometric score.

	Complete		Compatibility	
Fit	Full	Balanced	Full	Balanced
Grouped 0	47.78	29.71	39.64	23.62
Grouped 1	42.37	28.32	43.40	26.87
Grouped 2	44.09	28.35	44.04	28.13
Grouped mean	*44.74*	*28.79*	*42.36*	*26.21*
Local 0	34.85	17.66	34.78	18.00
Local 1	34.11	17.59	33.89	17.52
Local 2	34.04	17.82	35.09	18.16
Local mean	*34.33*	*17.69*	*34.59*	*17.89*
Dense Support	47.43	25.94	–	–

Across all queries, K_u has quartiles $0, 4, 48$ and range 0–1,020. A zero budget occurs when no candidate is represented; such queries remain wrong in both readouts. Among 2,966 candidate-present queries, the budget quartiles are $3, 16, 80$, and 489 have $K_u = 1$. After averaging repeats and fitted seeds, balanced Grouped–Local is $+11.10$ points [95% paired object-bootstrap CI $+9.03, +13.25$] under complete scoring and $+8.31$ $[+6.80, +9.98]$ under compatibility alone. Balanced Grouped–Dense is $+2.85$ $[+1.17, +4.61]$, compared with -2.69 $[-5.22, -0.17]$ under the full maximum. These intervals resample objects and condition on fitted runs and the 32 draws. The balanced intervention lowers all three patch-scoring methods markedly, consistent with both fewer maximum opportunities and less chance of sampling a true contact. It leaves the Grouped–Local feature-path contrast present but does not remove the separate patch-count association in candidate construction or establish structural semantics.

The direct paired contrast $(\text{Grouped} - \text{Local})_{\text{balanced}} - (\text{Grouped} - \text{Local})_{\text{full}}$ is $+0.69$ percentage points [95% paired object-bootstrap CI $-0.94, +2.32$] for complete scoring and $+0.54$ $[-0.93, +2.04]$ for compatibility scoring. Each query contributes its full-bank outcome and mean over the same 32 fixed score-subset repeats before object resampling. The intervals do not resolve a change in the size of the feature-path gain.

Table 16: *Complete-score equal-budget results by per-candidate pair budget.* All eligible queries remain in the corresponding stratum. Uniform is the expected accuracy from uniform choice among unchanged represented candidates, so its probability is unaffected by score subsampling. Grouped and Local average three fitted seeds; Dense is a fixed rule. Accuracy is in percent; differences in percentage points. Strata are descriptive, not separately powered tests.

K_u	Queries	Uniform	Grouped	Local	Dense
No candidate pair	1,103	0.00	0.00	0.00	0.00
1	489	15.34	19.63	11.92	6.99
2–4	605	15.32	24.67	12.67	14.58
5–16	420	27.34	45.22	28.61	26.00
17–64	653	34.85	52.98	34.25	47.40
> 64	799	40.10	48.86	30.17	64.37

D.3 Sampling, Direction and Group-ID Diagnostics

Component-covered intervention. We regenerate the same two selected PartNet-Mobility cohorts (512 and 419 objects) with their original preprocessing seeds, reference-pose meshes, 4,069 released-parent queries and 64-patch budget per object. The original geometry-only sampler has no per-component guarantee. The area-covered variant first assigns one anchor to every mesh component and allocates the remainder in proportion to component surface area; the uniform-covered variant allocates the remainder so component counts differ by at most one. The same anchor allocations and candidate lists are used by every scorer within a bank. Neither variant changes source training, checkpoints, labels, query eligibility or total patch count. Component membership is supplied by the

task and is used for anchor allocation and score aggregation, not as a neural input or as a parent label. This is a target-bank intervention after repeated inspection, not an independent validation set.

For each bank, three source-selected Full-context and self-only checkpoints score the newly sampled patches. Dense Support and Geometric NN use those same patch geometries. Direct full proxy additionally computes reference-mesh boundary and occupancy measurements. Most-patches chooses the largest candidate patch count; largest surface area chooses the largest candidate reference-mesh triangle area, breaking ties by component ID. The latter has whole-component geometric access and therefore diagnoses the target label distribution rather than matching patch-scorer computation. All reported accuracies are ungated on the same 4,069 queries; a missing gold candidate is incorrect. All 4,069 gold child–parent pairs occur in both covered banks.

Table 17: *Frozen-checkpoint outcomes by sampling bank.* ParentAcc (%) on the same objects and released parents. Original is the geometry-only bank; area and uniform each cover every mesh component at fixed total patch count. FC and SO denote Full-context and self-only. The largest-area rule accesses whole-component mesh area.

Scorer	Original	Area-covered	Uniform-covered
FC seed 0	47.78	67.02	47.78
SO seed 0	38.93	60.56	37.53
FC seed 1	42.37	66.55	43.57
SO seed 1	38.85	60.29	38.61
FC seed 2	44.09	62.55	43.23
SO seed 2	37.70	60.36	38.04
FC mean	44.74	65.37	44.86
SO mean	38.49	60.40	38.06
Dense Support	47.43	54.76	36.84
Geometric NN	42.91	54.44	34.01
Direct full proxy	46.72	58.86	38.66
Most parent patches	58.54	88.62	11.06
Largest mesh surface area	62.25	90.17	90.17

Fixed-candidate decomposition. Let R be the 2,824 queries whose gold child–parent pair is represented under original sampling, and let $Q = 4{,}069$. For $q \in R$, original candidates V_q^0 are contained in covered-bank candidates V_q^1; 691 queries in 153 objects gain competitors. For each applicable fit s, write $a_{q,s}^{ab}$ for Top-1 correctness using scores from sampling version $a \in \{0, 1\}$ on candidate set V_q^b, with the same ID tie rule and an absent finite score counted wrong. We evaluate the path $(S^0, V^0) \to (S^1, V^0) \to (S^1, V^1)$. Since original correctness is zero for $q \notin R$, the change in ParentAcc, in percentage points, is exactly

$$\Delta A = \frac{100}{QF} \sum_{s=1}^{F} \left[\sum_{q \in R} (a_{q,s}^{10} - a_{q,s}^{00}) + \sum_{q \in R} (a_{q,s}^{11} - a_{q,s}^{10}) + \sum_{q \notin R} a_{q,s}^{11} \right], \tag{38}$$

where $F = 3$ for the paired learned fits and $F = 1$ for deterministic geometric scorers. The terms are fixed-old-candidate score change, fixed-new-score candidate expansion and success on newly covered queries. For $q \in R$, fixed scores and tie-breaking imply $a_{q,s}^{11} \le a_{q,s}^{10}$: adding competitors cannot turn an incorrect old-candidate Top-1 into the correct parent. The area rule's fixed-candidate term is zero because its areas and tie rule do not change. This is accounting along a specified evaluation path, not a separate causal identification of local resampling, contextual processing or within-candidate aggregation. The last term is not a pure coverage treatment effect: new sampling, scores and candidates coexist there.

Table 18 reports the resulting terms. Under near-uniform allocation, Full-context's fixed-candidate change is -9.69 points [95% paired object-bootstrap CI $-11.74, -7.63$], compared with -0.86 $[-1.19, -0.55]$ from adding candidates; newly covered queries add $+10.67$ $[+7.53, +14.19]$. Under area allocation, these terms are $+2.30$ $[+0.32, +4.37]$, -0.52 $[-0.77, -0.31]$ and $+18.85$

Table 18: *Where covered-bank changes come from.* All entries are percentage points of the same 4,069-query total. On originally covered queries, the new bank's scores are first restricted to old candidates (V_0), then expanded to covered-bank candidates (V_1). The last term is success on newly covered queries. Net changes use unrounded values; paired object-bootstrap intervals for every cell are in the numerical audit.

New bank	Scorer	Fixed-V_0 change	Added-candidate change	New coverage	Net change
Area-covered	Full-context	+2.30	-0.52	+18.85	+20.63
	Self-only	+4.69	-0.59	+17.80	+21.91
	Dense Support	-1.79	-1.55	+10.67	+7.32
	Direct full proxy	+1.06	-1.52	+12.61	+12.14
	Largest mesh area	+0.00	-0.20	+28.12	+27.92
Uniform-covered	Full-context	-9.69	-0.86	+10.67	+0.11
	Self-only	-8.06	-0.88	+8.51	-0.43
	Dense Support	-14.87	-1.40	+5.68	-10.59
	Direct full proxy	-12.61	-1.55	+6.09	-8.06
	Largest mesh area	+0.00	-0.20	+28.12	+27.92

$[+15.03, +22.78]$. The largest-area rule has zero fixed-candidate change and -0.20 points from new competitors in either bank. Score changes here combine altered patch evidence and object context; this accounting does not assign separate causal effects to them. All 40 termwise intervals and deidentified query outcomes are in `c20_candidate_decomposition/`.

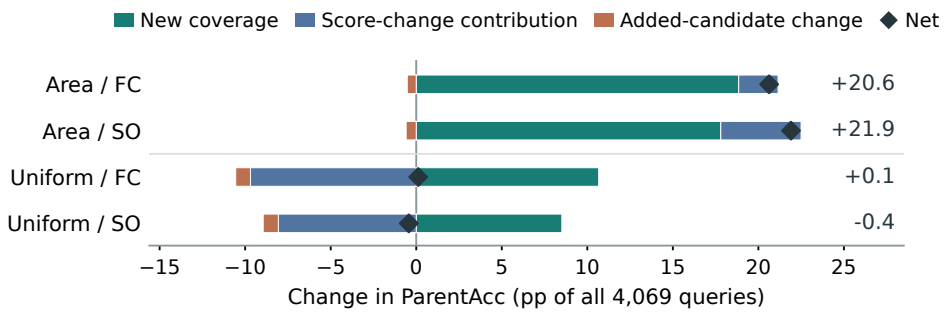


Figure 4: *Why full coverage can leave accuracy unchanged.* Changes in Full-context (FC) and self-only (SO) ParentAcc on all 4,069 queries. Resampling changes scores on the 2,824 previously covered queries; newly covered queries add correct predictions. Adding competitors cannot improve correctness at fixed scores. Diamonds show the net change before rounding.

Area-covered FC–SO is $+4.97$ percentage points [95% paired object-bootstrap CI $+1.35, +8.91$]; uniform-covered is $+6.80$ $[+3.80, +9.91]$. The paired changes in context gain relative to the original bank are -1.28 $[-5.13, +2.82]$ and $+0.55$ $[-2.70, +3.92]$, respectively. All intervals use 10,000 object-level resamples, condition on the three fitted seeds and the deterministic new banks, and do not measure uncertainty from a fresh target population or retraining. Area-proportional allocation makes patch count informative about component surface area: most-patches reaches 88.62%. Near-uniform allocation substantially reduces count variation and the rule falls to 11.06%; it does not prove count is uninformative. The largest-reference-mesh-area rule remains at 90.17% because all candidate components are now present. FC beats Dense Support and the direct full proxy by query-weighted accuracy in the uniform bank, but it remains far below the whole-mesh size rule. In the area bank, FC is also below patch count. The object-macro view below reverses FC's apparent lead over direct geometry. Neither protocol demonstrates higher overall accuracy than the area rule.

Across 931 equal-weighted objects, direct full proxy has higher accuracy than FC in both covered banks (74.83% versus 69.97%; 61.45% versus 57.20%). Across 42 equal-weighted categories, FC has higher accuracy in both banks (69.60% versus 64.97%; 55.23% versus 50.30%). The whole-mesh area rule has the highest point estimate under both summaries. These descriptive averages answer different weighting questions; no paired macro interval is claimed here.

Anchor assignment is not patch purity. For each of the 64 patch points used in this diagnostic, we compare the assigned anchor component with the component of the nearest face on the assembled

Table 19: *Object and category macro averages (%).* Each object or category contributes equally within a bank; query-level means are in Table 17. The area rule uses whole-component reference-mesh geometry.

	Area-covered		Uniform-covered	
Scorer	Object	Category	Object	Category
Full-context	69.97	69.60	57.20	55.23
Self-only	67.15	64.45	52.25	48.95
Dense Support	71.79	63.95	59.65	49.40
Direct full proxy	74.83	64.97	61.45	50.30
Most parent patches	85.41	82.81	19.32	16.22
Largest mesh surface area	85.79	83.98	85.79	83.98

reference mesh. The original bank matches on 94.68% of points, versus 88.80% for area-covered and 77.94% for uniform-covered sampling. Fully matching patches account for 80.36%, 70.92% and 58.87%, respectively. This is nearest-face agreement, not a guarantee of physical part purity: coincident or nearly touching surfaces can be ambiguous. The lower agreement under uniform allocation also limits interpretation of its attention contrast as a clean component-to-component information test.

Directed versus symmetric scoring. On the *original* bank, we re-infer the same three frozen Full-context checkpoints and compare $S(u, v)$ with $\max\{S(u, v), S(v, u)\}$. Candidate sets and the single released directed parent remain unchanged, unlike a multi-answer undirected endpoint. Mean ParentAcc changes from 44.74% to 47.05%, a paired $+2.30$ points $[+1.23, +3.49]$; 911 queries change their choice in at least one seed. Max-symmetrization performs better on this fixed bank. It also exposes each candidate to reverse-direction scores, increasing the opportunity for an extreme value; the contrast alone does not isolate direction semantics or establish which score better represents physical directionality.

Geometric-group numbering. The fixed-score component-ID relabeling in Section D.2 tests aggregation equivariance, not network invariance to numeric group labels. We therefore apply eight deterministic within-object bijections of geometric-group IDs and rerun each frozen Full-context network. Membership, geometry, edge types, physical-component IDs, candidates and gold parents remain fixed. Predicted parents change in 27.11% of the seed–permutation comparisons. Mean accuracy changes from 44.74% to 43.43%, a paired -1.32 points $[-2.66, +0.07]$. The interval crosses zero for accuracy, but the changed decisions rule out exact network-level invariance. No claim of permutation-invariant group semantics follows from the original fixed-score relabeling.

The anonymous numerical package provides per-query predictions, candidate IDs, checkpoint hashes, sampling-index hashes, patch-purity counts and paired-bootstrap point contrasts for both banks and the direction/group-ID interventions. Its standard-library evaluator verifies released numerical identities; third-party meshes and model weights are not included, so this is not a full training or inference reproduction.

Candidate availability and error overlap. The largest-area rule is recomputed from the same reference meshes in each bank. Its original 62.25% uses only patch-represented candidates, whereas both covered banks admit all supplied candidate components and make exactly the same 90.17% area-rule predictions query by query. The change is therefore in candidate availability, not in the rule's area measurements. Among 12,207 query–seed decisions (4,069 queries times three Full-context fits), Full-context alone is correct in 362 cases under area coverage and 369 under uniform coverage; the area rule alone is correct in 3,389 and 5,900 cases, respectively. These are paired descriptive counts, not independent decisions or evidence that the whole-mesh rule has matched patch-level information access. For each seed separately, an oracle choosing between that seed's Full-context Top-1 and the area-rule Top-1 using target correctness could gain 362/12,207 = 2.97 and 369/12,207 = 3.02 points over the area rule after averaging the three seedwise decisions. The corresponding post-hoc accuracies are 93.14% and 93.19%. These are bounds for that specific two-prediction choice, not for a multi-seed ensemble or candidate reranker. As one exploratory, fixed label-free check, we use the Full-context Top-1 only when all three fitted seeds agree and otherwise choose the largest-area candidate. This reaches 81.91% in the area-covered bank and 76.11% in the uniform-covered bank, respectively -8.26 [95% paired object-bootstrap CI $-11.67, -5.43$] and -14.06 $[-16.83, -11.41]$

points relative to the area rule. It evaluates three learned fits against one geometry rule on an inspected target cohort, with no source calibration or independent test; its failure does not exclude a different, properly selected rule. Per-query outcomes are in `c20_selector_probe/`. The deidentified candidate-gate rows and summary are in `c19_analysis/`.

D.4 Inference-Only Fusion With Patch Count

Table 20 tests whether frozen Full-context scores add value when combined with the sampled-patch-count rule. For query q, let V_q be the represented parent candidates, n_v their sampled patch counts and s_v the complete Full-context score after directed maximum aggregation. The count rule chooses $\arg\max_{v \in V_q}(n_v, -v)$, i.e., the lowest component ID on a count tie. Our confidence gate instead takes Full-context's Top-1 only when it differs from count's Top-1 and its softmax over finite s_v exceeds τ; otherwise it retains count's choice. We also test the log-linear rule $\arg\max_{v \in V_q}\{s_v + \lambda \log(1 + n_v)\}$ and a model-score tie-break restricted to the largest-count candidates. A further restricted ensemble normalizes candidate scores within each query and seed, averages the three normalized scores, and selects the highest-scoring candidate among those with $n_v \geq 0.75 \max_{w \in V_q} n_w$. Its count-fraction threshold was explored on the inspected target cohort, not selected on independent validation data. An empty candidate set remains incorrect. These are inference-only changes to the same three frozen source-selected checkpoints, 931 objects and 4,069 parent queries in each 64-patch bank. They use no target labels at prediction time and no retraining. The softmax uses unit temperature on the component-aggregated weighted log scores, not patch-pair logits; the switch condition is strictly $> \tau$. Exact count and score ties choose the lowest component ID. With no finite candidate, the prediction is absent and incorrect.

Table 20: *Frozen-score/count fusion.* Ungated ParentAcc (%). Except for the three-seed restricted ensemble, learned-score rows average over three Full-context seeds; count is repeated across seeds. The gate uses $\tau = 0.5$ and the log-linear rule uses $\lambda = 1$, both shown as fixed diagnostic settings. The restricted ensemble uses a target-inspected count fraction of 0.75; it is not an independently selected operating point. All three banks contain the same 4,069 inspected queries.

Inference rule	Original	Area-covered	Uniform-covered
Most patches	58.54	88.62	11.06
Full-context only	44.74	65.37	44.86
Count + confidence gate ($\tau = 0.5$)	56.40	86.97	19.13
Log-linear ($\lambda = 1$)	58.36	89.00	40.65
Count + model tie-break	58.71	89.01	35.84
Count-restricted ensemble (0.75)	59.01	89.31	44.63

Let $g_{q,s}(\tau)$ indicate that the model replaces count on query q and seed s, and let $Y^M_{q,s}$ and Y^C_q indicate correctness of the model and count choices. The exact percentage-point change is

$$\Delta A(\tau) = \frac{100}{3Q} \sum_{s=1}^{3} \sum_{q=1}^{Q} g_{q,s}(\tau)(Y^M_{q,s} - Y^C_q), \qquad Q = 4{,}069. \tag{39}$$

At $\tau = 0.5$, the gate changes 541 query–seed decisions on the original bank: 129 rescues, 390 harms and 22 cases where both choices are wrong. On the area-covered bank it changes 604 decisions: 201 rescues and 403 harms. These are repeated decisions over three fitted seeds, not $3Q$ independent samples. On the uniform-covered bank, count is a poor prior: the gate gains 8.07 points over count but remains 25.73 points below Full-context alone. Scanning the predeclared threshold grid from 0 to 1 does not improve over count on the original bank. There, $\tau = 0.55$ gives 58.27% (one rescue, 34 harms), while $\tau = 0.60$ gives 58.52% (no rescues, three harms). The highest softmax on a model–count disagreement is 0.6335, so $\tau \geq 0.64$ makes no override and exactly matches count. The largest area-bank gain on this grid is only 0.02 points at $\tau = 0.6$. This grid was evaluated on an already inspected target cohort and does not provide a valid threshold selection.

The model tie-break yields +0.17 points [95% paired object-bootstrap CI +0.03,+0.38] on original sampling and +0.39 [+0.07,+0.79] on area-covered sampling, but falls 9.02 points below Full-context on uniform-covered sampling. On the original bank, 111 queries from 46 objects have multiple candidates tied for the largest patch count. The tie-break changes only these queries; across the three seeds, its 190 changed decisions contain 24 rescues, three harms and 163 cases where both

answers are wrong. All non-tied predictions exactly match count. The restricted ensemble makes 52 rescues and 33 harms against count on the original bank, giving +0.47 points [95% object-bootstrap CI $-0.22, +1.22$]; the five-fold objectwise threshold-selection check reduces that difference to +0.15 [$-0.39, +0.65$]. On area-covered sampling it gains +0.69 points [+0.14,+1.29] over count, but its 89.31% accuracy remains below the 90.17% whole-mesh area rule. It also falls below Full-context alone on uniform-covered sampling (44.63% versus 44.86%). The ensemble requires three fitted models rather than one. Across the fixed log-linear weights $\lambda \in \{0, 0.25, 0.5, 1, 2, 5\}$, the best *same-target* area-bank result is 89.03%, only +0.41 points over count and below the 90.17% whole-mesh area rule. The apparent small gains do not establish a general or independently selected fusion benefit: the target queries and banks have been repeatedly inspected, and no separate calibration population exists. The configured s_v are weighted log scores, not calibrated parent logits; their softmax and the log-linear count weighting should not be interpreted as calibrated posterior probabilities. The separate anonymous numerical audit contains candidate scores, patch counts, point results, paired intervals and frozen-source hashes in `c22_fusion/`; its evaluator recomputes point results, not neural inference or bootstrap intervals.

E Sensitivity and Scope

E.1 Input-Only Shortcuts and Aggregation Sensitivity

The same 4,069 queries and represented candidate lists underlie every row below. The largest-ID rule uses the supplied component identifiers. The patch-count rule chooses the candidate parent with the most sampled patches; because all directed non-self patch pairs are scored, its ordering equals the ordering by child–parent pair count for a fixed child. Neither rule uses target labels for selection or a trained score. Both break ties by smallest component ID, as do the learned methods. Missing parents count as wrong.

Method	Query mean	Object macro	Category macro
Local, complete score	34.33	45.25	47.42
Grouped, complete score	44.74	58.26	57.35
Dense Support	47.43	64.83	60.06
Most parent patches	58.54	69.72	66.62
Largest component ID	62.60	74.06	69.19

Table 21: *Query weighting and shortcuts.* Percent ParentAcc. Query mean is the declared primary endpoint; object macro first averages within each of 931 objects, while category macro first pools queries within each of 42 categories. Learned rows average three fitted seeds before the macro calculation. Object/category macros are descriptive alternatives, not replacements selected on the target.

The largest-ID rule is correct on 2,547 of 2,824 represented-parent queries (90.19%). The patch-count rule is correct on 2,382 (84.35%) and exceeds Grouped by $+13.80$ points [95% paired object-bootstrap interval $+11.10, +16.68$] and Dense by $+11.11$ [$+8.24, +14.13$]. Intervals resample objects; comparisons involving Grouped additionally condition on its three fitted runs, while patch-count and Dense are fixed rules. Choosing the *smallest* eligible component ID instead reaches only 5.68% on all queries. Grouped seed 0 has no exact top-score ties and selects a largest-patch candidate on 73.94% of queries with candidates. Patch-count ties occur on 111 queries and use the same smallest-ID tie rule. Across candidate pairs of unequal size, the larger candidate has the larger Grouped score in 46,463 cases versus 16,992 in the opposite direction. This is an association with candidate size, not proof that the network explicitly uses patch count.

The evaluation bank's component-ID convention and sampled patch counts therefore carry unusually strong parent information. Appendix Section D.2 measures frozen identifier relabeling and equal-pair-score sampling directly. Neither analysis establishes why the ID convention arose or whether it persists in another dataset; both constrain what can be inferred from this bank's aggregate accuracy.

Post-hoc stratum	Queries	Grouped	Dense	Most patches	Largest ID
All queries	4,069	44.74	47.43	58.54	62.60
Gold-pair covered	2,824	64.47	68.34	84.35	90.19
Parent not largest ID	277	36.82	71.48	46.57	0
Parent not patch-count choice	442	19.98	54.30	0	66.52

Table 22: *Label-defined stress subsets.* Percent ParentAcc. The latter two subsets are defined using released parent labels *after* the overall analysis and are exploratory diagnostics, not source-selected deployment policies or new primary endpoints.

Category	Queries	Category	Queries
Box	49	Mouse	45
Bucket	4	Oven	105
Camera	249	Pen	2
Cart	153	Phone	169
Chair	14	Pliers	10
Clock	19	Printer	290
CoffeeMachine	305	Refrigerator	59
Dishwasher	15	Remote	615
Display	3	Safe	142
Door	40	Stapler	23
Eyeglasses	130	StorageFurniture	478
Fan	17	Suitcase	66
Faucet	46	Switch	86
FoldingChair	9	Table	104
Globe	9	Toaster	89
Kettle	16	Toilet	149
Keyboard	53	TrashCan	96
KitchenPot	5	USB	2
Knife	59	WashingMachine	106
Lamp	24	Window	130
Lighter	54	Microwave	30

Table 23: *Per-category query counts* in the fixed PartNet-Mobility bank. Categories contribute unequally to query-weighted ParentAcc; the macro analysis above gives each category equal weight after pooling its queries.

F Reproducibility and Assets

F.1 Result Identity and Numerical Recalculation

The released anonymous audit joins the fixed 4,069 query IDs, object IDs, gold parents, candidate IDs, saved outcomes, selected-checkpoint hashes and source-summary hashes. The checked primary registry generates Table 5 and the key fixed-candidate contrast; Figure 4 reads the same registry. The current-result audit checks checkpoint identity, table values and candidate-decomposition sums conditional on these frozen records. A standard-library bootstrap recomputes 60 paired intervals using 10,000 whole-object draws each. The larger numerical bundle also retains earlier project records for traceability; those are not additional evidence for the present claims. It does not contain third-party meshes, all preprocessing outputs or model weights, so it cannot independently rerun training or neural inference.

The candidate decomposition in Eq. (38) can be reused with two sampled score tables, their candidate IDs, the gold parent, object ID and fitted seed for each query. It outputs the fixed-original-candidate score change, fixed-new-score added-candidate change and newly covered contribution, each in percentage points of the full query denominator with object-cluster intervals. The audit verifies that the three terms recover the total accuracy change before rounding, that expanding a candidate superset at fixed new scores cannot improve an already covered Top-1 choice under consistent ties, and that the area rule has zero fixed-candidate score change. These are checks of a specified evaluation path, not causal identification of separate sampling mechanisms.

F.2 Assets, Limits and Release Scope

This study evaluates attachment-parent ranking on assembled meshes; a predicted parent is not a certified physical joint. The PartNet target bank has been inspected repeatedly. Exact source/target ID overlap was checked, but geometric near-duplicates were not audited. No independent Fusion

target validation was executed; its undirected adjacency endpoint would require a separate protocol before comparison with unique directed ParentAcc.

Table 24: *Third-party assets in this attachment-ranking study.* Raw meshes and annotations are not redistributed in the compact audit; users obtain them under upstream access terms.

Asset	Role	Release handling
ShapeNet (Chang et al., 2015)	Geometry pretraining, proxy targets and source checkpoint selection	Split IDs and configuration are documented; raw CAD assets require upstream access
PartNet-Mobility (Xiang et al., 2020)	Frozen independent-parent queries and supplied component mapping	Deidentified numerical outcomes are shared; meshes and released annotations remain upstream

Reproduction from raw assets requires the documented preprocessing and local-geometry initialization, the primary matched configuration in Table 8, frozen scoring of the same directed patch pairs, component aggregation, and paired evaluation on the fixed query set. The local training and inference scripts are retained in the project record; the compact anonymous upload currently supports numerical recalculation, not end-to-end raw-asset reproduction. Implementation and geometry evaluators are intended for release after de-anonymization subject to third-party licenses.