SIGMA: INTERACTION-CONDITIONED GEOMETRY TOKENIZATION FOR CROSS-SHAPE HUMANOID MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humanoid interaction with diverse objects depends on local surface structures that determine feasible contacts and motion strategies. Yet representing such geometry for policy learning remains challenging: compact global descriptors such as bounding boxes discard interaction-relevant surface details, while raw point-cloud inputs preserve fine-grained geometry but leave local surface structure implicit, increasing the burden on the policy to learn transferable control across shapes. To address this challenge, we propose **SIGMA** (**S**tructured **I**nteraction **G**eometry for humanoid **MA**nipulation), a geometry interface combining Structured Surface Attributes (SSA) with Interaction-Conditioned Geometry Compression (ICGC). SSA organizes object geometry into fixed-size local surface elements that encode position, normal, and represented area. ICGC aggregates these elements according to task compatibility, humanoid body–surface proximity, and surface-area priors, producing a geometry token that adapts to the current interaction state. We integrate SIGMA into a unified multi-task humanoid controller and evaluate frozen-policy transfer without target-object adaptation on objects with identical global extents but different local structures, unseen local geometric perturbations, and diverse furniture geometries. At 50k iterations, SIGMA achieves 58.33% average full-task success across the three evaluation settings, compared with 22.67% for PointNet and 6.76% for AABB-shape-token. On three held-out Sit object groups, SIGMA averages 89.03% target-entry success, compared with 82.52% for PointNet and 61.00% for AABB-shape-token.

1 INTRODUCTION

Recent advances in embodied intelligence have enabled humanoid robots to acquire increasingly diverse physical skills, including object manipulation, whole–body coordination, and human–scene interaction (Tessler et al., 2024; He et al., 2025; Pan et al., 2025; Xiao et al., 2024). As humanoid controllers evolve from task-specific policies toward unified frameworks that support multiple interaction behaviors, designing representations that generalize across diverse object shapes and interaction configurations becomes a central challenge (Tessler et al., 2024; Pan et al., 2025; Xiao et al., 2024). A key factor limiting cross–shape generalization is how object geometry is represented for physical interaction (Qin et al., 2023; Cho et al., 2024; Ze et al., 2024). Unlike conventional perception tasks that primarily require object recognition or coarse spatial understanding, humanoid interaction relies on local surface structures that determine feasible contacts and motion strategies (Xiao et al., 2024; Cho et al., 2024; Wang et al., 2025b). For example, handles provide grasping opportunities, openings constrain insertion behaviors, and narrow supporting regions influence whole–body balance. Although these structures may occupy only a small fraction of an object surface, they can fundamentally determine the success of an interaction, making them difficult to preserve under compact geometric representations dominated by global shape. Therefore, effective representations for humanoid control must shift from exhaustive reconstruction to targeted local precision—faithfully capturing sparse, affordance–critical structures while remaining compact enough for efficient policy learning.

Existing geometric representations face a fundamental dilemma: the cues dictating physical interaction reside in subtle, contact–critical surface geometries (Qin et al., 2023; Cho et al., 2024; Ze et al., 2024), whereas reactive controllers require low–dimensional, compact states for stable closed–loop

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

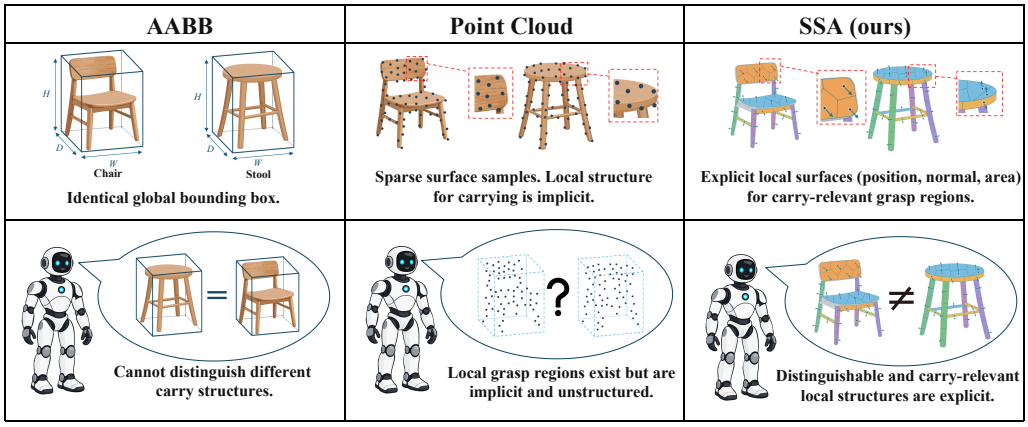


Figure 1: Comparison of geometry representations for humanoid interaction. AABBs capture only global extent, while point clouds preserve surface samples but leave local surface structure implicit. Our Structured Surface Attributes (SSA) explicitly encode surface position, normal, and area, providing compact interaction-relevant geometry.

execution (Kumar et al., 2021; Lee et al., 2020). Compact geometric descriptors have been widely adopted in physics–based HSI to provide efficient scene and interaction observations (Xiao et al., 2024; Wang et al., 2025a). Such compact representations, including 3D bounding–box descriptors, efficiently encode global object extent but inevitably collapse local surface variations, even when these variations are precisely what determine feasible contacts and motion strategies. This limitation becomes particularly problematic when different objects share similar global geometry but contain distinct local structures that determine feasible interactions. As shown in Figure 1, objects with different handles, openings, or supporting regions can be indistinguishable under bounding–box representations despite requiring different control strategies. Point-based representations alleviate this ambiguity by preserving surface samples (Qin et al., 2023; Cho et al., 2024; Ze et al., 2024), but represent geometry as unordered observations (Qi et al., 2017a), leaving local orientation, surface extent, and structural organization implicit. Consequently, forcing the controller to simultaneously infer local surface structure from unorganized samples and coordinate whole-body dynamics severely overburdens policy optimization, leading to prohibitive sample complexity and brittle generalization across novel shapes (Qin et al., 2023; Cho et al., 2024; Pan et al., 2025).

This bottleneck becomes more challenging because geometric relevance changes with the interaction context. Recent contact-conditioned HSI and manipulation methods already suggest that the useful part of an object's geometry depends on the current interaction objective and the configuration of the acting body (Xiao et al., 2024; Cho et al., 2024).

The same object can therefore require different surface information under different tasks and body configurations: a region useful for grasping or carrying may not be important for climbing, while a small supporting surface may become critical for maintaining balance or establishing stable contact. Thus, the challenge is not merely to describe object shape, but to construct a compact geometry interface that can preserve fine local structure under a fixed representation budget while adapting its emphasis to the current interaction state.

To address this challenge, we propose SIGMA (Structured Interaction Geometry for humanoid MAnipulation), a geometry interface that combines Structured Surface Attributes (SSA) with Interaction-Conditioned Geometry Compression (ICGC). SSA organizes object surfaces into a fixed number of local surface elements, where each element explicitly encodes its 3D position, surface orientation, and represented area. By transforming surface geometry from implicit samples into structured local elements, SSA provides the controller with explicit access to interaction-relevant geometric properties while maintaining a compact representation suitable for control.

Building upon SSA, we further propose **Interaction-Conditioned Geometry Compression (ICGC)** to adaptively compress structured surface geometry into a fixed-dimensional geometry token. ICGC determines the contribution of each local surface element according to task compatibility, humanoid body–surface proximity, and surface-area priors, allowing the geometry token to preserve interaction-

relevant structures under different tasks and humanoid configurations. As a result, the geometry token dynamically adapts to different tasks and body configurations, enabling the policy to preserve the surface structures that are most relevant to the ongoing interaction.

We integrate SIGMA into a unified multi-task humanoid controller and evaluate frozen-policy transfer across unseen surface configurations, local geometric perturbations, and diverse furniture objects. Our evaluation includes controlled geometry variations where objects share identical global spatial extents but differ in interaction-relevant surface structures, directly examining whether geometric representations preserve the information required for cross-shape control. Experimental results show that structured surface representations substantially improve generalization compared with coarse global geometry, while interaction-conditioned aggregation further improves performance over static surface pooling under the same geometric representation. These results demonstrate that adaptive geometry tokenization provides an effective interface between object geometry and embodied control. We additionally evaluate target-entry success on held-out Sit objects, extending the cross-shape comparison to a second interaction task.

Our contributions are summarized as follows:

- We identify geometry tokenization as an interface problem for unified humanoid control, showing that compact geometric representations must preserve interaction-relevant surface structures rather than only global shape statistics to enable cross-shape generalization.
- We introduce **Structured Surface Attributes (SSA)**, a compact surface-centric representation that turns surface samples into geometric elements with explicit position, orientation, and represented area, enabling policies to exploit fine-grained interaction geometry.
- We propose **Interaction-Conditioned Geometry Compression (ICGC)**, which dynamically aggregates structured surface elements according to task requirements and humanoid body configurations, producing adaptive geometry tokens for different interaction contexts.
- We integrate SSA and ICGC as SIGMA in a unified multi-task humanoid controller and demonstrate improved generalization across unseen surface configurations, local geometric variations, and diverse object scenarios.

2 Related Work

Physics-based humanoid interaction. DeepMimic and AMP learn physics-based skills from motion examples (Peng et al., 2018; 2021). InterPhys and UniHSI extend control to scene interactions through scene-conditioned imitation and contact prompts (Hassan et al., 2023; Xiao et al., 2024), while MaskedMimic and HOVER unify diverse control modes through motion inpainting and policy distillation (Tessler et al., 2024; He et al., 2025). SkillMimic and InterMimic learn dynamic object interactions from demonstrations (Wang et al., 2025b; Xu et al., 2025). We build on TokenHSI's task-tokenized controller (Pan et al., 2025) and study compact geometry tokens that preserve local interaction structure for frozen-policy transfer to unseen shapes.

Geometric representations for manipulation. PointNet, PointNet++, and Point Transformer develop point-cloud feature learning (Qi et al., 2017a;b; Zhao et al., 2021), while BPS encodes distances to fixed spatial probes (Prokudin et al., 2019). Geometry supports grasping in Dex-Net 2.0, DexPoint, and UniDexGrasp++ (Mahler et al., 2017; Qin et al., 2023; Wan et al., 2023), and visuomotor imitation in 3D Diffusion Policy (Ze et al., 2024). SSA associates surface elements with position, normal, and represented area before compression. Our AABB comparator encodes bounding-box corners, not BPS surface distances.

Contact- and task-conditioned geometry. Where2Act predicts local actionability, and CORN learns contact-based object representations (Mo et al., 2021; Cho et al., 2024). PerAct and RVT condition manipulation on task goals using voxelized observations and rendered views, respectively (Shridhar et al., 2023; Goyal et al., 2023). For whole-body control, ICGC reweights surface descriptors using task compatibility, body–surface proximity, and represented area. The comparison with a separately trained area-pooling baseline isolates the contribution of this conditioning.

Set aggregation and attention bottlenecks. Deep Sets formalizes permutation-invariant aggregation (Zaheer et al., 2017); Transformer, Set Transformer, and Perceiver develop content-dependent attention, learned-query pooling, and latent bottlenecks (Vaswani et al., 2017; Lee et al., 2019; Jaegle et al., 2021). ICGC draws on these principles with task queries, state-dependent body–surface distances, and an area prior. Our contribution is the surface-structured, interaction-conditioned geometry interface and its evaluation in cross-shape humanoid control, rather than a new attention operator.

3 METHOD

SIGMA combines structured surface representation and interaction-conditioned aggregation to produce a fixed-dimensional geometry token from an object mesh $\mathcal{S}$ and the interaction context. Structured Surface Attributes (SSA) represent the object surface as a fixed set of elements, each characterized by its 3D position, surface normal, and represented area. At each time step, the position and normal of each SSA element are expressed in the humanoid root-relative heading frame and encoded by a shared surface encoder, yielding one feature for each surface element.

Interaction-Conditioned Geometry Compression (ICGC) then aggregates these surface features using represented surface fraction, compatibility with the active task, and proximity to task-associated humanoid body parts, producing a geometry token $\mathbf{z}_t \in \mathbb{R}^d$. The resulting token is supplied to the unified controller as a dedicated input alongside the body and task representations.

3.1 STRUCTURED SURFACE ATTRIBUTES (SSA)

SSA describes an object mesh using M local surface regions. We summarize each region by its representative position, surface normal, and fraction of the object surface it represents; this summary is an *SSA element*. For a mesh $\mathcal{S}$, the object-frame representation is

$$\mathcal{E}^o(\mathcal{S}) = \{(\mathbf{p}_m^o, \mathbf{n}_m^o, a_m)\}_{m=1}^M, \qquad \sum_{m=1}^M a_m = 1, \tag{1}$$

where $\mathbf{p}_m^o \in \mathbb{R}^3$ is the representative position of region m, $\mathbf{n}_m^o \in \mathbb{R}^3$ is its surface normal, and a_m is its represented surface fraction.

To construct the SSA elements, we first draw N samples uniformly over the mesh surface area, denoted by $\{(\mathbf{u}_i, \boldsymbol{\nu}_i)\}_{i=1}^N$, where $\mathbf{u}_i \in \mathbb{R}^3$ denotes the sampled surface position and $\boldsymbol{\nu}_i \in \mathbb{R}^3$ its unit surface normal. We then select M anchors to represent surface regions with variation in both spatial position and surface orientation. Spatial proximity alone does not distinguish nearby surfaces with different orientations, such as opposite sides of a thin structure. We therefore define the joint position–orientation distance

$$\delta(i,j) = \frac{\|\mathbf{u}_i - \mathbf{u}_j\|_2^2}{D^2} + \lambda\left(1 - \boldsymbol{\nu}_i^\top \boldsymbol{\nu}_j\right), \tag{2}$$

where D is the diagonal length of the mesh's axis-aligned bounding box, which normalizes the spatial term across object scales, and λ controls the relative contribution of surface orientation. Farthest-point sampling under Eq. 2 is then used to select the M surface anchors.

The selected anchors partition the dense surface samples into M local surface regions. Let k_m denote the sample index of the m-th anchor. Each sample is assigned to its nearest anchor under the same position–orientation distance, yielding $\mathcal{C}_m = \{i \mid m = \arg\min_{\ell \in \{1,\dots,M\}} \delta(i, k_\ell)\}$, where $\mathcal{C}_m$ denotes the set of samples associated with the m-th surface region. Each anchor sample is assigned to its corresponding region, ensuring that every $\mathcal{C}_m$ is nonempty. For each region $\mathcal{C}_m$, we instantiate the corresponding SSA element in Eq. 1 as

$$\mathbf{p}_m^o = \mathbf{u}_{k_m}, \qquad \mathbf{n}_m^o = \frac{\sum_{i \in \mathcal{C}_m} \boldsymbol{\nu}_i}{\left\|\sum_{i \in \mathcal{C}_m} \boldsymbol{\nu}_i\right\|_2}, \qquad a_m = \frac{|\mathcal{C}_m|}{N}. \tag{3}$$

The selected anchor provides the representative position of the region, while its surface orientation is obtained by aggregating the sample normals within $\mathcal{C}_m$. If the summed normal has near-zero magnitude, we use the selected anchor normal $\boldsymbol{\nu}_{k_m}$ instead. Since the dense samples are drawn uniformly with respect to surface area, a_m estimates the fraction of the object surface represented by the region. This construction yields the object-frame SSA representation $\mathcal{E}^o(\mathcal{S})$ defined in Eq. 1.

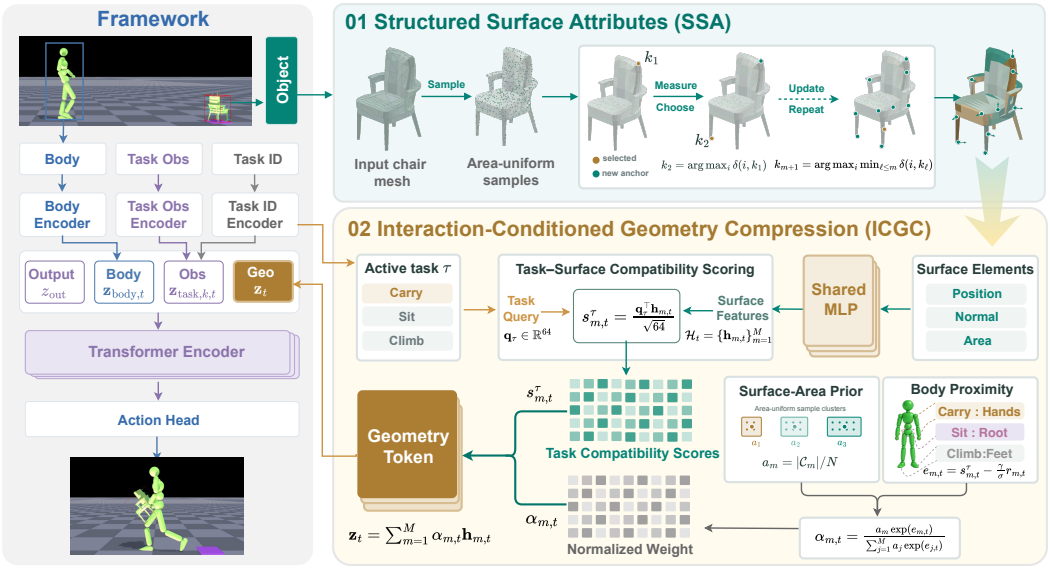


Figure 2: Overview of the proposed geometry interface. SSA converts object meshes into structured local surface elements with explicit position, surface orientation, and represented surface fraction. ICGC adaptively aggregates these elements according to task compatibility and humanoid body–surface proximity, using the represented surface fraction as an area-based reference weight.

Interaction-relative coordinates. We transform the object-frame SSA elements into the root-relative heading frame, expressing geometry relative to the humanoid root position and heading. Let $\mathbf{R}_{o,t}$ denote the object rotation, $\mathbf{t}_{o,t}$ the object translation, $\mathbf{t}_{r,t}$ the humanoid root position, and $\mathbf{R}_{h,t}$ the humanoid heading rotation. The position and normal of SSA elements are transformed as

$$\mathbf{p}_{m,t} = \mathbf{R}_{h,t}^{\top}\left(\mathbf{R}_{o,t}\mathbf{p}_m^o + \mathbf{t}_{o,t} - \mathbf{t}_{r,t}\right), \qquad \mathbf{n}_{m,t} = \mathbf{R}_{h,t}^{\top}\mathbf{R}_{o,t}\mathbf{n}_m^o. \tag{4}$$

The represented surface fraction a_m remains unchanged under this rigid transformation.

Local surface encoding. The root-relative SSA positions jointly describe the object's relative position to the humanoid and the spatial arrangement of its surface elements. We compute the area-weighted SSA centroid

$$\mathbf{c}_t = \sum_{m=1}^{M} a_m \mathbf{p}_{m,t}, \tag{5}$$

which provides an object-level position reference in the humanoid root-relative frame. The centered coordinate $\mathbf{p}_{m,t} - \mathbf{c}_t$ then describes the position of each surface element relative to this reference.

For each SSA element, we concatenate its centered position, surface normal, represented surface fraction, and object-level position reference, and map the resulting geometric descriptor to a d-dimensional feature using a shared encoder ϕ_θ,

$$\mathbf{h}_{m,t} = \phi_\theta\left(\left[\frac{\mathbf{p}_{m,t} - \mathbf{c}_t}{s_p}; \mathbf{n}_{m,t}; s_a a_m; \frac{\mathbf{c}_t}{s_c}\right]\right) \in \mathbb{R}^d, \tag{6}$$

where $[\cdot;\cdot]$ denotes concatenation, s_p and s_c are fixed normalization scales for the centered and object-level positions, respectively, and s_a is a fixed scaling factor for the represented surface fraction. The resulting features are collected as $\mathcal{H}_t = \{\mathbf{h}_{m,t}\}_{m=1}^{M}$, preserving the element-wise structure for subsequent interaction-conditioned aggregation.

3.2 Interaction-Conditioned Geometry Compression

Given the element-wise surface features $\mathcal{H}_t$, ICGC produces a d-dimensional geometry token by assigning an aggregation weight to each SSA element. The weighting combines the represented surface fraction with interaction-dependent task compatibility and body–surface proximity. For each task τ, we learn a task query $\mathbf{q}_\tau \in \mathbb{R}^d$. We express task-associated body keypoints in the same root-relative heading frame as the SSA elements: a world-frame keypoint $\mathbf{b}^w_{k,t}$ becomes $\mathbf{b}_{k,t} = \mathbf{R}^\top_{h,t}\left(\mathbf{b}^w_{k,t} - \mathbf{t}_{r,t}\right)$. Let $\mathcal{B}_{\tau,t}$ contain the left and right hands for Carry, the left and right feet for Climb, and the humanoid root for Sit. We define task compatibility, nearest-body distance, and their combined interaction score as

$$s^\tau_{m,t} = \frac{\mathbf{q}^\top_\tau \mathbf{h}_{m,t}}{\sqrt{d}}, \quad r_{m,t} = \min_{\mathbf{b}\in\mathcal{B}_{\tau,t}} \|\mathbf{p}_{m,t} - \mathbf{b}\|_2, \quad e_{m,t} = s^\tau_{m,t} - \frac{\gamma}{\sigma} r_{m,t}. \tag{7}$$

The task query is shared across states of the same task, while $\mathbf{h}_{m,t}$ varies with the object geometry and relative configuration. The proximity term uses the representative position $\mathbf{p}_{m,t}$ of each SSA element; γ controls its strength and σ its spatial scale.

The represented surface fraction a_m serves as the reference weight for aggregation. Normalizing $a_m \exp(e_{m,t})$ across surface elements gives the weights and resulting geometry token:

$$\alpha_{m,t} = \frac{a_m \exp(e_{m,t})}{\sum_{j=1}^M a_j \exp(e_{j,t})}, \quad \mathbf{z}_t = \sum_{m=1}^M \alpha_{m,t}\mathbf{h}_{m,t}. \tag{8}$$

The resulting weight $\alpha_{m,t}$ reweights the represented surface fraction according to task compatibility and the current body–surface configuration. A variational interpretation of this weighting rule is provided in Appendix A. Because the weights depend on the active task and humanoid configuration, the aggregation of the same object surface can vary across interaction contexts. When both task compatibility and body–surface proximity are omitted, $e_{m,t} = 0$ and the weights reduce to $\alpha_{m,t} = a_m$, recovering area-weighted aggregation as the context-independent case.

3.3 Unified Policy Integration and Training

We integrate the geometry token into the unified controller as a dedicated input token, alongside separate tokens for humanoid state and task-specific observations. Let $\mathbf{z}_{\mathrm{body},t}$ denote the body token encoding humanoid proprioception and $\mathbf{z}_{\mathrm{task},k,t}$ the observation token for task k. Together with the geometry token $\mathbf{z}_t$, the unified Transformer receives

$$\mathbf{X}_t = \left[\mathbf{z}_{\mathrm{out}}, \mathbf{z}_{\mathrm{body},t}, \mathbf{z}_t, \mathbf{z}_{\mathrm{task},1,t}, \ldots, \mathbf{z}_{\mathrm{task},T,t}\right], \tag{9}$$

where all tokens are embedded in $\mathbb{R}^d$, $\mathbf{z}_{\mathrm{out}}$ is a learned output token, and $T = 4$ corresponds to Trajectory Tracking (Traj), Sit, Carry, and Climb. The output and body tokens remain unmasked, while task tokens corresponding to inactive tasks are masked. The geometry token remains unmasked for object-interaction tasks with valid SSA elements and is masked for Traj.

This tokenization assigns humanoid state, task-specific observations, and interaction-conditioned geometry to dedicated token positions within the shared Transformer. The contextualized representation corresponding to $\mathbf{z}_{\mathrm{out}}$ is passed to the action head to produce the policy action.

The surface encoder, task queries, and unified policy are trained jointly from scratch using PPO with AMP. Object-frame SSA elements are precomputed and cached, while their root-relative coordinates, surface features, and interaction-conditioned aggregation weights are computed online.

4 Experiments

Our experiments evaluate two central questions. First, we examine whether structured local surface representation improves frozen-policy transfer across object shapes compared with coarse global geometry and raw surface points. Second, while holding the SSA representation design constant, we examine whether interaction-conditioned aggregation improves over static area-weighted pooling. We further examine the consistency of these effects across held-out objects and initial configurations, and assess multi-task compatibility and computational cost within the unified controller. We also evaluate held-out Sit objects to test whether the geometry interfaces transfer beyond Carry.

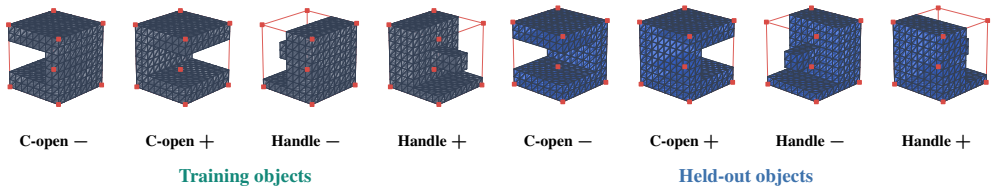


Figure 3: Construction of the Alias benchmark. Objects share the same global extent, mass, and inertia but differ in local interaction geometry. *C-open* and *Handle* denote structure types, while − and + indicate opposite orientations or configurations. Train and held-out objects vary these local structures within the same global spatial envelope.

4.1 Evaluation Protocol

Training and evaluation suites. All policies are trained from scratch on Trajectory Tracking (Traj), Sit, Carry, and Climb with identical non-geometric inputs, controller architecture, rewards, and optimization budget. We evaluate cross-shape Carry transfer using frozen policies without target-object adaptation, reporting 50k-iteration results unless stated otherwise. We use three held-out evaluation suites. *Alias* isolates local geometry: objects share an axis-aligned bounding box, mass, and inertia but differ in openings and handles (Figure 3). *Notch* tests unseen local surface perturbations. *Furniture* tests broader transfer on 48 assets from 3D-FUTURE and one YCB foam brick, without Furniture-specific Carry training. Furniture source models are disjoint from the Sit and Geometry-Alias Carry training splits; one overlaps with Climb, and Appendix E reports results excluding it. Evaluation sampling and pairing details are in Appendix B.

Geometry interfaces and metric. We compare four geometry interfaces: *AABB-shape-token* encodes eight box corners, and *PointNet* encodes sampled surface points. *SSA + Area Pooling* and *SIGMA* share the SSA representation and surface encoder; the former uses fixed represented-area weights, while the latter uses ICGC interaction-conditioned weights. Our primary metric is object-macro full-task success: an episode must complete pickup, transport, and stable goal placement. We average episode outcomes within each object, then average equally across objects and report the mean and sample standard deviation across training seeds. Exact success thresholds and statistical procedures are in Appendix B.

Held-out Sit evaluation. We evaluate three object groups absent from training: StraightChair, Stool, and Coffee Table. The frozen 50k policies use paired initial states and equal evaluation budgets across methods. An episode counts as a Sit target entry if the humanoid root comes within 0.20 m of the Sit target in 3D at any simulation step; we average these episode outcomes within each group and report mean and sample SD across three training seeds.

4.2 Structured Surface Representation for Cross-Shape Generalization

Table 1 compares SSA + Area Pooling with bounding-box and raw-point geometry without using ICGC. At 50k iterations, it exceeds AABB-shape-token by 61.16, 60.51, and 20.53 points and PointNet by 46.07, 34.99, and 13.39 points on Alias, Notch, and Furniture, respectively. PointNet also shows substantial between-seed variation on Alias and Notch (SD 28.89 and 33.11), supporting the benefit of structured surface attributes in this training setup.

4.3 Effect of ICGC

SIGMA and SSA + Area Pooling share the representation and encoder but use interaction-conditioned and fixed area weights, respectively. Percentage-point differences in this subsection use unrounded rates. At 50k iterations, SIGMA leads by 1.71, 5.08, and 5.73 points on Alias, Notch, and Furniture (Table 1). To probe its task and body signals in the same frozen policies, *No Task* removes task

Table 1: Cross-shape object-macro full-task success (%) at 50k: mean ± sample SD over three frozen policies. SSA + Area Pooling shares SIGMA's SSA encoder but uses fixed area weights.

Method	Alias	Notch	Furniture
AABB-shape-token	11.37 ± 11.35	7.31 ± 6.35	1.59 ± 1.55
PointNet	26.46 ± 28.89	32.83 ± 33.11	8.73 ± 10.76
SSA + Area Pooling	72.53 ± 3.22	67.82 ± 1.50	22.12 ± 10.39
SIGMA	$\mathbf{74.24 \pm 7.89}$	$\mathbf{72.89 \pm 13.86}$	$\mathbf{27.85 \pm 5.44}$

Table 2: Frozen-policy aggregation interventions: object-macro full-task success (%), mean ± SD over three policies. Full is a separate reevaluation of SIGMA checkpoints and serves as the paired reference; its rates differ slightly from Table 1.

Setting	Alias	Notch	Furniture
Full	$\mathbf{74.80 \pm 7.65}$	$\mathbf{72.89 \pm 13.86}$	$\mathbf{27.61 \pm 5.31}$
No Task	72.55 ± 8.11	70.81 ± 15.29	26.86 ± 5.18
No Body	70.42 ± 9.99	71.51 ± 15.43	25.89 ± 5.79
Area Only	69.42 ± 8.65	68.32 ± 15.62	24.46 ± 5.24

compatibility, *No Body* removes body proximity, and *Area Only* retains only represented-area weights. Each intervention lowers mean success on all three suites (Table 2); Area Only yields the largest drops, at 5.38, 4.58, and 3.15 points on Alias, Notch, and Furniture. These interventions show that the trained policy depends on both interaction signals.

To test whether the weights select policy-relevant surface information, we remove seven highest-weight, lowest-weight, or randomly selected SSA elements and renormalize the remaining weights, or replace Carry hand positions with feet in the proximity term. All trained parameters remain fixed. Top-7 removal hurts success much more than Bottom-7 or Random-7 removal across all suites; replacing hands with feet also causes large drops (Table 3). These effects associate high weights with policy-relevant geometry and show sensitivity to body–surface context.

Table 3: Frozen-policy surface-relevance interventions: changes in object-macro full-task success (points) relative to unmodified SIGMA.

Intervention	Alias	Notch	Furniture
Top-7 removal	-41.21	-25.59	-13.16
Bottom-7 removal	-0.39	$+3.32$	$+0.51$
Random-7 removal	-3.32	-4.10	-3.78
Hands→Feet	-28.52	-30.47	-11.56

4.4 Robustness Analysis of Cross-Shape Generalization

Held-out geometry breadth. Figure 4 compares seed-averaged success on each held-out geometry. SIGMA leads on all eight Notch objects, on 32 of 49 Furniture objects (12 losses, five ties), and on two of four Alias configurations. Thus, its Notch and Furniture gains extend across many objects.

Initial-state variation. We re-evaluate three policies per method at 50k on a second sample of initial states for the same 49 Furniture objects (Table 4). SIGMA's full-task advantage persists ($+5.73$ original, $+3.90$ new), including under stable termination ($+5.47$, $+3.14$ points). Its humanoid-fall rate is also higher in both samples ($+5.02$, $+4.88$ points). Checkpoint-level comparisons are in Appendix D.2.

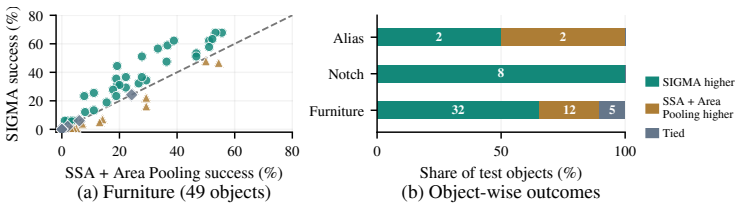


Figure 4: Object-wise SIGMA versus SSA + Area Pooling at 50k.

4.5 Interface-Level Zero-Shot Transfer on Sit

Table 5 evaluates interface-level zero-shot shape transfer on Sit: policies jointly trained on Sit and the other tasks are frozen and tested on three Sit object groups absent from training. This probes whether SIGMA's complete SSA+ICGC geometry interface transfers beyond Carry to a distinct whole-body support interaction. SIGMA averages 89.03% target-entry success across the three groups, versus 82.52% for point-cloud PointNet and 61.00% for AABB-shape-token. It leads PointNet by 11.29 points on Stool and 12.33 points on Coffee Table, while PointNet leads by 4.08 points on StraightChair. This comparison tests the complete interface rather than ICGC in isolation.

Table 4: Furniture rate gaps (SIGMA vs. SSA + Area Pooling, points); positive values favor SIGMA except falls.

Metric	Original sample	New sample
Full-task success	+5.73	+3.90
Full-task success with stable termination	+5.47	+3.14
Humanoid-fall rate	+5.02	+4.88

Table 5: Interface-level zero-shot transfer to unseen Sit objects at 50k: target-entry success (%), mean ± sample SD across three seeds.

Method	StraightChair	Stool	Coffee Table
AABB-shape-token	71.27 ± 18.74	61.11 ± 23.59	50.61 ± 36.36
PointNet	$\mathbf{93.75 \pm 4.01}$	70.14 ± 5.53	83.68 ± 10.30
SIGMA	89.67 ± 3.78	$\mathbf{81.42 \pm 5.06}$	$\mathbf{96.01 \pm 3.66}$

4.6 Multi-Task Compatibility and Computational Cost

Compatibility with the unified controller. We further examine whether replacing static aggregation with interaction-conditioned aggregation changes performance on the other jointly trained tasks. At 50k iterations, the area-pooling baseline and SIGMA achieve 99.67% and 99.67% on Trajectory Tracking, 99.39% and 99.41% on Sit, and 99.83% and 99.85% on Climb, respectively. Across the three non-Carry tasks, SIGMA and the area-pooling baseline differ by at most 0.02 percentage points, while their main performance difference occurs in cross-shape Carry.

Computational cost. The area-pooling baseline and SIGMA have the same actor-side parameter count of 1.263M. For a batch of 512 objects, geometry tokenization requires 1.011 ms for the area-pooling baseline and 1.120 ms for SIGMA, corresponding to an additional 0.109 ms (10.8%). These timings measure geometry tokenization only and exclude the policy Transformer and simulator. Under this benchmark, interaction-conditioned aggregation increases geometry-tokenization latency by 10.8% without increasing the actor-side parameter count.

5 Conclusion

We presented SIGMA, a geometry interface combining Structured Surface Attributes (SSA) and Interaction-Conditioned Geometry Compression (ICGC) for cross-shape humanoid manipulation. By organizing surface geometry into structured local elements and adaptively aggregating them via task and body-proximity cues, SIGMA preserves affordance-critical structures within a compact representation budget. Frozen-policy evaluations demonstrate substantial gains over bounding-box and point-cloud baselines, as well as consistent improvements over static pooling. Crucially, these benefits generalize across distinct physical interaction regimes, spanning dual-arm transport (Carry) and whole-body supporting balance (Sit). These findings highlight interaction-conditioned geometry tokenization as an effective interface for unified controllers, paving the way toward generalist humanoid interaction across diverse objects.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing parts of the manuscript, to provide feedback on experimental design and interpretation, and to check reported numerical summaries. The authors take responsibility for the final content, experimental results, and claims.

REPRODUCIBILITY STATEMENT

The primary Carry evaluation archive records checkpoint and configuration provenance, paired initial states, episode keys, and metric definitions. Appendix B documents evaluation and statistics; Appendix C reports per-seed Sit results; and Appendix G specifies implementation and training settings. Mathematical derivations appear in Appendix A.

REFERENCES

Yoonyoung Cho, Junhyek Han, Yoontae Cho, and Beomjoon Kim. CORN: Contact-based object representation for nonprehensile manipulation of general unseen objects. In *International Conference on Learning Representations*, 2024.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. RVT: Robotic view transformer for 3D object manipulation. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 694–710, 2023.

Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. doi: 10.1145/3588432.3591525.

Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, Linxi Fan, and Yuke Zhu. HOVER: Versatile neural whole-body controller for humanoid robots. In *IEEE International Conference on Robotics and Automation*, 2025. doi: 10.1109/ICRA55743.2025.11128549.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664, 2021.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Proceedings of Robotics: Science and Systems*, 2021. doi: 10.15607/RSS.2021.XVII.011.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020. doi: 10.1126/scirobotics.abc5986.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019.

Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems*, 2017.

Kaichun Mo, Leonidas J. Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2Act: From pixels to actions for articulated 3D objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6813–6823, 2021.

Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. TokenHSI: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5379–5391, 2025.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. AMP: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics*, 2021.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017a.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 2017b.

Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Hao Su, and Xiaolong Wang. DexPoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 594–605. PMLR, 2023.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-Actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 785–799, 2023.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. MaskedMimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics*, 43(6), 2024. doi: 10.1145/3687951.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3891–3902, 2023.

Wenjia Wang, Liang Pan, Zhiyang Dou, Jidong Mei, Zhouyingcheng Liao, Yuke Lou, Yifan Wu, Lei Yang, Jingbo Wang, and Taku Komura. SIMS: Simulating stylized human-scene interactions with retrieval-augmented script generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14117–14127, 2025a.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. SkillMimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17540–17549, 2025b.

Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *International Conference on Learning Representations*, 2024.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12266–12277, 2025.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3D diffusion policy: Generalizable visuomotor policy learning via simple 3D representations. In *Robotics: Science and Systems*, 2024.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021.

A Analysis of Interaction-Conditioned Aggregation

These derivations consider one forward pass over valid SSA elements with $a_m > 0$ and $\sum_m a_m = 1$. Surface-element features and task identity are held fixed when differentiating with respect to body–surface distances. For convenience, we define the aggregation logit

$$\ell_m = \log a_m + e_m,$$

such that

$$\alpha_m = \frac{\exp(\ell_m)}{\sum_j \exp(\ell_j)}.$$

Area-referenced variational form. Let Δ_M be the probability simplex and $\mathbf{a} = (a_1, \ldots, a_M)$. The normalized weights in Eq. 8 equivalently solve

$$\boldsymbol{\alpha} = \arg\max_{\mathbf{w} \in \Delta_M} \left\{ \sum_m w_m e_m - \mathrm{KL}(\mathbf{w} \| \mathbf{a}) \right\}. \tag{10}$$

Weights favor larger interaction scores while remaining referenced to the object's represented-surface-fraction distribution. For e_m defined in Eq. 7, write

$$\mathcal{L}(\mathbf{w}, \eta) = \sum_m w_m e_m - \sum_m w_m \log \frac{w_m}{a_m} + \eta \left(\sum_m w_m - 1 \right). \tag{11}$$

Stationarity gives $e_m - \log(w_m/a_m) - 1 + \eta = 0$, hence $w_m = a_m \exp(e_m + \eta - 1)$. Enforcing unit sum yields

$$w_m^* = \frac{a_m \exp(e_m)}{Z}, \qquad Z = \sum_j a_j \exp(e_j). \tag{12}$$

Equivalently, for any $\mathbf{w} \in \Delta_M$,

$$\sum_m w_m e_m - \mathrm{KL}(\mathbf{w} \| \mathbf{a}) = \log Z - \mathrm{KL}(\mathbf{w} \| \boldsymbol{\alpha}). \tag{13}$$

Nonnegativity of KL establishes the unique maximizer $\mathbf{w} = \boldsymbol{\alpha}$. This provides a pointwise interpretation of the aggregation rule as a softmax referenced by the represented surface fractions. Training still uses the original PPO+AMP objective without this expression as an auxiliary loss.

Distance sensitivity. The softmax Jacobian and the linear distance term imply

$$\frac{\partial \alpha_m}{\partial \ell_i} = \alpha_m(\mathbb{I}[m = i] - \alpha_i), \qquad \frac{\partial \alpha_m}{\partial r_i} = -\frac{\gamma}{\sigma} \alpha_m(\mathbb{I}[m = i] - \alpha_i). \tag{14}$$

Substituting into the weighted sum gives

$$\begin{aligned} \frac{\partial \mathbf{z}}{\partial r_i} &= \sum_m \mathbf{h}_m \frac{\partial \alpha_m}{\partial r_i} \\ &= -\frac{\gamma}{\sigma} \left(\alpha_i \mathbf{h}_i - \alpha_i \sum_m \alpha_m \mathbf{h}_m \right) = -\frac{\gamma}{\sigma} \alpha_i (\mathbf{h}_i - \mathbf{z}). \end{aligned} \tag{15}$$

For body keypoint $\mathbf{b}_k$, let $k(i)$ be the unique nearest keypoint to the i-th SSA element. Away from distance zero and nearest-keypoint ties,

$$\frac{\partial \mathbf{z}}{\partial \mathbf{b}_k} = -\frac{\gamma}{\sigma} \sum_{i:k(i)=k} \alpha_i (\mathbf{h}_i - \mathbf{z}) \frac{(\mathbf{b}_k - \mathbf{p}_i)^\top}{\|\mathbf{b}_k - \mathbf{p}_i\|_2}. \tag{16}$$

This $d \times 3$ Jacobian describes body-keypoint motion with the SSA-element features held fixed, including the root reference frame. Actual motion can also change those inputs. The displayed derivative applies within regions of fixed nearest-keypoint assignments and away from zero-distance configurations. The area-pooling baseline's geometry token has zero derivative with respect to these keypoints under the same fixed-input condition, although its downstream policy can still respond to body observations.

Element-order invariance. For a permutation P of SSA-element rows, the shared MLP, element-wise scores, and represented-surface-fraction weights permute together. With H the matrix of row embeddings, $H' = PH$ and $\boldsymbol{\alpha}' = P\boldsymbol{\alpha}$, so

$$\mathbf{z}' = (H')^\top \boldsymbol{\alpha}' = H^\top P^\top P\boldsymbol{\alpha} = H^\top \boldsymbol{\alpha} = \mathbf{z}. \tag{17}$$

Both aggregation rules are invariant to the ordering of SSA elements. Task and body conditioning preserve invariance to input row order while introducing sensitivity to interaction configuration.

A.1 Geometry in a Common Frame

We first characterize how the root-relative representation separates local surface structure from object placement and how it behaves under shared rigid transformations. We omit the time subscript t in this section for clarity.

Separating local structure from placement. With $\mathbf{d}_m = (\mathbf{p}_m - \mathbf{c})/s_p$ and $\bar{\mathbf{c}} = \mathbf{c}/s_c$, the coordinate features satisfy

$$\sum_m a_m \mathbf{d}_m = \mathbf{0}, \qquad \mathbf{p}_m = s_p \mathbf{d}_m + s_c \bar{\mathbf{c}}. \tag{18}$$

The first identity follows from $\mathbf{c} = \sum_m a_m \mathbf{p}_m$ and $\sum_m a_m = 1$. The second inverts the scaling and centering. Thus, the preprocessing preserves the original position information before the shared surface encoder.

Shared translation and yaw. The transform in Eq. 4 removes arbitrary shared translation and yaw. Let $\mathbf{Q}$ be a rotation about the vertical axis and $\mathbf{v}$ a translation applied to the entire configuration. The transformed quantities are

$$\begin{aligned} \mathbf{R}'_o &= \mathbf{Q}\mathbf{R}_o, \qquad \mathbf{R}'_h = \mathbf{Q}\mathbf{R}_h, \\ \mathbf{t}'_o &= \mathbf{Q}\mathbf{t}_o + \mathbf{v}, \quad \mathbf{t}'_r = \mathbf{Q}\mathbf{t}_r + \mathbf{v}, \quad (\mathbf{b}^w_k)' = \mathbf{Q}\mathbf{b}^w_k + \mathbf{v}. \end{aligned} \tag{19}$$

Substitution cancels this shared transform:

$$\begin{aligned} \mathbf{p}'_m &= (\mathbf{Q}\mathbf{R}_h)^\top \mathbf{Q}(\mathbf{R}_o \mathbf{p}^o_m + \mathbf{t}_o - \mathbf{t}_r) = \mathbf{p}_m, \\ \mathbf{n}'_m &= (\mathbf{Q}\mathbf{R}_h)^\top \mathbf{Q}\mathbf{R}_o \mathbf{n}^o_m = \mathbf{n}_m, \qquad \mathbf{b}'_k = \mathbf{b}_k. \end{aligned} \tag{20}$$

Consequently the encoder inputs, distances, and token agree. The resulting SSA coordinates, body keypoints, and aggregation distances are therefore invariant to a shared translation and yaw rotation of the complete object–humanoid configuration.

A.2 Query Response and Learning Gradients

The log-partition function. For fixed SSA features, represented surface fractions, and distances, define

$$\beta = \gamma/\sigma. \tag{21}$$

Direct differentiation yields

$$\nabla_{\mathbf{q}} \log Z = \frac{\mathbf{z}}{\sqrt{d}}, \qquad \nabla_{\mathbf{q}} \alpha_m = \frac{\alpha_m}{\sqrt{d}}(\mathbf{h}_m - \mathbf{z}), \qquad \nabla^2_{\mathbf{q}} \log Z = \frac{C_\alpha}{d}. \tag{22}$$

Here,

$$C_\alpha = \sum_m \alpha_m (\mathbf{h}_m - \mathbf{z})(\mathbf{h}_m - \mathbf{z})^\top \tag{23}$$

denotes the aggregation-weighted covariance matrix of the SSA features. Since $\nabla_{\mathbf{q}} \log Z = \mathbf{z}/\sqrt{d}$, differentiating once more gives

$$\frac{\partial \mathbf{z}}{\partial \mathbf{q}} = \frac{C_\alpha}{\sqrt{d}}. \tag{24}$$

For any vector $\mathbf{u}$,

$$\mathbf{u}^\top C_\alpha \mathbf{u} = \sum_m \alpha_m \left(\mathbf{u}^\top \mathbf{h}_m - \mathbf{u}^\top \mathbf{z}\right)^2 \geq 0. \tag{25}$$

The centered feature vectors $\mathbf{h}_m - \mathbf{z}$ have zero aggregation-weighted sum, so their span has dimension at most $M-1$. Define $D_H = \max_{i,j} \|\mathbf{h}_i - \mathbf{h}_j\|_2$. For unit $\mathbf{u}$, their scalar projections have range at most D_H. The variance of a scalar supported on $[l, u]$ is at most $(u-l)^2/4$, as follows from $\mathbb{E}[(X-l)(u-X)] \geq 0$. Therefore

$$\operatorname{rank}(C_\alpha) \leq \min(d, M-1), \qquad \left\| \frac{\partial \mathbf{z}}{\partial \mathbf{q}} \right\|_2 \leq \frac{D_H^2}{4\sqrt{d}}. \tag{26}$$

This rank statement applies to a fixed set of SSA features.

Gradients through the geometry-token pathway. Let $\mathcal{J}$ be any differentiable scalar objective downstream of the geometry token and $\mathbf{g} = \nabla_{\mathbf{z}} \mathcal{J}$. The contribution through this pathway satisfies

$$\frac{\partial \mathcal{J}}{\partial \ell_i} = \alpha_i \, \mathbf{g}^\top (\mathbf{h}_i - \mathbf{z}). \tag{27}$$

Treating the surface feature, represented surface fraction, query, and distance as intermediate variables gives

$$\begin{aligned} \nabla_{\mathbf{q}} \mathcal{J} &= \frac{C_\alpha \mathbf{g}}{\sqrt{d}}, \\ \nabla_{\mathbf{h}_i} \mathcal{J} &= \alpha_i \mathbf{g} + \frac{\alpha_i}{\sqrt{d}} \mathbf{q} \big[(\mathbf{h}_i - \mathbf{z})^\top \mathbf{g} \big]. \end{aligned} \tag{28}$$

The second term arises because the surface-element feature influences its own score as well as the pooled value. Under the baseline's fixed area-weighted aggregation, the corresponding embedding gradient is $a_i \mathbf{g}$. This gradient expression characterizes the geometry-token pathway under fixed intermediate variables.

A.3 Full Differential and Changes in Interaction Context

Write $\beta = \gamma/\sigma$. A general infinitesimal change in surface features, represented surface fractions, query, and distances gives

$$\begin{aligned} \mathrm{d}\mathbf{z} &= \underbrace{\sum_m \alpha_m \, \mathrm{d}\mathbf{h}_m}_{\text{feature change}} + \underbrace{\sum_m \alpha_m (\mathbf{h}_m - \mathbf{z}) \, \mathrm{d}\ell_m}_{\text{reweighting}}, \\ \mathrm{d}\ell_m &= \mathrm{d} \log a_m + \frac{\mathbf{q}_\tau^\top \mathrm{d}\mathbf{h}_m + \mathbf{h}_m^\top \mathrm{d}\mathbf{q}_\tau}{\sqrt{d}} - \beta \, \mathrm{d}r_m. \end{aligned} \tag{29}$$

ICGC responds to a local feature change through both the aggregated feature value and the corresponding aggregation score. Geometry perturbations may change several embeddings through the shared centroid and also change distances. Eq. 29 accounts for these coupled effects.

Writing $J_\alpha = \operatorname{diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$, the softmax differential is

$$\mathrm{d}\boldsymbol{\alpha} = J_\alpha \, \mathrm{d}\boldsymbol{\ell}, \qquad \mathrm{d}\alpha_m = \alpha_m \left(\mathrm{d}\ell_m - \sum_j \alpha_j \mathrm{d}\ell_j \right). \tag{30}$$

Applying the product rule to $\mathbf{z} = \sum_m \alpha_m \mathbf{h}_m$ proves Eq. 29. In particular, with other intermediate inputs fixed,

$$\frac{\partial \mathbf{z}}{\partial \mathbf{h}_i} = \alpha_i \left[I_d + \frac{(\mathbf{h}_i - \mathbf{z})\mathbf{q}^\top}{\sqrt{d}} \right], \qquad \mathrm{d}\mathbf{z}_{\text{static}} = \sum_m a_m \, \mathrm{d}\mathbf{h}_m + \sum_m \mathbf{h}_m \, \mathrm{d}a_m. \tag{31}$$

For the area-pooling baseline, the represented surface fractions satisfy $\sum_m a_m = 1$, and therefore $\sum_m \mathrm{d}a_m = 0$. A perturbation of raw geometry must propagate through all affected intermediates, including centering and the represented-surface-fraction feature. The embedding Jacobian alone is not that full geometric derivative.

For two interaction contexts with fixed SSA features and represented surface fractions, define $\Delta\mathbf{q} = \mathbf{q}' - \mathbf{q}$ and $\Delta r_m = r'_m - r_m$. Their weights obey the exact update

$$\alpha'_m = \frac{\alpha_m \exp(\delta_m)}{\sum_j \alpha_j \exp(\delta_j)}, \qquad \delta_m = \frac{\Delta\mathbf{q}^\top \mathbf{h}_m}{\sqrt{d}} - \beta \Delta r_m. \tag{32}$$

Taking pairwise log ratios yields

$$\log \frac{\alpha'_i}{\alpha'_j} - \log \frac{\alpha_i}{\alpha_j} = \frac{\Delta\mathbf{q}^\top (\mathbf{h}_i - \mathbf{h}_j)}{\sqrt{d}} - \beta(\Delta r_i - \Delta r_j). \tag{33}$$

Only relative score changes redistribute the aggregation weights. Adding the same constant to every logit leaves the weights unchanged. The identity separates task-query and body-distance effects algebraically. Their empirical roles are examined through the evaluation-time interventions in Sec. 4.3.

A.4 Aggregation Mass on a Surface Region

For a nonempty proper subset S of SSA elements, let $A_S = \sum_{m \in S} a_m$ denote the represented surface fraction associated with S. We define the area-weighted mean exponentiated score within S and its complement as

$$E_S = \frac{1}{A_S} \sum_{m \in S} a_m \exp(e_m), \qquad E_{\bar{S}} = \frac{1}{1 - A_S} \sum_{j \notin S} a_j \exp(e_j). \tag{34}$$

The exact total aggregation mass and a bound are

$$\begin{aligned} \alpha(S) &= \frac{A_S E_S}{A_S E_S + (1 - A_S) E_{\bar{S}}} \\ &\geq \frac{A_S \exp(\min_{m \in S} e_m)}{A_S \exp(\min_{m \in S} e_m) + (1 - A_S) \exp(\max_{j \notin S} e_j)}. \end{aligned} \tag{35}$$

Define the score margin between S and its complement as

$$\Delta_S = \min_{m \in S} e_m - \max_{j \notin S} e_j.$$

For a target aggregation mass $\rho \in (0, 1)$, a sufficient score margin is

$$\Delta_S \geq \log \frac{\rho(1 - A_S)}{(1 - \rho) A_S} \quad \Longrightarrow \quad \alpha(S) \geq \rho. \tag{36}$$

The bound characterizes the aggregation mass attainable for a surface region under a given score margin. It quantifies what the aggregation rule can represent when suitable scores are learned.

A.5 Finite Response to Body-Keypoint Motion

For fixed SSA features and task, the interaction sensitivity admits a finite bound. Let $D_H = \max_{i,j} \|\mathbf{h}_i - \mathbf{h}_j\|_2$ and $\varepsilon_B = \max_k \|\mathbf{b}'_k - \mathbf{b}_k\|_2$ for two matched body-keypoint configurations. Then

$$\|\mathbf{z}(\mathcal{B}'_\tau) - \mathbf{z}(\mathcal{B}_\tau)\|_2 \leq D_H \min\left\{1, \frac{\beta \varepsilon_B}{2}\right\}. \tag{37}$$

The proof below includes nearest-keypoint switches. The area-pooling baseline is constant under this fixed-geometry perturbation. The ICGC geometry token provides a bounded response to body-keypoint motion.

Distances remain bounded across nearest-keypoint switches. Consider matched keypoints with $\max_k \|\mathbf{b}'_k - \mathbf{b}_k\|_2 \leq \varepsilon_B$, while keeping the SSA features and task query fixed. The triangle inequality gives

$$\|\mathbf{p}_m - \mathbf{b}'_k\|_2 \leq \|\mathbf{p}_m - \mathbf{b}_k\|_2 + \varepsilon_B \quad \Longrightarrow \quad |r'_m - r_m| \leq \varepsilon_B, \tag{38}$$

where the reverse inequality follows by swapping the configurations. Hence $\delta_m = \ell'_m - \ell_m = -\beta(r'_m - r_m)$ has range at most $2\beta\varepsilon_B$.

Softmax redistributes bounded mass. Interpolate logits by $\boldsymbol{\ell}(s) = \boldsymbol{\ell} + s\boldsymbol{\delta}$, $s \in [0, 1]$. With expectation under the corresponding $\boldsymbol{\alpha}(s)$,

$$\begin{aligned}\left\| \frac{\mathrm{d}\boldsymbol{\alpha}(s)}{\mathrm{d}s} \right\|_1 &= \mathbb{E}_{\alpha(s)} \left\| \delta - \mathbb{E}_{\alpha(s)} \delta \right\| \\ &\leq \sqrt{\mathrm{Var}_{\alpha(s)}(\delta)} \leq \frac{\max_m \delta_m - \min_m \delta_m}{2}.\end{aligned} \tag{39}$$

Integration and the maximum distance between probability vectors give

$$\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\|_1 \leq \min\left\{2, \frac{\max_m \delta_m - \min_m \delta_m}{2}\right\} \leq \min\{2, \beta\varepsilon_B\}. \tag{40}$$

Finally, let $\eta = \frac{1}{2}\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\|_1$. Positive and negative changes in weights have equal total mass η, so for $\eta > 0$,

$$\mathbf{z}' - \mathbf{z} = \eta(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-), \qquad \boldsymbol{\mu}_\pm \in \mathrm{conv}\{\mathbf{h}_1, \ldots, \mathbf{h}_M\}. \tag{41}$$

The convex hull has diameter D_H, giving

$$\|\mathbf{z}' - \mathbf{z}\|_2 \leq \frac{D_H}{2}\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\|_1 \leq D_H \min\left\{1, \frac{\beta\varepsilon_B}{2}\right\}. \tag{42}$$

For $\eta = 0$ the bound is immediate. This proves Eq. 37 even when the nearest keypoint changes. The feature diameter is configuration-dependent and is not assumed uniformly bounded over all states. The bound applies to the geometry encoder under fixed SSA features and task query.

B Evaluation Details, Metrics, and Data Provenance

Training and pairing. The three multi-seed primary arms, AABB-shape-token, SSA + Area Pooling, and SIGMA, are trained from scratch on the Geometry Alias multi-task protocol without warm-starting, bridge training, or distillation. The implementation label `bps` denotes eight AABB corners. It does not refer to surface-distance BPS. The raw-episode audit covers all 45 paired evaluations of SIGMA and the area-pooling baseline at 10k, 20k, 30k, 40k, and 50k across the three evaluation suites and three training seeds. Episodes are sorted by repeat and environment ID and checked for identical object identity, evaluation seeds, initial object and humanoid poses, target, physical properties, and success thresholds. Trajectory, sitting, and climbing retain their task-specific success criteria.

Evaluation set sizes. Alias contains four held-out configurations, Notch contains eight held-out objects, and Furniture contains 49 evaluation objects: 48 assets from 3D-FUTURE and one YCB foam brick. Each policy–suite cell uses 512 parallel environments and three measured repeats after one warm-up repeat, for 1,536 measured episodes. The policy rate is 30 Hz and the primary evaluation seed is 0.

Statistical procedure. For each trained policy, episode outcomes are first averaged within each evaluation object and then averaged equally across objects. For multi-seed comparisons, the resulting object-macro scores are averaged across independently trained policies, and sample standard deviations are computed across training seeds.

Metric definitions. Full-task success requires completion of the Carry sequence through pickup, transport, and stable goal completion. Pickup requires the object to be raised at least 0.10 m above its initial height while the nearest hand remains within 0.20 m of an object-surface anchor for five consecutive policy steps. After the first pickup, transport requires the horizontal object–goal distance to decrease by at least $\max(0.50\ \mathrm{m}, 0.5d_{\mathrm{pickup}})$ while the hand remains near the object. Stable goal completion requires full 3D object–goal position error at most 0.20 m, linear speed at most 0.20 m/s, and angular speed at most 1.0 rad/s for 30 consecutive policy steps. Full-task success is the conjunction of the pickup, transport, and stable-goal flags. For the Furniture robustness analysis, we additionally require *stable termination*: the episode must end normally (the reset flag is set and the humanoid-fall termination flag is not set), the object must remain within the goal tolerance with linear and angular speeds below the evaluation thresholds at the terminal step, and the stable-goal

dwell requirement must already have been satisfied. This supplemental metric is the conjunction of `full_task_success` and `end_success`; it excludes episodes that completed the task earlier but no longer satisfy the stability conditions when the episode ends. Humanoid fall is recorded when the environment's fall detector triggers after the first two simulation steps: at least one non-foot body has a contact-force component exceeding 0.1 in magnitude, and at least one non-foot body lies below its prescribed height threshold. Normal episode timeouts are not counted as falls.

Metric consistency checks. All 9,216 terminal trace records from the new scenes agree with their episode summaries. Among the 2,215 full-success episodes, none first reaches the stable-goal flag before the transport flag. Under the additional terminal requirements used in the Furniture diagnostics, SIGMA retains higher mean full-task success than the area-pooling baseline. Additional stage-progression and terminal-metric analyses are reported in Appendix F.

Data sources. The compact evaluation archive records input hashes and source provenance for the reported results. Primary comparisons and per-object analyses are computed from the paired evaluation records, while the continuation curves and corresponding table are derived from the supplied rounded evaluation matrix, with identical duplicate entries counted once. Later aggregate records do not extend the earlier raw-episode pairing audit. Furniture scene replication and terminal-condition checks are based on the dedicated follow-up evaluations, and localized intervention results are reported separately. The held-out Sit evaluation is a separate experiment; its per-seed rates are given in Appendix C.

C Held-Out Sit Transfer

Table 6 lists the per-seed rates underlying Table 5. The three object groups were absent from training. Within each group, the methods use paired initial states and equal evaluation budgets. Target entry is recorded if the humanoid root comes within 0.20 m of the Sit target at any simulation step. The main table averages the three seed rates within each group and reports their sample SD; its three-group average gives each group equal weight.

Table 6: Held-out Sit target-entry success (%) by training seed at 50k iterations.

Method	Seed	StraightChair	Stool	Coffee Table
SIGMA	0	89.58	86.98	99.48
SIGMA	1	85.94	77.08	96.35
SIGMA	2	93.49	80.21	92.19
PointNet	0	94.79	68.49	95.57
PointNet	1	97.14	65.62	77.86
PointNet	2	89.32	76.30	77.60
AABB-shape-token	0	50.52	66.93	48.96
AABB-shape-token	1	86.98	35.16	15.10
AABB-shape-token	2	76.30	81.25	87.76

D Additional Training Results

D.1 Per-Seed Primary Results

Table 7: 50k object-macro full-task success (%). Each row represents one independently trained policy.

Method	Seed	Alias	Notch	Furniture
PointNet	0	21.88	32.29	5.34
PointNet	1	0.13	0.00	0.07
PointNet	2	57.36	66.21	20.77
SSA + Area Pooling	0	75.46	69.47	34.09
SIGMA	0	81.90	80.47	32.71
SSA + Area Pooling	1	73.05	67.45	16.69
SIGMA	1	74.67	81.32	28.87
SSA + Area Pooling	2	69.08	66.54	15.57
SIGMA	2	66.15	56.90	21.97

D.2 Training Persistence

We next examine whether SIGMA's advantage over the area-pooling baseline persists beyond the primary 50k checkpoint.

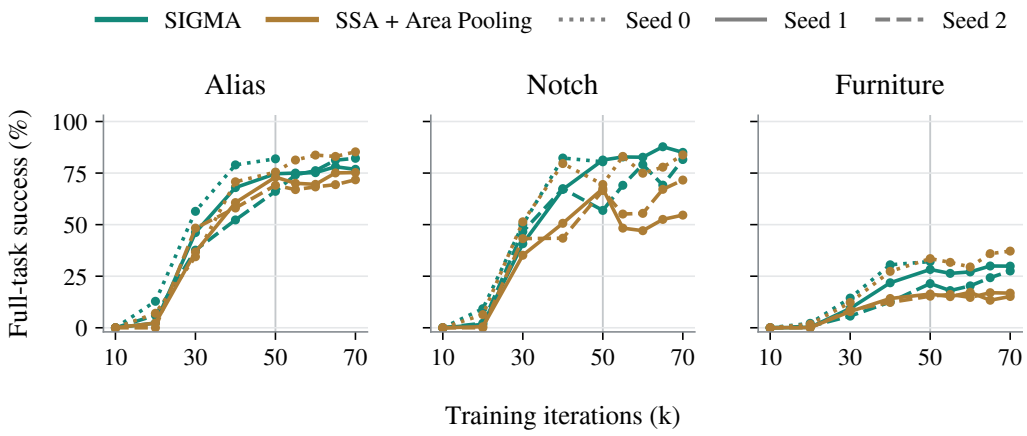


Figure 5: Episode-micro full-task success across the available SIGMA and SSA + Area Pooling checkpoints. All available runs are shown. SIGMA seed 0 ends at 50k, so comparisons from 55k to 70k use seeds 1 and 2.

For both seeds with continuation checkpoints, SIGMA remains above the area-pooling baseline on all three evaluation suites at every evaluated checkpoint from 55k to 70k. Averaging the paired seed differences at each checkpoint and then averaging across the 55k–70k continuation window yields $+6.56$, $+23.16$, and $+9.75$ percentage points on Alias, Notch, and Furniture, respectively. These continuation results show that SIGMA's advantage over the area-pooling baseline persists beyond the 50k checkpoint. The primary 50k comparison uses object-macro success, whereas the continuation curves report episode-micro success.

E METRIC SCOPE AND FURNITURE SUBSET CHECK

Table 8: Furniture object-macro rates (%) on the original scene sample, averaged over three Geometry-Alias-trained seeds.

Method	Legacy goal entry	Full task	Humanoid fall
AABB-shape-token	2.99	1.59	4.06
SSA + Area Pooling	34.54	22.12	9.45
SIGMA	41.82	27.85	14.47

Legacy goal entry only requires the object center to enter the goal threshold, whereas full-task success requires pickup, transport, and stable-goal progression. Excluding the one Furniture source-model ID that occurs in the climbing training split preserves SIGMA's advantage over the area-pooling baseline, changing it from $+5.73$ to $+5.62$ points on the original scenes and from $+3.90$ to $+4.05$ on the new scenes.

F ADDITIONAL FURNITURE DIAGNOSTICS

F.1 BEHAVIORAL PROGRESSION

Aggregate success alone does not indicate at which stage the two policies diverge. We therefore analyze the progression of Furniture episodes through pickup, transport, and final task completion.

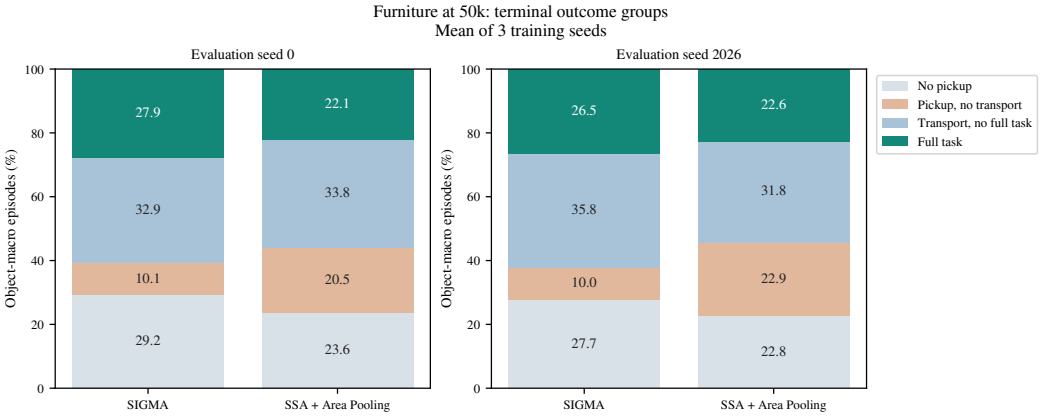


Figure 6: Furniture outcome stages under the original and replicated initial-state samples. SIGMA shows lower pickup frequency, fewer episodes terminating after pickup and before transport, and higher full-task completion.

Across both initial-state samples, SIGMA reaches pickup less frequently, but fewer episodes terminate after pickup and before transport, while a larger fraction reaches full-task completion. The outcome distribution therefore shifts toward later stages of the Carry sequence under SIGMA.

F.2 TERMINAL-METRIC SENSITIVITY

The completion difference remains positive when full-task success additionally requires stable termination. The conditional drop-rate difference changes sign across the two initial-state samples, while the higher humanoid-fall rate under SIGMA is observed in both. These results characterize the accompanying changes in Furniture episode outcomes beyond the primary full-task-success metric.

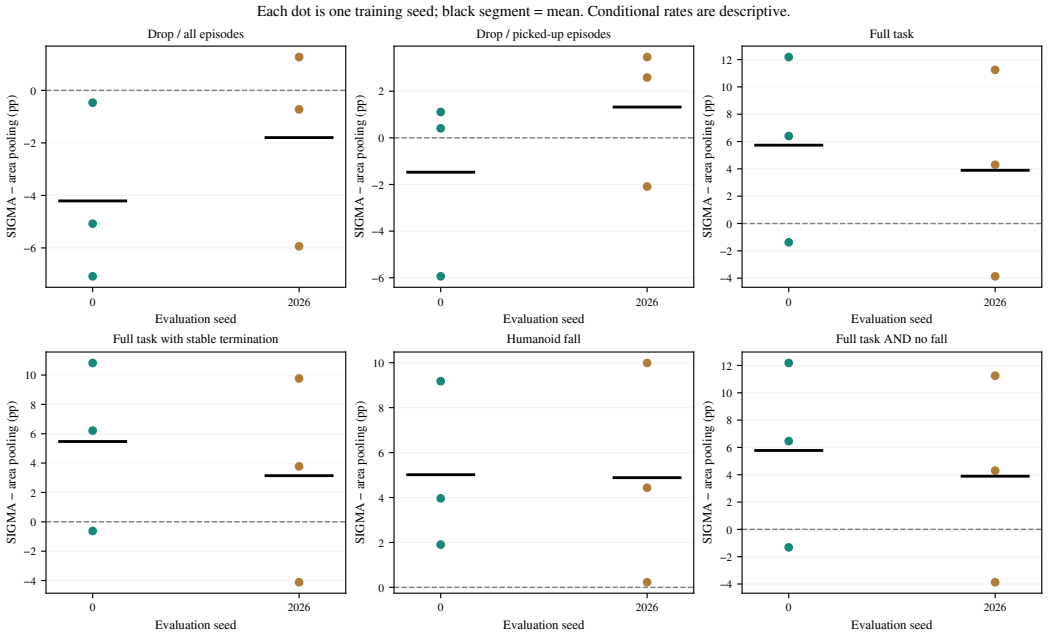


Figure 7: Furniture terminal-metric differences across the two initial-state samples. SIGMA's completion advantage over the area-pooling baseline remains positive under the stable-termination criterion. SIGMA also exhibits a higher humanoid-fall rate under both samples, while the conditional drop-rate difference changes sign across samples.

G Implementation Details and Hyperparameters

The mesh sampler uses a fixed seed, $N = 2048$ dense samples, $M = 26$ active anchors, and normal-distance weight $\lambda = 0.1$. Each anchor retains its source sample under degenerate nearest-anchor ties, ensuring that every SSA region is nonempty and has positive represented surface fraction. SSA normals are computed from normalized sums of the sample normals assigned to each region. If the summed normal has norm at most 10^{-12}, we use the anchor normal. The cached tensor contains $K = 32$ storage slots, including $M = 26$ active SSA elements and six padded slots. The main text defines the geometric descriptor over valid SSA elements. The fixed-slot implementation additionally appends a binary validity indicator, yielding an 11-dimensional encoder input. Valid elements use a validity value of one, while padded slots use zero and are masked during aggregation.

PointNet baseline. For each object, we sample 64 surface points uniformly by mesh area offline. Each point contains only its 3D coordinates, without normals, area, or color. We transform the points from the object frame into the humanoid-root-relative heading frame and divide their coordinates by 0.5; we do not normalize each object to a unit sphere. A shared 3–256–128–64 MLP applies ReLU after its first two linear layers, followed by channel-wise max pooling over the 64 points to produce one 64-dimensional geometry token. The three linear layers contain $(3 \times 256 + 256) + (256 \times 128 + 128) + (128 \times 64 + 64) = 42{,}176$ trainable parameters, excluding the shared policy network. This is our PointNet baseline; it does not use T-Net alignment. Its downstream Transformer has token dimension 64, four layers, two attention heads, and feed-forward dimension 512, with positional encoding disabled, as in the other methods.

All methods use one GPU and 4096 parallel environments per training run, with the same PPO/AMP settings and policy framework. Checkpoints are saved every 10k training iterations, and the primary Carry comparisons use the 50k checkpoint.

PPO / AMP backbone (shared by all arms):

SSA geometry encoder with 26 active elements stored in 32 fixed slots:

Hyperparameter	Value
Discount factor / GAE parameter	0.99 / 0.95
Learning rate (constant)	2×10^{-5}
ϵ_{clip}	0.2
Horizon / minibatch / mini-epochs	32 / 16384 / 6
Critic coeff. / value clip	5 / off
Grad-norm clip / entropy coeff.	1.0 / 0.0
Disc batch / minibatch	512 / 4096
Disc grad penalty / logit reg. / coeff.	5 / 0.01 / 5
AMP buffer size / keep-prob	200k / 0.01
Task-reward w / disc-reward w	0.5 / 0.5

Table 9: PPO/AMP training hyperparameters.

Hyperparameter	Value
Storage slots K / active SSA elements M	32 / 26
Dense samples per mesh / normal weight λ	2048 / 0.1
Position / centroid scales s_p, s_c	0.5 / 5.0
Surface-fraction scale s_a	32
Surface encoder MLP	11–256–128–64
Token dimension d	64
Transformer layers / heads	4 / 2
Feed-forward / extra MLP	512 / [1024, 512]
Task queries	4
Aggregation (SIGMA)	interaction-conditioned
Aggregation (SSA + Area Pooling)	represented-surface-fraction weighted
Proximity strength γ / distance scale σ	1.0 / 0.5

Table 10: Geometry-encoder and unified-policy hyperparameters.

Arm	Actor-side count	Counted MLP/query MACs per object	Batch latency (ms)
AABB-shape-token	1,266,112	47,104	0.151
SSA + Area Pooling	1,263,040	1,400,832	1.011
SIGMA	1,263,040	1,402,880	1.120

Table 11: Object-geometry tokenization cost. Actor counts include scalar entries in actor-side checkpoint tensors, retaining allocated query parameters in both surface variants and excluding critic and discriminator tensors. MAC estimates count MLP linear layers over all allocated rows and the ICGC query dot products. They exclude normalization, distances, softmax, and weighted aggregation. Timings use FP32 on one CUDA device with a batch size of 512 objects and report mean batch latency over 500 timed iterations after warm-up, with GPU synchronization immediately before and after the timed loop. The archived timing record does not retain the device model or warm-up count. Timings cover only the geometry encoder, excluding the policy Transformer and physics simulation.