

Born to Learn: Behaviour as a Design Axis for Embodied Self-Supervised Learning

Danila Kozlov, Mikhail Rudchenko,
Yukun Wang, Payton Liao, Kevin Xie

Department of Computer Science, University College London
Supervisor: Dr. Lorenzo Jamone

Abstract—Does the way an embodied agent moves shape the body schema representation it learns under self-supervised objectives? The answer determines where the lever for better embodied AI sits: if behaviour shapes representation, progress on self-supervised body models requires designing how agents move, not only larger architectures and more data. We address this question with a three-stage pipeline. In the first stage we train two behaviour-generation methods that produce different self-touch behaviours on the same body: a per-region mixture of experts with one specialist per body group, and a single skill-conditioned actor trained under a quality-diversity objective. In the second stage we use these methods to construct a collection of behaviour-conditioned trajectory datasets, and we summarise each dataset with a fixed set of behavioural descriptors. In the third stage we train one body schema model from scratch on each dataset using a single architecture and analyse the resulting body schemas with six metrics grounded in the cognitive neuroscience of body representation and in modern representation learning, tested on both an in-distribution test split and a shared held-out test distribution. We find that the behavioural composition of training data measurably shapes the body schema a self-supervised learner extracts from it.

Index Terms—body schema, self-touch, self-supervised learning, intrinsic motivation, quality-diversity, developmental robotics, BabyBench.

I. INTRODUCTION

A. Background & Problem Statement

A body schema is the sensorimotor, action-oriented internal model an embodied agent maintains of its own body’s geometry, posture, and effector reach [1]. In humans it is not innate; instead, it is constructed in early life from sensorimotor activity, and self-touch is one of the densest sources of evidence an infant has about which parts of its body exist and where those parts are relative to each other [2], [3]. Developmental robotics inherits the same constructive assumption [4], and recent representation-learning systems recover body-level structure from sensorimotor data [5] as well as 3D body geometry and kinematics from visual observation [6].

Developmental psychology also observes that infants do not move uniformly. Self-touch is densely distributed over the body from the first weeks of life, and the per-region composition shifts with motor development: head and trunk are contacted at comparable rates early on, with touches to the trunk increasing in the weeks approaching head control [3]. If those behavioural patterns are a necessary scaffold for body schema acquisition, then any embodied learner that ingests sensorimotor data should be sensitive to the same scaffold,

and the right axis along which to improve embodied self-supervised representations is behavioural composition. If they are a peculiarity of biological development, the right axes may instead be data scale and architecture. The two readings have different implications for where future embodied self-supervised learning effort should go.

The aim of this project is to answer the following question:

Does the way an embodied agent moves, in particular what self-touch it produces, shape the body schema representation it learns under modern self-supervised learning frameworks?

To answer this question, we isolate behaviour as the independent variable in the body schema acquisition process of an agent controlling a simulated infant body (namely, MIMo [7]). We assess the effect behaviour has on the quality of learned body schemas. No prior work carries out this isolation to the best of our knowledge.

B. Approach & Contributions

We answer the question in three stages. **(1) Behaviour generation:** we train two behaviour-generation models that produce qualitatively different self-touch on the same body. **(2) Dataset generation:** we use these models to generate a family of behaviourally distinct datasets, each summarised by a fixed set of behavioural descriptors. **(3) Body schema training and evaluation:** we train one body schema model from scratch on each dataset under an identical recipe and score it along six complementary axes of body schema quality, on both a held-out split of its own training distribution and a shared leak-free held-out distribution. Because the learning recipe is fixed across datasets, the only variable is behaviour: we correlate dataset descriptors against schema-quality axes to test whether, and along which axes, behaviour shapes the learned body schema. Section III describes each stage in detail.

We make three contributions. **First**, two methods for behaviour-controllable self-touch on a simulated infant body; datasets can be produced with specifiable behavioural signatures using these methods. **Second**, a body schema evaluation framework whose six probes are grounded in the cognitive neuroscience of body representation and in modern representation learning, all detailed in Section IV-C; the framework is intended to be reusable beyond this work. **Third**, empirical evidence that the behavioural composition of training

data measurably shapes the body schema a self-supervised learner extracts from it, broken down per evaluation axis, with implications for how embodied self-supervised representation learning should approach behaviour-side design.

II. PROJECT MANAGEMENT

A. Individual Contributions

Individual contributions are summarised in Table I.

TABLE I: Individual contributions.

Member	Contribution
Mikhail	Per-region mixture-of-experts; dataset generation; 4 body schema evaluation metrics; dataset descriptors; report writing.
Danila	Skill-conditioned QDAC actor; dataset generation; 2 body schema evaluation metrics; evaluation and results analysis; report writing.
Kevin	Quality-diversity metric design; minor report writing.
Payton, Yukun	VAE-training experiments; minor report writing.
Danila, Mikhail, Kevin	Team coordination.

B. Tasks, Milestones, and Timeline

The project ran from October 2025 to the Report 2 deadline (22 May 2026). The main tasks and their schedule were:

- **Oct 2025–Jan 2026:** literature survey and initial brainstorming (Report 1).
- **Feb 2026:** re-scoping and design of the methodology and metric suite.
- **Mar–May 2026:** behaviour-generation models and metrics development and testing.
- **May 2026:** Mid-project review, cross-team integration, and metric-suite audit.
- **by 22 May 2026:** Report 2 write-up and submission.

C. Risks and Mitigation

Table II lists the main risks anticipated at the start and the mitigations applied. The behaviour-generation and skill- and VAE-collapse risks materialised and were handled as described; inter-team communication broke down and was only partially mitigated; the compute risk was avoided by capping training budgets up front.

III. METHODOLOGY AND SYSTEM DESIGN

In this section we describe the technical design of key components of the pipeline in detail.

TABLE II: Key project risks, status, and mitigations.

Risk	Status	Mitigation
A single goal-conditioned policy cannot cover all body regions.	Materialised	The unified policy collapsed onto the most reachable region; replaced by a per-region mixture of experts with no parameter or gradient sharing above the encoder, plus a single skill-conditioned actor as a complementary single-model approach.
VAE representation collapse during body schema training.	Materialised	Hyperparameter sweeps, gradient clipping, and a four-step VAE loop (pretrain on babbling, train specialists, refit on rollouts, retrain).
Insufficient compute for the MIMO body.	Avoided	Training capped at 3×10^6 steps for the skill-conditioned actor and $5 \times 1.5 \times 10^5$ per generation for the MoE; the skill-conditioned actor completes in ≈ 11 h on one RTX 4070 Ti SUPER.
Communication breakdown within the team.	Materialised	Partially mitigated through frequent meetings.

A. Approach

The pipeline has three stages (Figure 1).

- 1) **Behaviour generation.** We present two methods that produce different self-touch patterns on the same MIMO body (Figure 4). The first is a per-region mixture of experts (MoE): five independent SAC agents, one for each of head, upper torso, lower torso, legs, and arms, each rewarded for driving the infant’s hands onto its assigned region; the MoE is composed at inference by switching between specialists. The second is a single skill-conditioned actor trained under a quality-diversity objective, conditioned on a continuous five-dimensional skill prompt. The two methods span a useful axis of design: many small specialists composed by switching, against one larger model with continuous skill control.
- 2) **Dataset generation.** Each behaviour model is rolled out under several composition rules to produce a family of trajectory datasets. For the MoE we generate per-region, multi-region, and within-episode-switching variants; for the QDAC actor we sweep the skill manifold and sample uniformly. Random motor babbling and scripted programmed-motion trajectories are included as exploration-only controls. Each dataset is summarised by a fixed set of behavioural descriptors so that downstream representation effects can be linked back to measurable behavioural differences in the data.
- 3) **Body schema training and evaluation.** For each dataset we train a body schema model from scratch across multiple seeds under a fixed recipe: a dual VAE that encodes proprioception and touch into a single latent, plus a residual forward model in latent space. Architecture, optimiser, sample budget and seed schedule are held constant across datasets; only the training data changes. Each trained schema is then evaluated by a six-metric suite on two held-out tiers (Section III-I). The core instrument is the relationship between each dataset’s behavioural descriptors

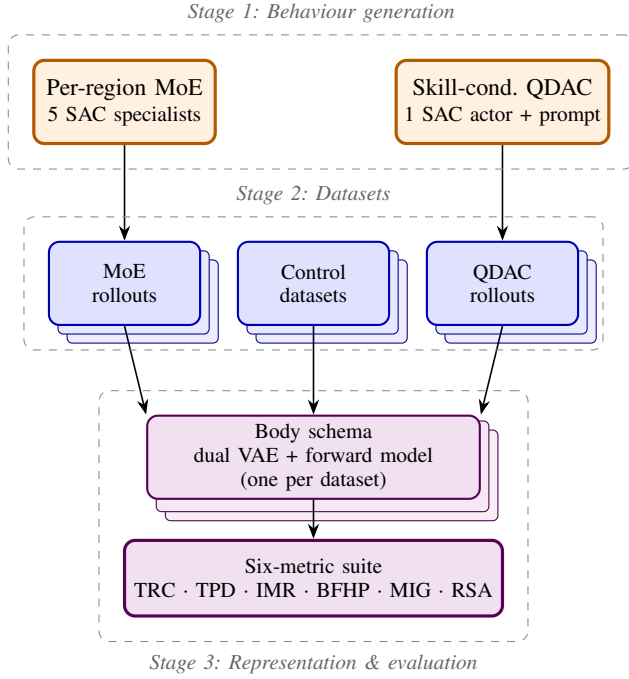


Fig. 1: End-to-end three-stage pipeline. *Stage 1* trains two behaviour-generation methods: a per-region mixture of experts (five SAC specialists) and a single skill-conditioned QDAC actor. *Stage 2* rolls each out under several composition rules to produce many trajectory datasets (shown as card stacks), together with random-babbling and scripted programmed-motion control datasets; the full list and composition rules are in Section IV. *Stage 3* trains one body schema (dual VAE plus residual forward model) per dataset under an identical recipe and evaluates them with the six-metric suite (Section IV-C). Holding all non-data variables fixed makes across-dataset metric differences attributable to behaviour.

and the schema-quality metrics of the model trained on it; Section IV defines the metrics in detail.

The next sections describe the components of the pipeline in detail.

B. Body and Observations

We use the MIMo body [7], a developmentally-realistic humanoid infant simulated in MuJoCo with anatomically plausible morphology. Two raw observation streams are exposed at each environment step:

- **Proprioception** $\mathbf{p}_t \in \mathbb{R}^{380}$: the concatenation, over all 47 robot joints, of joint position, joint velocity, joint torque and a joint-limit proximity indicator, plus a 176-D muscle-state block containing per-muscle activations and per-muscle isometric forces for the 44 actuated joints. The vector is therefore not pure kinematics; roughly half its dimensions are muscle state.
- **Touch** $\mathbf{c}_t \in \mathbb{R}^{17175}$: the flattened readings of 5,725 surface taxels distributed across all body parts, each emitting a three-channel signed force vector (F_x, F_y, F_z) in the taxel’s own local body frame.

The agent emits an action $\mathbf{a}_t \in [0, 1]^{88}$ that drives 44 antagonist muscle pairs (one pair per actuated joint). We use the muscle actuation model rather than a torque controller because muscle activations are biomechanically more realistic than direct torque commands, which matters for any developmental claim that hinges on the agent’s motor model resembling that of a human infant.

The body is partitioned into five target regions for the per-region MoE: head, upper torso, lower torso, legs, and arms. Hands are excluded from the target set: they are the toucher, never the touched, and rewarding hand contact would create a degenerate hand-to-hand pathology. Feet are excluded from the evaluation analysis on the basis that more than 25 pilot runs in the supine scene produced no foot contact, which we attribute to biomechanical reach in that posture rather than to policy failure. Full simulation specifications (scene, frame rate, episode length, sensor configuration) are deferred to Section IV.

C. Sensory Encoder

To make reinforcement learning tractable over the 17,555-D raw observation, both behaviour-generation methods consume a compressed 88-D state produced by a frozen dual VAE [8], following Yang & Jamone’s body schema architecture [9]. The proprioceptive encoder $\Phi_p : \mathbb{R}^{380} \rightarrow \mathbb{R}^{64}$ and the tactile encoder $\Phi_c : \mathbb{R}^{17175} \rightarrow \mathbb{R}^{24}$ are two separate VAEs; the fused state seen by the policy is the concatenation $\mathbf{s}_t = [\Phi_p(\mathbf{p}_t); \Phi_c(\mathbf{c}_t)] \in \mathbb{R}^{88}$ (Figure 4). At inference each Φ returns the deterministic mean of its latent posterior. We keep the two modalities in separate encoders so that each can be reconstructed by its own decoder due to the very different statistics of proprioception (continuous, dense) and touch (sparse, near-zero most of the time).

Each encoder is a two-hidden-layer MLP of width 512 with ReLU activations, paired with a symmetric decoder. Training minimises the standard ELBO,

$$\mathcal{L}_{\text{VAE}}(\mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 + \beta \text{KL}(q_\psi(\mathbf{z} | \mathbf{x}) \| p(\mathbf{z})) \quad (1)$$

with $\beta = 0.01$ and a unit Gaussian prior. The KL weight is deliberately small so the latent remains informative; we do not require disentangled latents at this stage because the SAC policies treat $\mathbf{s}_t$ as opaque. Both encoders are pretrained on 50,000 samples drawn from random motor babbling under Adam with learning rate 3×10^{-4} and batch size 256. In the MoE pipeline we additionally refit them on policy-generated rollouts after the first round of specialist training; the details are given in Section III-E.

D. Choice of SAC

Both behaviour-generation methods use Soft Actor-Critic [10] as their base reinforcement-learning algorithm. SAC’s off-policy structure suits an expensive simulator since every transition collected from MuJoCo is reused for many gradient updates from a replay buffer, which gives better sample efficiency per environment step than an on-policy method

such as PPO. SAC’s automatically tuned entropy temperature also pairs naturally with the count-based novelty bonus in the MoE reward (Section III-E) by keeping the action distribution spread out independently of state-space coverage. Across both methods we use ReLU activations, $\log \sigma$ clipped to $[-20, 2]$, twin Q-critics with the minimum-of-two trick for Bellman targets, Polyak target smoothing at $\tau=0.005$, and automatically tuned entropy temperature α with a half-width target entropy ($-\text{action_dim}/2$). The same loop is applied once for the QDAC actor and five times in parallel for the MoE specialists.

A behaviour-conditioned generative VAE and goal-conditioned SAC with hindsight experience replay were both considered as alternatives and rejected for reasons given in Appendix A.

E. Per-Region Mixture of Experts

a) Architecture: The MoE consists of five independent SAC agents, one per target region. Each specialist owns its own actor, twin Q-critics, target Q-networks, log-temperature $\log \alpha$ and a 500,000-transition replay buffer; there is no parameter sharing above Φ . The actor is an MLP of width 128 (two hidden layers) producing a tanh-squashed Gaussian over $\mathbb{R}^{88}$; the two Q-critics are MLPs of the same width over the state-action concatenation. We add several stabilisation tweaks on top of vanilla SAC: small-init Q heads to suppress early Q overestimation, an α floor at 0.03 so the entropy coefficient cannot collapse, TD3-style delayed actor updates (critics every step, actor and temperature every other step), Adam weight decay (10^{-5} on actor, 10^{-4} on critics), and gradient-norm clipping at 1.0. The first 1,000 environment steps are filled with uniform random actions before SAC updates begin. Each specialist is trained for 150,000 environment steps; the five are trained in parallel as independent processes.

At inference the actor’s output $\tanh(\mu + \sigma\epsilon) \in [-1, 1]^{88}$ is multiplicatively scaled by 0.5 before being passed to the environment and is then clipped against the muscle range $[0, 1]^{88}$. The negative half of the tanh range therefore collapses to zero on contact with the actuator and the effective control surface is the positive half of the policy’s sampled distribution. This is intentional: muscle activations are inherently non-negative, and the clip plays the role of a rectifier rather than a hard constraint.

b) Reward: The shaped reward sums six additive terms, all with non-zero weight in the canonical configuration. For specialist with target region g^* , at step t with action $\mathbf{a}_t$ and previous action $\mathbf{a}_{t-1}$:

$$\begin{aligned} r_t = & c_{\text{tgt}} \phi_t[g^*] - c_{\text{oth}}(t) \sum_{g \neq g^*} \phi_t[g] \\ & - c_{\text{dst}} \min(d_t, d_{\text{max}}) + c_{\text{bil}} \mathbf{1}[\text{both hands on } g^*] \\ & + c_{\text{mov}} \|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2 + c_{\text{nov}} (n_t(\mathbf{s}_t) + 1)^{-1/2}. \end{aligned} \quad (2)$$

The components are:

- $\phi_t \in \mathbb{R}^5$ is the force-weighted touch distribution over the five body groups at step t . For each group g we sum the magnitudes of contact forces between any hand geometry and any body geometry in g , then normalise so $\sum_g \phi_t[g] = 1$ whenever there is any hand-on-body contact, and $\phi_t =$

0 otherwise. $\phi_t[g^*]$ is therefore the share of total contact force going to the target region at step t . Force-weighting rather than binary contact gives a dense Bellman signal whenever *any* contact is made on the prompted region and prevents the “drag a finger across everything” pathology that a binary indicator allows.

- d_t is the mean Euclidean distance from each hand centroid to the target group’s centroid in world coordinates, capped at $d_{\text{max}} = 1$ m. It provides a dense gradient toward the target group before any contact is made.
- $\mathbf{1}[\text{both hands on } g^*]$ is the bilateral-touch indicator: 1 if both the left and right hand make any contact with the target region in step t , 0 otherwise. The bonus is small and acts as a mild incentive toward two-handed engagement, which is developmentally plausible and discourages fixation on a single dominant hand.
- $\|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2$ is the L2 norm of the action delta between consecutive steps. The movement bonus pays for *changing* the muscle command rather than for absolute muscle activation, which prevents the SAC actor from collapsing to a frozen near-zero pose under sparse contact reward, especially before contact is first achieved.
- $n_t(\mathbf{s}_t)$ is the visit count of the current 88-D latent state under a 32-bit SimHash (32 Gaussian random projections, sign-thresholded). The hash buckets together latent states that are structurally similar, so the bonus encourages coverage of the latent state-space without an additional learned network.
- The off-target coefficient $c_{\text{oth}}(t)$ follows a curriculum: zero for the first 25,000 environment steps, then linearly ramped to its terminal value over the next 25,000 steps. The curriculum exists to avoid a penalty trap; without it, a still-exploring policy is punished for any contact it stumbles onto before it has learned to target the right region, and converges to a no-contact pose.

Canonical coefficients are $c_{\text{tgt}} = 5.0$, $c_{\text{oth}}(\infty) = 0.1$, $c_{\text{dst}} = 0.1$, $c_{\text{bil}} = 0.2$, $c_{\text{mov}} = 0.005$, $c_{\text{nov}} = 0.1$. The target term dominates by an order of magnitude whenever contact is achieved; the smaller shaping terms exist to make the contact-free regime differentially navigable rather than to compete with the contact signal.

c) Four-stage VAE-policy loop: The frozen encoder Φ must be in-distribution for the states the specialists actually visit, but the initial Φ is trained on random motor babbling, which rarely produces contact. We therefore alternate encoder fitting and policy training across four stages:

- 1) Pretrain Φ on 50,000 random-babbling samples (Section III-C).
- 2) Train a first generation of five specialists against this Φ for 150,000 env steps each.
- 3) Refit Φ on the union of 50,000 random-babbling samples and the rollouts produced by the first-generation specialists, under the same ELBO objective and Adam settings as the original pretraining.
- 4) Train a fresh second generation of five specialists against the refit Φ for 150,000 env steps each.

Per-dimension tactile reconstruction error on rollout states drops by approximately an order of magnitude between the first and second VAE fits, confirming that the original encoder was substantially out-of-distribution for the contact-bearing states the specialists learned to visit.

d) In-training evaluation: Every 25,000 environment steps we roll out five evaluation episodes of 1,000 steps under both stochastic and deterministic actor modes and record per-step mean $\phi_t[g^*]$ (the share of total hand-to-body contact force on the target region) and $\sum_{g \neq g^*} \phi_t[g]$ (the share on the other regions). This force share on the target is the primary success metric we used to gauge whether a specialist was actually targeting its assigned region during training. We retain a single best checkpoint per specialist based on $\max(\text{stoch}, \text{det})$ target force share.

F. Skill-Conditioned QDAC Actor

a) Architecture: The pilot for this route used a one-hot region prompt and collapsed to head-only contact in 8/8 seeds. The reason is biomechanical: most one-hot directions are infeasible (the actor cannot reach its leg with its head), so the policy converges to the only direction the body actually permits. We replace the one-hot prompt with a Gaussian density over body-group anchors along a continuous rostro-caudal scalar $z \in [0, 1]$:

$$\mathbf{g} = \left(\mathcal{N}(z; \mu_r, \sigma^2) \right)_{r=1}^5 \in \mathbb{R}^5, \quad (3)$$

so every prompt has support somewhere biomechanically reachable. A single SAC actor of shape $93 \rightarrow 256 \rightarrow 256$ takes the concatenation $[\mathbf{s}_t; \mathbf{g}_t]$ as input, and a successor-features critic $\psi_\eta(\mathbf{s}, \mathbf{g}) \in \mathbb{R}^5$ feeds the QDAC [11] diversity term with a Lagrange-coupled dual.

b) Fixes that resolved head-only collapse: Five concrete changes (continuous rostro-caudal prompt, force-weighted descriptors, learning-progress curriculum, capped hindsight relabelling, and a standing scene), detailed in Appendix B, turned the original 8/8 head-collapse into all-region coverage in 10^6 steps. The strongest run reaches a sum of positive prompt-to-realised correlations of +1.49 across the four reachable groups (legs alone +0.91, the strongest single-group controllability we have measured), against ~ 0 for the prior pilot.

c) Checkpoint selection: Reward-greedy selection returns an uncontrollable policy because SAC return keeps rising while prompt-following plateaus. We instead pick the checkpoint maximising the sum of positive per-group prompt-to-realised contact correlations across a fixed grid of skill prompts (five one-hot specialists plus balanced and two-region mixtures), i.e. the +1.49 number above.

d) Quality-diversity archive: The descriptor-space coverage metric (Section IV-C) is read off a quality-diversity archive built over the QDAC rollouts. The archive is a fixed discretisation of the policy’s behaviour-descriptor space into a 20×20 grid over two axes (movement frequency on one axis, touch coverage on the other), giving 400 cells in total. Each rollout episode is mapped to one cell determined by its realised descriptor pair, and the archive records one episode per filled

cell. Descriptor-space coverage is then the fraction of cells with at least one episode, that is, filled-cells over total. We report it for QDAC sources only because the other generators do not target a skill manifold.

G. Generating Trajectory Datasets

Once the behaviour generators are trained, we use them to produce a family of trajectory datasets. Each dataset records the full raw observation streams $(\mathbf{p}_t, \mathbf{c}_t)$ alongside the executed action $\mathbf{a}_t$ at every step, together with several additional signals used downstream by the evaluation metrics; the resulting per-episode tables are written to HDF5 and consumed by the body schema training pipeline described in Section III-H. The dataset families used in the evaluation, including episode counts and composition rules, are detailed in Section IV.

After generation, every dataset is characterised by a fixed set of behavioural descriptors so that the differences between datasets are quantified rather than assumed: *self-touch entropy*, *self-touch fraction*, *bimanual fraction*, *pose vigour*, *movement coherence*, and, for the QDAC datasets only, *descriptor-space coverage*. These descriptors are the behaviour-side variables that the evaluation correlates against body schema quality; their precise definitions are given in Section IV.

H. Body Schema Model

For each dataset we train a new body schema model from scratch under the same architecture, optimiser, split rule, and training budget. The only intended variable is the data produced by Stage 2. Each model is trained independently under several random seeds; the train/test split are detailed in Section IV.

The representation model is composed of a dual VAE and a residual forward model in latent space.

The dual VAE follows Yang & Jamone [9] and is architecturally identical to the Stage-1 encoders explained in Section III-C, while the forward model is an MLP with two ReLU-activated hidden layers of width 512, which takes the current latent $\mathbf{z}_t$ and the current action $\mathbf{a}_t$ as input and predicts the change in the latent state:

$$\hat{\mathbf{z}}_{t+1} = \mathbf{z}_t + f_\theta(\mathbf{z}_t, \mathbf{a}_t), \quad (4)$$

where f_θ is the forward model and $\hat{\mathbf{z}}_{t+1}$ is the predicted latent at the next time step.

The full training loss is:

$$\mathcal{L} = \lambda_p \mathcal{L}_{\text{VAE}}(\mathbf{p}_t) + \lambda_c \mathcal{L}_{\text{VAE}}(\mathbf{c}_t) + \lambda_f \log(\text{MSE}(\hat{\mathbf{z}}_{t+1}, \mathbf{z}_{t+1}) + \epsilon), \quad (5)$$

with all three weights set to one. We wrap the dynamics MSE in a logarithm so that the forward term stays on a scale comparable to the two reconstruction losses during joint training; lower values still correspond to better forward prediction, with the caveat that the loss is negative whenever the underlying MSE is below one.

Non-VAE alternatives we considered (JEPA, LeJEPA, SimCLR, VQ-VAE) are discussed in Appendix A; we preferred a VAE for its decoder, which allows reconstruction to be debugged directly.

I. Body Schema Evaluation

After training, each body schema model is frozen and evaluated with a six-metric suite explained in detail in Section IV-C: touch-region correctness (TRC), two-point discrimination (TPD), inverse-model recoverability (IMR), body-frame hand-pose prediction (BFHP), mutual-information gap (MIG), and representational similarity analysis (RSA).

Each schema is evaluated on two tiers: one measures how well it fits its own training distribution, the other how well it generalises to a shared held-out distribution it did not train on. The construction of both tiers is detailed in Section IV-B.

IV. EXPERIMENTAL SETUP

A. Simulator and Hardware

All experiments run in BabyBench [12] on the MIMo infant body [7] in MuJoCo, simulated at 50 Hz with 1000-step episodes. Training uses one NVIDIA RTX 4070 Ti SUPER (16GB) GPU per run; the skill-conditioned policy completes 3×10^6 environment steps in ≈ 11 h. The MoE family uses $5 \times 1.5 \times 10^5$ steps per generation across two generations, i.e. 1.5×10^6 total. Including ablations that did not become deliverables, we ran roughly 5×10^7 environment steps over about five days of compute time.

B. Datasets

Each rollout episode has two phases. The first 20 steps execute uniformly random actions before the policy takes over. This random burst is recorded in addition to the policy-driven phase that follows and prevents each episode from starting from the same pose, which would cause the first few policy steps to be near-identical across episodes; the burst injects per-episode pose variability so that the body schema model is trained on a wider effective state distribution than the environment’s default reset alone provides. The policy-driven phase records up to 250 further steps, capped to encourage diversity over depth. Each episode thus contributes up to 270 recorded transitions, smaller than the 1,000-step in-training evaluation episodes.

The MoE supports flexible composition: a single rollout episode can be driven by one specialist (yielding a region-targeted dataset), by several specialists rotating in round-robin or random order at a configurable within-episode cadence (yielding multi-region datasets with controlled inter-region transitions), or by per-specialist mode and encoder overrides (which allows a single dataset to combine specialists from different stages of the four-stage loop running in different stochastic or deterministic modes). The QDAC actor is rolled out under chosen skill prompts $\mathbf{g}$: one-hot vectors mimicking the per-region specialists, fixed mixture skills, a within-episode schedule across rostral-caudal anchors (Eq. 3), or uniform samples from the 5-D simplex via Dirichlet(1) for broad coverage. We additionally collect random motor babbling and scripted programmed-motion datasets as exploration-only controls.

Altogether, 32 dataset sources were produced, spanning various behaviour axes, from contact tightly concentrated on a single region, through switched multi-region contact, to diffuse exploration with no targeting; the controls ensure the low end of every behavioural descriptor (Section IV-C) is sampled rather than left empty. Per-family source counts and episode budgets are reported in Table VI.

A separate body schema model is trained per source under the fixed recipe of Section III-H, repeated across up to five random seeds; every reported metric is the mean over the seeds available for that source. Each (source, seed) pair uses a deterministic episode-level 80/20 train/test split, so the same source and seed always yield the same held-out episodes.

Schemas are evaluated on two tiers. In *Tier A* a schema is scored on the held-out 20% of its own dataset, so Tier A is in-distribution. In *Tier B* every schema is scored on one shared set, pooled from a single seed’s test split of all sources. No schema was trained on any of these episodes, so Tier-B scores are directly comparable across schemas and chiefly measure how well a schema generalises to behaviour it did not see during training.

C. Evaluation Metrics

We evaluate with two families of metrics. Dataset descriptors summarise the behaviour in each generated dataset; body schema metrics measure the quality of the representation a schema learned from that dataset. We correlate the two groups of metrics.

We do not use the encoder’s reconstruction or forward-prediction loss as the schema-quality measure. Yang & Jamone [9] read a falling VAE reconstruction loss and forward-model loss as evidence that a body schema has formed. A low reconstruction loss proves only that the latent compresses and predicts the sensory stream; it does not prove that the latent is organised as a body schema, because a VAE can drive reconstruction loss down with an entangled code that supports no recoverable geometry, posture, effector pose, or action grounding. We therefore measure schema quality with six probes that test those properties directly.

Dataset descriptors: Every source is summarised by the six descriptors below. The three contact descriptors are built on the BabyBench self-touch detector [12], which records every hand-to-body contact in simulation: each contact is attributed to the body geometry it touched, so environment contact is excluded by construction and contact is resolved at body-geometry granularity. The posture descriptors are computed from the recorded proprioception.

- *Self-touch entropy:* the normalised Shannon entropy of the contact distribution over body geometries. This is the central behavioural-composition axis: low entropy is contact concentrated on one part of the body, high entropy is contact spread across it, and it asks whether the breadth of self-touch matters.
- *Self-touch fraction:* the fraction of steps with any hand-to-body contact. It measures the amount of self-touch

independently of where it lands, so an effect of self-touch entropy cannot be explained as merely more contact.

- *Bimanual fraction*: the fraction of steps in which both hands contact the body at once. A bimanual self-touch proxy in the spirit of the double-touch cue, it asks whether coordinated two-handed contact, which carries relational body information, helps.
- *Pose vigour*: the average distance the body’s pose moves between consecutive steps, that is the mean over all steps of the Euclidean norm of the change in the 47 joint angles. It asks whether more movement matters.
- *Movement coherence*: the fraction of total proprioceptive variance captured by the five leading principal components of the full proprioception vector. It is high when the movement is structured and low-dimensional and low when the movement is diffuse and high-dimensional, so it asks whether coherent rather than undirected movement matters.
- *Descriptor-space coverage*: for the QDAC sources only, the fraction of skill-archive cells filled (see Section III-F for the archive’s axes and grid). It quantifies how much of the intended skill-descriptor space the policy realised.

Body schema metrics: With the presented probes, we intend to address the different aspects of the definition of good body schema (Section I).

Touch-side probes first reduce the flat 17,175-D vector to per-taxel activations: taxel k aggregates flat channels $3k$, $3k+1$, $3k+2$ and is active if any exceeds $\tau = 0.01$ in absolute value. Precomputed per-taxel torso-frame coordinates $\mathbf{X} \in \mathbb{R}^{5725 \times 3}$ and a region map $R(k)$ assigning each taxel to one of the five groups support the geometric probes. Every metric is computed with the encoder frozen.

a) Touch-Region Correctness (TRC): Tests whether the schema reconstructs contact at the right place on the body, the touch-region-identity property. From the latent we decode the touch reconstruction, binarise it and the true touch at τ , reduce both to per-taxel activations, and aggregate to the five body groups via $R(k)$. We report the macro-averaged per-group F_1 between predicted and true active regions (TRC- F_1). Because some regions are in near-constant passive contact in the supine pose, the score is read against a per-group base-rate predictor rather than in absolute terms. Scoring at body-region rather than individual-taxel granularity is deliberate and biologically motivated: human tactile spatial acuity is coarse away from the extremities, so getting the region right matters more than the exact location [13]. A schema that decodes touch onto the correct regions scores high; one that smears or misplaces contact scores low.

b) Two-Point Discrimination (TPD): Tests whether the latent preserves the metric structure of the body surface, the somatotopic property: nearby touches should map to nearby latents, mirroring the topographic organisation of the cortical body map [14] and the two-point discrimination threshold that quantifies tactile spatial acuity in humans [13]. We sample 5×10^4 random pairs of test frames that each have active contact. For each pair we compute a body-surface distance (the symmetric Chamfer distance, in metres, between the

two frames’ active-taxel sets in $\mathbf{X}$ coordinates) and a latent distance $\|\mathbf{z}_i - \mathbf{z}_j\|_2$. The headline is the Spearman correlation ρ between the two across all pairs. A random or collapsed latent gives $\rho \approx 0$; a somatotopically organised latent gives $\rho > 0$.

c) Inverse-Model Recoverability (IMR): Tests whether the action that produced a transition is recoverable from the latent change, the action-grounding property: a body schema links motor commands to their sensory consequences [15], and a representation that has captured this link makes the executed action decodable from the transition it caused [16]. On the train split we fit a small one-hidden-layer MLP $h : [\mathbf{z}_t; \mathbf{z}_{t+1}] \mapsto \hat{\mathbf{a}}_t$ (Adam, batch 256, up to 500 epochs, early stopping on a 10% validation split). On the test split we report the action R^2 averaged over the 88 action dimensions, with zero-variance dimensions contributing $R^2 = 0$; action MSE is a secondary number. The pairing uses $\mathbf{a}_t$, the action that drove $\mathbf{z}_t \rightarrow \mathbf{z}_{t+1}$. A latent that does not encode the consequences of action gives $R^2 \approx 0$.

d) Body-Frame Hand-Pose Prediction (BFHP): Tests whether the latent localises the effectors on the body, the effector-reach property, isolating “where are my hands on my body” from “where am I in the world”. Body-part position is coded in body-centred coordinates in primate parietal and premotor cortex, and that code reorganises with experience [17], [18]; a body schema should make the same body-frame hand position recoverable. The target is the left and right hand positions in the torso frame, a 6-D vector invariant to whole-body pose. On the train split we fit ridge regression $r : \mathbf{z} \mapsto \mathbf{y}$; on the test split we report total MSE (lower is better). A low MSE means hand pose is linearly recoverable from the latent; an MSE at the predict-the-mean level means effector pose is not encoded.

e) Mutual-Information Gap (MIG): Tests whether the latent is organised rather than merely compressed: a body schema should localise each body factor to a single latent dimension instead of smearing it across an entangled code. For each factor (the 47 MIMo joint angles and the contacted-region class), we compute the R^2 that each latent dimension achieves in predicting it, take the gap between the best and second-best latent dimension, and average that gap across factors. This is an R^2 -based approximation of the mutual-information gap [19], a standard disentanglement measure, used here noting that unsupervised disentanglement is not identifiable without an inductive bias [20]. We also report the mean best-dimension R^2 so a near-zero gap caused by a factor not being encoded can be told from one caused by redundant encoding.

f) Representational Similarity Analysis (RSA): Tests whether the latent preserves the metric structure of body posture, the configuration of the kinematic chain, complementary to TPD’s touch-surface geometry. On a 1500-frame subsample we form the pairwise Euclidean distance matrices of the latent and of the 47-dimensional joint-angle block, and report the Spearman correlation of their off-diagonal entries, following representational similarity analysis [21]. A latent whose geometry tracks posture gives a high positive value.

V. RESULTS AND DISCUSSION

A. Stage 1: Touch-Generation Quality

We evaluate both data factories on two common axes: *Stage-2 dataset descriptors* and *per-region success rate*, defined as the fraction of evaluation episodes in which the agent achieved contact on region r when prompted (skill-conditioned) or dispatched (MoE) to that region. A uniform-random policy provides the baseline here.

a) Per-region success rates: We score each policy by per-region rollout contact rate on *strict* sub-regions of the body that exclude the parts producing passive self-contact at MIMo’s resting pose: the hip (excluded from lower torso), the feet (excluded from legs), and the forearms (excluded from arms). Restricting to these strict sub-regions isolates voluntary reaching from the rest-pose floor. Table III reports the resulting per-region contact rate for the random baseline and for the MoE and skill-conditioned QDAC families ($n=30\text{ k}–60\text{ k}$ timesteps per cell), together with the five dataset descriptors over the Stage-2 trajectory datasets each family produces (Section III-G). The two halves of the table summarise the two faces of the same generator: how often it hits each body region (targeting accuracy) and how it moves while doing so (trajectory structure).

TABLE III: Generator behaviour summary. *Top half:* per-region contact rate on strict sub-regions during post-training evaluation. *Bottom half:* Stage-2 dataset descriptors (see Section IV-C) averaged over the trajectory datasets each family produces. Random is a single uniform-action rollout; MoE and Skill-cond. values are aggregated across their respective trained policy families ($n=13$ MoE datasets).

	Random	MoE	Skill-cond.
<i>Per-region contact rate</i>			
Head	0.17	1.00	1.00
Upper torso	0.04	0.55	0.48
Lower torso (abdomen)	0.00	0.35	0.46
Legs (thighs / shins)	0.01	0.34	0.79
Arms (upper arm)	0.09	0.46	0.76
Mean	0.06	0.54	0.70
<i>Dataset descriptor (Stage-2 trajectories)</i>			
Self-touch entropy	0.21	0.46	0.26
Self-touch fraction	0.81	0.67	0.57
Bimanual fraction	0.66	0.28	0.22
Pose vigour	0.91	0.37	0.89
Movement coherence	0.50	0.75	0.53

Both trained families exceed the random floor on every region, with mean voluntary contact rate at $9\times$ above random for MoE and $12\times$ above for skill-conditioned QDAC. Random’s values stay ≤ 0.17 across the board. The descriptor half exposes the qualitative behaviour difference that the contact-rate half does not: MoE and Skill-cond. are within $1.2\times$ of each other on self-touch fraction and bimanual fraction, but Skill-cond. moves $2.4\times$ as energetically (pose vigour 0.89 vs. 0.37) while producing markedly less structured motion (movement coherence 0.53 vs. 0.75). MoE has both broader contact spread (self-touch entropy 0.46 vs. 0.26) and lower-dimensional posture trajectories; Skill-cond. hits the lower-

body regions more often but at higher motor cost and without the postural structure. This is the behavioural asymmetry that may explain why Skill-cond. wins Stage-1 targeting but MoE wins Stage-3 representation quality (Section V-D).

b) Per-region success heatmaps: The body diagrams in Figure 5 (Appendix C) colour each reachable region by its success rate.

c) From expert name to realised label: Table III’s row labels are each expert’s *target* region, not its realised contact region. On a dense rollout of 50 deterministic episodes per expert, two of the five MoE specialists miss their target: the `head` expert reaches the head in only 50% of episodes (hands spend most of each episode on the upper torso), and the `legs` expert reaches the legs in 0/50 episodes (hands collapse onto head contact). The `arms` expert is a tooling rather than a training issue: arms is absent from MIMo’s `BODY_GROUPS` taxonomy, so its contacts do not register on the per-group readout. We therefore re-label every Stage-1 episode by its dominant realised region before passing it to Stage 3, so the downstream dataset labels reflect what the body actually touched rather than which expert was running. Composition rules are described in Section III-G.

B. Stage 2: Realised Datasets

Each Stage-1 policy produces several datasets, and Stage-3 trains one body schema per source. We report the realised properties of the 32 sources used downstream (Table VI, Appendix C).

Two observations follow. First, head contact is uninformative on its own: it fires in $\geq 99.9\%$ of frames in every dataset because MIMo lies in a supine crib pose and is in passive contact with the pillow regardless of policy. The dominant-region label must therefore ignore the head channel, which the strict-sub-region definition of Section V-A already does. Second, the arms region is the most-touched body group across heterogeneous policies (random, scripted, three of four holdouts), which is consistent with MIMo’s arm geometry making contact along the torso the path of least resistance: this is a biomechanical artefact rather than evidence of arm-directed behaviour. Both observations motivate the per-region paired comparisons in Section V-D, which control for the rest-pose floor explicitly.

C. Stage 3: Per-Region Representation Quality

a) Selected suite on both tiers: Table IV reports the body schema probe scores across 32 schemas (108 Tier A schema-seeds, 31 leak-free Tier B schema-seeds). On Tier A every load-bearing probe separates structured from baseline at $|d| \geq 0.52$, with RSA-joint and MIG-joint showing the largest effects ($d=+3.26$ and $+2.03$). On Tier B the structured/baseline gap attenuates sharply on TPD and IMR, and TRC saturates against its per-group F_1 base-rate bound; we unpack the attenuation below.

TABLE IV: Body schema probe scores. Each family column is the mean over schemas in that family. On Tier A, MoE $n=13$, QDAC $n=8$, Random/Babble $n=7$. On Tier B, only schemas with a seed-42 checkpoint contribute, so the family n 's shrink: MoE $n=6$ (specialists and aggregates lack a seed-42 schema), QDAC $n=8$, Random/Babble $n=4$ (babbling sources have no seed-42 schema). Yang is a single-run external reference (Section V-C), not re-evaluated on Tier B. The d columns are Cohen's d for structured (MoE + QDAC pooled) vs. baseline (random + babbling pooled): $|d| \approx 0.2$ small, ≈ 0.5 medium, ≥ 0.8 large. A positive d means structured is better; BFHP is an MSE so a negative d is the structured-wins direction. MIG-joint is the lite (linear- R^2) variant.

Metric	Tier A					Tier B			
	MoE	QDAC	Yang	Rand/Babb	d	MoE	QDAC	Rand/Babb	d
TRC (F_1)	0.789	0.718	0.579	0.687	+1.39	0.737	0.737	0.737	-0.80 [†]
TPD (ρ)	0.512	0.381	0.138	0.138	+2.17	0.222	0.224	0.214	+0.37
IMR (R^2)	0.545	0.435	0.381	0.372	+0.93	0.172	0.114	0.113	+0.37
BFHP (MSE)	0.0036	0.0026	0.0076	0.0041	-0.52*	—	—	—	— [‡]
MIG-joint (lite)	0.045	0.033	—	0.019	+2.03	0.029	0.039	0.038	-0.30
RSA-joint (ρ)	0.717	0.526	—	0.204	+3.26	0.732	0.772	0.618	+4.05

* BFHP Tier A fit on the subset of schema-seeds within the compute budget; other schemas excluded from this row. [†] Tier B TRC cluster on three unique values near the per-group F_1 base-rate bound; flat reading is base-rate saturation, not a transfer claim. [‡] BFHP Tier B is structurally skipped because the Tier B held-out h5 does not carry hand-position trajectories (the BFHP target).

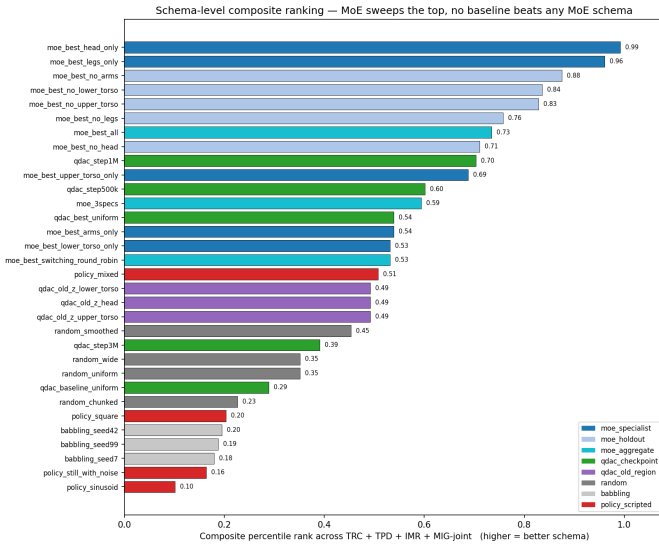


Fig. 2: Tier A schema-level composite ranking: mean percentile across TRC, TPD, IMR and MIG-joint. Bars are coloured by family.

b) Behaviour predicts representation: Table V reports the Spearman rank correlations between Stage-1 dataset descriptors and the selected probes across the 32 schemas. Movement coherence is the strongest behaviour $\rightarrow$ representation predictor ($\rho=+0.75$ with RSA-joint, $+0.70$ with MIG-joint, $+0.63$ with TPD), while pose vigour and bimanual fraction correlate negatively with the same probes (ρ as low as -0.61 on RSA-joint). Self-touch entropy is weakly positive overall ($\rho=+0.44$ on TRC, $+0.46$ on RSA-joint) but flips to a strong negative on BFHP.

c) Schema-level composite ranking: Figure 2 ranks each schema by its mean percentile across TRC, TPD, IMR and MIG-joint. MoE schemas occupy the top of the ranking, QDAC checkpoints occupy the middle, and the random, babbling and scripted families occupy the bottom.

d) Tier A $\rightarrow$ Tier B: On Tier A the family ranking is consistent across the discriminative probes (Figure 6). Tier B's leak-free held-out set covers TRC, TPD and IMR for the seed-42

schemas ($n=31$), plus MIG-joint and RSA-joint for the $n=18$ subset re-evaluated on the held-out h5. The five probes then tell five different transfer conclusions (Figure 7, Table IV). TPD and IMR both attenuate, indicating that action-grounded and two-point-discrimination signals do not survive distribution shift; the IMR collapse is asymmetric by family, with MoE schemas retaining modest recovery while QDAC schemas drop to the baseline level. TRC reads as flat, but every Tier-B value is on one of three numbers near the per-group F_1 base-rate, so the flat reading is base-rate saturation rather than transfer. MIG-joint reverses direction: the linear- R^2 disentanglement signal does not survive distribution shift even directionally. The one probe that transfers fully, and in fact sharpens, is RSA-joint: posture-geometry similarity is preserved out of distribution.

D. Discussion

a) Self-touch is necessary but not sufficient: The Yang column of Table IV reports a single-run external Yang et al. self-touch baseline [9] evaluated under our Stage-3 protocol. It improves over random/babbling on IMR but does not close the gap to MoE or QDAC; its TRC sits below the random/babbling mean and its MIG-joint score is at the random/babbling level. The diagnosis is that self-touch alone is not enough: the policy also has to move in a structured way. The Yang reference makes reliable contact, but its movement coherence is only 0.39, close to babbling (0.38) and far below the MoE (0.75) and QDAC (0.53) averages. Contact without structured motion produces a body schema roughly at the babbling level.

b) Behaviour shapes representation: magnitude and structure: Across the full six-metric suite (TRC, TPD, IMR, BFHP, MIG-joint, RSA-joint) structured exploration produces a measurably better body schema than random or babbling actions, with RSA-joint and MIG-joint showing the largest gaps ($d=+3.26$ and $+2.03$) and IMR the smallest among the load-bearing probes ($d=+0.93$). BFHP separates structured from baseline in the same direction but at a smaller magnitude ($d=-0.52$; lower MSE is better). The ranking is the same across metrics

TABLE V: Spearman rank correlations between Stage-1 dataset descriptors (rows) and selected VAE probes (columns), $n=32$ schemas, Tier A; BFHP sign-flipped so higher = better. BFHP $n=18$ (recorded hand-position trajectories not available for every source). Red positive, blue negative; saturation $\propto |\rho|$.

Dataset descriptor	TRC	TPD	IMR	BFHP	MIG-joint	RSA-joint
Self-touch entropy	+0.44	+0.18	+0.35	−0.63	+0.26	+0.46
Self-touch fraction	−0.10	−0.36	−0.04	−0.07	−0.48	−0.41
Bimanual fraction	−0.24	−0.47	−0.21	−0.15	−0.46	−0.61
Pose vigour	−0.22	−0.33	−0.21	+0.12	−0.55	−0.60
Movement coherence	+0.43	+0.63	+0.38	−0.07	+0.70	+0.75

(Figure 6) and at the schema level (Figure 2): MoE schemas occupy the top of the composite, QDAC checkpoints occupy the middle, and the baselines occupy the bottom. The research question therefore has a positive answer: behaviour shapes representation, and at the 150-episode budget used here the gap between the best and worst behaviour is several standard deviations on every probe in the suite.

c) *The body-frame channel is the weakest link*: BFHP, the only kinematic probe in the suite, separates structured from baseline at $d=-0.52$: correct direction, but a fraction of the magnitude of the touch- and action-side probes. The matching kinematic signal on Tier B (IMR, action-from-latent-change) also collapses ($d: +0.93 \rightarrow +0.37$, $R^2: 0.50 \rightarrow 0.15$). The latent encodes touch geometry well but loses kinematic detail. Touch and proprio carry equal nominal weight in the loss (Eq. 5), but touch is $\sim 45\times$ higher-dimensional (17,175 vs. 380), so its reconstruction error dominates the gradient and the proprio side is under-fit. BFHP additionally covers fewer schema-seeds than the other probes due to compute limits.

d) *Structured movement, not motor effort, drives representation quality*: As shown in Table V: movement coherence is the only descriptor positively and consistently correlated with every probe in the suite; pose vigour and the contact-coverage pair (self-touch fraction and bimanual fraction, collinear at $\rho=+0.84$) are negative. Because movement coherence is the fraction of proprioceptive variance in the leading principal components, a high value means the motion is structured and low-dimensional; a low value means it is diffuse and high-dimensional. The interpretation is therefore: what shapes the body schema is whether the agent moves in a coordinated, structured way, not how energetically it actuates or how much skin it contacts. The babbling, random and scripted families sit at the bottom of movement coherence and at the bottom of every selected probe; the MoE specialists, which produce coordinated region-targeted movement, sit at the top of both.

e) *MoE versus QDAC: structured movement outweighs targeting accuracy*: MoE wins five of six probes; QDAC wins only BFHP. This inverts Stage 1, where QDAC has the higher per-region contact rate (Table III). The descriptor side resolves the inversion: QDAC moves vigorously inside a narrow postural corridor (high pose vigour, low movement coherence), while MoE moves with coordinated, lower-dimensional structure. BFHP is the lone QDAC win because it reads hand position rather than joint configuration, which QDAC’s vigorous reaching is well-suited to produce; every other probe scores configuration structure, where MoE wins. The takeaway is that data structure matters more than targeting accuracy on every

probe except the kinematic one.

f) *QDAC training trajectory: post-peak representational collapse*: On every discriminative probe the QDAC checkpoint trajectory traces a U-curve (Figure 3): representation quality climbs through 1M steps, plateaus near 1.5M, and collapses back to baseline by 3M. Per- z controllability falls in parallel ($2/5 \rightarrow 1/5$ reachable groups between 1.5M and 3M, Section V-A), so the policy is narrowing to a single mode while its SAC return continues to climb. This is the same long-horizon collapse that motivates pretrained-foundation-model approaches such as Meta Motivo [22], which trains its policy on a large behaviour buffer rather than from a single online reward signal. The practical consequence: a reward-greedy checkpoint selector picks a model 1–2M steps past the representational peak, whereas the controllability-grid selector (Section III-F0c) picks within the peak window.

E. Strengths, Limitations, and Challenges

a) Strengths:

- *Controlled comparison*. Every dataset is consumed by an identical training and evaluation pipeline, so differences in Table IV are attributable to the data rather than to training-side confounds.
- *Two independent data sources reduce bias*. The per-region MoE and the skill-conditioned QDAC actor are trained with different objectives, yet both feed the same downstream pipeline; the behaviour→representation map (Table V) is computed across both families.
- *Explicit data-dependence quantification*. The behaviour→representation map is measured rather than asserted (Table V), so the behavioural axes that drive representation quality are identified rather than guessed.
- *Reward-peak $\neq$ controllability-peak*. On QDAC the controllability-grid selector picks the checkpoint inside the representational peak window, where reward-greedy selection picks one past the post-peak collapse (Figure 3).

b) Limitations:

- *Single SSL family*: only the multi-branch VAE was evaluated; no contrastive, forward-only, or masked-prediction baseline.
- *$n=2$ VAE seeds per dataset*; SE bands across seeds are wide and statistical claims at the per-source level are tentative.
- *Feet are excluded from the contact regions, and joint cross-region contact is under-sampled in every dataset*.

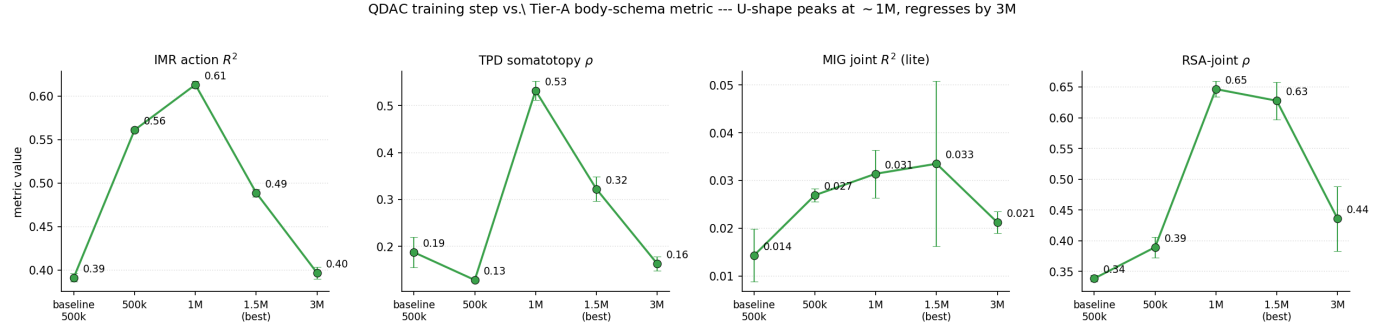


Fig. 3: QDAC body-schema metric vs. training step (tier2_seed7). All four probes trace the same U-shape: performance climbs from baseline through the 1M checkpoint, plateaus near the peak at 1.5M (the controllability-grid “best”), and regresses by 3M. MIG-joint here is the lite (linear- R^2) variant, consistent with Table IV.

- Region-based dataset labels are a coarse axis; within-region structure (e.g. proximal vs. distal arm) is not resolved.

c) Challenges and negative results:

- Vanilla RL on the MIMo body is hard; the per-region decomposition was motivated by this difficulty.
- QD descriptors (ϕ_1, ϕ_2) failed to separate body-schema-relevant variance, forcing the pivot to region-based dataset labels.
- Pilot one-hot QDAC prompt collapsed to head-only contact in 8/8 seeds; fixed by a continuous rostro-caudal Gaussian prompt (Section III-F).
- Reward-greedy QDAC checkpoint selection returns an uncontrollable policy; replaced by a 50-episode controllability-grid selector (Section III-F0c).
- Offline FB-CPR [22] on the BabyBench buffer collapsed to single-region attractors with diverging Q-values; behavioural-foundation-model recipes designed for clean locomotion do not transfer to contact-rich bodies in pure-offline form.

VI. CONCLUSIONS

Summary

We built a three-stage pipeline that asks whether the behaviour an embodied infant produces shapes the body-schema representation a self-supervised model learns from it. Stage 1 generates self-touch with a mixture-of-experts and a skill-conditioned QDAC actor; Stage 2 emits one dataset per Stage-1 policy, labelled by realised body region; Stage 3 trains a fresh dual-VAE body schema on each dataset under one fixed recipe and evaluates with a six-metric suite (TRC, TPD, IMR, BFHP, MIG-joint, RSA-joint) on Tier A (in-distribution) and leak-free Tier B. Across the 32 schemas, structured exploration beats random and babbling baselines on every selected probe, movement coherence is the strongest behaviour-side predictor of representation quality, and the QDAC body-schema metric collapses back to baseline well before SAC return plateaus. A controlled Yang et al. self-touch reference [9] is right between babbling and the structured families, confirming that reliable contact alone is not enough.

Insights

- *Behaviour shapes representation, measurably.* Structured exploration beats random and babbling baselines on every load-bearing probe, and the schema-level composite ranking puts every MoE schema above every baseline schema. The result is robust to which probe carries the headline.
- *Behaviour as a data-efficiency lever.* Every source in the comparison consumes the same Stage-3 training recipe and a comparable episode budget (~ 50 – 200 per source); the only moving part is the behaviour that produced the trajectories. Choosing a structured behaviour delivers a body schema several standard deviations better than a random or babbling one at the same data budget, with no architecture, loss, or compute change. For embodied self-supervised learning, designing how the agent moves is therefore a cheaper lever than scaling the data factory.
- *Structured movement wins, vigorous motion does not.* Movement coherence (the fraction of proprioceptive variance in the top principal components, i.e. how low-dimensional and coordinated the motion is) is the strongest positive predictor across the suite; pose vigour, the mean per-step joint-angle change, is the strongest negative one. The lever for better body schemas is whether the body moves in a coordinated, structured way, not how hard it moves.
- *Self-touch alone is not enough.* The Yang et al. self-touch reference baseline is stronger than pure babbling on some representation axes, but it still falls far short of MoE/QDAC on TPD, MIG-joint and RSA-joint. Reliable contact is therefore not the whole story; the policy also needs broad, varied proprioceptive coverage.
- *Reward-peak $\neq$ controllability-peak.* On QDAC, SAC return keeps climbing while prompt-following and downstream representation quality regresses, so the reward-greedy checkpoint is essentially uncontrollable. A controllability-grid selector picks inside the representational peak window (Figure 3).
- *Cross-distribution transfer is metric-specific.* On Tier B, the action-grounded probes (TPD, IMR) attenuate sharply, TRC saturates at the per-group F_1 base-rate, and MIG-joint even reverses direction. The one channel that transfers cleanly, and in fact sharpens, is RSA-joint: posture-

geometry similarity is what survives distribution shift, whereas the action-grounded and disentanglement-style readouts do not.

Limitations and Lessons Learned

Two lessons we would carry into a follow-up. First, pick the downstream metric before designing the data factory: we chose Stage-1 descriptors first and only later realised the evaluation suite required region structure, which made much of the QD descriptor search redundant. Second, once a pipeline is locked, compute is better spent on additional seeds than on further ablations; an addition body schema seed across every dataset would have moved several per-source effects out of the noise floor. The headline methodological lesson is that reward-peak does not coincide with controllability-peak: the best touch-generator is not the best representation source.

Future Work

A natural next step is to repeat the comparison under non-reconstruction objectives such as JEPA and contrastive methods, to see whether the finding generalises beyond the VAE. Stage-2 descriptors could also extend beyond touch region into the temporal and frequency domains. The same pipeline could eventually move from MIMo to a real platform, where the morphological-adaptability argument of Section I becomes testable rather than motivational.

REFERENCES

- [1] M. Hoffmann, H. G. Marques, A. Hernández Arieta, H. Sumioka, M. Lungarella, and R. Pfeifer, “Body schema in robotics: A review,” *IEEE Trans. Auton. Ment. Develop.*, vol. 2, no. 4, pp. 304–324, 2010.
- [2] P. Rochat, “Self-perception and action in infancy,” *Experimental Brain Research*, vol. 123, no. 1–2, pp. 102–109, 1998.
- [3] A. DiMercurio, J. P. Connell, M. Clark, and D. Corbetta, “A naturalistic observation of spontaneous touches to the body and environment in the first 2 months of life,” *Frontiers in Psychology*, vol. 9, p. 2613, 2018.
- [4] P. D. H. Nguyen, Y. K. Georgie, E. Kayhan, M. Eppe, V. V. Hafner, and S. Wermter, “Sensorimotor representation learning for an “active self” in robots: A model survey,” *KI – Künstliche Intelligenz*, vol. 35, pp. 9–35, 2021.
- [5] F. Gama and M. Hoffmann, “The homunculus for proprioception: Toward learning the representation of a humanoid robot’s joint space using self-organizing maps,” *arXiv preprint arXiv:1909.02295*, 2019.
- [6] S. L. Li, A. Zhang, B. Chen, H. Matusik, C. Liu, D. Rus, and V. Sitzmann, “Controlling diverse robots by inferring Jacobian fields with deep networks,” *Nature*, vol. 643, no. 8070, pp. 89–95, 2025.
- [7] D. Mattern, P. Schumacher, F. M. López, M. C. Raabe, M. R. Ernst, A. Aubret, and J. Triesch, “MIMo: A multi-modal infant model for studying cognitive development,” *IEEE Trans. Cogn. Develop. Syst.*, vol. 16, no. 4, pp. 1291–1301, 2024.
- [8] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *ICLR*, 2014.
- [9] L. Yang and L. Jamone, “A novel computational framework based on curiosity and self-touch rewards for multimodal body schema learning,” 2025, first Place, BabyBench Competition at IEEE ICDL 2025, Prague.
- [10] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *ICML*, 2018, pp. 1861–1870.
- [11] L. Grillotti, M. Faldor, B. G. León, and A. Cully, “Quality-diversity actor-critic: Learning high-performing and diverse behaviors via value and successor features critics,” in *ICML*, 2024.
- [12] F. M. López, V. Marcel, X. Hinaut, J. Triesch, and M. Hoffmann, “BabyBench: A multimodal benchmark of infant behaviors for developmental AI,” 2025, IEEE ICDL 2025 Competition, Prague. [Online]. Available: <https://babybench.github.io/2025/>
- [13] F. Mancini, A. Bauleo, J. Cole, F. Lui, C. A. Porro, P. Haggard, and G. D. Iannetti, “Whole-body mapping of spatial acuity for pain and touch,” *Annals of Neurology*, vol. 75, no. 6, pp. 917–924, 2014.
- [14] W. Penfield and E. Boldrey, “Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation,” *Brain*, vol. 60, no. 4, pp. 389–443, 1937.
- [15] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, “An internal model for sensorimotor integration,” *Science*, vol. 269, no. 5232, pp. 1880–1882, 1995.
- [16] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, “Curiosity-driven exploration by self-supervised prediction,” in *ICML*, 2017.
- [17] M. S. A. Graziano, “Where is my arm? the relative role of vision and proprioception in the neuronal representation of limb position,” *Proceedings of the National Academy of Sciences*, vol. 96, no. 18, pp. 10418–10421, 1999.
- [18] A. Iriki, M. Tanaka, and Y. Iwamura, “Coding of modified body schema during tool use by macaque postcentral neurones,” *NeuroReport*, vol. 7, no. 14, pp. 2325–2330, 1996.
- [19] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, “Isolating sources of disentanglement in variational autoencoders,” in *NeurIPS*, 2018.
- [20] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, “Challenging common assumptions in the unsupervised learning of disentangled representations,” in *ICML*, 2019.
- [21] N. Kriegeskorte, M. Mur, and P. A. Bandettini, “Representational similarity analysis – connecting the branches of systems neuroscience,” *Frontiers in Systems Neuroscience*, vol. 2, p. 4, 2008.
- [22] A. Tirinzoni, A. Touati, J. Farebrother, M. Guzek, A. Kanervisto, Y. Xu, A. Lazaric, and M. Pirota, “Zero-shot whole-body humanoid control via behavioral foundation models,” in *ICLR*, 2025, arXiv:2504.11054.
- [23] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, “Hindsight experience replay,” in *NeurIPS*, 2017, pp. 5048–5058.
- [24] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, “Self-supervised learning from images with a joint-embedding predictive architecture,” in *CVPR*, 2023, pp. 15 619–15 629.
- [25] R. Balestriero and Y. LeCun, “LeJEPA: Provable and scalable self-supervised learning without the heuristics,” 2025, arXiv:2511.08544.
- [26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *ICML*, 2020, pp. 1597–1607.
- [27] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in *NeurIPS*, 2017, pp. 6306–6315.

APPENDIX A ALTERNATIVES CONSIDERED

For the behaviour-generation algorithm: Two alternatives were considered and rejected. A behaviour-conditioned generative VAE (trajectories in and out, latent selecting region) would have replaced the Stage-1 RL stack with a single decoder, but the unsupervised latent has no guarantee of organising around touch regions rather than other features, and the quality of such a model would also require a good dataset which is absent in our setting. Goal-conditioned SAC with hindsight experience replay [23] is the other natural baseline, but our shaped reward remains reachable via random movements, so the reward itself is not truly sparse and HER is not a good application here.

For the body-schema model family: Several non-VAE families were considered. We preferred VAE for one practical reason: its decoder lets us debug reconstruction directly, whereas encoder-only models require a dedicated probe for every property of interest. JEPA [24] is encoder-only and additionally requires a stop-gradient and EMA target encoder to avoid collapse, both of which add complexity and data demand. Regularisation alternatives such as LeJEPA [25] have been proposed but are recent and not yet broadly tested. Contrastive

methods such as SimCLR [26] are similarly reasonable, but the augmentation design for paired proprio/taxel observations is non-trivial and implementing it was beyond the scope of this work. VQ-VAE [27] was not attempted but is the most interesting next baseline, codebook quantisation could in principle align discrete codes with body regions automatically, which we shall leave for future investigation.

APPENDIX B

FIXES THAT RESOLVED QDAC HEAD-ONLY COLLAPSE

Five fixes together turn the original 8/8 head-collapse into all-region coverage in 10^6 steps:

- Continuous rostro-caudal prompt (Eq. 3): every prompt has reachable support.
- Force-weighted descriptors: a ϕ -style dense Bellman signal in place of binary contact.
- Learning-progress curriculum across the prompt axis.
- Capped 50/50 hindsight relabelling [23] against the achieved descriptor; uncapped relabelling leaked Q-values across prompts.
- Standing scene where random actions actually produce contact, so warm-up exploration has gradient signal at all.

APPENDIX C

SUPPLEMENTARY FIGURES AND TABLES

This section gathers the five floats deferred from the main text: the Stage-1 encoder-actor architecture (Fig. 4), the per-region contact-rate heatmaps for the three policy families (Fig. 5), the Stage-2 dataset-size summary (Table VI), the per-family Tier-A means on TPD, IMR and MIG-joint (Fig. 6), and the structured-versus-baseline comparison on Tier A and Tier B (Fig. 7).

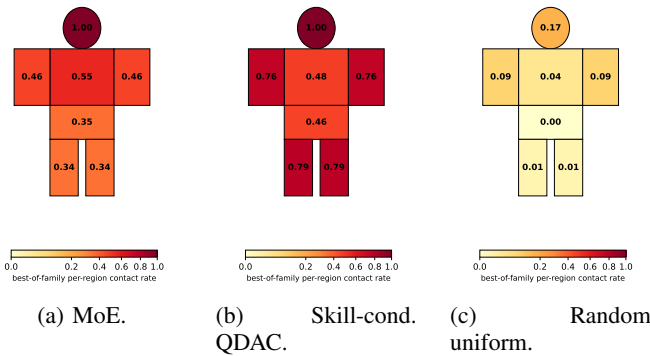


Fig. 5: Per-region contact rate on strict sub-regions of the frontal MIMo silhouette. Colour scale is PowerNorm ($\gamma=0.55$); values in cells.

TABLE VI: Stage-2 dataset sizes (Tier A). n_{src} is the number of source datasets per family (one Stage-3 schema is trained per source). n_{ep} is the Stage-3 episode count per source; each episode is 1000 steps.

Family	n_{src}	$n_{\text{ep}}/\text{src.}$
MoE specialist	5	50
MoE holdout	5	52
MoE aggregate	3	50–200
QDAC checkpoint	5	150
QDAC old (per-region)	3	150
Random	4	150
Babbling	3	50
Policy scripted	4	150

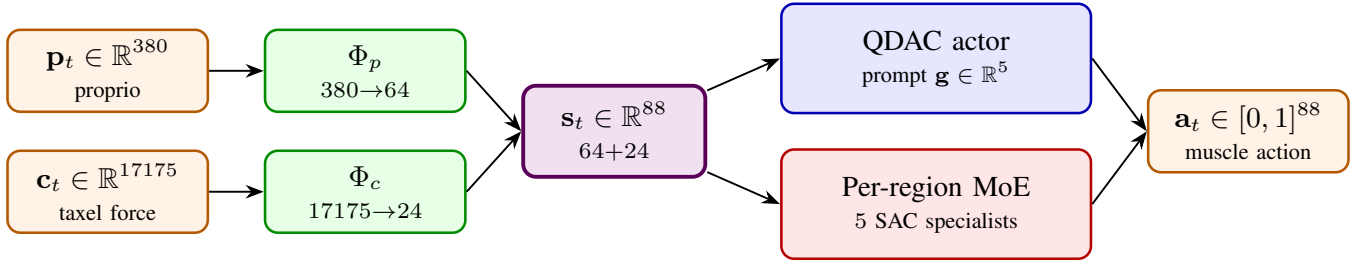


Fig. 4: Behaviour-generation architecture. The frozen multimodal encoder $\Phi = (\Phi_p, \Phi_c)$ compresses raw proprioception and tactile force readings into a fused 88-D state s_t , which is consumed in parallel by the QDAC actor and by the per-region mixture of experts. Both emit an 88-D muscle action a_t .

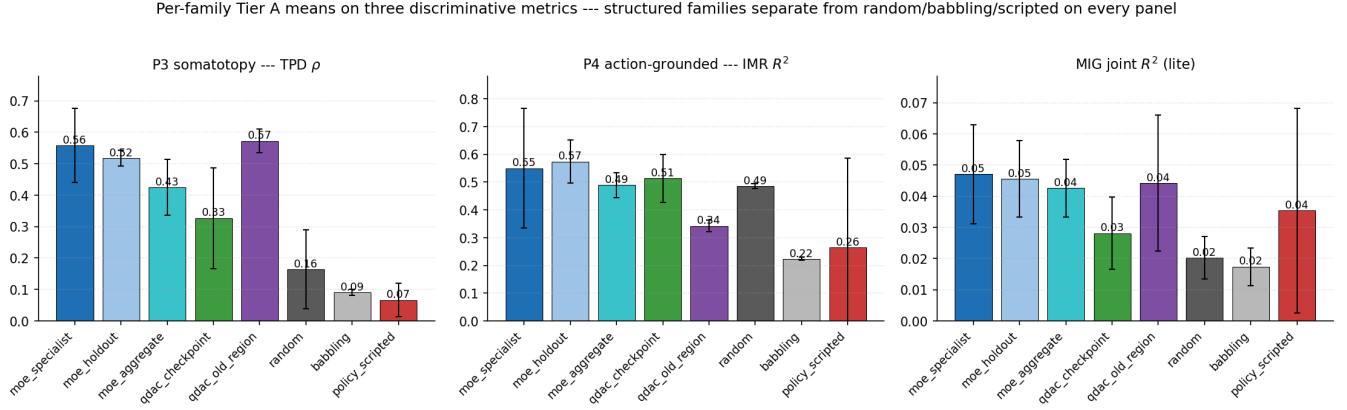


Fig. 6: Per-family Tier A means on TPD, IMR and MIG-joint (lite); error bars ± 1 s.d. across schema-seeds. The four MoE/QDAC families sit above the random / babbling / scripted floor on TPD and IMR. On MIG-joint (lite) the structured/baseline gap is preserved but compressed, so the metric is less discriminating between adjacent structured families.

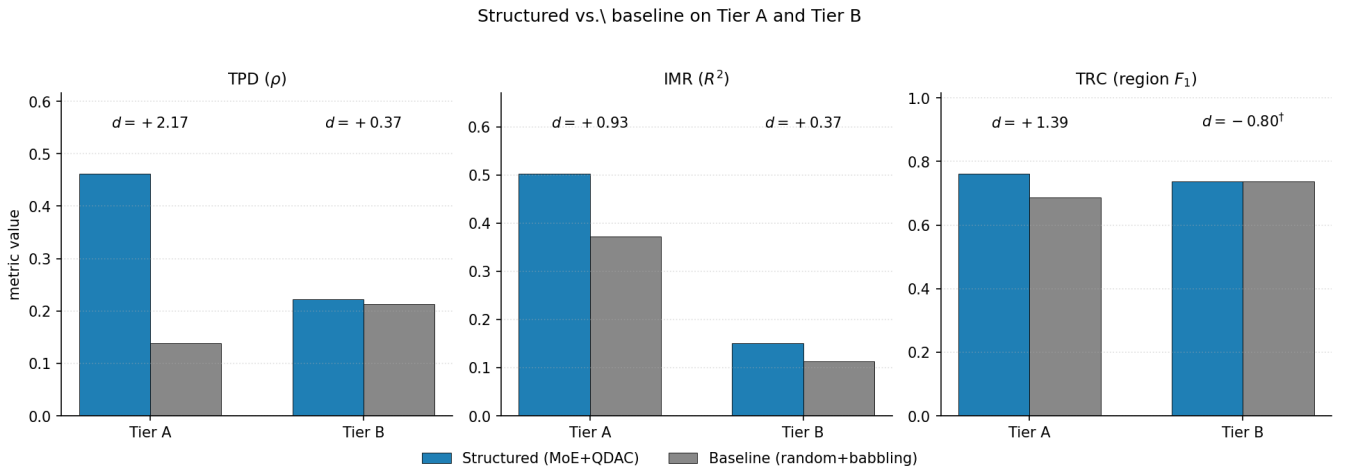


Fig. 7: Structured vs. baseline on Tier A and Tier B for the three probes with leak-free Tier B coverage. TRC Tier B is base-rate-saturated ([†]); the TPD and IMR collapses are the transfer signal.