Disentangled Skill Representations for Predictive Human Modeling

Mariah Schrum, Deepak Gopinath, Srijan Srivatsa, Guy Rosman, Tiffany Chen

Toyota Research Institute
Los Altos, California, USA

Abstract

Understanding human skill is important for AI systems that collaborate with, coach, or assist people. Unlike typical latent variable estimation problems which rely on single observations, skill is a persistent, compositional, and behaviorally grounded construct that must be inferred from patterns over time. We introduce Skill Abstraction with Interpretable Latents (**SAIL**), a method for modeling human skill as an interpretable, multi-dimensional construct inferred from naturalistic behavior. Our approach produces a skill embedding that is robust to transient performance fluctuations and learns a transferable representation of human subskills. Furthermore, **SAIL** supports skill-informed behavior prediction that generalizes across a variety of in-domain contexts. We represent each individual with a persistent skill embedding that controls a blend between expert and novice bases and is trained using counterfactual subskill swaps for disentanglement. This design encourages representations that are both robust to performance variation and structured for interpretability. We demonstrate across racing and baseball that **SAIL** achieves strong predictive performance and consistently improves behaviorally grounded disentanglement over the evaluated baselines, while also improving downstream AI coaching performance.

1 Introduction

AI systems that support, collaborate with, or coach humans must reason about human skill to personalize instruction, anticipate behavior, and adapt assistance over time. Unlike many latent variables in machine learning, human skill cannot be inferred from individual actions or outcomes. Instead, it is a persistent, behaviorally grounded, and compositional construct that must be inferred from patterns across repeated interactions while accounting for noise, variability, and changing task conditions (Iso-Ahola 2024; Langley, Cummings, and Shapiro 2004).

Human skill differs from many notions of "skill" used in machine learning. In robotics and reinforcement learning, skill often refers to reusable action primitives or policies for task execution (Lesort et al. 2018). In contrast, we model human skill as a persistent, participant-level construct composed of multiple interpretable subskills (Newell 1991; Ericsson, Krampe, and Tesch-Römer 1993). We further distinguish skill from performance: performance reflects trial-specific outcomes influenced by situational factors such as fatigue or risk-taking, whereas skill represents stable abilities that generalize across contexts (Iso-Ahola 2024; Fitts and Posner 1967). Conflating the two can lead to inaccurate assessment and inappropriate interventions.

We propose that an effective skill representation should satisfy three desiderata:

(1) **Construct Validity** (Messick 1995): representations should remain stable across sessions and robust to trial-level noise.

(2) **Predictive Utility**: representations should support accurate behavior prediction across in-domain contexts.

(3) **Interpretability**: representations should decompose into disentangled subskills that correspond to human-recognizable aspects of expertise.

To satisfy these desiderata, we introduce **S**kill **A**bstraction with **I**nterpretable **L**atents (**SAIL**), a computational framework for representing human skill. Rather than learning an unconstrained latent embedding and hoping it reflects skill, SAIL incorporates inductive biases inspired by theories of skill acquisition: skill is persistent across observations, expressed through behavior, progresses relative to expertise, and is composed of interpretable subskills.

Intuitively, we ask how an individual's behavior differs from characteristic novice and expert behaviors, rather than asking the representation to explain every observed trajectory directly. This constrains the representation to encode structured, skill-relevant variation instead of behavioral fluctuations. To encourage interpretability, we supervise subskill-specific latent slices with behaviorally grounded metrics and introduce a counterfactual training procedure that encourages each latent slice to represent a distinct subskill. Importantly, behavior prediction serves as a supervision signal for learning the representation rather than the primary objective.

In this work we contribute the following:

1. We formulate human skill modeling as a representation learning problem and identify construct validity, predictive utility, and interpretability as key desiderata.
2. We propose **SAIL**, a participant-level skill representation that combines participant embeddings, novice–expert basis blending, and counterfactual supervision to learn stable, predictive, and interpretable skill representations.
3. We demonstrate across racing and baseball that **SAIL** outperforms strong baselines and improves downstream instructor-feedback prediction by 10%.

2 Related Work

Human skill has been studied across education, sports science, robotics, and human–AI interaction (Anderson 2014; Ericsson, Krampe, and Tesch-Römer 1993). Unlike task performance, skill is a persistent, latent construct that must be inferred from behavior accumulated over time rather than individual outcomes (Newell 1991; Schmidt et al. 2018). Traditional measures such as completion time or accuracy (Fitts and Posner 1967) are highly context dependent and often reflect transient *performance* rather than underlying skill. Psychometric approaches, including Item Response Theory and Bayesian Knowledge Tracing, estimate related latent constructs (Embretson and Reise 2013; Corbett and Anderson 1994; Piech et al. 2015), but are designed for discrete responses rather than continuous behavioral trajectories.

Trajectory-based approaches, including clustering and inverse reinforcement learning, infer latent structure from demonstrations (Ziebart et al. 2008; Abbeel and Ng 2004). Likewise, work in robot teaching and reinforcement learning often represents "skills" as reusable action primitives or control policies (Argall et al. 2009; Cakmak and Thomaz 2012; Hausman et al. 2018; Petangoda et al. 2019; Dave and Rueckert 2025). While effective for policy learning, these methods do not model skill as a persistent, interpretable construct that generalizes across repeated observations.

Representation learning methods, including autoencoders, variational autoencoders, and contrastive learning, have been widely used to encode human behavior (Kingma and Welling 2013; van den Oord, Li, and Vinyals 2018; Zhang et al. 2019). More recently, participant-level representations have been explored to model persistent latent characteristics across repeated interactions, supporting personalization, adaptive human–AI interaction, and human behavior modeling (Jacques et al. 2019; Gopinath, Jain, and Argall 2017; DeCastro et al. 2024; Jeon, Losey, and Sadigh 2020; Schrum, Hedlund-Botti, and Gombolay 2023). Disentangled representation learning further seeks to recover interpretable latent factors (Higgins et al. 2017; Chen et al. 2016; Kim and Mnih 2018), although purely unsupervised objectives do not guarantee semantic alignment or identifiability (Locatello et al. 2019). While these methods demonstrate the value of participant-specific representations and disentangled latent spaces, they generally optimize downstream prediction or personalization rather than explicitly modeling human skill as a persistent, interpretable construct.

Our work differs by modeling *human skill* as a persistent participant-level representation that is explicitly optimized for construct validity, predictive utility, and interpretable subskill decomposition. By combining participant-specific embeddings, novice–expert basis blending, and counterfactual subskill supervision, **SAIL** learns representations that are stable across repeated observations, predictive across contexts, behaviorally interpretable, and useful for downstream personalization tasks.

3 Approach

Problem Formulation: We aim to learn a latent representation of human *skill* from behavioral data. Let $\mathcal{D} = \{\tau_1, \ldots, \tau_N\}$ denote a set of trajectories, where each $\tau_i = \{x_i^t\}_{t=1}^{T_i}$ is a sequence of feature vectors $x_i^t \in \mathbb{R}^D$ executed in a task context $c \in \mathcal{C}$ (e.g., racetrack or batting condition). We assume contexts c_i are observed at both train and test time and that an individual's skill is transferable across contexts, while its behavioral expression depends on c. Trajectories may include multimodal features such as vehicle telemetry, gaze, or body kinematics.

Our goal is to infer an individual-specific skill embedding $z_s \in \mathbb{R}^d$ that is stable across trajectories and transferable across in-domain contexts (task instances drawn from the same domain). We distinguish *skill*, a persistent construct, from *performance* (Iso-Ahola 2024), which reflects trial-specific outcomes and is sensitive to situational factors. We represent skill as *compositional* in which z_s decomposes into interpretable subcomponents $z_s^{(k)}$ corresponding to distinct subskills, consistent with motor learning theories that describe skill as arising from multiple interacting components (Newell 1991; Anderson 1982).

To connect subskills with behavior, we use *skill metrics* $m \in \mathcal{M}$. Skill metrics provide noisy behavioral proxies. These metrics are derived from trajectories, expert annotations, or auxiliary tasks. Multiple metrics may map to the same subskill and they provide supervision for learning structured representations of z_s (described in Sec 3.3). Our approach is detailed in Alg. 1 and described below.

3.1 Participant-Specific Skill Embedding

Human skill is a persistent characteristic of an individual rather than a single behavioral observation. Inferring skill independently from each trajectory therefore conflates stable ability with trial-specific factors such as fatigue, measurement noise, environmental variation, and strategy. Moreover, trajectory-level embeddings are not inherently tied to the individual who produced them.

To model persistence, **SAIL** assigns each training participant a learnable skill embedding $z_s \in \mathbb{R}^d$, optimized jointly with the model parameters (Alg. 1, Line 1). A single embedding is shared across all trajectories from the same participant, pooling evidence across repeated observations to capture stable behavioral tendencies despite trial-to-trial variability. This is conceptually similar to participant embeddings used in recommender systems and speaker recognition (Koren, Bell, and Volinsky 2009; Snyder et al. 2018).

However, persistence alone does not imply skill: a participant embedding could simply summarize average behavior. In the following section, we introduce an additional inductive bias by constraining behavior to be generated relative to canonical novice and expert behavior bases, encouraging z_s to encode expertise rather than arbitrary behavioral variation.

At test time, the model parameters (g_ϕ, q_θ, and h_ψ) are frozen, and only the participant embedding is optimized using the trajectory reconstruction loss.

To prevent collapse of the embedding, we introduce an auxiliary network q_θ that reconstructs z_s from generated trajectories. This provides a variational lower bound on the mutual information between z_s and predicted behavior and encourages the embedding to encode information that is both

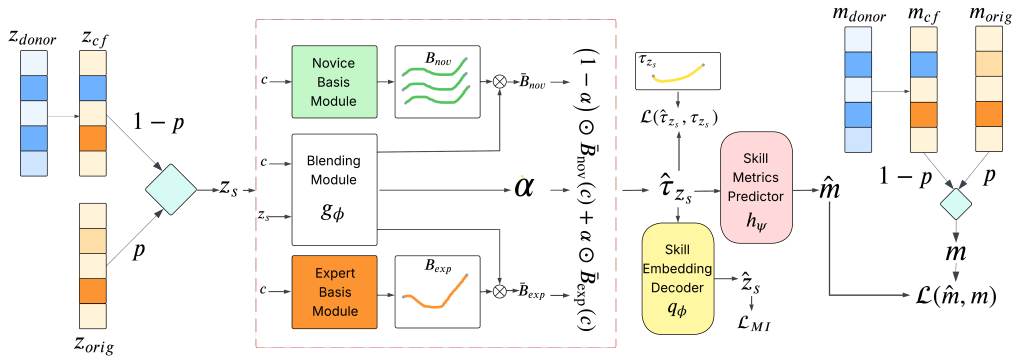

Figure 1: Overview of **SAIL**. Each participant is associated with a persistent skill embedding z_s learned across multiple behavioral observations. Rather than decoding trajectories directly, z_s predicts behavior by blending canonical novice and expert basis trajectories, encouraging the representation to capture stable skill-related variation instead of transient behavioral fluctuations. The embedding is partitioned into subskill-specific slices that are supervised using behaviorally grounded skill metrics and disentangled through counterfactual subskill swaps.

Algorithm 1: Training **SAIL**

Require: Trajectories τ_i, contexts c_i, subskill metrics m_i
Initialize participant embeddings $z_{s,i} \sim \mathcal{N}(0, 0.1)$
for each training iteration **do**
  Sample batch of participants and trajectories
  Predict behavior $\hat{\tau}_{z_s}$ via expert–novice blending (Sec. 3.2)
  Decode predicted subskill metrics $\hat{m} = h_\psi(\hat{\tau}_{z_s})$
  Compute total loss $\mathcal{L} = \lambda_{\text{traj}}\mathcal{L}_{\text{traj}} + \lambda_{\text{metric}}\mathcal{L}_{\text{metric}} + \lambda_{\text{MI}}\mathcal{L}_{\text{MI}}$
  if CF step (probability $1-p$) **then**
    Swap $(z^{(k)}_{\text{orig}}, m^{(k)}_{\text{orig}}) \leftarrow (z^{(k)}_{\text{donor}}, m^{(k)}_{\text{donor}})$
    Reconstruct CF trajectory $\tilde{\tau}_{\text{orig}}$ from $\tilde{z}_{\text{orig}}$
    Skip trajectory reconstruction loss; apply metric loss only for swapped subskill k
  end if
  Update model parameters and participant embeddings jointly via back-propagation
end for

behaviorally meaningful and recoverable from observed trajectories (Kingma and Welling 2013; Chen et al. 2016).

3.2 Skill Representation via Novice–Expert Basis Blending

A participant-specific embedding provides a persistent representation of an individual, but persistence alone does not imply that the embedding represents *skill*. Without additional inductive structure, the embedding may instead encode an individual's average behavior, preferred driving style, or other participant-specific characteristics unrelated to expertise. The challenge is therefore to constrain the representation so that it explains the stable behavioral variation associated with skill while remaining insensitive to transient performance fluctuations.

A straightforward approach is to decode trajectories directly from the learned skill embedding. However, this requires the embedding to account for every aspect of the observed behavior, including variability arising from fatigue, measurement noise, environmental conditions, and idiosyncratic execution. Consequently, the learned representation is encouraged to memorize trajectories rather than isolate the latent factors responsible for expertise.

Instead, we model behavior *relative to canonical novice and expert behaviors*. Rather than asking the embedding to generate a trajectory from scratch, we ask it to explain where an individual's behavior lies relative to representative novice and expert executions for the task context. This imposes an inductive bias: the embedding need only encode deviations associated with expertise, while the basis trajectories explain common behavioral structure shared across participants. As a result, transient variation and stylistic differences are less likely to be absorbed into the skill representation.

For each task context c (e.g., a racetrack or batting condition), we define sets of canonical novice and expert basis trajectories,

$$B_{\text{exp}}(c) = \{B^{(i)}_{\text{exp}}(c)\}_{i=1}^{M}, \qquad B_{\text{nov}}(c) = \{B^{(i)}_{\text{nov}}(c)\}_{i=1}^{K},$$

where each basis trajectory, $B^{(i)} \in \mathbb{R}^{T\times D}$, represents a characteristic mode of behavior observed near the extremes of the skill distribution. The basis trajectories may be obtained from demonstrations, learned jointly with the model, or generated by an optimal controller. In our implementation, we define the expert basis using trajectories from the demonstrator with the strongest domain-specific performance measure and derive the novice basis by applying principal component analysis (PCA) to novice trajectories. The resulting novice bases capture the dominant modes of variation among inexperienced participants, while the expert basis provides a canonical target behavior.

The participant embedding is mapped by g_ϕ to expert and novice basis weights, $w_{\text{exp}}(z_s, c)$ and $w_{\text{nov}}(z_s, c)$, together with an interpolation coefficient $\alpha(z_s, c) \in [0, 1]^{T\times D}$. The basis weights are constrained to the simplex so that the predicted behavior is expressed as an element-wise convex interpolation of the expert and novice bases (Fig. 1):

$$\bar{B}_\bullet(z_s, c) = \sum_j w_\bullet^{(j)}(z_s, c) B_\bullet^{(j)}(c), \quad \bullet \in \{\text{exp}, \text{nov}\},$$
$$\hat{\tau}_{z_s} = \alpha \odot \bar{B}_{\text{exp}}(z_s, c) + (1 - \alpha) \odot \bar{B}_{\text{nov}}(z_s, c). \tag{1}$$

Although behavior is expressed as a blend of novice and expert bases, our formulation does not assume that skill lies on a single linear axis. Multiple novice bases capture diverse low-skill strategies (e.g., overcautious, inconsistent, or poorly timed behavior), while multiple expert bases can represent distinct high-skill styles. Furthermore, each subskill independently modulates its own blending coefficients, enabling complex, nonlinear representations of skill.

Unlike direct trajectory decoding, this formulation encourages the embedding to explain behavior in terms of deviations from canonical novice and expert behaviors rather than memorizing every trajectory detail. Behavior prediction serves as a supervision signal that encourages the embedding to capture stable, skill-related structure while remaining predictive of behavior.

3.3 Counterfactual Training for Subskill Disentanglement

Human skill is inherently *compositional*: coaches reason about performance in terms of multiple interacting subskills and design interventions that target individual deficiencies (Ericsson, Krampe, and Tesch-Römer 1993; Newell 1991; Wulf 2016; Anderson 1982). Accordingly, we partition the latent representation into subskill-specific components that should independently influence the behaviors associated with each subskill. This requires both *disentanglement* (independent latent factors) and *identifiability* (each factor corresponds to a human-recognizable subskill).

Existing disentanglement methods (e.g., InfoGAN, β-VAE, FactorVAE) encourage statistical independence but do not ensure that latent dimensions correspond to meaningful subskills or support selective behavioral interventions (Higgins et al. 2017; Kim and Mnih 2018; Locatello et al. 2019). Conditional supervision associates latent dimensions with labels (Kingma et al. 2014; Sohn, Lee, and Yan 2015), but changing a supervised latent need not produce the expected behavioral change.

To address this limitation, we explicitly partition the embedding into subskill-specific slices and train the representation using counterfactual interventions. During training, one subskill slice is replaced with that of another participant while the remaining slices are held fixed. The model is then required to produce behavior that reflects only the substituted subskill, encouraging both disentanglement and identifiability. The embedding space is partitioned as

$$z_s = \left[\, z_s^{(1)},\ z_s^{(2)},\ \ldots,\ z_s^{(K)} \,\right],$$

where each slice $z_s^{(k)} \in \mathbb{R}^{d_k}$ is intended to represent subskill k, and $\sum_k d_k = d$.

Reconstructed trajectories $\hat{\tau}_{z_s}$ are passed through a predictor network h_ψ to produce subskill metrics $\hat{m}$ (Fig. 1) that serve as behaviorally grounded supervision signals during training. Each subskill metric is defined in collaboration with domain experts. Skill metrics reflect a measurable behavioral quantity that serves as a proxy for an underlying subskill (e.g., steering smoothness for control or gaze dispersion for visual attention). These metrics provide weak yet semantically meaningful supervision that anchors each subskill dimension to interpretable aspects of human behavior.

To enforce CF consistency, we perform subskill swaps between a randomly chosen pair of training examples: an *original* sample (the one being modified) and a *donor* sample (the one borrowed from). For a subskill k, we replace the k-th slice of the original embedding with that of the donor:

$$\tilde{z}_{\text{orig}}^{(k)} = z_{\text{donor}}^{(k)}, \quad \tilde{z}_{\text{orig}}^{(\ell)} = z_{\text{orig}}^{(\ell)} \ \ \forall \ell \neq k,$$

and apply the same operation to the associated skill metrics to ensure supervision remains consistent:

$$\tilde{m}_{\text{orig}}^{(k)} = m_{\text{donor}}^{(k)}, \quad \tilde{m}_{\text{orig}}^{(\ell)} = m_{\text{orig}}^{(\ell)} \ \ \forall \ell \neq k.$$

In practice, we interleave CF and standard training. With probability p, a batch is trained using the regular reconstruction and metric objectives, and with probability $(1 - p)$, a batch is trained with CF swaps (Alg. 1, Lines 7–8). This procedure creates CF examples where the original z_s retains all except one subskill slice which is borrowed from the donor. Doing so allows the model to learn how isolated subskills should influence predicted behavior and skill metrics.

This approach encourages reconstruction fidelity while also promoting disentanglement. Since no ground-truth trajectory exists for this CF, we do not apply a reconstruction loss to $\hat{\tau}_{z_s}$ for the swapped items (Alg. 1 Line 10). Instead, the predictor network, h_ψ, is required to output the swapped metric for subskill k, thus forcing the model to adjust behavior in a way that matches the intervention.

Unlike approaches that impose constraints directly on the latent space (Lin et al. 2020), our method encourages disentanglement through behavior. By requiring reconstructed trajectories to predict subskill metrics during CF swaps, each latent slice is forced to encode its designated subskill.

3.4 Modeling Details and Losses

The overall training objective encourages predictive accuracy, semantic alignment, and disentanglement:

$$\mathcal{L} = \lambda_{\text{traj}} \mathcal{L}_{\text{traj}} + \lambda_{\text{metric}} \mathcal{L}_{\text{metric}} + \lambda_{\text{MI}} \mathcal{L}_{\text{MI}}. \tag{2}$$

Here, $\mathcal{L}_{\text{traj}}(\hat{\tau}_{z_s}, \tau)$ is a trajectory reconstruction loss between the predicted trajectory $\hat{\tau}_{z_s}$ and the observed trajectory τ. $\mathcal{L}_{\text{metric}}(h_\psi(\hat{\tau}_{z_s}), m)$ supervises behaviorally grounded subskill metrics by comparing predicted metrics $h_\psi(\hat{\tau}_{z_s})$ to targets m. Finally, $\mathcal{L}_{\text{MI}} = -\mathbb{E}_{\hat{\tau} \sim p_\phi(\hat{\tau} | z_s, c)} \left[\log q_\theta(z_s \mid \hat{\tau})\right]$ is a mutual-information objective. The mutual-information term is implemented by re-encoding the predicted trajectory $\hat{\tau}_{z_s}$ through q_θ to obtain $\hat{z}_s$, and encourages the embedding z_s to be recoverable from generated behavior. During CF training steps, $\mathcal{L}_{\text{traj}}$ is omitted since no ground-truth trajectory exists for the swapped embedding, and only the metric loss for the swapped subskill is applied.

Our model integrates skill embeddings with trajectory and context encoders from established sequence architectures. The trajectories are predicted via two decoders, which produce elementwise blending weights for basis blending. h_ψ uses an LSTM to predict the skill metrics from $\hat{\tau}_{z_s}$.

4 Domains and Datasets

We evaluate **SAIL** in two domains with substantially different movement dynamics and subskill structure: high-performance racing and baseball batting. Both domains require coordinated mastery of multiple interacting subskills and provide measurable behavioral outcomes, making them suitable testbeds for evaluating human skill representations.[1]

4.1 High-Performance Racing

High-performance racing is a compelling domain for studying skill because it requires the integration of multiple subskills to achieve mastery. We focus on six core subskills identified by expert coaches and prior work (Schrum et al. 2025): (i) vehicle handling, (ii) gaze control, (iii) know-how, (iv) control inputs, (v) physical ability, and (vi) perceptual ability. These subskills correspond to how professional coaches diagnose driver weaknesses and training interventions.

Each trajectory τ_i consists of vehicle pose, speed, and control signals downsampled to 100 points per track segment. The context c for this dataset refers to the racetrack that the trajectory was performed on. We collected a dataset of racing trajectories from 95 participants spanning novices to experts, using a driving simulator. We collected data in two phases: 70 participants each completed at least ten laps on a single track modeled after a nearby raceway, and 25 participants completed four laps on each of four distinct tracks at the same venue. This design provided both breadth (a large participant pool) and depth (multiple laps and multiple contexts). In total we collected 1545 laps.

To connect observed behavior to underlying subskills, we used a set of behaviorally grounded skill metrics $m \in \mathcal{M}$, defined in collaboration with expert coaches in prior work (Schrum et al. 2025). Each metric is derived from a task designed to probe a specific subskill. For example, peak lateral g-force in a skidpad drill reflects vehicle handling, gaze fixation during driving sessions reflects gaze policy, and written test scores reflect know-how of racing lines and other HPD techniques. These metrics (among others) provide partial, noisy evidence about latent subskills and provide the supervision signals necessary for learning disentangled representations of z_s.

4.2 Baseball Hitting

We applied **SAIL** to a supplemental dataset of baseball hitting collected from 13 players on a competitive adult team in a semi-professional league. While all participants were experienced players, they were not at the level of an expert benchmark and thus exhibited substantial variation across subskills. In collaboration with a coach, one highly skilled participant was identified as an expert and used to define the canonical expert basis for blending, while the remaining players provided a diverse set of trajectories. In total, 74 batting trials were recorded, across both pitching machine sessions and tee batting conditions. Whole-body kinematics of swing motions were captured using an optical motion capture system. The coach identified three core subskills and associated metrics of hitting: (i) the *kinematic chain*, or the sequential transfer of momentum across body segments; (ii) *pelvis pausing*, or the ability to momentarily stabilize the pelvis to build rotational power; and (iii) *thigh pausing*, or the controlled deceleration of the lead thigh. The contexts c are tee batting and machine-pitch batting.

To address the limited size of the dataset, we generated synthetic participants by applying trajectory augmentations (time warping, noise injection, and scaling) to the data. For players with both tee and machine-pitch trials, we estimated a global offset between conditions and used it to synthesize additional regular swings. This produced artificial batting trials that preserved the underlying structure while introducing diversity.

To our knowledge, there are no existing datasets that capture multimodal behavioral signals and skill metrics that are comparable in richness to our racing dataset. Unlike the racing dataset, the baseball dataset is smaller, narrower in subskill coverage, and augmented with synthetic trials. We therefore treat this baseball dataset as a supplemental, secondary domain to test the generality of **SAIL**.

5 Results

Since our contribution is a representation learning method rather than a policy-learning or imitation-learning algorithm, we compare **SAIL** against established representation learning baselines designed to evaluate latent representations. Composite scores are shown for Racing (Fig. 2), our primary domain, while Baseball results are reported in Table 1 as a supplemental domain.

SimCLR (contrastive baseline). A self-supervised method that uses contrastive losses to encourage invariance within an individual. We adapt SimCLR to trajectory data to test whether a contrastive objective is sufficient for extracting skill-relevant embeddings (Chen et al. 2020).

β-VAE (disentanglement baseline). An extension of the VAE with stronger KL regularization that encourages factorized latents. We include β-VAE as a disentanglement method to test whether standard disentanglement approaches yield interpretable subskills (Higgins et al. 2017).

AE (autoencoder baseline). A standard trajectory autoencoder (Hinton and Salakhutdinov 2006) that captures per-trial variability but is not designed to model persistent skill or subskill structure.

AE-LC (AE with linear constraints). An extension of the AE framework that incorporates linear constraints derived from subskill metrics to encourage semantically meaningful and identifiable latents. We include this method to test whether metric-based structure alone can recover interpretable subskills compared to **SAIL** (Lin et al. 2020).

Ablation: without CF training (SAIL w/o CF). This ablation removes the CF swap objective and trains only with behavioral prediction via expert–novice basis blending to isolate the contribution of CF supervision.

Ablation: without expert–novice basis and CF training (SAIL w/o basis). This ablation decodes trajectories directly from the skill embedding without basis blending or CF supervision to test whether the basis decomposition is necessary for isolating skill-related variation from transient factors.

[1]The human-subjects data collection protocol was approved by WCG IRB in June 2023.

Table 1: Results in Racing (R) and Baseball (B). Higher is better for ↑, lower is better for ↓. Bold = best. Values report mean (standard error) across evaluation folds.

	SAIL (ours)		**SAIL w/o CF**		**SAIL w/o basis**		**SimCLR**		**β-VAE**		**AE**		**AE-LC**	
	R	B	R	B	R	B	R	B	R	B	R	B	R	B
Construct Validity														
Silhouette (↑)	.72 (.08)		**.77 (.16)**		.67 (.07)		.40 (.44)		.74 (.16)		.75 (.18)		.67 (.18)	
Test–retest similarity (↑)	**.995 (.003)**	**1.000 (.003)**	.995 (.001)	.998 (.001)	.990 (.004)	.995 (.003)	.928 (.10)	.958 (.119)	.928 (.02)	.498 (.330)	.839 (.12)	.979 (.090)	.891 (.22)	.998 (.001)
Overall Construct Score (↑)	1.86	**1.00**	**2.0**	.994	1.69	.992	.57	.919	1.49	0.00	.95	.961	1.06	.996
Predictive Utility														
Behavior prediction (RMSE ↓)	**2.75 (.12)**	.161 (.070)	2.75 (.12)	**.155 (.075)**	5.05 (.26)	.176 (.088)	4.15 (.99)	.204 (.087)	4.48 (.36)	.479 (.158)	4.61 (.38)	.257 (.117)	4.87 (.55)	.191 (.093)
OOC generalization (RMSE ↓)	**6.37 (3.0)**	.120 (.053)	6.50 (3.1)	**.115 (.061)**	12.77 (3.1)	.115 (.059)	10.27 (5.0)	.395 (.090)	12.51 (3.5)	.768 (.128)	12.42 (3.1)	.152 (.087)	14.12 (2.8)	.153 (.094)
Overall Predictive Score (↑)	**2.0**	1.94	1.98	**2.00**	.17	1.89	.89	.98	.46	0.00	.41	.81	.08	.95
Disentanglement & Interpretability														
Alignment Ratio (AR ↑)	**3.25 (.79)**	**0.758 (.025)**	1.24 (.083)	0.416 (.023)	1.11 (.21)	0.301 (.057)	1.73 (.78)	0.087 (.074)	1.12 (.11)	0.046 (.052)	1.05 (.14)	0.387 (.030)	2.40 (.17)	0.724 (.029)
Targeted Change Index (TCI ↑)	**.93 (.06)**	**0.146 (.019)**	.89 (.10)	0.139 (.013)	.73 (.05)	0.128 (.010)	.45 (.10)	0.132 (.011)	.63 (.03)	0.140 (.021)	.78 (.06)	0.134 (.022)	.63 (.03)	0.144 (.023)
Relative Influence Ratio (RIR ↑)	**2.11 (.46)**	**1.244 (.054)**	1.76 (.19)	1.129 (.177)	1.67 (.22)	1.104 (.107)	1.85 (1.14)	1.101 (.076)	1.86 (.23)	0.731 (.267)	1.61 (.32)	1.203 (.049)	1.56 (.28)	1.122 (.072)
Overall Interpretability Score (↑)	**3.0**	**2.85**	1.37	1.82	.81	1.08	.83	.97	.95	1.00	.78	1.68	1.20	2.46

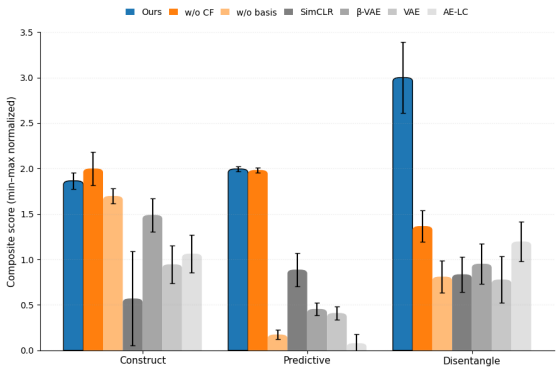

Figure 2: Composite scores across the three desiderata in Racing. Bars show performance of (**SAIL**), ablations, and baselines. Higher is better for all desiderata.

For all baselines that operate at the trial level (SimCLR, AE, β-VAE, AE-LC), we extract embeddings per trajectory and pool across laps for each participant which produces a participant-level embedding comparable to our method. We evaluate our approach and baselines along the three desiderata introduced in Section 1: (1) construct validity, (2) predictive utility , and (3) disentanglement and interpretability. For each desideratum, we compute a composite score by min–max normalizing each metric across methods, reversing lower-is-better metrics, and summing the normalized values. We also evaluate whether the learned representation improves a downstream AI coaching model and validate the representation against a professional coach's ratings.

5.1 Construct Validity

We evaluate *construct validity* by measuring whether the learned embedding captures stable, skill-relevant structure rather than transient fluctuations (Table 1). We operationalize construct validity as stable within-participant and discriminative across-skill representations. Because no coaching occurred during data collection, we assume participants' underlying skill remained approximately constant. We evaluate construct validity using silhouette score and test–retest similarity.

- **Silhouette score** (↑): clustering quality by skill group.
- **Test–retest similarity** (↑): stability of embeddings across repeated trials.

Discussion: As shown in Figure 2 and Table 1, among the evaluated methods, **SAIL** achieves strong overall construct validity across both racing and baseball. **SAIL** produces highly stable embeddings (test–retest similarity of 0.995 in racing and 1.000 in baseball), indicating that z_s captures persistent aspects of skill rather than trial-level variability. The no-CF ablation performs similarly, suggesting that counterfactual supervision preserves construct validity while primarily benefiting interpretability. In contrast, removing the novice–expert basis reduces clustering quality, likely because the embedding captures more trial-specific variation. Overall, these results suggest that participant-level embeddings and basis blending contribute to learning stable skill representations. We do not report silhouette scores for baseball because discrete skill labels are unavailable.

5.2 Predictive Utility

We next investigate *predictive utility* by evaluating if the learned skill embeddings support accurate trajectory prediction within and across contexts. Predictive utility is a key desideratum, because it indicates whether **SAIL** can be used to anticipate behavior for a given skill and how behavior will change under novel conditions. We evaluate predictive utility using in-context (trained and tested on the same set of racetracks, with held-out trials) and out-of-context (trained on one racetrack, tested on different track) prediction metrics.

- **In-context prediction (RMSE ↓):** trajectory accuracy within the same context.
- **Out-of-context prediction (RMSE ↓):** generalization to novel contexts.

Discussion: As shown in Figure 2 and Table 1, **SAIL** achieves the best predictive performance in racing and remains competitive in baseball. Removing counterfactual (CF) supervision has little effect on prediction, indicating that CF primarily improves interpretability. In contrast, removing novice–expert basis blending nearly doubles prediction error in racing, suggesting that the basis is the primary source of predictive generalization by separating stable skill from transient variation. Together, these results indicate that participant-level embeddings and basis blending drive predictive utility, while CF selectively improves disentanglement.

Table 2: Downstream coaching-instruction prediction (mean $\pm$ SE over 15 participant-held-out folds). **SAIL** significantly outperforms the trial-time baseline on weighted F1 (paired $t(14)$=2.50, p=.025) and accuracy (p=.028); on macro F1, **SAIL** is the only condition that significantly improves over no conditioning (p<.001).

Conditioning	Weighted F1 ↑	Macro F1 ↑	Acc. ↑
None	.541 (.015)	.504 (.018)	.538 (.014)
Trial time	.573 (.014)	.520 (.020)	.567 (.014)
SAIL (ours)	**.595 (.014)**	**.537 (.018)**	**.589 (.013)**

5.3 Disentanglement and Interpretability

Finally, we evaluate whether the representation decomposes into interpretable subcomponents that correspond to distinct subskills via alignment ratio, targeted change index, and relative influence ratio metrics (Table 1): Together, these metrics evaluate disentanglement along three complementary axes: semantic alignment (AR), selective intervention effects (TCI), and relative influence on outputs (RIR).

- **Alignment Ratio (AR ↑):** measures how well each subskill slice $z_s^{(k)}$ predicts its intended metrics compared to non-target ones, indicating subskill–metric correspondence (Eastwood and Williams 2018).
- **Targeted Change Index (TCI ↑):** operationalizes the idea of intervention selectivity described in Bengio et al. (2019) and Schölkopf et al. (2021) and quantifies the effect of CF swaps by checking whether trajectory changes are concentrated in the targeted features, with higher values reflecting more selective control.
- **Relative Influence Ratio (RIR ↑):** complements TCI by perturbing one subskill at a time and measuring how much this changes an expected behavioral feature, relative to the change induced by perturbing other subskills.

Discussion: Figure 2 and Table 1 show that **SAIL** consistently achieves the strongest disentanglement across both domains. Removing counterfactual (CF) supervision substantially reduces all interpretability metrics, demonstrating that CF training is the primary mechanism for learning semantically meaningful subskill representations. Although AE-LC incorporates explicit metric supervision, it consistently underperforms **SAIL**, indicating that supervision alone is insufficient to produce behaviorally grounded, selectively controllable representations.

5.4 Skill-Informed Coaching

Finally, we evaluate the downstream utility of **SAIL** by incorporating the learned skill representation into a previously proposed imitation-learning model for predicting instructor feedback (Gopinath et al. 2025). We augment the original model with a frozen participant embedding computed from the participant's previous four laps, and compare against conditioning on a scalar baseline (trial time) over the same window. We evaluate on two previously collected simulator coaching datasets comprising 38 participants (Sumner et al. 2026; Schrum et al. 2026). As shown in Table 2, conditioning on the **SAIL** embedding yields a 10.0% relative improvement in weighted F1 over the unconditioned model and significantly outperforms trial-time conditioning on weighted F1 and accuracy. Unlike trial time, **SAIL** also significantly improves macro F1, suggesting that the representation captures information beyond overall ability and improves prediction across both common and infrequent instruction categories. These results suggest that access to z_s enables more accurate prediction of both when and what instructors will coach.

5.5 External Validation by Professional Coach

To assess whether the learned representation aligns with expert human judgment, we compare **SAIL** skill estimates against independent ratings from a professional driving coach who evaluated 22 held-out participants over the course of a coaching study. These ratings were collected independently of the drill-based metrics used for training supervision. To obtain a scalar overall skill estimate from **SAIL**, we project each participant's embedding onto the direction between the mean novice and expert embeddings and convert the resulting position to a percentile, where 0 and 100 correspond to the novice and expert reference points, respectively. The coach's overall skill ratings agree strongly with **SAIL**'s overall skill estimate (Spearman $\rho = 0.81$, $p < .001$, 95% bootstrap CI $[0.56, 0.94]$, $n = 22$), indicating that the embedding aligns well with human expert judgments of skill. Because these coach ratings and downstream coaching labels were not used as supervision for learning the representation, these results provide evidence that SAIL captures information beyond the behaviorally grounded metrics used during training.

6 Limitations

Our evaluation is limited by dataset scale and scope, particularly in the baseball domain where data are small and augmented. The method also depends on noisy, predefined subskill metrics, and assumes a smooth novice–expert continuum that may miss certain strategies. While metrics are defined in collaboration with domain experts, they may be incomplete or biased. Our evaluation focuses on practical properties of a useful skill representation—stability, predictive utility, interpretability, and agreement with expert judgment—rather than establishing a unique or complete computational definition of human skill. Future work should investigate additional forms of construct validation and longitudinal studies of skill acquisition.

References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 1–8.

Anderson, J. R. 1982. Acquisition of cognitive skill. *Psychological Review*, 89(4): 369–406.

Anderson, J. R. 2014. *Learning and memory: An integrated approach*. John Wiley & Sons.

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A Survey of Robot Learning from Demonstration. *Foundations and Trends in Robotics*, 1(4): 1–157.

Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A.; and Pal, C. 2019. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. *arXiv preprint arXiv:1901.10912*.

Cakmak, M.; and Thomaz, A. L. 2012. Designing Robot Learners that Ask Good Questions. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 17–24.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*, 1597–1607.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*.

Corbett, A. T.; and Anderson, J. R. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4: 253–278.

Dave, V.; and Rueckert, E. 2025. Skill Disentanglement in Reproducing Kernel Hilbert Space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16153–16162.

DeCastro, J.; Silva, A.; Gopinath, D.; Sumner, E.; Balch, T. M.; Dees, L.; and Rosman, G. 2024. Dreaming to Assist: Learning to Align with Human Objectives for Shared Control in High-Speed Racing. In *Conference on Robot Learning (CoRL)*.

Eastwood, C.; and Williams, C. K. I. 2018. A Framework for the Quantitative Evaluation of Disentangled Representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Embretson, S. E.; and Reise, S. P. 2013. *Item response theory for psychologists*. Psychology Press.

Ericsson, K. A.; Krampe, R. T.; and Tesch-Römer, C. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3): 363–406.

Fitts, P. M.; and Posner, M. I. 1967. *Human performance*. Brooks/Cole.

Gopinath, D.; Cui, X.; DeCastro, J.; Sumner, E.; Costa, J.; Yasuda, H.; Morgan, A.; Dees, L.; Chau, S.; Leonard, J.; Chen, T.; Rosman, G.; and Balachandran, A. 2025. Computational Teaching for Driving via Multi-Task Imitation Learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 7019–7027.

Gopinath, D.; Jain, S.; and Argall, B. 2017. Human-in-the-Loop Optimization of Shared Autonomy in Assistive Robotics. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 3925–3932.

Hausman, K.; Springenberg, J. T.; Wang, Z.; Heess, N.; and Riedmiller, M. 2018. Learning an Embedding Space for Transferable Robot Skills.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786): 504–507.

Iso-Ahola, S. E. 2024. A theory of the skill-performance relationship. *Frontiers in Psychology*, 15: 1296014.

Jacques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P. A.; Strouse, D.; Leibo, J. Z.; and de Freitas, N. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 3040–3049. PMLR.

Jeon, H. J.; Losey, D. P.; and Sadigh, D. 2020. Shared autonomy with learned latent actions. *arXiv preprint arXiv:2005.03210*.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*.

Kingma, D. P.; Rezende, D. J.; Mohamed, S.; and Welling, M. 2014. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. In *Computer*, volume 42, 30–37. IEEE.

Langley, P.; Cummings, K.; and Shapiro, D. 2004. Hierarchical Skills and Cognitive Architectures. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 779–784.

Lesort, T.; Díaz-Rodríguez, N.; Goudou, J.-F.; and Filliat, D. 2018. State representation learning for control: An overview. *Neural Networks*, 108: 379–392.

Lin, X.; Thekumparampil, K. K.; Fanti, G.; and Oh, S. 2020. Learning Semantically Meaningful Embeddings Using Linear Constraints. In *International Conference on Learning Representations (ICLR)*.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 4114–4124.

Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9): 741–749.

Newell, K. M. 1991. Motor skill acquisition. *Annual Review of Psychology*, 42(1): 213–237.

Petangoda, J. C.; Gammulle, H.; Denman, S.; Fookes, C.; Sridharan, S.; et al. 2019. Disentangled skill embeddings for reinforcement learning. *arXiv preprint arXiv:1906.09223*.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, volume 28.

Schmidt, R. A.; Lee, T. D.; Winstein, C.; Wulf, G.; and Zelaznik, H. N. 2018. *Motor Learning and Performance: From Principles to Application*. Human Kinetics.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5): 612–634.

Schrum, M.; Morgan, A.; Gopinath, D.; Costa, J.; Sumner, E.; Rosman, G.; and Chen, T. 2025. A Data-Driven Framework for Skill Representation. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI) Workshops*. LEAP-HRI Workshop paper.

Schrum, M. L.; Hedlund-Botti, E.; and Gombolay, M. 2023. Reciprocal MIND MELD: Improving Learning From Demonstration via Personalized, Reciprocal Teaching. In Liu, K.; Kulic, D.; and Ichnowski, J., eds., *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, 956–966. PMLR.

Schrum, M. L.; Srivatsa, S.; Dees, L.; Dixon, E.; Reyes Gomez, P.; Gopinath, D.; Sumner, E. S.; Rosman, G.; and Chen, T. L. 2026. Skill Modulates Coaching Language in Embodied Motor Learning. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, 1–6. New York, NY, USA: Association for Computing Machinery.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. IEEE.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Sumner, E.; Gopinath, D. E.; Dees, L.; Reyes Gomez, P.; Cui, X.; Silva, A.; Costa, J.; Morgan, A.; Schrum, M.; Chen, T. L.; Balachandran, A.; and Rosman, G. 2026. SimCoach-Corpus: A Naturalistic Dataset with Language and Trajectories for Embodied Teaching. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Wulf, G. 2016. Attentional focus and motor learning: a review of 15 years. *International Review of Sport and Exercise Psychology*, 9(1): 77–104.

Zhang, S.; Li, H.; Gan, H.; Xu, L.; and Zhang, X. 2019. Self-supervised learning for human activity recognition using 700,000 accelerometer records. *IEEE Transactions on Mobile Computing*, 20(9): 2424–2437.

Ziebart, B. D.; Maas, A.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *AAAI*, 1433–1438.