GENERATIVE INTERACTIONS: WEAVING MULTIPARTY HUMAN MOTION WITH BILEVEL LATENT DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human social behaviour is not a collection of independent motions, but a jointly organised process in which group dynamics and individual variation continuously shape one another. Yet existing social motion models often prioritise plausible trajectories while leaving interaction state implicit, limiting their ability to transfer across groups, tasks, and partial-observation regimes. To address this gap, we introduce Bilevel Representations for Agent Interaction Dynamics (BRAID), a hierarchical sequential latent-variable model for generative multi-person interaction. BRAID explicitly formulates social motion generation as a meta-transfer learning problem: shared interaction priors are learned across datasets and adapted through arbitrary context sets of observed people and joints. The model represents each scene through a group-level latent state that captures shared interaction dynamics and person-level latent states that capture individual behaviour conditioned on the evolving group context. This modelling choice enables coherent generation under full, sparse, or partial observations while exposing compact social-state vectors that can serve as an interface for downstream embodied-agent systems. We evaluate BRAID under a unified SMPL-based representation on social forecasting, tracking and in-filling, and response generation, using metrics that assess not only reconstruction accuracy but also realism, diversity, temporal alignment, and interpersonal coordination. We further analyse the hierarchical latent space, showing that it captures separable group- and individual-level structure. Code is given here.

1 INTRODUCTION

Human social interaction couples group-level coordination with individual behavioural variation. Embodied AI systems that perceive or participate in these interactions must track how a situation is unfolding, how participants influence one another, and which continuations are plausible. This motivates learning structured social-state representations alongside realistic motion generation. These representations must account for non-verbal signals, including posture, head nods, gestures, and gaze (Sacks et al., 1974; Terry et al., 1999; Oers et al., 2005; Kleef et al., 2007; Matsumoto, 2007; Smith & Louis, 2009; Pruitt & Riechert, 2011; Levinson & Torreira, 2015; Zanlungo et al., 2017; Raman et al., 2023; Rathbone et al., 2023; Markhorst et al., 2026), while accommodating multiple valid responses shaped by individual traits, group roles, cultural contexts, and situational norms.

Learning transferable social structure is difficult with scarce and fragmented interaction data. Motion capture datasets such as Panoptic (Joo et al., 2016), DnD (Mughal et al., 2024), and Embody3D (McLean et al., 2025) cover only a fraction of real-world situations and interaction styles. As Raman et al. (2023) argue, the challenge is therefore not only data scarcity but transfer: models must learn structure that generalises across datasets, contexts, and observation regimes.

Existing human motion generators (Tevet et al., 2022a;b; Guo et al., 2022; Diomataris et al., 2024) provide strong foundations for individual movement. Social motion methods address interaction, but often leave the interplay between shared dynamics and individual traits implicit (Wang et al., 2021; Chew et al., 2025; Lin et al., 2026), with limited adaptability to varying context. The central gap is a reusable social-state representation that jointly captures group and individual dynamics, supports generation under full or partial observations, and transfers across datasets. These capabilities are rarely addressed together: plausible trajectories remain the primary endpoint, while compact,

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

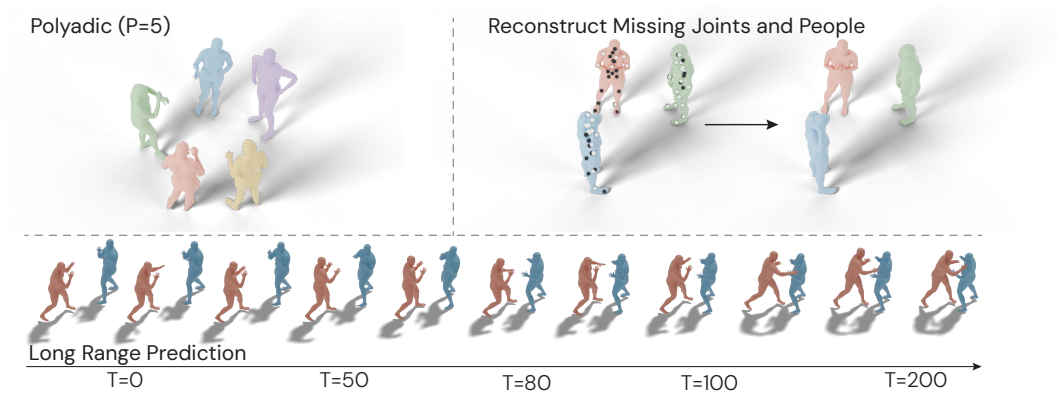


Figure 1: **A generative model for multi-agent interaction.** We propose Bilevel Representations for Agent Interaction Dynamics (BRAID), a unified model of multi-party group behaviour.

inspectable state representations could also inform policies, planners, and memory systems that condition on ongoing interaction.

To address this gap, we propose BRAID (*Bilevel Representations for Agent Interaction Dynamics*), a hierarchical sequential latent-variable model for multi-person social behaviour. A group latent captures shared interaction dynamics, while person latents encode individual behaviour conditioned on the group state. Their coupled temporal evolution supports coherent generation, and arbitrary context sets accommodate full, sparse, or partial observations. We learn shared priors through cross-dataset meta-transfer learning across conversational, dance, and boxing data. The resulting group and person latents expose compact, inspectable social-state vectors for downstream agents.

Our main contributions can be summarised as: (i) A generative state space model for multi-person social behaviour under full, partial, or absent observations. (ii) A hierarchical social-state representation of shared group dynamics and individual behavioural variation. (iii) Evaluation across datasets and context regimes, including partially observed interaction partners in a scene, and (iv) analysis of group- and person-level latent structure, informed by cognitive science.

2 Related Work

Table 1: Comparison of methods across interaction capabilities. Latent space denotes an explicit compact latent representation of behaviour; polyadic denotes support for $P \geq 3$ participants.

Method	Missing/ Noisy data	Latent space	Partner inpaint	Partner predict	Long motion	Agentic generation	Joint future prediction	Polyadic $P \geq 3$
DuoLando (2024)	×	×	✓	×	×	×	×	×
ReMoS (2023)	×	×	✓	×	×	×	×	×
ReGenNet (2024)	×	×	✓	×	×	×	×	×
Human-X (2025)	×	×	✓	×	×	×	×	×
ARFlow (2025)	×	×	×	✓	✓	×	×	×
Ready-to-React (2025)	×	×	×	✓	✓	×	×	×
MAGNet (2025)	×	×	✓	✓	✓	✓	✓	✓
Social Processes (2023)	×	✓	×	✓	×	×	✓	✓
BRAID	✓	✓	✓	✓	✓	✓	✓	✓

Human motion generation. Recent motion generation models produce high-quality single-person motions from text, action labels, or goals (Tevet et al., 2022b;a; Guo et al., 2022; Diomataris et al., 2024). These systems provide strong priors for individual behaviour, but they generally treat the person as the unit of generation. Social settings require a different abstraction: a pause, gesture, or

step is meaningful because it is coordinated with other participants. BRAID builds on this progress but models the interaction, rather than an isolated body.

Social motion generation. Social motion methods explicitly target interaction through multi-person prediction, inpainting, response generation, or specialised domains such as gesture, dance, and reactive character motion (Wang et al., 2021; Tanke et al., 2023; Mughal et al., 2024; Siyao et al., 2024; Maluleke et al., 2025; Cen et al., 2025; Chew et al., 2025). Adjacent to this literature, Social Processes (Raman et al., 2023) forecast nonverbal social cues in conversational groups by meta-learning group-specific dynamics, but do not address full-body social motion generation. Together, these approaches have substantially broadened the modelling of social interaction, but most still treat plausible future cues or trajectories as the primary endpoint. Group-level state is either absent, implicit in symmetric multi-agent architectures, or embedded inside a denoising process that does not naturally expose a compact representation for inspection or downstream integration (Preechakul et al., 2021; Pandey et al., 2022; Chen et al., 2022; Tanke et al., 2025). BRAID instead learns explicit group and person-level latents, z_t^g and z_t^p, so that shared interaction dynamics can condition individual behaviour while remaining available as reusable social-state representations.

Neural Processes and meta-learning. Neural Processes model distributions over functions using a latent variable conditioned on a context set (Garnelo et al., 2018b;a; Kim et al., 2019). This makes them well suited to few-shot generalisation and arbitrary missing-observation patterns. Sequential Neural Processes add temporal dependence between latent states (Singh et al., 2019), and Social Processes (Raman et al., 2023) show that this family can model conversational group dynamics from sparse interaction data. BRAID extends this line by replacing the single latent state with a temporal hierarchy: a group latent for shared interaction structure and person latents for individual behaviour conditioned on that structure.

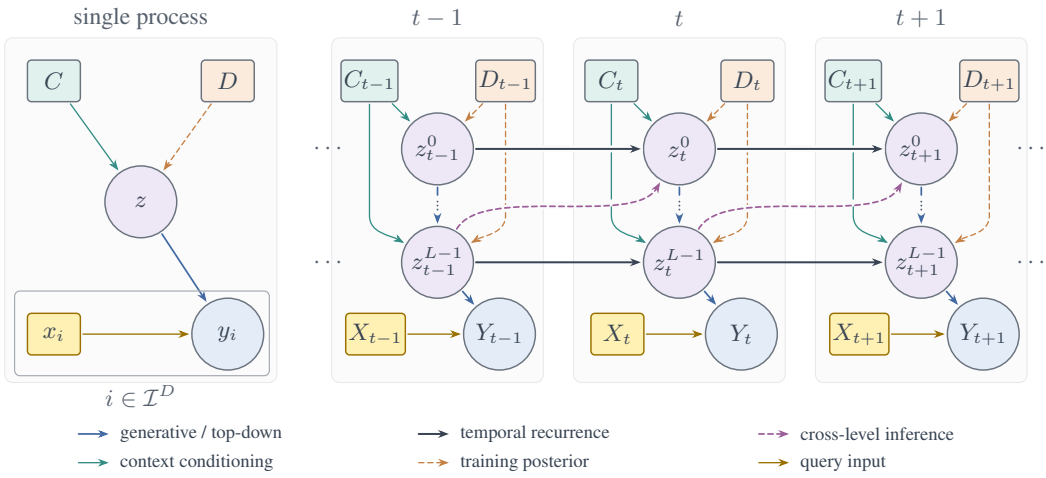
Hierarchical latent variable models. Hierarchical VAEs use top-down structure to let higher-level latents shape lower-level ones, often yielding richer multi-scale representations (Sønderby et al., 2016). Sequential variants couple these hierarchies to recurrent state models (Gregor et al., 2018), and precision-weighted or product-of-experts posteriors combine bottom-up evidence with top-down priors (Hinton, 2002; Sønderby et al., 2016). BRAID instantiates these ideas in the social motion domain: the higher-level latent captures emergent interaction properties, while person-level latents encode individual variation conditioned on that shared context. Our staged KL annealing and free-bits terms follow established practice for stabilising hierarchical VAEs (Kingma et al., 2016; Sønderby et al., 2016; Vahdat & Kautz, 2020), with a warm-up order chosen to stabilise group structure before person-level variation.

3 Problem Statement

Consider a temporal Neural Process setting with stochastic processes $\mathcal{P}_1, \ldots, \mathcal{P}_T$. At each time step t, we observe a possibly empty context set $C_t := \{(x_t^i, y_t^i)\}_{i \in \mathcal{I}_t^C}$ and define a target set $D_t := \{(x_t^i, y_t^i)\}_{i \in \mathcal{I}_t^D}$, with $\mathcal{I}_t^C \subseteq \mathcal{I}_t^D$. Let $X_t := \{x_t^i\}_{i \in \mathcal{I}_t^D}$ and $Y_t := \{y_t^i\}_{i \in \mathcal{I}_t^D}$ denote the corresponding target inputs and outputs; the held-out elements $D_t \setminus C_t$ are the query locations and target values to be predicted.

A standard Neural Process uses a single latent z_t conditioned on the current context C_t, but does not model dependencies between successive processes. BRAID instead introduces L latent levels $z_t^{(0)}, \ldots, z_t^{(L-1)}$, conditioned on the context, latent history, and lower levels in the hierarchy. Writing C, D, X, Y for the full temporal sequences, $Z^{(k)} := (z_1^{(k)}, \ldots, z_T^{(k)})$, and $Z_{<t} := \{z_{<t}^{(k)}\}_{k=0}^{L-1}$, the generative model factorises as:

$$p(Y, Z^{(0)}, Z^{(1)}, \ldots, Z^{(L-1)} \mid X, C) = \prod_{t=1}^{T} p_\theta(Y_t \mid X_t, z_t^{(0)}, \ldots, z_t^{(L-1)})\, p_\theta(z_t^{(0)} \mid Z_{<t}, C_t) \prod_{k=1}^{L-1} p_\theta(z_t^{(k)} \mid Z_{<t}, z_t^{(k-1)}, C_t). \tag{1}$$



(a) Neural Process: one latent for the observed context.

(b) BRAID: a hierarchy of latents linked across time. Each column is one time step; vertical dots indicate intermediate levels.

Figure 2: Neural Process and BRAID. C: observed context; $D \supseteq C$: full target set; X: query inputs; Y: outputs. Dashed arrows denote inference-only dependencies.

The decoder likelihood factorises over target elements, and we initialise $z_0^{(k)} := \text{null}$ for all $k \in [0, L-1]$. In the two-level social-motion model introduced in Section 4, $z_t^{(0)}$ is the group latent and $z_t^{(1)}$ is the structured collection of person-level latents.

4 Method

While BRAID admits a L-level hierarchy, our approach instantiates the simplest case of a two-level latent hierarchy. A group latent z_t^g captures shared interaction state, such as phase, tempo, and proxemics, while person latents $\{z_t^p\}_{p=1}^P$ capture individual behaviour conditioned on that group context. This mirrors accounts of social interaction as a jointly maintained process in which group-level coordination constrains individual action (Sacks et al., 1974; Chartrand & Bargh, 1999; Hoehl et al., 2020; Hale et al., 2020; Pötschulat, 2026). We describe the scene representation, latent dynamics, context encoder, and learning objective below.

4.1 Motion Representation

We represent a scene as T time steps with P people. Each person has SMPL-X shape $\beta^p \in \mathbb{R}^{16}$, 6D joint rotations $\Theta_t^p \in \mathbb{R}^{J\times 6}$ (Zhou et al., 2018), and root-relative joints $\tilde{K}_{t,j}^p$. We additionally store canonical root motion g_t^p, using floor-anchored, yaw-normalised SE(3) quantities similar to Holden et al. (2016; 2017); Yi et al. (2024); Maluleke et al. (2025). To retain spatial awareness, we encode pairwise canonical-frame transforms $\mathbf{T}_t^{p\to q} \in \mathbb{R}^9$ with an MLP and mean-pool them over observed partners before fusing them into the context representation. Throughout, we write $\mathcal{M} = \{\Delta\mathbf{T}^{\text{can}}, \mathbf{T}^{\text{can}\to\text{root}}, \mathbf{T}^{p\to q}\}$ for the three canonical transforms the model predicts—the frame-to-frame canonical displacement, the canonical-to-root transform, and the pairwise partner transform—each carrying a rotation R_m and a translation t_m.

Following section 3, each observation is represented as a context point $(x_t^{(p,j)}, y_t^{(p,j)})$, where $x_t^{(p,j)}$ identifies a person and joint/root slot. For $j \geq 1$, $y_t^{(p,j)} = [\Theta_{t,j}^p, \tilde{K}_{t,j}^p]$; for $j = 0$, $y_t^{(p,0)} = g_t^p$. We define the complete target set and observed context set using index sets $\mathcal{I}_t^D = [1, P] \times [0, J]$ and $\mathcal{I}_t^C \subseteq \mathcal{I}_t^D$: $D_t := \{(x_t^{(p,j)}, y_t^{(p,j)})\}_{(p,j)\in\mathcal{I}_t^D}$, $C_t := \{(x_t^{(p,j)}, y_t^{(p,j)})\}_{(p,j)\in\mathcal{I}_t^C}$. Missing joints or people are simply omitted from C_t, so forecasting, tracking/in-filling, and response generation differ only in which indices are observed.

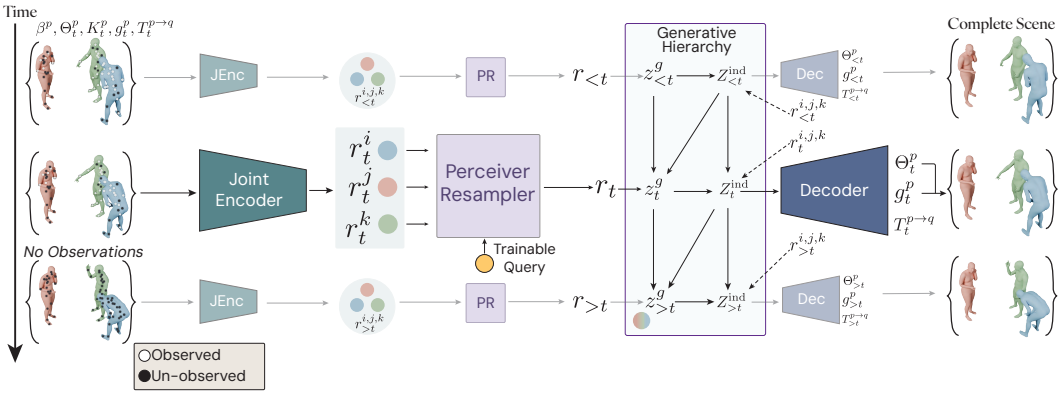


Figure 3: High-level architectural view of the model. Illustration shows a interacting triad. On the left the model receives partial observations at various time steps and it generates the complete scene including each person $p \in \{i, j, k\}$

4.2 Latent Encoding and Generative Model

We instantiate the hierarchy from eq. (1) with a group latent z_t^g and person latents $Z_t^{\text{ind}} = \{z_t^p\}_{p=1}^P$. Given C_t, the model samples the group latent, samples person latents conditioned on it, and decodes all target motion quantities (dropping index j for simplicity):

$$p_\theta(Y, Z^g, Z^{\text{ind}} \mid X, C) = \prod_{t=1}^{T} p_\theta(z_t^g \mid z_{<t}^g, Z_{<t}^{\text{ind}}, C_t) \prod_{p=1}^{P} p_\theta(z_t^p \mid z_{<t}^p, z_t^g, C_t)\, p_\theta(y_t^p \mid x_t^p, z_t^p), \tag{2}$$

The decoder factorises over target indices $(p, j) \in \mathcal{I}_t^D$, predicting g_t^p for $j = 0$ and root-relative rotations/keypoints for $j \geq 1$.

Pretrained motion decoder. Body pose is decoded through a temporal motion VAE (Kingma & Welling, 2013) that is pretrained on complete motion and then frozen. A trainable adapter maps the latents, person histories, and canonical root motion to the VAE's codes (one per four frames), which the frozen decoder turns into joint rotations; separate heads predict the canonical and partner transforms. The VAE thus supplies short-horizon body kinematics, while BRAID models the social structure that selects among them. Each decoded frame therefore depends on a short window of z^p and z^g around t (Appendix D).

4.3 Context Encoding

Both the generative and inference networks condition on deterministic context representations: a group summary r_t and per-person summaries r_t^p. We embed observed joint features with joint tokens, use cross-attention to relate context joints to target slots within each person, and add a continuous RoPE-inspired person embedding (Su et al., 2021) so person slots remain distinguishable under masking. Temporal attention and cross-person attention then capture short-range dynamics and coordination, and a perceiver-style resampler (Jaegle et al., 2021) compresses per-person tokens into the group representation r_t.

4.4 Inference and Learning

Learning requires posterior inference over the latent social state implied by eq. (2). Since the true posterior $p(Z^g, Z^{\text{ind}} \mid C, D)$ is intractable, we train with a variational approximation (Feynman, 1955; 1972; Peterson & Anderson, 1987; Kingma & Welling, 2013; Rezende et al., 2014; Blei et al., 2016). We introduce an amortised variational distribution that mirrors the temporal and hierarchical



Figure 4: Samples from BRAID for dyadic prediction, in-betweening, and polyadic prediction, spanning boxing and conversational interactions. Frames in Pink are given as input. Qualitative videos are in Supplementary.

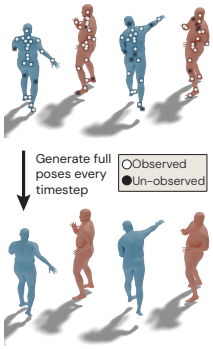


Figure 5: Full-pose generation with missing joints.

structure of the generative model:

$$q_\phi(Z^g, Z^{\text{ind}} \mid C, D) = \prod_{t=1}^{T} q_\phi(z_t^g \mid z_{<t}^g, Z_{<t}^{\text{ind}}, C, D) \prod_{p=1}^{P} q_\phi(z_t^p \mid z_{<t}^p, z_t^g, C, D). \tag{3}$$

Temporal Parameterisation. The conditional distributions above are parameterised by recurrent networks that encode their stated latent histories. The group network encodes $(z_{<t}^g, Z_{<t}^{\text{ind}})$, while the person network encodes $z_{<t}^p$; both also receive the corresponding context representations from section 4.3. Their LSTM activations are deterministic implementation details, rather than additional variables in the probabilistic model, so the factorisations are written directly in terms of latent histories and observed sets.

Precision-Weighted Posterior. Rather than parametrising the posterior directly, we compute it as a precision-weighted combination of two signals: a *bottom-up* proposal driven by the full target data and a *top-down* term from the prior, based on prior work by Sønderby et al. (2016). For the group latent, let $H_t^g = (z_{<t}^g, Z_{<t}^{\text{ind}})$, and define $(\mu_{\text{bu}}, \sigma_{\text{bu}}) = f_\phi^{\text{bu}}(H_t^g, C, D)$ and $(\mu_{\text{td}}, \sigma_{\text{td}}) = f_\theta^{\text{td}}(H_t^g, C_t)$, where f^{bu} and f^{td} are the bottom-up and top-down encoder networks respectively. The fused posterior parameters are obtained by precision merging:

$$\sigma_q^{-2} = \sigma_{\text{bu}}^{-2} + \sigma_{\text{td}}^{-2}, \qquad \mu_q = \sigma_q^2\big(\sigma_{\text{bu}}^{-2}\,\mu_{\text{bu}} + \sigma_{\text{td}}^{-2}\,\mu_{\text{td}}\big). \tag{4}$$

This can be viewed as a product-of-experts combination (Hinton, 2002) of two diagonal Gaussians, where each expert's influence is proportional to its precision. The same merging scheme is applied for each person-level latent z_t^p: the corresponding networks condition on $(z_{<t}^p, z_t^g, C, D)$ and $(z_{<t}^p, z_t^g, C_t)$, respectively. Along with sharing information from the generative and inference paths, this precision-weighted formulation allows the posterior to smoothly interpolate between data-driven evidence and the learned prior, which is particularly beneficial when context is sparse or absent.

Training Objective. Combining the generative model in eq. (2) with the approximate posterior in eq. (3) yields the following weighted variational objective:

$$\begin{aligned}\text{ELBO}_\beta := &\sum_{t=1}^{T} \Big\langle \ln p_\theta(Y_t \mid X_t, z_t^g, Z_t^{\text{ind}}) \Big\rangle_{q_\phi} \\ &- \beta_g \sum_{t=1}^{T} \Big\langle \mathbb{KL}\big(q_\phi(z_t^g \mid z_{<t}^g, Z_{<t}^{\text{ind}}, C, D) \,\|\, p_\theta(z_t^g \mid z_{<t}^g, Z_{<t}^{\text{ind}}, C_t)\big) \Big\rangle_{q_\phi} \\ &- \beta_p \sum_{t=1}^{T} \sum_{p=1}^{P} \Big\langle \mathbb{KL}\big(q_\phi(z_t^p \mid z_{<t}^p, z_t^g, C, D) \,\|\, p_\theta(z_t^p \mid z_{<t}^p, z_t^g, C_t)\big) \Big\rangle_{q_\phi},\end{aligned} \tag{5}$$

where expectations include latent histories and, for person KLs, the current group latent. For $\beta_g = \beta_p = 1$, this is a lower bound on $\ln p_\theta(Y \mid X, C)$; Appendix A gives the derivation and

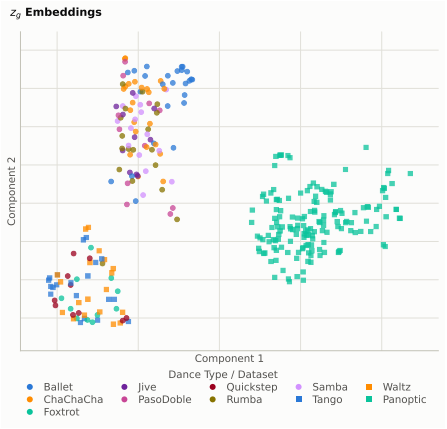


Figure 6: Z^g PACMAP (Wang et al., 2020) embeddings over DD100, and Panoptic Haggling.

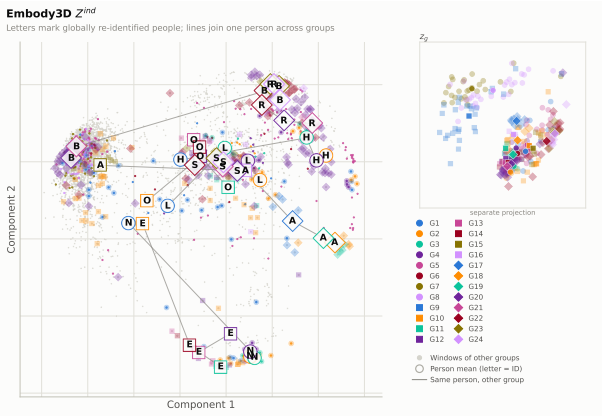


Figure 7: Z^{ind} PCA embeddings of Embody3D with common individuals across groups. Embeddings show the model latent space organising itself to cluster individuals.

weighting details. We maximise eq. (5); equivalently, we minimise $\mathcal{L}_{\text{ELBO}} \coloneqq -\text{ELBO}_\beta$. We also supervise joint rotations, root-relative joint positions, and canonical transforms with reconstruction losses $\mathcal{L}_{\text{pose}}$, $\mathcal{L}_{\text{key}}$, and $\mathcal{L}_{\text{root}}$, respectively. The full training objective is

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{ELBO}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{key}}\mathcal{L}_{\text{key}} + \lambda_{\text{root}}\mathcal{L}_{\text{root}}. \tag{6}$$

Reconstruction-loss definitions and further training details are given in Appendix C.

5 Experiments

Datasets. To evaluate the performance of BRAID, we conduct a number of experiments using the Haggling (Joo et al., 2016), DnD (Mughal et al., 2024), DD100 (Siyao et al., 2024), DuoBox (Cen et al., 2025) and Embody3D (McLean et al., 2025) datasets. Unless otherwise stated, models have approximately 16M parameters and use AdamW (Loshchilov & Hutter, 2017) with a learning rate of $1e-4$. Training details and model-size comparisons are given in Appendices C and E.

Metrics. Multiparty generation requires more than agreement with a single recorded future: valid continuations can differ in timing, role assignment, and interaction style. Standard reference errors (MPJPE, MPJVE) penalise such alternatives without separating them from socially incoherent motion, so low error alone does not establish the quality or variety of generated interactions. Likewise, diversity (DIV) is only meaningful alongside distributional fit (Fréchet distance, FD), and foot skating (FS) and interpenetration (IP) flag specific artefacts without establishing physical plausibility. We report these for comparability with prior work, but prioritise interaction-level structure following Shirekar et al. (2025): synchrony, via cross-recurrence quantification analysis as deviations in recurrence rate (ΔRR) and determinism (ΔDET) from the ground truth, and structural similarity, via Soft-DTW (SDTW) and its cross-person variant (Cr. SDTW), which compare trajectories up to natural timing variation. All metrics are lower-is-better except DIV; Appendix H gives definitions and interpretation.

5.1 Results

A shared social-state formulation supports both forecasting future behaviour and completing missing observations. Tables 2 to 4 evaluate the reusable representation motivated in Section 1 through forecasting, inpainting, and partner/dyadic prediction. The same group and person latent formulation handles each setting by treating observed joints and people as context and unobserved ones as targets. We first examine what social information the latent state retains, then assess how this shared representation supports generation across tasks and observation regimes.

Group latents organise activity structure, while person latents retain participant information across changing groups. Figure 6 shows that the group latent Z^g forms structured clusters across

DD100, Panoptic Haggling, and within DD100 it separates dances based on whether they belong to the Latin or Western ballroom dance styles. In contrast, Figure 7 shows that Z^{ind} preserves person-specific structure: embeddings of the same participant remain close even across different Embody3D groups. In the full latent space, same-person distances have a median of 6.36, versus 16.22 for different-person pairs (ratio 0.392); the median same-person pair lies at the 15.7th percentile of the different-person distance distribution. Nearest-neighbour retrieval, excluding candidates from the query's group, recovers the same individual in $99/239$ queries (41.4%); Appendix B gives the protocol and detailed results. This persistence across groups supports the social-state aim in Section 1: retaining information about who is participating as the interaction context changes. Together with the activity structure in Z^g, it provides evidence that the latents carry inspectable social information alongside their use for generation. Hierarchy ablations and qualitative probes in Appendix G further connect these representations to generation: *the full hierarchy improves cross-person alignment over either single-level variant, while varying person latents or changing the group-latent source produces different interaction realisations.*

Table 2: Social forecasting by observability and group size N. Bold: best; underlined: second best per setting, excluding RS and NN. Models trained without the motion VAE (Appendix D). †RS and NN ranked, as no learned baseline is available.

N	Method	ΔRR↓	ΔDET↓	SDTW↓	Cr. SDTW↓	IP↓	FD↓	DIV↑	FS↓	MPJPE↓	MPJVE↓
Full Observability											
2	RS	0.013	0.131	8.741	2.028	<0.005	5.485	90.58	0.011	1.471	0.013
	NN	0.013	0.093	8.292	2.059	<0.005	0.319	–	<0.001	0.898	0.014
	MAGNet	**0.012**	0.350	**6.041**	**1.348**	0.024	0.490	0.524	0.570	0.595	0.185
	BRAID	**0.012**	**0.168**	7.416	1.620	**<0.005**	**0.239**	**2.316**	**0.001**	**0.305**	**0.013**
3	RS	0.013	0.166	8.533	1.898	0.222	2.131	80.86	0.012	1.647	0.010
	NN	0.013	0.149	8.342	2.194	0.177	1.808	–	0.018	1.120	0.009
	MAGNet	0.013	0.335	**7.318**	**1.531**	0.189	0.378	0.586	0.452	0.755	0.437
	BRAID	**0.012**	**0.202**	8.025	1.964	**0.005**	**0.221**	**1.974**	**0.071**	**0.260**	**0.012**
4	RS	0.013	0.22	8.184	2.470	0.354	6.01	65.94	0.01	1.63	0.01
	NN	0.013	0.18	8.324	2.194	0.177	1.80	–	0.013	1.47	0.01
	MAGNet	0.013	0.391	**6.996**	**1.731**	0.232	0.768	0.710	0.481	0.937	0.624
	BRAID	**0.012**	**0.197**	7.771	1.919	**0.022**	**0.216**	**3.19**	**0.099**	**0.353**	**0.015**
5†	RS	0.0128	0.174	8.279	**1.542**	< 0.005	1.348	**21.61**	**0.0**	0.751	0.011
	NN	0.0129	**0.160**	**8.160**	1.549	0.001	**1.291**	–	**0.0**	0.742	0.011
	BRAID	**0.012**	0.228	8.725	2.186	**<0.001**	1.766	0.039	**0.0**	**0.217**	**0.008**
Partial Observability											
2	RoHM	0.013	0.241	**4.552**	**0.883**	0.147	0.406	0.093	0.167	0.489	0.043
	BRAID	**0.012**	**0.168**	7.420	1.60	**0.0013**	**0.126**	**2.51**	**0.0012**	**0.400**	**0.016**
3	RoHM	**0.013**	0.314	**4.666**	**0.848**	0.168	0.631	0.081	**0.173**	0.444	0.041
	BRAID	**0.013**	**0.200**	8.06	1.934	**0.003**	**0.141**	**2.398**	0.235	**0.297**	**0.015**
4	RoHM	0.013	0.302	**4.593**	**0.862**	0.125	0.701	0.101	0.186	0.461	0.041
	BRAID	**0.012**	**0.2**	7.80	1.89	**0.016**	**0.20**	**3.59**	**0.096**	**0.33**	**0.015**
5	RoHM	0.014	**0.206**	**8.130**	**1.019**	0.3626	**0.32**	**0.107**	0.131	0.533	0.028
	BRAID	**0.012**	0.226	8.722	2.164	**0.0005**	1.73	0.039	**0**	**0.213**	**0.008**

Table 3: Partner inpainting on Panoptic and DD100. Bold: best; underlined: second best per setting, excluding RS and NN.

	Method	ΔRR↓	ΔDET↓	SDTW↓	Cr. SDTW↓	IP↓	FD↓	DIV↑	FS↓	MPJPE↓	MPJVE↓
Panoptic	RS	0.013	0.142	8.069	2.399	0.167	22.692	68.505	0.462	1.371	0.017
	NN	0.013	0.129	8.483	1.932	0	4.203	–	0.587	1.080	0.010
	Social Processes (2023)	**0.009**	**0.156**	8.952	3.689	**0**	0.951	0.120	0.385	0.247	**0.006**
	BRAID	**0.009**	0.196	**7.490**	**1.260**	**0**	**0.139**	**1.525**	**<0.005**	**0.217**	0.009
Partial Observability											
	RoHM (2024)	0.013	0.221	8.252	**1.008**	0.179	0.255	0.030	0.172	0.417	0.022
	BRAID	**0.009**	**0.195**	**7.493**	1.262	**0.0**	**0.130**	**3.640**	**0.016**	**0.239**	**0.018**
DD100	RS	0.010	0.063	8.595	3.377	0.175	11.600	88.595	0.228	1.761	0.042
	NN	0.009	0.044	7.673	2.018	0.302	1.420	–	0.021	1.642	0.049
	Duolando (2024)	–	–	–	–	–	18.180	0.000	1.880	1.680	0.070
	MAGNet (2025)	0.013	0.184	5.646	2.022	0.169	0.42	0.71	0.873	1.187	0.036
	Social Processes (2023)	**0.007**	0.180	8.328	1.937	0.332	2.05	1.42	0.097	1.062	0.033
	BRAID	0.008	**0.06**	**4.590**	**1.527**	**0.117**	**0.075**	**7.72**	**0.012**	**1.050**	**0.028**
Partial Observability											
	RoHM (2024)	0.013	0.289	7.017	**1.494**	0.241	0.173	0.08	0.311	1.053	0.047
	BRAID	**0.008**	**0.066**	**4.623**	1.514	**0.118**	**0.075**	**7.49**	**0.013**	**1.050**	**0.028**

Under partial observation, forecasting retains sample diversity and temporal structure while improving distributional fit. *Partial observability* here denotes joints missing through self-occlusion (Sigal & Black, 2006; Yao et al., 2022) or occlusion by other participants (Joo et al., 2016; Raman et al., 2022), simulated by masking context joints as in prior work (Zhang et al., 2024); occluding all of a person's joints extends this to a missing person. Comparing BRAID's full and partial observation results in Table 2, FD decreases at every group size, including 0.239 to 0.126 for two people and 0.221 to 0.141 for three. DIV increases for two to four people and is unchanged for five; SDTW changes by less than 0.5%, and cross-person SDTW improves throughout. This stability supports the introductory goal of a social state that sustains generation as available evidence changes. The robustness is metric-specific: three-person FS rises from 0.071 to 0.235, and MPJPE increases for two and three people. Recorded Embody3D motion has zero FS, so the nonzero FS for three and four people is an artefact. Its interpenetration rate is also low (0.0002–0.0007); BRAID's partial-observation IP (0.0013–0.016) remains above this level, though 8–110 times lower than RoHM's. Against baselines, BRAID has the lowest FD for two to four people within each observation regime; its lowest MPJPE at every group size further supports reference recovery.

For five-person DnD, participants mostly stand around a table playing the game. BRAID's zero FS matches the recorded motion, which also has zero FS, and its IP (at most 0.001) stays below the recorded rate of 0.0016. DIV 0.039 should be interpreted in this activity context, as the dataset involves five people standing still quite often. FD favours NN under full observations and RoHM under partial observations, identifying distributional matching as a direction for improvement in this setting. Activity and group size vary together, precluding a group-size-only explanation.

Partner inpainting combines greater sample variation with better temporal alignment and less foot skating than RoHM. Under partial observations, BRAID improves DIV, FS, FD, and SDTW over RoHM on both inpainting datasets (Table 3). On DD100, DIV rises from 0.080 to 7.49, while FS falls from 0.311 to 0.013 and SDTW from 7.017 to 4.623. Recorded motion has zero FS on both

Table 4: Partner and dyadic forecasting with a missing partner on DuoBox. Bold: best; underlined: second best per setting, excluding RS and NN.

Method	Partner Motion Prediction										Dyadic Motion Prediction									
	ΔRR↓	ΔDET↓	SDTW↓	Cr. SDTW↓	IP↓	FD↓	DIV↑	FS↓	MPJPE↓	MPJVE↓	ΔRR↓	ΔDET↓	SDTW↓	Cr. SDTW↓	IP↓	FD↓	DIV↑	FS↓	MPJPE↓	MPJVE↓
RS	0.010	0.064	7.702	2.554	0.001	0.310	15.9	0.049	1.071	0.041	0.011	0.066	7.698	2.552	0.002	0.119	16.232	0.047	1.074	0.042
NN	0.009	0.064	6.598	2.525	0.001	0.766	–	0.026	0.880	0.040	0.009	0.065	6.604	2.528	0.002	0.295	–	0.026	0.895	0.041
R2R (2025)	–	–	–	–	–	0.181	**0.318**	0.255	0.580	0.029	–	–	–	–	–	0.337	**0.395**	0.249	**0.624**	**0.029**
MAGNet (2025)	0.007	0.069	1.560	0.444	0.001	**0.057**	0.034	**0.071**	**0.125**	**0.022**	0.012	0.120	4.564	1.886	0.004	**0.102**	0.370	**0.087**	0.697	0.032
Partial Observability																				
RoHM (2024)	0.014	0.166	**3.457**	1.551	0.095	0.325	0.026	0.638	**0.660**	0.041	0.0134	0.231	6.914	1.760	0.210	0.325	0.026	0.638	**0.660**	0.041
BRAID	**0.007**	**0.087**	6.360	**0.971**	**0.009**	**0.116**	**2.39**	**0.172**	0.912	**0.030**	**0.008**	**0.081**	**5.743**	**0.945**	**0.009**	**0.115**	**2.37**	**0.157**	0.773	**0.031**

datasets, so BRAID's FS is also closer to the data. Interpenetration must be read against the recorded rate. Panoptic participants never touch, and BRAID likewise produces no interpenetration, whereas RoHM reaches 0.179. DD100 dancers in close hold overlap in 29% of recorded frames (IP 0.288), so BRAID's lower IP (0.118 versus 0.241 for RoHM) reflects looser partner coupling rather than fewer artefacts. Lower MPJPE additionally supports reference recovery, while RoHM's lower cross-person SDTW on both datasets identifies closer interpersonal alignment as its complementary strength.

A social state can represent varied futures while preserving distributional fit under partial observation. Relative to RoHM, BRAID produces more diverse samples with lower FD in both DuoBox prediction settings under partial observations (Table 4). In dyadic prediction, DIV rises from 0.026 to 2.37 and FD falls from 0.325 to 0.115. Foot skating must be read against the data: fast boxing footwork gives the recorded motion an FS of 0.276 (0.283 in partner prediction). RoHM exceeds this rate (0.638), whereas BRAID falls below it (0.157 and 0.172), consistent with generated joint speeds of roughly half the recorded ones. Dyadic IP (0.009) remains about ten times the recorded rate of 0.0008. RoHM tracks the recorded future more closely by MPJPE (0.660 versus 0.773). For the social-state motivation, this supports representing possible continuations beyond a single recorded trajectory. Such samples could supply candidate futures to a agent's planner.

The social state carries structure from context, even across domains. A model trained only on DD100 dance, never shown conversation, still keeps a plausible conversational formation when given 40 frames of held-out Panoptic context. What it lacks is the activity's own dynamics: it drifts toward dance-like rotations, whereas adding conversational data lowers cross-person SDTW from 1.424 to 0.511 and FD from 1.94 to 0.27 (Appendix F).

6 Discussion and Conclusion

BRAID frames multi-person motion as conditional generation over arbitrary context sets: a single joint-level interface covers full or sparse observation, missing joints or people, and future rollout without task-specific heads. Across forecasting, partner inpainting, and dyadic prediction, it retains distributional fit and temporal alignment as evidence is removed, and produces varied completions rather than a single averaged future. Its latent social state is also inspectable: without supervision, Z^g organises interactions by activity and dance style, and Z^{ind} recovers the same participant across different groups. The model also generalises formation structure across domains (Appendix F).

These two levels of social state open several directions. Because Z^g summarises the ongoing interaction, it could serve as a compact belief state for agents that observe or join a group, conditioning their policies on the phase of an interaction or detecting when it shifts. Because Z^{ind} persists across groups, it could support memory of who someone is across encounters and adaptation to an individual's interaction style. More broadly, both states could be supplied to foundation models, vision-language-action models, and robot controllers as a compact summary of human interaction, helping them gauge how a group is coordinated and how each person tends to behave without reasoning over raw pose streams. Editing or exchanging the two levels, as in the group-latent switch of Appendix G, suggests controllable synthesis of social scenarios for training and evaluating agents. The evidence covers the evaluated social-motion datasets and depends on SMPL-X preprocessing quality. Diversity is low in the largely stationary five-person DnD scenes, and pose and timing errors can accumulate over long open-loop rollouts; since new observations can enter the context at any time, however, the model can correct its predictions as evidence arrives, as in tracking. The latent analyses show useful structure but not complete disentanglement; incorporating BRAID in a simulator or agent world model is the natural next step.

AI USE STATEMENT

We have not used generative AI tools for proposing or refining hypotheses, generating synthetic data sets, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis and providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. We used generative AI tools for modifying scientific figures, creating or editing software code, and editing the research paper to improve readability. We have reviewed all AI-assisted work.

REFERENCES

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *arXiv*, 2016. doi: 10.48550/arxiv.1601.00670.

Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Ready-to-React: Online Reaction Policy for Two-Character Interaction Generation. *arXiv*, 2025. doi: 10.48550/arxiv.2502.20370.

Tanya L Chartrand and John A Bargh. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893–910, 1999. ISSN 0022-3514. doi: 10.1037/0022-3514.76.6.893. URL https://doi.apa.org/doi/10.1037/0022-3514.76.6.893.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your Commands via Motion Diffusion in Latent Space. *arXiv*, 2022. doi: 10.48550/arxiv.2212.04048.

Jouh Yeong Chew, Zhi-Yi Lin, and Xucong Zhang. SBM: Social Behavior Model for Human-Like Action Generation. *Companion Proceedings of the 27th International Conference on Multimodal Interaction*, pp. 32–36, 2025. doi: 10.1145/3747327.3763038.

I Coco, Moreno, Dan Mønster, Giuseppe Leonardi, Rick Dale, and Sebastian Wallot. Unidimensional and multidimensional methods for recurrence quantification analysis with crqa. *The R Journal*, 13(1):145, 5 2021. ISSN 2006.0195. doi: 10.32614/rj-2021-062. URL https://journal.r-project.org/archive/2021/RJ-2021-062/index.html.

Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J Black. WANDR: Intention-guided human motion generation. *arXiv*, 4 2024. doi: 10.48550/arxiv.2404.15383. URL http://arxiv.org/abs/2404.15383.

R. P. Feynman. Slow electrons in a polar crystal. *Physical Review*, 97(3):660–665, 1955. ISSN 0031-899X. doi: 10.1103/physrev.97.660.

R. P. Feynman. *Statistical mechanics: a set of lectures*. Frontiers in physics : a lecture note and reprint series. Sarat Book Distributors, 1972. ISBN 9788187169970. URL https://books.google.nl/books?id=Vflsca1UFZoC.

Marta Garnelo, Dan Rosenbaum, Chris J Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J Rezende, and S M Ali Eslami. Conditional neural processes. *arXiv*, 2018a. doi: 10.48550/arxiv.1807.01613.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, S M Ali Eslami, and Yee Whye Teh. Neural Processes. *arXiv*, 7 2018b. doi: 10.48550/arxiv.1807.01622. URL http://arxiv.org/abs/1807.01622.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. ReMoS: 3d motion-conditioned reaction synthesis for two-person interactions. *arXiv [cs.CV]*, 11 2023. ISSN 2311.1705. URL http://arxiv.org/abs/2311.17057.

Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, and Theophane Weber. Temporal Difference Variational Auto-Encoder. *arXiv*, 2018. doi: 10.48550/arxiv.1806.03107.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating Diverse and Natural 3D Human Motions from Text. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 00:5142–5151, 2022. doi: 10.1109/cvpr52688.2022.00509.

Joanna Hale, Jamie A Ward, Francesco Buccheri, Dominic Oliver, and Antonia F de C Hamilton. Are you on my wavelength? interpersonal coordination in dyadic conversations. *Journal of nonverbal behavior*, 44(1):63–83, 2020. ISSN 0191-5886. doi: 10.1007/s10919-019-00320-3. URL http://dx.doi.org/10.1007/s10919-019-00320-3.

Geoffrey E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 2002. ISSN 0899-7667. doi: 10.1162/089976602760128018.

Stefanie Hoehl, Merle Fairhurst, and Annett Schirmer. Interactional synchrony: signals, mechanisms and benefits. *Social Cognitive and Affective Neuroscience*, 16(1-2):5–18, 2020. ISSN 1749-5016. doi: 10.1093/scan/nsaa024.

Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. ISSN 0730-0301. doi: 10.1145/2897824.2925975.

Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073663.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General Perception with Iterative Attention. *arXiv*, 2021. doi: 10.48550/arxiv.2103.03206.

Kaiyang Ji, Ye Shi, Zichen Jin, Kangyi Chen, Lan Xu, Yuexin Ma, Jingyi Yu, and Jingya Wang. Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis. *arXiv*, 2025. doi: 10.48550/arxiv.2508.02106.

Wentao Jiang, Jingya Wang, Kaiyang Ji, Baoxiong Jia, Siyuan Huang, and Ye Shi. ARFlow: Human action-reaction flow matching with physical guidance. *arXiv*, 2025. doi: 10.48550/arxiv.2503.16973.

Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *arXiv*, 12 2016. ISSN 1612.0315. doi: 10.48550/arxiv.1612.03153. URL http://arxiv.org/abs/1612.03153.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive Neural Processes. *arXiv*, 2019. doi: 10.48550/arxiv.1901.05761.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv [stat.ML]*, 12 2013. ISSN 1312.6114. doi: 10.48550/arxiv.1312.6114. URL http://dx.doi.org/10.48550/arxiv.1312.6114.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving Variational Inference with Inverse Autoregressive Flow. *arXiv*, 2016. doi: 10.48550/arxiv.1606.04934.

Gerben A Van Kleef, Wolfgang Steinel, Daan Van Knippenberg, Michael A Hogg, and Alicia Svensson. Group member prototypicality and intergroup negotiation: How one's standing in the group affects negotiation behaviour. *British Journal of Social Psychology*, 46(1):129–152, 3 2007. ISSN 0144-6665. doi: 10.1348/014466605x89353. URL http://dx.doi.org/10.1348/014466605X89353.

Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731, 6 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00731. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4464110.

Zhi-Yi Lin, Thomas Markhorst, Jouh Yeong Chew, and Xucong Zhang. PolySLGen: Online Multimodal Speaking-Listening Reaction Generation in Polyadic Interaction. *arXiv*, 2026. doi: 10.48550/arxiv.2604.08125.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv*, 2017. doi: 10.48550/arxiv.1711.05101.

Vongani H. Maluleke, Kie Horiuchi, Lea Wilken, Evonne Ng, Jitendra Malik, and Angjoo Kanazawa. Diffusion forcing for multi-agent interaction sequence modeling. *arXiv*, 2025. doi: 10.48550/arxiv.2512.17900.

Thomas Markhorst, Zhi-Yi Lin, Jouh Yeong Chew, Jan van Gemert, and Xucong Zhang. MuPPet: Multi-person 2D-to-3D Pose Lifting. *arXiv*, 2026. doi: 10.48550/arxiv.2604.09715.

David Matsumoto. Culture, context, and behavior. *Journal of personality*, 75(6):1285–1319, 12 2007. ISSN 1467-6494. doi: 10.1111/j.1467-6494.2007.00476.x. URL https://www.ncbi.nlm.nih.gov/pubmed/17995466.

Claire McLean, Makenzie Meendering, Tristan Swartz, Orri Gabbay, Alexandra Olsen, Rachel Jacobs, Nicholas Rosen, Philippe de Bree, Tony Garcia, Gadsden Merrill, Jake Sandakly, Julia Buffalini, Neham Jain, Steven Krenn, Moneish Kumar, Dejan Markovic, Evonne Ng, Fabian Prada, Andrew Saba, Siwei Zhang, Vasu Agrawal, Tim Godisart, Alexander Richard, and Michael Zollhoefer. Embody 3d: A large-scale multimodal motion and behavior dataset. *arXiv*, 2025. doi: 10.48550/arxiv.2510.16258.

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. ConvoFusion: Multi-modal conversational diffusion for co-speech gesture synthesis. *arXiv*, 3 2024. doi: 10.48550/arxiv.2403.17936. URL http://arxiv.org/abs/2403.17936.

Kees van Oers, Margreet Klunder, and Piet J Drent. Context dependence of personalities: risk-taking behavior in a social and a nonsocial situation. *Behavioral ecology: official journal of the International Society for Behavioral Ecology*, 16(4):716–723, 7 2005. ISSN 1045-2249. doi: 10.1093/beheco/ari045. URL http://academic.oup.com/beheco/article/16/4/716/214729/Context-dependence-of-personalities-risktaking.

Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents. *arXiv*, 2022. doi: 10.48550/arxiv.2201.00308.

Carsten Peterson and James R Anderson. A mean field theory learning algorithm for neural networks. *Complex systems*, 1(5), 1987.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. *arXiv*, 2021. doi: 10.48550/arxiv.2111.15640.

Jonathan N Pruitt and Susan E Riechert. How within-group behavioural variation and task efficiency enhance fitness in a social group. *Proceedings. Biological sciences*, 278(1709):1209–1215, 4 2011. ISSN 0962-8452. doi: 10.1098/rspb.2010.1700. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049074.

Maike Pötschulat. Transcript frame analysis: Thinking with goffman about interview data. *Qualitative Research*, 2026. ISSN 1468-7941. doi: 10.1177/14687941251398982.

Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and H Hung. ConfLab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild. *Neural Information Processing Systems*, 35:23701–23715, 10 May 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/95f9ad2e251e9014697589037450f9bb-Abstract-Datasets_and_Benchmarks.html.

Chirag Raman, Hayley Hung, and Marco Loog. Social processes: Self-supervised meta-learning over conversational groups for forecasting nonverbal social cues. In *ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022*, Lecture Notes in Computer Science, pp. 639–659. Springer Nature Switzerland, Cham, 2023. ISBN 9783031250651. doi: 10.1007/978-3-031-25066-8_37. URL https://link.springer.com/10.1007/978-3-031-25066-8_37.

Joanne A Rathbone, Tegan Cruwys, Mark Stevens, Laura J Ferris, and Katherine J Reynolds. The reciprocal relationship between social identity and adherence to group norms. *The British journal of social psychology*, 62(3):1346–1362, 7 2023. ISSN 2044-8309. doi: 10.1111/bjso.12635. URL https://www.ncbi.nlm.nih.gov/pubmed/36786397.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv*, 2014. doi: 10.48550/arxiv.1401.4082.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696, 12 1974. ISSN 0097-8507. doi: 10.2307/412243. URL https://www.jstor.org/stable/412243?origin=crossref.

Ojas Shirekar, Wim Pouw, Chenxu Hao, Vrushank Phadnis, Thabo Beeler, and Chirag Raman. Multimodal Quantitative Measures for Multiparty Behavior Evaluation. *Proceedings of the 27th International Conference on Multimodal Interaction*, pp. 249–264, 8 2025. doi: 10.1145/3716553.3750752.

L Sigal and M J Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*. IEEE, 2006. ISBN 9780769525976. doi: 10.1109/cvpr.2006.180. URL https://ieeexplore.ieee.org/document/1641003.

Gautam Singh, Jaesik Yoon, Youngsung Son, and Sungjin Ahn. Sequential neural processes. *arXiv [cs.LG]*, 24 June 2019. URL http://arxiv.org/abs/1906.10264.

Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower GPT with off-policy reinforcement learning for dance accompaniment. *arXiv*, 2024. doi: 10.48550/arxiv.2403.18811.

Joanne R Smith and Winnifred R Louis. Group norms and the attitude-behaviour relationship: Group norms and attitude-behaviour relations. *Social and personality psychology compass*, 3(1):19–35, 1 2009. ISSN 1751-9004. doi: 10.1111/j.1751-9004.2008.00161.x. URL https://compass.onlinelibrary.wiley.com/doi/full/10.1111/j.1751-9004.2008.00161.x.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv*, 2021. doi: 10.48550/arxiv.2104.09864.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *arXiv*, 2016. doi: 10.48550/arxiv.1602.02282.

Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 00:9567–9577, 10 2023. doi: 10.1109/iccv51070.2023.00880. URL http://dx.doi.org/10.1109/ICCV51070.2023.00880.

Julian Tanke, Takashi Shibuya, Kengo Uchida, Koichi Saito, and Yuki Mitsufuji. Dyadic mamba: Long-term dyadic human motion synthesis. *arXiv [cs.CV]*, 14 May 2025. doi: 10.48550/arXiv.2505.09827. URL http://arxiv.org/abs/2505.09827.

Deborah J Terry, Michael A Hogg, and Katherine M White. The theory of planned behaviour: Self-identity, social identity and group norms. *The British journal of social psychology*, 38(3):225–244, 9 1999. ISSN 2044-8309. doi: 10.1348/014466699164149. URL https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/014466699164149.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. *arXiv [cs.CV]*, 3 2022a. ISSN 2203.0806. URL http://arxiv.org/abs/2203.08063.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human Motion Diffusion Model. *arXiv*, 9 2022b. ISSN 2209.1491. doi: 10.48550/arxiv.2209.14916. URL http://arxiv.org/abs/2209.14916.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *arXiv*, 2020. doi: 10.48550/arxiv.2007.03898.

Sebastian Wallot. Multidimensional cross-recurrence quantification analysis (MdCRQA) - a method for quantifying correlation between multivariate time-series. *Multivariate behavioral research*, 54(2):173–191, 3 2019. ISSN 0027-3171. doi: 10.1080/00273171.2018.1512846. URL https://www.ncbi.nlm.nih.gov/pubmed/30569740.

Sebastian Wallot and Giuseppe Leonardi. Analyzing multivariate dynamics using cross-recurrence quantification analysis (CRQA), diagonal-cross-recurrence profiles (DCRP), and multidimensional recurrence quantification analysis (MdRQA) - a tutorial in r. *Frontiers in psychology*, 9:2232, 12 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.02232. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.02232/full.

Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-Person 3D Motion Prediction with Multi-Range Transformers. *arXiv*, 2021. doi: 10.48550/arxiv.2111.12073.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *arXiv*, 2020. doi: 10.48550/arxiv.2012.04456.

Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. ReGenNet: Towards human action-reaction synthesis. *arXiv [cs.CV]*, pp. 1759–1769, 18 March 2024. doi: 10.48550/arXiv.2403.11882. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xu_ReGenNet_Towards_Human_Action-Reaction_Synthesis_CVPR_2024_paper.html.

Chun-Han Yao, Jimei Yang, Duygu Ceylan, Yi Zhou, Yang Zhou, and Ming-Hsuan Yang. Learning visibility for robust dense human body estimation. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, pp. 412–428. Springer Nature Switzerland, Cham, 2022. ISBN 9783031197680,9783031197697. doi: 10.1007/978-3-031-19769-7_24. URL http://dx.doi.org/10.1007/978-3-031-19769-7_24.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv*, 2024. doi: 10.48550/arxiv.2410.03665.

Francesco Zanlungo, Zeynep Yücel, Dražen Brščić, Takayuki Kanda, and Norihiro Hagita. Intrinsic group behaviour: Dependence of pedestrian dyad dynamics on principal social and personal features. *PloS one*, 12(11):e0187253, 11 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0187253. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667819.

Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. RoHM: Robust human motion reconstruction via diffusion. *arXiv*, 2024. doi: 10.48550/arxiv.2401.08570.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *arXiv*, 2018. doi: 10.48550/arxiv.1812.07035.

A Derivation of the Evidence Lower Bound

We derive the objective in eq. (5) by introducing the approximate posterior into the marginal likelihood, applying Jensen's inequality, and separating the temporal and hierarchical factors. We first take $\beta_g = \beta_p = 1$, then describe the weighted training objective.

Notation and factorisation. Let $Z = (Z^g, Z^{\mathrm{ind}})$ denote the full latent trajectory and $H_t = (z^g_{<t}, Z^{\mathrm{ind}}_{<t})$ its history before time t. The observed context C and full target data $D = (X, Y)$ are fixed throughout the derivation. To keep the algebra readable, write the conditional densities as

$$\begin{aligned} q^g_t &:= q_\phi(z^g_t \mid H_t, C, D), & p^g_t &:= p_\theta(z^g_t \mid H_t, C_t),\\ q^p_t &:= q_\phi(z^p_t \mid z^p_{<t}, z^g_t, C, D), & p^p_t &:= p_\theta(z^p_t \mid z^p_{<t}, z^g_t, C_t),\\ \ell^p_t &:= p_\theta(y^p_t \mid x^p_t, z^p_t). \end{aligned}$$

As in eq. (2), the joint-slot index is suppressed in ℓ^p_t; the likelihood factorises over the target slots for each person. The generative and variational distributions are therefore

$$p_\theta(Y, Z \mid X, C) = \prod_{t=1}^{T} p^g_t \prod_{p=1}^{P} p^p_t \ell^p_t, \qquad q_\phi(Z \mid C, D) = \prod_{t=1}^{T} q^g_t \prod_{p=1}^{P} q^p_t.$$

In particular, there is one group factor per time step and one person factor per person and time step. Below, q abbreviates $q_\phi(Z \mid C, D)$.

From marginal likelihood to a lower bound. Multiplying and dividing the integrand by q, and using the concavity of the logarithm, gives

$$\begin{aligned} \ln p_\theta(Y \mid X, C) &= \ln \int p_\theta(Y, Z \mid X, C)\, \mathrm{d}Z \\ &= \ln \mathbb{E}_q\left[\frac{p_\theta(Y, Z \mid X, C)}{q_\phi(Z \mid C, D)}\right] \\ &\geq \mathbb{E}_q\left[\ln \frac{p_\theta(Y, Z \mid X, C)}{q_\phi(Z \mid C, D)}\right] =: \mathrm{ELBO}_1. \end{aligned} \tag{7}$$

The gap is $\ln p_\theta(Y \mid X, C) - \mathrm{ELBO}_1 = \mathbb{KL}(q_\phi(Z \mid C, D) \| p_\theta(Z \mid X, Y, C)) \geq 0$, so equality holds when the variational distribution equals the true posterior.

Separating time, people, and hierarchy. Substituting the factorisations into eq. (7), expanding the logarithm of each product, and using linearity of expectation yields

$$\begin{aligned} \mathrm{ELBO}_1 &= \mathbb{E}_q\left[\ln \frac{\prod_{t=1}^{T} p^g_t \prod_{p=1}^{P} p^p_t \ell^p_t}{\prod_{t=1}^{T} q^g_t \prod_{p=1}^{P} q^p_t}\right] \\ &= \sum_{t=1}^{T}\sum_{p=1}^{P} \mathbb{E}_q[\ln \ell^p_t] - \sum_{t=1}^{T} \mathbb{E}_q\left[\ln \frac{q^g_t}{p^g_t}\right] - \sum_{t=1}^{T}\sum_{p=1}^{P} \mathbb{E}_q\left[\ln \frac{q^p_t}{p^p_t}\right]. \end{aligned} \tag{8}$$

The first term scores reconstruction of each person's target motion. The remaining terms compare the target-informed posterior with the context-conditioned prior at the group and person levels.

Conditional KL terms. Let $q_{<t}$ be the marginal distribution of H_t under q, and let $q_{<t,g}$ be the marginal distribution of (H_t, z^g_t). For a fixed history, integrating the group log-ratio over z^g_t gives a KL divergence. For the person log-ratio, we also fix z^g_t before integrating over z^p_t. The law of iterated expectation thus gives

$$\begin{aligned} \mathbb{E}_q\left[\ln \frac{q^g_t}{p^g_t}\right] &= \mathbb{E}_{q_{<t}}\left[\mathbb{E}_{q^g_t}\left[\ln \frac{q^g_t}{p^g_t}\right]\right] = \mathbb{E}_{q_{<t}}[\mathbb{KL}(q^g_t \| p^g_t)], \\ \mathbb{E}_q\left[\ln \frac{q^p_t}{p^p_t}\right] &= \mathbb{E}_{q_{<t,g}}\left[\mathbb{E}_{q^p_t}\left[\ln \frac{q^p_t}{p^p_t}\right]\right] = \mathbb{E}_{q_{<t,g}}[\mathbb{KL}(q^p_t \| p^p_t)]. \end{aligned}$$

Future latents and other current person latents integrate out. The outer expectations remain because the conditional factors depend on sampled histories and the person factors also depend on the sampled group latent. Substituting into eq. (8) gives

$$\begin{aligned}\mathrm{ELBO}_1 = &\sum_{t=1}^{T} \mathbb{E}_q[\ln p_\theta(Y_t \mid X_t, z_t^g, Z_t^{\mathrm{ind}})] \\ &- \sum_{t=1}^{T} \mathbb{E}_{q_{<t}}[\mathbb{KL}(q_t^g \| p_t^g)] - \sum_{t=1}^{T}\sum_{p=1}^{P} \mathbb{E}_{q_{<t,g}}[\mathbb{KL}(q_t^p \| p_t^p)]. \end{aligned} \tag{9}$$

Although the likelihood is written using both latent levels, the decoder in eq. (2) depends directly on the person latents; the group latent influences it through those latents. The group KL is counted once per time step, rather than once per person.

Weighting and evaluation during training. Denote the expected reconstruction term in eq. (9) by $\mathcal{R}$, and its summed group and person KL terms by $\mathcal{K}_g$ and $\mathcal{K}_p$. Introducing the training weights gives exactly eq. (5):

$$\mathrm{ELBO}_\beta = \mathcal{R} - \beta_g \mathcal{K}_g - \beta_p \mathcal{K}_p = \mathrm{ELBO}_1 - (\beta_g - 1)\mathcal{K}_g - (\beta_p - 1)\mathcal{K}_p.$$

Thus $\beta_g = \beta_p = 1$ recovers the standard ELBO. If both weights are at least one, the objective remains a lower bound, but arbitrary weights (for example, values below one during KL warm-up) do not in general preserve that guarantee. The weights are training choices, not a consequence of Jensen's inequality. In eq. (5), the common expectation notation $\langle\cdot\rangle_{q_\phi}$ implicitly uses the appropriate marginals specified above.

For the diagonal Gaussian factors used here, each conditional KL can be evaluated analytically at a sampled history (and sampled group latent for the person terms). The remaining expectations over latent trajectories and reconstruction can be estimated with reparameterised posterior samples. The auxiliary reconstruction losses in eq. (6) are additional supervision and are not part of this likelihood-bound derivation.

B Cross-group person re-identification

Protocol. We analyse the model trained jointly on the two-, three-, and four-person subsets of Embody3D. The analysis split contains 87 groups, 1,600 sequences of 200 frames, and 91 individuals. We first average each participant's inferred person latent over the frames of a sequence, then average across sequences containing that same person in the same group. This produces 268 vectors, one per (person, group) pair. Distances are Euclidean distances in the full latent space; the two-dimensional PCA projection is used only for visualisation. Identity labels are used only for evaluation: training includes neither identity supervision nor a re-identification objective.

Cross-group similarity. The median same-person distance is 6.36, compared with 16.22 for different people (ratio 0.392). To make these distances easier to interpret, we rank each same-person pair within the different-person distance distribution. The median pair lies at the 15.7th percentile. Different-person pairs may come from the same group, so the reference distribution also includes people who share an interaction context.

- Top left: the displayed different-person control pairs ($n = 150$) approximately follow the uniform percentile reference.
- Bottom left: same-person pairs ($n = 399$) instead pile up near 0, with a median percentile of 16: the typical person is closer to themselves in another group than 84% of stranger pairs are to each other. The shaded strip holds 38 relatively distant same-person pairs above the 80$^{\text{th}}$ percentile, showing that this structure is not equally strong for every pair.
- Right: the cumulative view of the same percentiles; the diagonal is the uniform reference, and the shaded area shows the excess of close same-person pairs.

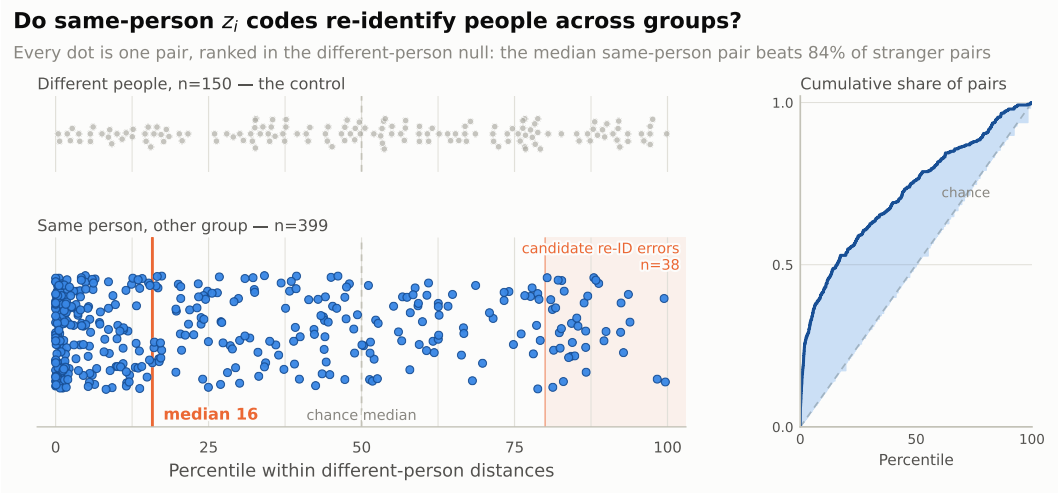


Figure 8: The individual latent z^i remembers who someone is, across different groups. Each person gets one z^i vector per group they appear in. We measure the distance between two vectors of the same person seen in two different groups, and report it as a percentile of the distances between different people: percentile 20 means that pair is closer together than 80% of pairs of strangers. If z^i carried no identity-related distance structure, the reference would be a uniform percentile distribution with median 50.

Nearest-neighbour retrieval. For each vector whose person appears in at least two groups, we rank vectors from other groups by distance and test whether the nearest neighbour belongs to the same individual. This recovers the correct identity in 99 of 239 eligible queries (41.4% top-1 accuracy), without metric learning or re-ranking. The 239 retrieval queries and 399 same-person pairs in Figure 8 count different units: one is a query vector, the other a cross-group pair. Excluding the query's group tests transfer across group contexts, although different groups can still share partners or activities. Together with the distance analysis, retrieval supports persistent participant-specific information in the person latent, while leaving room for context-dependent variation; it does not establish a context-free or causally identifiable identity representation.

C Training Details

Data preparation. We train on the Haggling, DnD, DD100, and DuoBox datasets using the train/validation/test partitions described in the respective papers. For Embody3D we use the split provided by MAGnet authors. Further, broken clips from Haggling are filtered out. All datasets are converted to the common motion representation of Section 4.1. Motion is resampled to 30 fps and divided into windows of 200 frames, using a stride of 150 frames during training. We normalise all inputs using statistics computed on the training split only. We apply person order augmentaiton and joint masking.

Training-example construction. Each training example contains between 2 and 5 participants. To train a single model for forecasting, in-filling, missing-joint completion, and response generation, we sample a context pattern for every clip. For forecasting examples, the observed prefix contains 30 to 70 frames; for temporal in-filling, observations are retained *[give the interval or mask distribution]*; and for joint- or person-level completion, we hide 30% of the joints randomly or completely hide one or more participants randomly. The observed context varies throughout the clip and across training samples, so the model learns to adapt to arbitrary context sets. Masks are regenerated at every update with a step-dependent seed. The target is always the full window of 200 frames.

Architecture. Unless otherwise stated, models contain approximately 16M trainable parameters. The group and person latent variables have dimensions 32 and 16, respectively, and the person-history LSTM has hidden width 256. The context encoder uses 3 layers with hidden width 256, 8

attention heads, and a perceiver resampler. Dedicated heads predict canonical root motion and partner transforms, while body pose is decoded through a frozen, pretrained motion VAE (Appendix D). Dropout is set to 0.1. Unless noted otherwise, these settings are shared across all datasets and evaluation tasks.

Reconstruction losses. In addition to the ELBO, we supervise the decoded body directly to stabilise training and keep rotations, joint positions, and global placement mutually consistent. The reconstruction terms in eq. (6) use the geodesic distance between rotations on $\mathrm{SO}(3)$ and the smooth L1 (Huber) penalty:

$$d_R(\hat{R}, R) = \arccos\big((\mathrm{tr}(\hat{R}^\top R) - 1)/2\big), \qquad \rho(x;\beta) = \begin{cases} x^2/2\beta, & |x| < \beta, \\ |x| - \beta/2, & \text{otherwise.} \end{cases} \tag{10}$$

With these, we supervise per-joint rotations (weighted by α_j), root-relative joint positions, and the canonical transforms of $\mathcal{M}$ defined in section 4.1:

$$\begin{aligned} \mathcal{L}_{\text{pose}} &= \mathbb{E}\Big[\textstyle\sum_j \alpha_j\, \rho\big(d_R(\hat{R}^j, R^j);\, 1\big)\Big], & \mathcal{L}_{\text{key}} &= \mathbb{E}\big[\rho(\hat{K} - K;\, 0.05)\big], \\ \mathcal{L}_{\text{root}} &= \sum_{m \in \mathcal{M}} \mathbb{E}\big[\rho\big(d_R(\hat{R}_m, R_m);\, 1\big) + \rho(\hat{t}_m - t_m;\, 1)\big], \end{aligned} \tag{11}$$

Each term is averaged over valid (non-masked) entries and summed over time and people.

Optimisation and objective. We optimise all models with AdamW (Loshchilov & Hutter, 2017), an initial learning rate of $1\mathrm{e}-4$ with a cosine scheduler. Weight decay $1\mathrm{e}-4$. The base model is trained for approximately 50 passes over the training windows, with the cosine schedule spanning the full run; gradients are clipped to a maximum norm of 1. In Equation (6), we use $\lambda_{\text{pose}} = 20$, $\lambda_{\text{key}} = 10$, and $\lambda_{\text{root}} = 1$, with unit weight on the rotation and translation of each transform in $\mathcal{M}$. We additionally use auxiliary terms with the following weights: joint velocity 10 (Huber, $\beta = 0.05$); root position 1; canonical trajectory position and yaw 10 each; canonical translation and rotation velocity 10 and 5; rotation-velocity matching 5; partner consistency 5; foot contact 1; canonical-trajectory negative log-likelihood 1; and body negative log-likelihood 0.001. The group and person KL terms use $\beta_g = 0.01$ and $\beta_p = 0.05$, respectively. During initial training, both coefficients are multiplied by $\min(s/1000, 1)$, where s is the optimisation step; continuation runs retain the full coefficients without restarting warm-up. The main posterior objective applies free bits of 0.05 nats per latent dimension, using $\max(\mathrm{KL}_d, 0.05)$ before aggregation. The posterior-dropout auxiliary objective uses the same KL coefficients without free bits.

D Model implementation

Training has two parts: a temporal motion VAE, pretrained once and then frozen, and the hierarchical latent-variable model, trained on top of the frozen VAE decoder. The hierarchical model is first trained from scratch and then refined through a sequence of continuation stages (Table 5).

Scope. The motion VAE and the staged training below are used for the DD100 and DuoBox models. The Embody3D and DnD models behind the $N = 2$–5 results in Table 2 were trained with the base BRAID objective only (Appendix C), without the pretrained motion VAE or the continuation stages due to the lack of resources. We will update these results with models trained through the full pipeline. **Likewise, the analyses and ablation tables in Appendices E to G and I use BRAID without the motion VAE.**

D.1 Temporal motion VAE

Architecture. The VAE is a temporal convolutional encoder–decoder of width 256 and depth 3, with $4\times$ temporal compression: each person receives one code $c_k^p \in \mathbb{R}^{128}$ per block of four frames. Its input per frame is the 6D rotation of each of the $J = 21$ non-root body joints together with the canonical-to-root transform $\mathbf{T}^{\text{can}\to\text{root}} \in \mathbb{R}^9$, giving 135 channels. Encoder and decoder are both conditioned on $u_t^p = [\Delta\mathbf{T}_t^{\text{can}}, \beta^p] \in \mathbb{R}^{25}$, the canonical frame-to-frame increment and the 16 shape coefficients; the increment of the first frame is set to the identity. Poses, conditions and codes are

standardised with fixed training-set statistics. The encoder outputs a diagonal Gaussian posterior whose standard deviation is bounded as $\sigma = 0.1 + 0.9\,\mathrm{sigmoid}(\cdot)$, used consistently for sampling and for the KL term. A separate head predicts foot-contact logits. During pretraining, the encoder and decoder see the complete motion of every present person and the true canonical trajectory; only padding is masked.

Objective. With $\hat{\cdot}$ denoting reconstructions, the VAE minimises

$$\mathcal{L}_{\mathrm{VAE}} = \mathrm{MSE}(\hat{\Theta}, \Theta) + 10\,\|\hat{K} - K\|_1 + 10\,\|\hat{t}^{\mathrm{root}} - t^{\mathrm{root}}\|_1 + 0.1\,\rho\big(\hat{\dot{K}}^{\mathrm{w}} - \dot{K}^{\mathrm{w}};\, 0.2\big) + 0.1\,\mathrm{BCE}(\hat{m}, m) + 0.5\,\mathcal{L}_{\mathrm{foot}} + \beta_{\mathrm{VAE}}\,\mathbb{KL}\big(q_\psi(c \mid \Theta, u)\,\|\,\mathcal{N}(0, I)\big), \tag{12}$$

where the MSE is taken over standardised poses, K are root-relative joint positions and t^{root} root positions (both in metres), $\dot{K}^{\mathrm{w}}$ are world-frame joint velocities, ρ is the Huber penalty of eq. (10), m are foot-contact labels, and $\mathcal{L}_{\mathrm{foot}}$ is a Huber stance/skating penalty. The KL is averaged per latent dimension, and β_{VAE} is ramped linearly from 0 to 10^{-3} over the first 1,000 updates.

Optimisation and freezing. We use AdamW with weight decay 10^{-4} and global-norm clipping at 1, batch size 16, and a learning rate warmed up from 10^{-5} to $2\mathrm{e}{-4}$ over 200 steps and then cosine-decayed to $2\mathrm{e}{-5}$, for approximately 100 passes over the training windows. After pretraining, the VAE weights are excluded from all optimiser updates; gradients still propagate through the frozen decoder to its inputs. A digest of the VAE parameters is re-verified at the end of every later stage.

D.2 Coupling the hierarchy to the motion VAE

Latent update rates and posterior merging. The group latent $z^g \in \mathbb{R}^{32}$ is updated every five frames and held constant in between; person latents $z_t^p \in \mathbb{R}^{16}$ are sampled every frame, conditioned on the current group sample. Each person carries a recurrent history h_t^p (LSTM, width 256), and the group transition receives a masked max-pool over people of features computed from the updated h_t^p and the previous z_{t-1}^p. In the precision merge of eq. (4), the prior precision is capped at the bottom-up precision, $\sigma_{\mathrm{td}}^{-2} \leftarrow \min(\sigma_{\mathrm{td}}^{-2}, \sigma_{\mathrm{bu}}^{-2})$, so the posterior cannot be dominated by the prior when target evidence is available. During training, the bottom-up evidence is produced by the context encoder of section 4.3 applied to the full target, followed by a backward LSTM smoother.

Trajectory and interaction heads. Dedicated heads predict the canonical increments $\Delta\mathbf{T}_t^{\mathrm{can}}$ and the partner transforms $\mathbf{T}_t^{p\to q}$ from z^g, z_t^p, and h_t^p.

Body path. Person latents are first smoothed with a fixed $[1, 2, 1]/4$ temporal filter at strength 0.5. A trainable temporal adapter (four-frame packing, width 256, three dilated residual blocks) maps $[z^g, z^p, \phi(h^p), \bar{u}^p]$, where $\phi(h^p)$ are history features and $\bar{u}^p$ the normalised canonical increments and shape coefficients, to standardised 128-dimensional codes $\hat{c}_k^p$. These codes are de-standardised and decoded by the frozen VAE decoder into joint rotations and $\mathbf{T}^{\mathrm{can}\to\mathrm{root}}$, which are then supervised with the reconstruction losses of eq. (11). The decoder condition uses the recorded $\Delta\mathbf{T}^{\mathrm{can}}$ with probability 0.5 during training (teacher forcing) and the predicted increment otherwise; at generation it always uses the prediction.

Generation. Generation reads only the observed context: latents are sampled from the priors, and neither target motion nor target codes are accessed. World-frame roots are anchored to the observed roots.

D.3 Training stages

Stage 1: base model. The hierarchical model is initialised from scratch around the frozen VAE and trained with the objective and schedule of Appendix C at batch size 4. In this stage, gradients from the trajectory losses to z^p through its direct input to the trajectory head are stopped.

Continuation recipe (Stages 2–4). Each continuation restores model weights, AdamW moments, the step counter, and the random-number state from its parent's final checkpoint. We verify that

Table 5: Training stages. The motion VAE is frozen after Stage 0. Each continuation restores the full state of its parent's final checkpoint.

Stage	Initialisation	Trainable	Change relative to parent
0. Motion VAE	Scratch	VAE	–
1. Base	Scratch; VAE from 0	All but VAE	–
2. Posterior dropout	Stage 1	All but VAE	Dual-rollout posterior dropout; trajectory losses reach z^p directly
3. Code supervision	Stage 2	All but VAE	Frozen-encoder code loss on the dense branch
4. VAE-informed posterior	Stage 3	All but VAE, + projection	Frozen-encoder codes as additional z^p posterior evidence

a save/restore round trip is bitwise exact and that the restored model reproduces the parent's final held-out evaluation before training resumes. Continuations replace the parent's schedule with a constant learning rate of $3\mathrm{e}{-5}$, keep weight decay and clipping unchanged, fix the KL multiplier at 1 (no annealing), and run for half the number of updates of Stage 1.

Stage 2: dual-rollout posterior dropout and trajectory coupling. Each update evaluates two independent recurrent rollouts with the same weights. The *dense* branch samples every frame from the posterior and uses the full Stage-1 objective, including teacher forcing and free bits. The *mixed* branch draws a per-frame mask $b_t \sim \mathrm{Bernoulli}(0.5)$, shared across people and the minibatch, and samples frame t from the posterior if $b_t = 1$ and from the prior otherwise. Only posterior-sampled frames are scored in this branch, using the likelihood and the raw KL (the group KL only on its update frames); normalisers use the full count of eligible frames, geometry, velocity and foot terms are omitted, and the body decoder always uses the predicted trajectory. The two gradients are summed with unit weight, clipped jointly, and applied as a single AdamW step. In both branches the posterior retains access to the full target: posterior dropout changes where samples are drawn from, not the evidence available to the posterior. This stage also removes the Stage-1 stop-gradient, so trajectory losses update z^p through its direct input to the trajectory head.

Stage 3: frozen-encoder code supervision. In the dense branch only, we add

$$\mathcal{L}_{\mathrm{code}} = \frac{1}{|\mathcal{V}|\, D} \sum_{(b,k,p)\in\mathcal{V}} \sum_{d=1}^{D} \left(\frac{\hat{c}^{p}_{b,k,d} - \mathrm{sg}\big[\mu_\psi(Y,u)^{p}_{b,k,d}\big]}{s_d} \right)^2 , \tag{13}$$

where $\hat{c}$ are the adapter's codes in the dense posterior rollout, $\mu_\psi(Y,u)$ is the frozen VAE encoder's posterior mean for the complete training motion under the true conditions, sg denotes stop-gradient, s_d is the fixed training-set standard deviation of code channel d, $D = 128$, and $\mathcal{V}$ is the set of valid (batch, code position, person) indices. Targets are packed exactly as in VAE pretraining. The weight $\lambda_{\mathrm{code}} = 3.5$ was fixed before training on the first four continuation batches, such that the gradient norm of the code term over the adapter parameters equals half that of the dense-branch objective. A rule based on the whole-model gradient norm was not used, since that norm is dominated by trajectory and hierarchy gradients by roughly three orders of magnitude.

Stage 4: VAE-informed posterior. Stage 4 continues Stage 3 with the same recipe and λ_{code}. Frozen-encoder mean codes of the posterior-visible target frames are standardised, repeated over each four-frame block, passed through a new zero-initialised linear layer ($128 \to 80$, 10,320 parameters), and added to the per-person evidence entering the bottom-up person posterior $q_{\mathrm{bu}}(z^p_t \mid h^p_t, \cdot, z^g)$. The group posterior, all priors, and the generation path are unchanged; the codes are detached and the VAE remains frozen. Before training we verify that installing the branch leaves the dense rollout bit-identical, that generation is unchanged when hidden target frames are corrupted, and that existing AdamW moments are preserved while the new parameters' moments start at zero. The DD100 and DuoBox results use the final Stage 4 checkpoint.

Implementation. All models are implemented in Jax/Flax NNX.

E Model Size Comparisons

We vary model capacity while holding the training data, learning-rate schedule, and epoch budget fixed. This comparison evaluates model selection under a common training budget; it does not establish a scaling limit. Larger models may require different optimisation settings or more data, which this sweep does not test.

The 16M model attains the lowest SDTW, cross-person SDTW, and FD on both Panoptic and DD100 (Tables 6 and 7). Larger variants improve determinism error and produce greater sample spread, but their higher cross-person SDTW and FD show that this additional variation does not translate into closer interaction structure or distributional agreement under the tested setup. The 16M choice therefore provides a useful balance for our emphasis on coordinated, distributionally faithful generation, rather than a claim of superiority on every diagnostic.

The best result in each column is shown in bold, and the second-best distinct result is underlined. Tied values receive the same formatting.

Table 6: Model-size comparison on Panoptic.

Model	CRQA (RR)	CRQA (DET)	SDTW ($\times 10^3$)	xPerson SDTW ($\times 10^3$)	FD	DIV	FS	MPJPE	MPJVE
4M	**0.011**	0.33	3.146	0.774	0.86	0.18	**0.52**	0.16	0.013
16M	**0.011**	0.28	**3.061**	**0.511**	**0.27**	0.42	0.59	**0.13**	**0.012**
25M	**0.011**	**0.21**	8.596	1.489	0.34	1.40	0.55	0.20	0.015
32M	**0.011**	**0.21**	8.600	1.371	0.34	**1.47**	0.61	0.20	0.013

Table 7: Model-size comparison on DD100.

Model	CRQA (RR)	CRQA (DET)	SDTW ($\times 10^3$)	xPerson SDTW ($\times 10^3$)	FD	DIV	FS	MPJPE	MPJVE
4M	0.012	0.45	2.818	1.072	2.51	12.00	0.14	0.83	**0.04**
16M	**0.010**	0.36	**2.390**	**0.785**	**0.63**	5.81	**0.13**	**0.72**	0.05
25M	0.012	0.23	6.358	1.526	0.92	33.39	0.16	1.36	0.05
32M	0.012	**0.22**	6.202	1.463	1.04	**35.47**	0.15	1.34	0.05

F Cross-dataset and cross-domain generalisation

Setup. We distinguish generalisation to unseen groups from transfer to an unseen activity domain. Model A is trained jointly on Panoptic Haggling (three-person conversational interaction) and DD100 (partnered dance), whereas Model B is trained only on DD100. Both are evaluated on the same held-out Panoptic groups, using 40 frames of conversational motion as context and generating the following 60 frames. Evaluation covers 200 batches of 32 sequences. Thus, both models encounter unseen groups, while conversational interaction is an unseen training domain only for Model B. Adaptation uses the supplied context without test-time parameter updates.

Table 8: Cross-domain forecasting on held-out Panoptic Haggling groups, with 40 context frames and 60 generated frames. Bold: best; underlined: second best.

			SDTW↓	Cr. SDTW↓					
Training data	ΔRR↓	ΔDET↓	$\times 10^3$	$\times 10^3$	FD↓	DIV↑	FS↓	MPJPE↓	MPJVE↓
Panoptic + DD100	**0.011**	**0.28**	**3.061**	**0.511**	**0.27**	0.42	0.59	**0.13**	**0.012**
DD100 only	**0.011**	0.31	4.473	1.424	1.94	**1.10**	**0.24**	0.28	0.026

Activity-specific dynamics within a shared model. Joint training improves cross-person SDTW from 1.424 to 0.511 and FD from 1.94 to 0.27 (Table 8), alongside closer temporal alignment and reference recovery. In the qualitative comparison (Figure 9), the jointly trained model maintains the conversational formation and generates gestures appropriate to that activity. The dance-only model retains plausible body poses and a sensible group formation, but also produces spinning motions

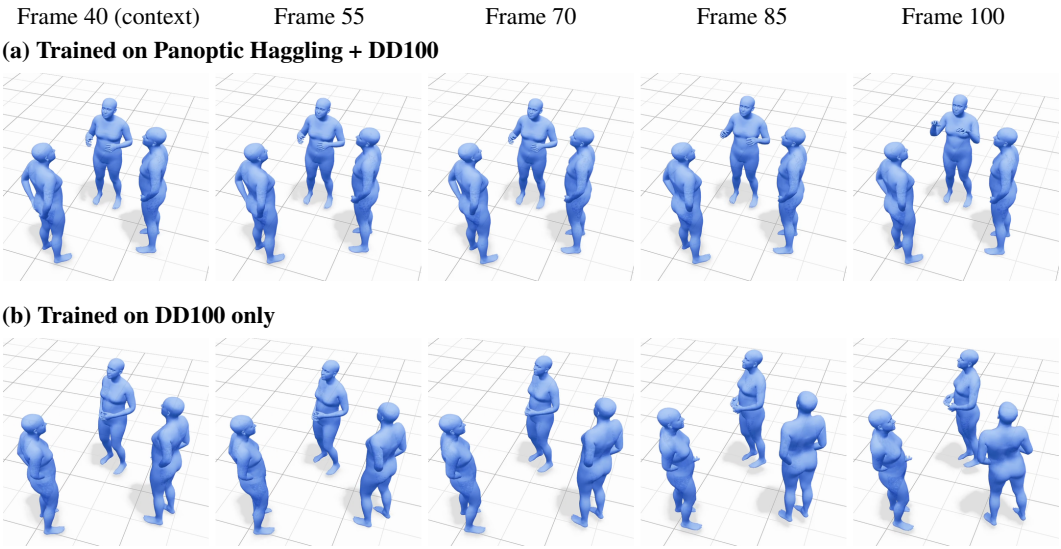


Figure 9: Cross-domain forecasting on a held-out Panoptic Haggling group, using frames from the supplementary videos. Both models observe the same 40 frames of conversational motion and generate the next 60. Context frames are rendered from each model's own reconstruction, so the frame-40 panels differ slightly between rows. (a) The jointly trained model keeps a stable conversational formation while the central participant continues to gesture. (b) The DD100-only model, which never sees conversational data during training, also preserves plausible poses and a coherent conversational formation, indicating that it carries forward the group structure given in the context. However, the right-hand participant turns away from the group, a rotation characteristic of partnered dance rather than conversation.

characteristic of its training domain. This separates transfer of broad pose and formation structure from adaptation to activity-specific interaction dynamics. The result supports learning multiple domains with shared weights and conditioning on new groups; it does not establish equivalent performance on activities absent from training.

The dance-only model has higher DIV and lower FS, illustrating why sample spread and foot stability should be interpreted alongside activity-appropriate alignment and distributional fit. The group-latent structure in Figure 6 complements this comparison: corpus and dance-family organisation is consistent with the inferred social state retaining activity information within a shared model.

G Contribution of the bilevel hierarchy

Variants. All variants in this section, including the full model, are trained without the motion VAE (Appendix D). We compare the full group-and-person hierarchy with two removals, keeping the remainder of the architecture fixed. The *group-only* variant removes person latents, so individual variation must be represented through the shared z^g. The *person-only* variant removes social context: the person axis is folded into the batch, and each participant is modelled independently through z^p. These comparisons test the functional contributions of the two levels alongside the latent-space analysis in Appendix B.

Table 9: Bilevel-hierarchy ablations on Panoptic Haggling. Bold: best; underlined: second best.

Latent structure	ΔRR↓	ΔDET↓	SDTW↓ $\times 10^3$	Cr. SDTW↓ $\times 10^3$	FD↓	DIV↑	FS↓	MPJPE↓	MPJVE↓
z_g and z_i	**0.011**	<u>0.28</u>	**3.061**	**0.511**	<u>0.27</u>	0.42	**0.59**	**0.13**	**0.012**
z_g	**0.011**	0.38	<u>3.193</u>	<u>1.119</u>	0.51	<u>0.67</u>	0.69	0.21	**0.012**
Person only (no social context)	<u>0.012</u>	**0.19**	8.530	1.305	**0.18**	**1.28**	<u>0.61</u>	<u>0.18</u>	<u>0.014</u>

Individual flexibility supports interpersonal alignment. On Panoptic, removing person latents changes SDTW modestly, from 3.061 to 3.193, but more than doubles cross-person SDTW, from 0.511 to 1.119 (Table 9). Removing social context increases cross-person SDTW further to 1.305, while SDTW rises to 8.530. The same cross-person ordering appears on DD100: 0.785 for the full hierarchy, 0.920 for group only, and 1.534 for person only. On DD100, removing social context also raises SDTW from 2.39 to 6.35 and MPJPE from 0.72 to 1.30. Together, these results suggest that separate person latents help express individual behaviour in relation to the group, and that social context contributes to both interpersonal alignment and individual motion prediction.

Complementary metrics reveal the contribution of social structure. The person-only model attains lower FD on Panoptic (0.18 versus 0.27) and lower determinism error on both datasets. These are complementary strengths rather than evidence of uniform superiority by either variant. FD over per-person frame features evaluates marginal motion statistics, while determinism summarises one aspect of recurrence structure; neither alone establishes that the generated participants remain aligned with one another. The full hierarchy's advantage in cross-person SDTW therefore provides more specific evidence for the social-state motivation: modelling participants jointly helps recover the structure of their interaction. These architectural ablations support distinct functional contributions of group and person latents, without asserting causal identifiability of the learned representation.

Conditional person-level variation. We also examine generation while keeping the observed context and group latent fixed and varying only the person latents (Figure 10a). In the DuoBox example, three samples produce different poses and orientations from the same context. Their relative arrangements also vary: the person latents express individual realisations of an interaction rather than motion details that leave group geometry invariant. This complements the re-identification result in Appendix B: participant-specific information can persist across contexts while allowing multiple ways to participate in the current interaction.

Changing the group state during generation. In a second experiment, we generate 50 frames using a group-latent trajectory from one source sequence, then switch to the trajectory from another sequence containing the same individuals in a different interaction. Person latents are sampled conditionally on the active group latent, so they respond to the switch through the hierarchy. The selected frames in Figure 10b show a change in participants' postures and group arrangement. This illustrates that the group latent affects the generated interaction through its conditional person-level realisations. Since the person latents also change, the experiment tests the hierarchy's response to a change of group state, rather than isolating a direct effect of the group latent with every other latent held fixed.

H Evaluation metrics and interpretation

Our evaluation follows the multiparty behaviour framework of Shirekar et al. (2025) and combines interaction structure, distributional agreement, sample variation, and physical diagnostics. These quantities address complementary questions; no single score establishes that a continuation is both plausible and socially appropriate.

Baselines. RS and NN are naive, non-learned reference baselines: RS returns a randomly sampled sequence from the training data, and NN returns the training sequence closest to the observed context. They indicate the scores reachable by replaying recorded motion without modelling the interaction, and are therefore excluded from the best/second-best ranking in the main tables. MAGNet (Maluleke et al., 2025) was not evaluated on five-person interactions in its original work, so we do not report it for DnD ($N = 5$) in Table 2; for that setting under full observation, where no learned baseline remains, the ranking includes RS and NN.

Recurrence and temporal coordination. Cross-recurrence quantification analysis (CRQA) compares participants' motion states across time. A recurrence occurs when two states are within a dataset-specific distance radius, calibrated to a 2% recurrence rate and then held fixed for analysis (Wallot & Leonardi, 2018; Wallot, 2019; Coco et al., 2021; Shirekar et al., 2025). Recurrence rate (RR) is the fraction of state pairs marked as recurrent. Determinism (DET) is the fraction of

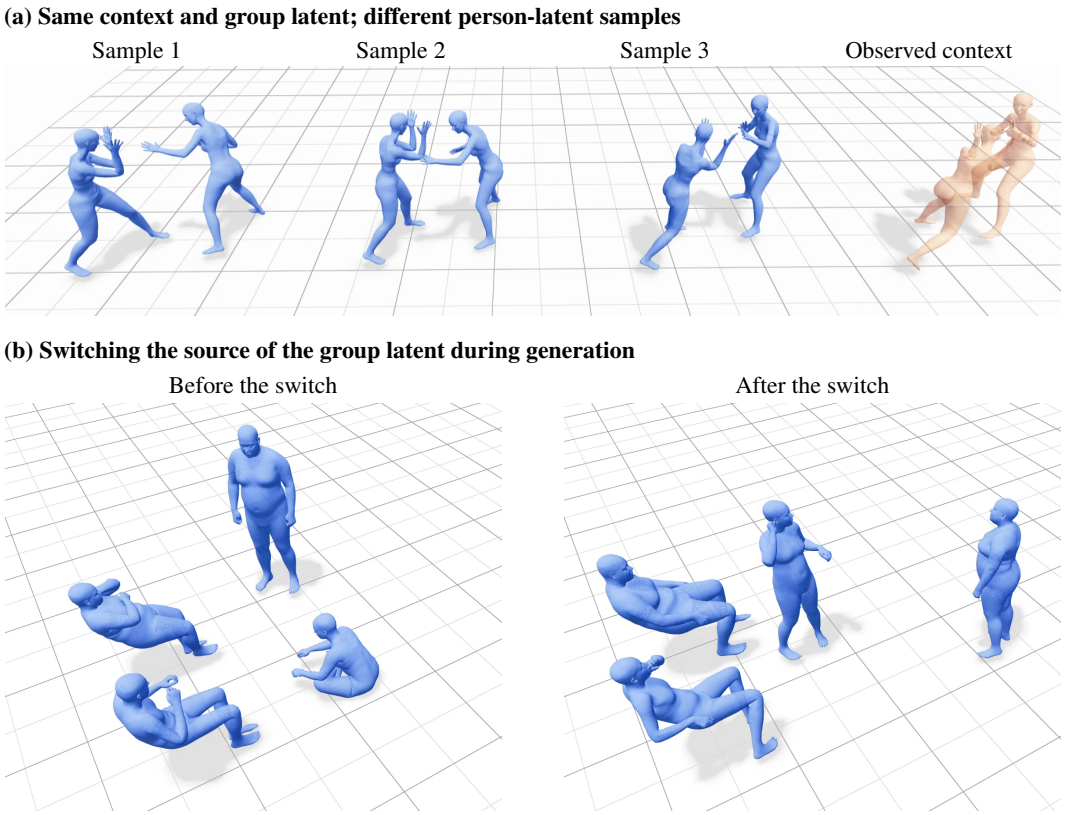


Figure 10: Qualitative probes of the latent hierarchy, using frames from the supplementary videos. (a) Three generated dyads (blue) share the same context (peach, right) and group latent but use different person-latent samples. (b) Selected states before and after changing the group-latent source; person latents are sampled conditionally on the active group state.

recurrent points that belong to diagonal line structures, reflecting sustained sequences of similar states. Off-diagonal recurrences allow temporal offsets, so the comparison can capture leading and following rather than requiring participants to move identically at the same instant. We report absolute deviations of RR and DET from their ground-truth values, denoted $\Delta\mathrm{RR}$ and $\Delta\mathrm{DET}$; lower values indicate closer agreement with the recorded recurrence statistics.

Trajectory structure with timing flexibility. Soft dynamic time warping (SDTW) compares motion trajectories while allowing temporal alignment. It therefore tolerates timing shifts that framewise position error penalises directly. The cross-person variant applies this comparison to relations between participants' motion signals, testing whether the generated interaction preserves their relative temporal structure. Lower values are better in both cases. Cross-person SDTW is especially relevant to our social-state motivation: individually plausible motions can still form a poorly aligned group. Where a table header specifies $\times 10^3$, each displayed entry represents that many thousands in the underlying SDTW score.

Distributional fit and variation. Fréchet distance (FD) compares generated and reference motion distributions through the means and covariances of their evaluation features. Lower FD indicates closer agreement in that feature space. For the hierarchy ablations, these are per-person frame features: a good marginal motion distribution can coexist with errors in how people move together. Diversity (DIV) measures the average pairwise distance between generated samples. Higher DIV indicates more variation, but does not by itself demonstrate coverage of valid interactions. We consequently interpret DIV alongside FD, temporal structure, and the activity: a relatively stationary conversation need not exhibit the same motion spread as dancing or boxing.

Table 10: Additional baselines on Panoptic Haggling: 70 observed and 130 predicted frames, with 20% missing context joints. Bold: best; underlined: second best; –: unreported.

Model	ΔRR↓	ΔDET↓	SDTW↓ $\times 10^3$	Cr. SDTW↓ $\times 10^3$	FD↓	DIV↑	FS↓	MPJPE↓	MPJVE↓
BRAID	**0.011**	0.28	**3.061**	**0.511**	**0.27**	0.42	0.59	**0.13**	0.012
SNP-style	**0.011**	0.38	3.193	1.119	0.51	**0.67**	0.69	0.21	0.012
SP-style	**0.011**	**0.20**	4.530	1.940	0.92	0.11	0.31	0.21	**0.006**
RoHM	0.013	0.21	8.251	1.007	1.285	–	**0.17**	0.442	0.052

Table 11: Additional baselines on DD100: 70 observed and 130 predicted frames, with 20% missing context joints. Bold: best; underlined: second best; –: unreported.

Model	ΔRR↓	ΔDET↓	SDTW↓ $\times 10^3$	Cr. SDTW↓ $\times 10^3$	FD↓	DIV↑	FS↓	MPJPE↓	MPJVE↓
BRAID	**0.010**	0.36	**2.390**	**0.785**	**0.63**	**5.81**	0.13	0.72	0.05
SNP-style	0.012	0.38	2.620	0.920	1.45	3.53	0.12	**0.71**	0.04
SP-style	**0.010**	**0.28**	4.303	1.504	2.07	1.28	**0.08**	0.90	**0.03**
RoHM	0.013	0.29	7.017	1.493	2.29	–	0.31	1.03	0.11

Reference recovery and physical artefacts. Mean per-joint position error (MPJPE) averages the distance between predicted and recorded joint positions; mean per-joint velocity error (MPJVE) similarly measures velocity disagreement. Lower values indicate closer recovery of the particular recorded continuation. Because other continuations may also be valid, these errors complement the generative metrics rather than defining overall sample quality. Foot skating (FS) measures sliding during detected ground contact, and interpenetration (IP) measures body overlap. Lower values indicate fewer of these specific artefacts; they provide useful physical checks without establishing physical plausibility in every respect. Both are also non-zero in recorded motion for some activities, for example when dancers in close hold overlap under the capsule test or boxers move their feet quickly, so values below the recorded rate need not indicate better motion. Where relevant, we quote these recorded rates alongside the results.

I Additional neural-process baseline comparisons

Protocol and baselines. We report an additional forecasting comparison using 70 observed frames with 20% of joints randomly missing and a 130-frame prediction horizon. These results belong to this specific protocol and should be read separately from the main evaluation tables. We compare BRAID with SP-style (Raman et al., 2023) and SNP-style (Singh et al., 2019) implementations and RoHM (Zhang et al., 2024). The SP-style implementation is adapted to accept missing joints; the SNP-style variant uses a single group latent without the person-level hierarchy.

A consistent gain in trajectory and interaction structure. BRAID achieves the lowest SDTW, cross-person SDTW, and FD on both datasets in this comparison (Tables 10 and 11). The result links the hierarchy's benefit to both individual trajectory structure and interpersonal alignment, alongside better distributional agreement. The advantage is not uniform across all diagnostics: SP-style has lower determinism and velocity errors, and lower foot skating; SNP-style yields greater diversity on Panoptic and slightly lower position error on DD100. On DD100, BRAID combines the highest reported diversity with the lowest FD. Read together, these measures support richer samples that remain close to the reference distribution under this protocol, while the physical and reconstruction scores identify complementary strengths of the baselines.