Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dynamic Networks of Social Interaction: Reconfiguration of Joint Attention across Time, Context, and Individuals

Jan-Marten Ohlenmacher

ohlenmacher@med.uni-frankfurt.de

University Hospital Frankfurt, Goethe University

David Goretzko

Goethe University Frankfurt

Rosa Weber

University Hospital Frankfurt, Goethe University

Paula R. Neubauer

University Hospital Frankfurt, Goethe University

Munib Uddin

University Hospital Frankfurt, Goethe University

Merlinda Bracht

University Hospital Frankfurt, Goethe University

Elena Maher

University of Brescia

Leonie Polzer

University Hospital Frankfurt, Goethe University

Naisan Raji

University Hospital Frankfurt, Goethe University

Luisa Schnettler

University Hospital Frankfurt, Goethe University

Solvejg Kleber

University Hospital Frankfurt, Goethe University

Christian Lemler

University Hospital Frankfurt, Goethe University

Janina Kitzerow-Cleven

University Hospital Frankfurt, Goethe University

Ziyon Kim

University Hospital Frankfurt, Goethe University

Karoline Teufel

University Hospital Frankfurt, Goethe University

Christine M. Freitag

University Hospital Frankfurt, Goethe University

Nico Bast

University Hospital Frankfurt, Goethe University

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11677308/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dynamic Networks of Social Interaction: Reconfiguration of Joint Attention across Time, Context, and Individuals

Jan-Marten Ohlenmacher[1], David Goretzko[2], Rosa Weber[1], Paula R. Neubauer[1], Munib Uddin[1], Merlinda Bracht[1], Elena Maher[3], Leonie Polzer[1], Naisan Raji[1], Luisa Schnettler[1], Solvejg Kleber[1], Christian Lemler[1], Janina Kitzerow-Cleven[1], Ziyon Kim[1], Karoline Teufel[1], Christine M. Freitag[1] & Nico Bast[1]

Affiliations

[1] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Autism Research and Intervention Center of Excellence, University Hospital Frankfurt, Goethe University Frankfurt, Frankfurt am Main, Germany.

[2] Department of Psychological Methods, Goethe University Frankfurt, Frankfurt am Main, Germany.

[3] Unit of Child and Adolescent Neuropsychiatry, Department of Clinical and Experimental Sciences, ASST Spedali Civili, University of Brescia, Brescia, Italy.

Correspondence

Corresponding author: Jan-Marten Ohlenmacher

Email: ohlenmacher@med.uni-frankfurt.de

Abstract

Social interaction is an avenue of social learning that defines human development. Joint attention emphasizes this social learning but its conditional unfolding within dyadic interaction is largely unknown. Here, we derive atomic interaction states from videotaped social encounters and model them as a dynamic network based on a Bayesian hierarchical semi-Markov framework. This quantifies how person-level features affect state durations and transition probabilities, while modelling temporal and contextual dependencies. We benchmarked this in preschool autism, where joint attention is heterogenous. Across 25,682 interaction states from 28 children at two timepoints, recent joint engagement raised transition rates while elapsed duration shifted destinations. Task context defined most transitions, whereas higher nonverbal cognitive ability was associated with less disengagement, faster recovery from aversion, and both quicker and more probable responding to referential bids. The framework characterizes ecologically valid social interaction as a dynamic network and quantifies the factors that reconfigure it.

Social interaction likely propelled human development from evolution to culture (Tomasello et al., 2005). Central to this capacity is joint attention (JA), the triangulation of attention with an interaction partner on a shared object or context (Carpenter et al., 1998). By coordinating focus with another person, JA decouples cognition from perception (Sodian & Kristen-Antonow, 2015) in the development of a Theory of Mind (Charman et al., 2000). This predicts language acquisition (Bottema-Beutel, 2016) and executive control (Vaughan Van Hecke et al., 2012). Quantifying joint attention is therefore key to understanding different social interaction.

Such non-normativity is common across mental health conditions. In adults, reduced social interaction has been linked to a detachment spectrum overlapping with anxiety and depression (Kotov et al., 2017). In childhood, neurodevelopmental conditions including autism (Falck-Ytter et al., 2023) and attention deficit hyperactivity disorder with co-occurring conduct problems are associated with impaired social interactions (Fairchild et al., 2019). These childhood-onset conditions often precede adulthood anxiety and depression (Erskine et al., 2016; Lai et al., 2019), which suggests overlapping mechanisms of social interaction problems across conditions and development. Autistic preschoolers show heterogeneous JA development (Mundy, 2018), which may offer a model for characterizing how social interaction varies between individuals.

In normative development, JA occurs as a switch from dyadic to triadic social interaction within the first year of life (Scaife & Bruner, 1975). Building on face preference and attentional orienting (Portugal et al., 2024), infants first follow referring cues and associate the words expressed by the interaction partner with the jointly attended object (Lasch et al., 2023). This shared attention remains reactive until, around nine months, infants begin to actively initiate joint attention themselves through pointing and gazing (Falck-Ytter et al., 2023). Across these components, JA unfolds as a continual alternation of gaze between the partner and the object, which describes a sequence of atomic interaction states (Fan et al., 2019) rather than a single behavior. Thus, the durations of these states and the transitions between them constitute a network of social interaction.

Yet, current measures quantify JA through aggregated scores, which collapse temporal and functional interdependencies into summary counts and proportions. For example, the Early Social Communication Scales (ESCS) utilize structured behavioral observations across different tasks to differentiate initiating versus reactive JA (Mundy et al., 2003). Such measures are clinically established but discard more fine-grained social interaction dynamics. This reductionism may carry practical consequences, as meta-analyses of early autism interventions report only modest increases in traditional JA outcomes, despite plausible changes in the underlying interaction (Sandbank et al., 2023; Tiede & Walton, 2019). Consequently, aggregated measures might be insensitive to the micro-level quality at which social interaction development occurs.

Capturing this change requires modelling the transitions between interaction states directly. In eye-tracking research, transition models represent gaze as movement through

a discrete state space, described either as probabilities in discrete time (Masedu et al., 2022), or as intensities, the instantaneous transition rates between states in continuous time (Jackson, 2011). Applied to autism, a continuous-time model found lower transition intensities toward and among socially salient regions in autistic than in non-autistic children (Bruschetta et al., 2025). These findings, however, derive from screen-based paradigms, in which a child passively watches a recorded interaction, removing the reciprocal contingency central to social interaction (Redcay & Schilbach, 2019). In contrast, few studies preserve this contingency by recording gaze during live interaction (Falck-Ytter et al., 2015; Thorup et al., 2016; Tönsing et al., 2025). Yet, these studies quantify gaze with summary measures, rather than modelling the transitions between states as a process with its own durations and destinations. One study has modelled transitions in live dyadic gaze based on annotations, but it restricted the state space to looking at the partner versus looking away (Alie et al., 2011), collapsing the components of JA into a single partner-directed state. Studying these components in real-world interaction thus requires both a state space rich enough to distinguish them and an explicit model of the transitions among them.

To this end, we operationalize JA as a network of atomic interaction states (Fan et al., 2019) in social dyads (C = Child, T = Tester). These states entail mutual eye contact (Mutual-CT), shared attention to a common object (Share-CT), referring and following by either partner (Refer-C/Refer-T and Follow-C/Follow-T), and gaze aversions (Avert-C/Avert-T). Within this state space, nine focal transitions are pre-specified and mapped to five theoretically derived domains of social interaction (Table 1). The first two correspond to initiating and responding components of joint attention (Mundy et al., 2009). Next, disengagement and reengagement represent leaving and returning to joint engagement; the disruption and repair of interaction described in the mutual-regulation tradition (Tronick, 1989). Lastly, synchrony captures transitions in which the child and tester switch between mutual eye contact and shared attention simultaneously. Interpersonal synchrony has been previously defined to admit short lags (Georgescu et al., 2020), whereas we require strict simultaneity and assign the lagged case to responding.

Table 1

Focal state transitions of interest and associated social interaction domains.

Transition	Description	Domain
Mutual-CT → Refer-C	The child refers to an object after having established eye contact with the tester	Initiating
Refer-T → Follow-C	The child reacts to the tester's referring by following their gaze to the target object	Responding
Follow-C → Share-CT	The child's gaze following is continued into shared attention on the object with the tester	Responding
Mutual-CT → Avert-C	The child averts its gaze from eye contact with the tester	Disengagement
Share-CT → Avert-C	The child averts its gaze from shared attention with the tester	Disengagement
Avert-C → Mutual-CT	The child's gaze returns to eye contact with the tester after a period of disengagement	Reengagement
Avert-C → Share-CT	The child's gaze returns to shared attention on an object after a period of disengagement	Reengagement
Mutual-CT → Share-CT	The child and tester switch simultaneously from eye contact to shared attention on an object	Synchrony
Share-CT → Mutual-CT	The child and tester switch simultaneously from shared attention on an object to eye contact	Synchrony

To characterize these transitions, we model their intensities and extend this approach in two ways. First, an intensity is the standard quantity in continuous-time transition models. It combines how long a state is held with which state follows it, so that a group difference like a reduced social orienting reported in autism (Bruschetta et al., 2025), cannot be attributed to either. We therefore decompose each transition into two interpretable quantities, the sojourn (the time spent in a state) and the next-state probability (which state follows). Together, these yield a directed, weighted network of social interaction, in which the interaction states constitute the nodes and the next-state probabilities the weighted edges, each node carrying its sojourn duration.

Second, a continuous-time Markov process rests on assumptions that live interaction is likely to violate (Aalen et al., 2008; Cook & Lawless, 2018). The Markov property states that a future state depends only on the present state. This restricts sojourns to exponential shapes, so that a state's rate of leaving cannot change with its duration. We give each state a Weibull sojourn instead, yielding a semi-Markov process in which that rate may rise or fall with elapsed time (Asanjarani et al., 2022). Further, this semi-Markov formulation holds next-state probabilities independent of elapsed duration, a dependence we estimate rather than assume. To address the influence of previous states, we add a history term for each domain (Andersen & Keiding, 2002; Kay, 1986), estimating whether joint engagement (Mutual-CT or Share-CT) as the preceding state alters following transitions. Lastly, time homogeneity would hold the transition intensities constant throughout, so we enter session phase and current task as time-varying predictors, which allow the dynamics to differ across the session and across activities (Andersen & Keiding, 2002). Together, this represents a Bayesian hierarchical semi-Markov model, which has been applied to behavioral observations for quantifying multistate transitions of emotion expression (Dagne & Snyder, 2009), but not to dyadic gaze.

We apply this model to 25,682 atomic interaction states, annotated from ESCS observations of 28 autistic preschoolers assessed at two timepoints within a randomized controlled trial (Freitag et al., 2025; Kitzerow et al., 2020). A single hierarchical model establishes the average structure of this network, providing its sojourn durations, baseline transition intensities, and next-state probabilities. Temporal and structural dependencies then capture their dynamic variation within a session through task and phase contrasts, a recent-engagement history term, and a dependence of the next-state probability on the current sojourn duration. Person-level covariates (age, sex, nonverbal cognitive ability, timepoint, and intervention arm) enter as multiplicative effects on the focal transition intensities, complemented by subject-level random effects, so that social interaction is characterized by both its intra- and interindividual variability.

Results

Durations of and transitions between atomic interaction states

Descriptive data of atomic interaction states are displayed below (Table 2). Social encounters were dominated by shared object-focused attention. Child gaze aversion was the second most prominent state with shorter durations. Mutual eye contact was relatively frequent but brief. Transient states like following and referring were of short median durations. "Other" collapsed states not of interest to dynamic social interaction including single fixations of the child and parent-related behaviors. "Break" marked intervals in which coding was not possible (e.g., obstructed camera view). Other and Break states were retained to preserve an event-time sequence (see Pre-processing in Methods).

Table 2

Duration of atomic interaction states by manual annotation with sojourn descriptives across sessions and individuals for child (C) and tester (T).

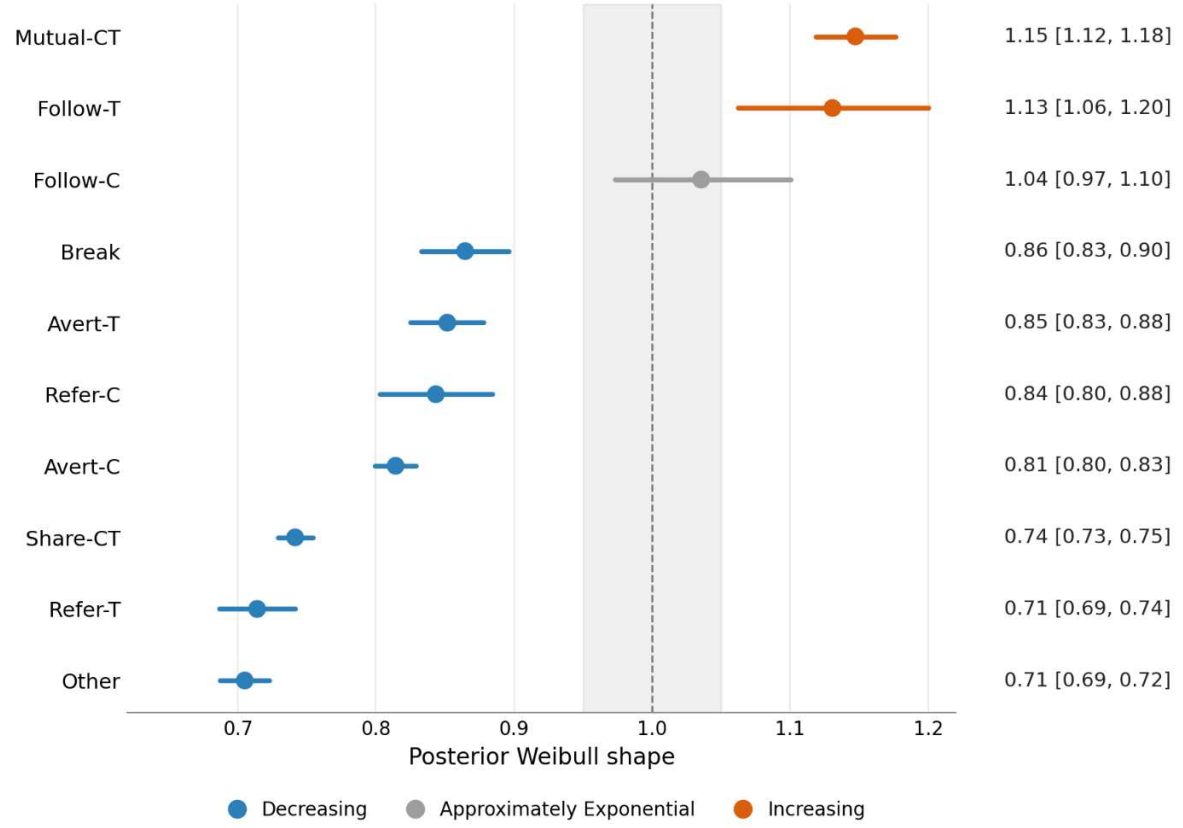
State	*n* Sojourns	Median Duration (s) [IQR]	Time Occupancy (%)
Share-CT	7,376	2.68 [1.00, 7.84]	59.4
Avert-C	6,172	1.08 [0.40, 2.72]	16.6
Other	3,359	1.27 [0.44, 3.32]	11.1
Break	1,320	2.04 [1.04, 4.24]	6.2
Mutual-CT	3,013	0.60 [0.36, 1.04]	3.0
Refer-T	1,418	0.48 [0.16, 1.56]	2.4
Refer-C	800	0.60 [0.28, 1.24]	0.9
Follow-C	475	0.20 [0.10, 0.44]	0.2
Avert-T	1,302	0.04 [0.04, 0.04]	0.1
Follow-T	447	0.12 [0.08, 0.20]	0.1

Note. States are ordered by time occupancy (descending). *n* Sojourns is the number of state occurrences, which are occupancies that ended in an observed transition within the session. Median duration and interquartile range (IQR) are computed on those sojourns, each measured from the onset of the origin state to the onset of the destination state. Time occupancy is the share of total recorded time spent in the state, computed over all analyzed segments including those right-censored at a phase boundary or at the end of a session.

The hazard of leaving a state, the instantaneous rate at which the state ends, depended on how long the state had already lasted, as captured by non-exponential Weibull sojourn-shape parameters (Figure 1). Seven of the ten interaction states had shape parameters credibly below 1, indicating that the hazard of leaving the state decreased the longer it had persisted, including each of the most prevalent states (Share-CT, Avert-C, Other, and Break). Mutual eye contact and tester gaze following were the only states with shape parameters credibly above 1, describing cases in which the hazard increased with elapsed duration, while child gaze following was approximately exponential.

Figure 1

State-dependent persistence of gaze behaviors represented by posterior Weibull sojourn-shape parameters.



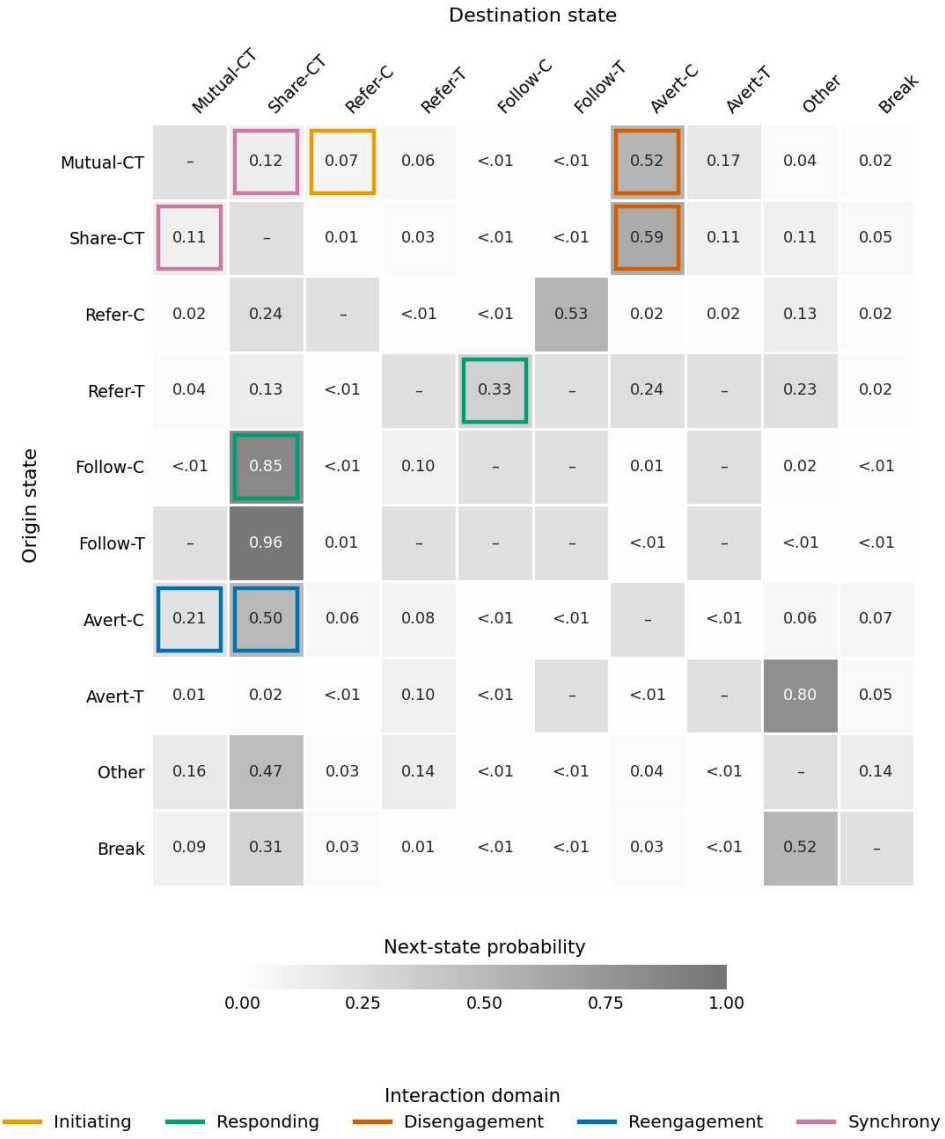
Note. Posterior mean (points) and 95% credible intervals (lines) of the Weibull shape parameter (α) for each state's sojourn-time distribution from the full hierarchical semi-Markov model. Each state's sojourn is the time until exit to any destination, so the relevant hazard aggregates the competing transition intensities into the state's exit rate. α determines how that hazard changes as the sojourn continues, with the exit rate fixing its overall level and α its trajectory: α < 1, decreasing hazard (blue); α > 1, increasing hazard (orange); α ≈ 1, approximately constant (grey). The dashed line marks α = 1 and the shaded band (0.95–1.05) the approximately-exponential region. As a state-level parameter, α does not vary with covariates and is pooled across all observations.

State transitions represent the sample-averages from the full model (see Figure 2). It summarizes the probability of reaching the destination state for each origin state given that a transition occurred. The focal transitions of interest (mapped to the five interaction domains in Table 1) are embedded. Children's disengagement was the most probable outcome after joint engagement, i.e., after mutual eye contact, children were more likely to transition into aversion (Mutual-CT → Avert-C, 0.52 [0.50, 0.54]) than into shared object focus (Mutual-CT → Share-CT, 0.12 [0.11, 0.13]), and the same held after shared attention (Share-CT → Avert-C, 0.59 [0.58, 0.60]) relative to a direct return to eye contact (Share-CT → Mutual-CT, 0.11 [0.10, 0.12]). Following disengagement, children more often returned directly to a shared object focus rather than re-establishing eye contact first (Avert-C → Share-CT, 0.50 [0.48, 0.52]; Avert-C → Mutual-CT, 0.21 [0.20, 0.22]). The two responding transitions differed notably in strength: tester's referring was followed by child following

with probability 0.33 [0.30, 0.36] (Refer-T → Follow-C), yet once a child followed, the sequence most often resolved into shared attention (Follow-C → Share-CT, 0.85 [0.80, 0.89]). On average, coordinated social interaction was rather interspersed with aversion states than maintained or directly restored, with re-engagement proceeding largely through shared object focus rather than renewed eye contact.

Figure 2

Next-state probability transition matrix of atomic interaction states.



Note. Next-state probabilities from an origin state (row) to a target state (column) of the full hierarchical semi-Markov model (n = 25,682 states), summarized from 12,000 posterior draws across 28 children (56 child × timepoint units). For each draw, a transition-intensity matrix was reconstructed for every observed child and timepoint at its covariate values and posterior random effects, and the resulting next-state probabilities were averaged across the sample. Shading encodes the next-state probability. Colored borders display the respective interaction domains. Empty grey cells are transitions that do not occur in the data, comprising self-transitions on the diagonal and off-diagonal state pairs never observed in the sample; <.01 denotes a transition with a non-zero probability below 0.01.

Temporal and contextual dynamics in social interaction domains

Within-session dynamics for the social interaction domains (see Table 1) were modelled by adding four predictors to the focal transitions in the full model (Table 3). Phase divided each session into early (as reference), mid, and late segments. History captured the change in transition intensity when the immediately preceding sojourn was an engagement state (Mutual-CT or Share-CT). Sojourn duration captured how the odds of exiting to a focal rather than a non-focal destination changed as the current state persisted. In addition, task captured how the intensities of each domain differed across the eight ESCS activities (Supplementary Figure S1). Phase, history and task effects are reported as hazard ratios (HR), and the elapsed-duration effect as an odds ratio (OR), as it acts on which destination is favored rather than on how quickly the state is left. An effect is described as credible when its 95%-interval excludes one.

Phase effects were modest and credible in four of the ten domain-by-phase contrasts with downward shifts over the session: the mid phase reduced responding and disengagement intensities (HR 0.91 [0.85, 0.97] and 0.73 [0.64, 0.83]), and the late phase reduced initiating and synchrony intensities (HR 0.77 [0.63, 0.94] and 0.78 [0.68, 0.88]). Concerning history, recent engagement had a stronger and more consistent influence: when the preceding sojourn entailed engagement (Mutual-CT or Share-CT), responding, synchrony, and reengagement intensities all increased (HR 2.01 [1.60, 2.48], 1.61 [1.43, 1.82], and 1.41 [1.25, 1.60]). Longer current sojourns shifted exits toward focal destinations in four domains (reengagement OR 1.59 [1.49, 1.69], synchrony 1.47 [1.38, 1.57], disengagement 1.30 [1.24, 1.36], and initiating 1.26 [1.09, 1.45]), but not responding (OR 0.95 [0.84, 1.07]). Together, within-session position mattered modestly, whereas recent engagement and elapsed duration shaped both the intensity and the direction of subsequent transitions.

Task effects were larger but varied widely across the eight ESCS tasks (Supplementary Figure S1). Engaging tasks (Turn-Taking, Plastic Jar, Object Spectacle) elevated initiating intensities by factors of 1.76 to 2.34, whereas the Social Interaction and Response to Invitation tasks suppressed them (HR 0.33 [0.21, 0.47] and 0.49 [0.35, 0.67]). The Social Interaction task also elevated synchrony (HR 2.25 [1.99, 2.53]), and the Gaze Following task suppressed responding intensities by nearly threefold (HR 0.35 [0.29, 0.41]). Across the 40 task-by-domain contrasts, 30 were credible, indicating that the model captured substantial contextual dependency in interaction dynamics. Importantly, this was accounted for when interpreting person-level effects below.

Table 3

Temporal and contextual effects on focal transition rates by joint-attention domain.

Domain	Phase (mid vs early) HR [95% CrI]	Phase (late vs early) HR [95% CrI]	History HR [95% CrI]	Sojourn duration OR [95% CrI]
Initiating	1.34 [0.99, 1.74]	0.77 [0.63, 0.94]*	1.16 [0.89, 1.48]	1.26 [1.09, 1.45]*
Responding	0.91 [0.85, 0.97]*	1.02 [0.95, 1.09]	2.01 [1.60, 2.48]*	0.95 [0.84, 1.07]
Disengagement	0.73 [0.64, 0.83]*	0.89 [0.62, 1.24]	0.98 [0.89, 1.07]	1.30 [1.24, 1.36]*
Reengagement	0.91 [0.75, 1.08]	0.97 [0.90, 1.04]	1.41 [1.25, 1.60]*	1.59 [1.49, 1.69]*
Synchrony	0.97 [0.90, 1.05]	0.78 [0.68, 0.88]*	1.61 [1.43, 1.82]*	1.47 [1.38, 1.57]*

Note. Effects are reported as posterior mean with 95% equal-tailed credible interval [CrI]. Phase (early, mid, and late session part) and history effects are hazard ratios (HR) on focal transition intensities. Phase contrasts use the early phase as reference. History encodes whether the immediately preceding sojourn was an engaged state (Mutual-CT or Share-CT). The sojourn-duration effect is an odds ratio (OR) on the relative odds of a focal versus non-focal exit destination, per one standard deviation increase in the standardized log of elapsed sojourn time. Asterisks (*) denote 95% credible intervals excluding 1.

Changes of focal state transitions by person-level features

Person-level covariates were estimated for each of the nine focal transitions in the full model: age at enrollment, sex, nonverbal cognitive ability, intervention group (Early Intervention as Usual [EIAU] versus Frankfurt Early Intervention Program for Autism Spectrum Disorder [A-FFIP]), timepoint (baseline [T2] versus follow-up after 9 months [T5]), and their interaction. Hazard ratios are per one standard deviation for the continuous variables. Group and time effects are reported below as contrasts within each arm and at each time point rather as the model includes their interaction. Sixteen of the 54 covariate-by-transition coefficients were credible (Table 4).

Age and sex effects largely affected reengagement and responding. Older children showed stronger transitions from gaze aversion back to joint engagement (Avert-C → Share-CT, HR = 1.20 [1.05, 1.35]; Avert-C → Mutual-CT, HR = 1.47 [1.28, 1.68]). Compared with girls, boys were less likely to respond to tester referential looks with gaze following (Refer-T → Follow-C, HR = 0.57 [0.34, 0.89]) but more likely to transition from gaze aversion to shared attention (Avert-C → Share-CT, HR = 1.34 [1.02, 1.70]).

Intervention-related contrasts were credible across social interaction domains. At baseline, EIAU showed higher intensities than A-FFIP for synchrony (Mutual-CT → Share-CT, HR = 1.62 [1.02, 2.42]) and responding (Refer-T → Follow-C, HR = 2.12 [1.33, 3.20]), and both differences were of similar magnitude at follow-up (HR = 1.50 [0.93, 2.26] and 2.44 [1.50, 3.74]) with no credible change in either. Two transitions increased in both arms: reengagement after aversion (Avert-C → Mutual-CT, HR = 1.40 [1.20, 1.61] and 1.44 [1.23, 1.66]) and the switch from eye contact to shared attention (HR = 1.36 [1.07, 1.70] and 1.26 [1.02, 1.53]). Two interactions were credible: disengagement increased in A-FFIP but not EIAU (HR = 1.49 [1.28, 1.72] versus 1.11 [0.95, 1.31]; interaction 0.75 [0.60, 0.92]), whereas child referring increased in EIAU but not A-FFIP (HR = 2.01 [1.42, 2.78] versus 1.12 [0.78, 1.55]; interaction 1.84 [1.14, 2.83]). The groups therefore shared a common trajectory of improved reengagement and synchrony while differing on transitions out of mutual eye contact.

Nonverbal cognitive ability showed the most coherent covariate pattern. Higher cognitive ability was associated with lower transition intensities from joint engagement into aversion (Mutual-CT → Avert-C, HR = 0.80 [0.66, 0.97]; Share-CT → Avert-C, HR = 0.82 [0.68, 0.98]), stronger reengagement after aversion (Avert-C → Mutual-CT, HR = 1.73 [1.53, 1.94]; Avert-C → Share-CT, HR = 1.20 [1.07, 1.34]), and stronger responding to referential bids (Refer-T → Follow-C, HR = 1.89 [1.45, 2.43]). As each transition intensity reflects both, the direction and the timing of state changes, a hazard ratio on its own cannot indicate which of these components drives the association. The following section therefore decomposes the cognitive ability association into next-state probabilities and state durations, separating the directional from the temporal contribution and translating the estimated intensities into dynamics of social interaction.

Table 4

Reconfiguration of focal state transitions by person-level features as significant covariate effects.

Covariate Class	Variable	Focal transition (domain)	HR	95% CrI
Demographics	Baseline age (per 1 SD)	Avert-C → Share-CT (Reengagement)	1.20	[1.05, 1.35]
	Baseline age (per 1 SD)	Avert-C → Mutual-CT (Reengagement)	1.47	[1.28, 1.68]
	Sex (male vs female)	Refer-T → Follow-C (Responding)	0.57	[0.34, 0.89]
	Sex (male vs female)	Avert-C → Share-CT (Reengagement)	1.34	[1.02, 1.70]
Intervention	EIAU vs A-FFIP at T2	Mutual-CT → Share-CT (Synchrony)	1.62	[1.02, 2.42]
	EIAU vs A-FFIP at T2	Refer-T → Follow-C (Responding)	2.12	[1.33, 3.20]
	T5 vs T2 (in A-FFIP)	Mutual-CT → Share-CT (Synchrony)	1.36	[1.07, 1.70]
	T5 vs T2 (in A-FFIP)	Avert-C → Mutual-CT (Reengagement)	1.40	[1.20, 1.61]
	T5 vs T2 (in A-FFIP)	Mutual-CT → Avert-C (Disengagement)	1.49	[1.28, 1.72]
	Group × Time	Mutual-CT → Avert-C (Disengagement)	0.75	[0.60, 0.92]
	Group × Time	Mutual-CT → Refer-C (Initiating)	1.84	[1.14, 2.83]
Cognitive Ability	Nonverbal Cog. Ability (per 1 SD)	Mutual-CT → Avert-C (Disengagement)	0.80	[0.66, 0.97]
	Nonverbal Cog. Ability (per 1 SD)	Share-CT → Avert-C (Disengagement)	0.82	[0.68, 0.98]
	Nonverbal Cog. Ability (per 1 SD)	Avert-C → Share-CT (Reengagement)	1.20	[1.07, 1.34]
	Nonverbal Cog. Ability (per 1 SD)	Avert-C → Mutual-CT (Reengagement)	1.73	[1.53, 1.94]
	Nonverbal Cog. Ability (per 1 SD)	Refer-T → Follow-C (Responding)	1.89	[1.45, 2.43]

Note. Effects with 95% equal-tailed credible intervals (CrI) excluding 1 are shown (16 of 54, full set in Supplementary Table S1). Effects are hazard ratios (HR) on focal transition intensities: posterior means with 95% CrI. Continuous covariates (baseline age, nonverbal cognitive ability) are scaled per one standard deviation; categorical covariates are indicators (sex: male = 1; intervention: EIAU = 1; time point: T5 = 1). As the model includes a group × time-point interaction, the intervention and time-point main effects are interpreted within their reference category (A-FFIP or T2). HR = hazard ratio; CrI = credible interval. A-FFIP = Frankfurt Early Intervention Program for Autism Spectrum Disorder. EIAU = Early Intervention as Usual.

Reconfiguration of the social interaction network - exemplified by cognitive ability

To quantify how cognitive ability reshapes social interaction, we evaluated the full model at low and high values on a sample-average basis (see Methods). To assess robustness, contrasts were computed at four in-sample widths (interquartile range, ±1 SD, 10^{th}-90^{th} percentile, and observed minimum–maximum). The effects were consistent in direction and credibility across these widths, differing only in magnitude. In-text values are reported at ±1 SD (nonverbal cognitive ability: 40.5 versus 79.9), and Figure 3 depicts the observed minimum–maximum contrast (32 versus 108). Full results for all transitions and states across all contrasts are provided in Supplementary Tables S2-S3.

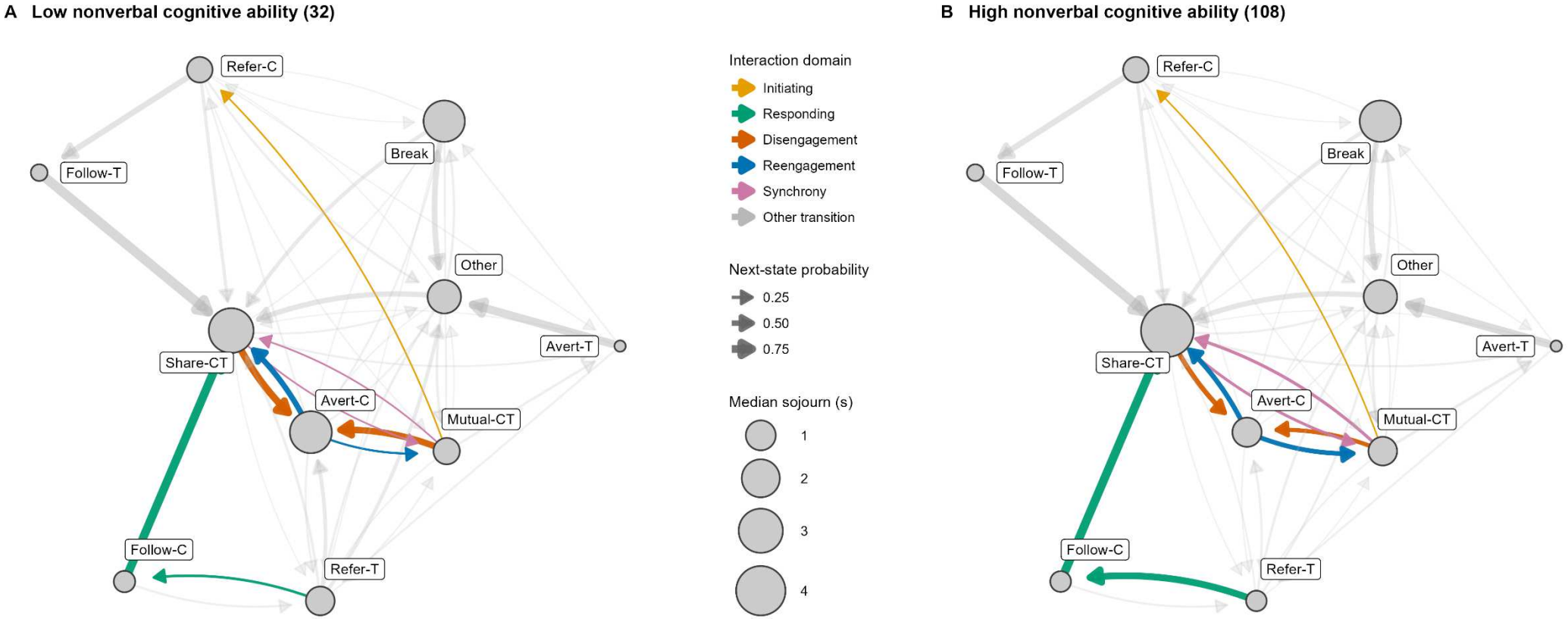
Higher cognitive ability was credibly associated with six of the nine (see Table 1) focal next-state probabilities at ±1 SD, indicating that cognitive ability was related to the majority of the modelled interaction pathways (Supplementary Table S2). The dominant pattern was a shift away from disengagement and toward sustained and re-established joint engagement. The probability of moving from joint engagement into aversion was lower at higher cognitive ability for both disengagement transitions (Mutual-CT → Avert-C, .58 to .46, and Share-CT → Avert-C, .65 to .53), indicating that higher-ability children were less likely to disengage from social interaction. When disengagement occurred, re-engagement was more likely (Avert-C → Mutual-CT, .14 to .28). In addition, the probability of responding to tester referential bids more than doubled (Refer-T → Follow-C, .21 to .45), and both synchrony transitions became more likely (Mutual-CT → Share-CT, .08 to .16, and Share-CT → Mutual-CT, .08 to .14), reflecting more fluid alternation between the two engagement states.

Cognitive ability was also associated with shorter durations of specific states. Two median sojourns were credibly shorter at higher scores, gaze aversion (Avert-C, 2.40 s to 1.48 s) and tester's referring (Refer-T, 0.83 s to 0.51 s). Combined with the probability shifts, this indicates that higher-ability children were less likely to disengage and recovered from aversion more quickly, while reacting faster and more successfully to the tester's referential bids.

These transition- and duration-level associations corresponded to a reorganization of the network. Incoming next-state probability was higher at greater cognitive ability for mutual eye contact (Mutual-CT in-strength, +0.18 [+0.13, +0.24]) and child following (Follow-C, +0.24 [+0.14, +0.32]), and lower for gaze aversion (Avert-C, −0.33 [−0.50, −0.15]). Higher cognitive ability reconfigures the social interaction network towards centralizing mutual gaze and child following, while marginalizing disengagement as gaze aversion.

Figure 3

Model-inferred gaze-transition networks at low and high nonverbal cognitive ability.



Note. Networks show the inferred interaction-state dynamics with cognitive ability set to the lowest (panel A) and highest (panel B) observed values (32 versus 108), with all other covariates and subject random effects held at their sample values (see Methods). Nodes are the ten interaction states, sized by model-implied median sojourn duration. Directed edges are next-state probabilities, with edge-width encoding probability (edges below 0.02 omitted). Edge color encode the interaction domains. Node positions are fixed across both panels, set by a force-directed layout (ggraph), and node size and edge width are on a common scale, so that differences between A and B directly reflect the estimates. *n* = 28 children (56 sessions); estimates summarize 12,000 posterior draws.

Model evaluation of dynamic social interaction

Predictive accuracy increased at each layer of the model. Adding the temporal and contextual predictors improved the expected log predictive density by 745.2 (SE = 43.1) over a baseline retaining only the sojourn distributions and random effects, and adding the person-level covariates by a further 130.3. Domain-specific Bayesian R^2 was highest for initiating (0.47) and responding (0.42) and lowest for reengagement (0.15). Full model comparison, convergence diagnostics, out-of-sample predictive accuracy, and posterior predictive checks are provided in Supplementary Table S4 and Supplementary Results 1.

Between-child variation was substantial and differed markedly across domains. The posterior standard deviation of the subject-by-domain random intercepts, on the log-intensity scale, was 1.06 [0.72, 1.53] for initiating, 0.59 [0.42, 0.81] for synchrony, 0.53 [0.33, 0.81] for responding, 0.43 [0.32, 0.59] for disengagement, and 0.25 [0.18, 0.35] for reengagement, corresponding to between 2.9-fold and 1.3-fold differences in baseline transition intensity between children. Accounting for this variation was crucial for valid inference. Omitting the random effects reduced predictive accuracy substantially and more than doubled the number of covariate effects judged credible, from 16 to 30, which indicates inflated inferences when between-child heterogeneity is ignored. The cost was greatest for initiating, the most heterogeneous domain, whose R^2 fell from 0.47 to 0.13. Thus, random effects captured genuine between-child structure rather than noise.

Discussion

We developed a Bayesian hierarchical semi-Markov framework that translates videotaped social encounters to a network of atomic interaction states and estimates how person-level features reconfigure that network. Applied to 25,682 interaction states annotated from 28 autistic preschoolers assessed at two timepoints, the model characterizes three aspects of social interaction concurrently: the average architecture of JA, the temporal and contextual dependencies that govern its unfolding, and the interindividual differences that reshape it. Joint engagement was punctuated by gaze aversion rather than sustained, the process carried memory of both the preceding state and the time already spent in the current state, and nonverbal cognitive ability produced the most coherent reconfiguration among the person-level features examined.

The average network characterizes the punctuated JA. After mutual eye contact, children moved into gaze aversion more than four times as often as into shared attention on an object, and this asymmetry was even more pronounced following shared attention. Aversion was therefore the ordinary destination of joint engagement rather than a marker of its failure. What distinguished re-engagement was its route. On leaving aversion, children returned to a shared object focus more than twice as often as to eye contact, which indicates that the object compared to the partner's face is more central to joint engagement. This accords with evidence from typical development that one-year-olds

and their parents coordinate attention largely through objects and manual action rather than through gaze following or the partner's face attention (C. Yu & Smith, 2013). Our sample differs in age and is neurodiverse, but an object-mediated route might be a normative feature of JA development, which is poorly captured by face-centered JA paradigms. A complementary asymmetry appeared in responding. Only about a third of the tester's referential bids were followed by the child, yet once following had begun it largely resolved into shared attention. Therefore, responding failed on initiation rather than completion, a distinction that aggregate response rates cannot detect.

We provided novel evidence on how these pathways unfold dynamically in time. Nine of the ten states departed from the constant-hazard exponential form that continuous-time Markov models impose. Seven of them, including the four most dominant states, became more persistent with duration. In contrast, mutual eye contact behaved in the opposite way, its persistence decreased the longer it lasted. Together with its brief median duration of 0.60s, this marks mutual gaze as a short and self-limiting hub through which interaction passes rather than a state in which it settles. Moreover, recent engagement raised the intensity of responding, synchrony and reengagement transitions, so that the process retained information about the preceding state and not solely about the current one. Elapsed duration in turn shifted the destination, with longer sojourns favoring focal exits in four of the five domains. This second dependence relaxes a restriction that multistate models ordinarily entail (Jackson, 2025). Estimating it as a separate coefficient makes it interpretable, and it matters substantively, since how long a child has been looking away is informative about where they will look next.

Context accounted for a further substantial share of the variation. Thirty of the forty task-by-domain contrasts were credible, while the direction of these effects tracked the demands of the activity: turn-taking and object-spectacle tasks raised initiating intensities. Tasks that place no object within the child's reach afford few opportunities to initiate, so this pattern also served as a check that the model captured the elicitation structure of the protocol. Surprisingly, the gaze following task designed to elicit responding suppressed it threefold, which possibly reflects the child’s body turn to follow the tester’s referring to a target placed behind rather than our encoded shift of gaze. Position within the session mattered less, with four of ten phase contrasts credible and all of them modest. Thus, JA scores rather are a statement about the situation in which it was elicited, and because task and phase entered our model as time-varying predictors, the person-level effects reported here are estimated net of that context.

Against this background, nonverbal cognitive ability was credibly associated with six of the nine modelled pathways, shifting the network away from disengagement and towards sustained and re-established engagement, and shortening both the child’s gaze aversion and the tester's referring. Decomposing each intensity into a duration and a next-state probability enables interpretability. Since cognitive ability entered the model only through the nine focal transitions, the shortening of each state can be traced to the respective

transition. Aversion had two such exits (i.e., both re-engagement transitions), so a shorter aversion reflects that children returned to joint engagement sooner. The tester's referring showed a single exit, the child's following response, so its shortening indicates that children answered the bid more quickly. At the level of the whole network, mutual eye contact and child following gained incoming probability while aversion lost it. Higher cognitive ability was thus associated with social interaction that is better maintained and faster re-engaged. This extends previous literature that has related higher cognitive ability to higher aggregate JA scores using clinician-rated and task-based measures (Mundy et al., 1994; Sano et al., 2021; Thurman & Dimachkie Nunnally, 2022). The present results add a within-session dimension by locating the association on specific pathways. Notably, cognitive ability was not credibly related to child-initiated referring, the component most closely aligned with initiating in conventional JA assessment (Billeci et al., 2016).

Other studies have moved beyond static scoring by modelling gaze as movement through a state space, using discrete- and continuous-time Markov chains (Bruschetta et al., 2025; Masedu et al., 2022) alongside evidence that social attention in autism declines over prolonged viewing (Del Bianco et al., 2021; Hedger & Chakrabarti, 2021). The present study extends this work in three respects. First, those paradigms were passive and screen-based, so that traced gazes index where the child looks at rather than what the look accomplishes between two people (Bruschetta et al., 2025). Our state space is defined over the dyad, which makes responding to and initiating of JA observable within a reciprocal exchange (Redcay & Schilbach, 2019). Second, a continuous-time Markov chain treats sojourns as memoryless, so that neither elapsed time nor prior states influence what follows, whereas our results indicate that both do. Third, that work contrasted diagnostic groups, whereas the present design characterizes heterogeneity within autism as dynamic network of social interaction. The difficulty in maintaining and restoring engagement that previously separated autistic from non-autistic individuals appears here as heterogeneity among autistic children that covaries with cognitive ability.

Conventional scoring makes such characterizations difficult to observe. The Early Social Communication Scales, from which our recordings derive, quantify JA as the frequency of initiating bids and the proportion of successful responses to cues (Mundy et al., 2003). Latency does not enter this scoring, so a child who follows a bid immediately and a child who follows after a delay receive equal credit, whereas disengagement, which occupies a large part of the estimated network, is not represented at all. Scoring is further anchored to the episodes designed to elicit each construct, whereas the present model uses the whole session including spontaneous behavior while adjusting for the activity in progress. The current framework extends established scales by retaining their elicitation structure while recovering temporal information their summary scores discard.

The intervention comparison illustrates both the sensitivity and the limits of this approach. The investigated subsamples of the two arms were found to differ before

intervention. Thus, intervention effects as group-by-time interaction remain difficult to interpret, but have been investigated for predefined outcomes in a clinical study (Freitag et al., 2025). Nonetheless, the observed pattern indicates specific pathways on which an adequately powered confirmatory test could concentrate.

Between-child variation was itself informative. On the log-intensity scale, the spread of subject-level intercepts corresponded to an approximately threefold range for initiating and to a 1.3-fold range for reengagement. Initiating was thus simultaneously the most heterogeneous component and the one least accounted for by the covariates. It suggests that drivers of a child's initiating JA is largely not captured by age, sex, cognitive ability or intervention arm. Accounting for this heterogeneity was necessary for valid inference, as a non-hierarchical analysis would have reassigned between-child variance to systematic covariate effects and overstated the evidence for them.

These conclusions are limited. The analysis includes 28 children assessed twice, a sample constructed to prioritize dense within-child sequences over interindividual variance. This design supports the temporal structure the model estimates but constrains generalization (e.g., sex effects rest on seven girls and should be treated as preliminary). Furthermore, the interaction was represented as a single stream of mutually exclusive states anchored on the child as individual of primary interest. Thus, moments in which the partners attend to different targets are not represented. Finally, model-implied median sojourns were over-predicted by potentially 10-30 percent, as a single Weibull shape per state must accommodate sojourns generated under different exit rates. The resulting marginal distribution is a mixture of Weibull distributions sharing one shape and is more right-skewed than individual distributions, which overestimates the median. This bias is multiplicative and shared across covariate scenarios, so the within-state ratios reported above remain unaffected.

Future work might address these limitations by estimating the Weibull shape per transition rather than per state (Dagne & Snyder, 2009). In addition, representing each state's duration as a mixture of gamma kernels (Wu et al., 2024; Wu & Sarkar, 2024) could remove the median bias. Annotating each partner separately and constructing the dyadic state from their combination would permit a coupled two-stream model that estimates autonomous dynamics of each partner alongside their coupling, thus resolving the child-anchored perspective. JA also unfolds across gesture, vocalization and affect, which are present in the videos and could be further explored as additional streams of the network. Replication in an independent sample, and with non-autistic comparison groups, would establish which features of this architecture are general or specific to autism.

A final constraint is scalability, since manual annotation at this density is costly. Current computer-vision work on children's gaze concentrates on estimating where a person is looking (Ryan et al., 2025; Tafasca et al., 2023, 2024), whereas the quantities modelled here require sematic knowledge. Progress towards automated segmentation of interaction states with visual language models might remove this bottleneck. Vice versa,

manually annotated corpora of the present kind supply the semantically labelled sequences such systems require for training and validation. The network-level quantities provide a benchmark against which automated pipelines could be evaluated.

Taken together, these results characterize JA in videotaped behavioral observation as a dyadic network within a temporal and contextual variable space. JA is punctuated rather than continuous, it carries memory of both the preceding state and elapsed time, it depends strongly on the activity in which it is embedded, and it varies substantially between children. Higher cognitive ability was associated with a shift of this network towards maintenance and repair, expressed as less disengagement, faster recovery, and quicker and more probable responding. The presented framework enhances computational modelling of social interaction by enabling the investigation of specific pathways being reconfigured by a person-level feature, a developmental process, or an intervention. It thereby offers an ecologically grounded account of how JA is built, broken and restored in real-world social encounters.

Methods

Procedure

Data were drawn from the A-FFIP randomized controlled trial, a multi-center, observer-blind, parallel-group trial comparing an autism-specific early intervention (Kitzerow et al., 2020) against early intervention as usual in toddlers and preschoolers with autism spectrum disorder (ASD; Freitag et al., 2025). Eligible children were aged 24-66 months at baseline and met DSM-5 criteria for ASD, confirmed by the Autism Diagnostic Interview-Revised (Bölte et al., 2006; Kim et al., 2013) and the Autism Diagnostic Observation Schedule-2 (Lord, 2015). Key exclusion criteria were a nonverbal developmental quotient at or below 30, a nonverbal mental age at or below 12 months, and uncorrected sensory, neurological, or severe psychosocial conditions interfering with intervention (full criteria in Freitag et al., 2025). Children were randomly allocated 1:1 to a manualized, low-intensity naturalistic developmental behavioral program delivered at approximately two hours per week (A-FFIP) or to early intervention as usual (EIAU), comprising the standard community-based early support available in the participating regions. The trial was approved by the Ethics Committee of the Medical Faculty of Goethe University Frankfurt (No. 10/18) and pre-registered with the German Clinical Trials Register (DRKS00016330). Parents or legal guardians gave written informed consent, including for video recording and further data processing.

The present study is a secondary analysis combining two data streams from this trial. Participant-level covariates (age, sex, cognitive ability, and intervention arm) were taken directly from the trial's enrolment and assessment records, whereas the social interaction data were derived from video recordings of the Early Social Communication Scales (ESCS; Mundy et al., 2003), a semi-structured joint-attention assessment

administered as part of the trial's standard assessment battery at two timepoints approximately nine months apart, at baseline (T2) and follow-up (T5). Rather than applying the conventional ESCS scoring, we used these recordings as a standardized elicitation context and applied a fine-grained scheme to derive atomic interaction states.

Sample

To maximize standardization of the recording setup, the present analysis was restricted to children enrolled at the Frankfurt site, where ESCS sessions were recorded under uniform conditions. From this pool, a sub-sample was selected for fine-grained manual annotation on the basis of three considerations: the availability of usable ESCS recordings at both T2 and T5 (sessions with a missing timepoint, non-standard frame rate, or a deviating camera angle were not eligible), coverage of a broad range of nonverbal cognitive ability, and balance between the A-FFIP and EIAU arms.

The resulting sample comprised 28 children (14 A-FFIP, 14 EIAU), each assessed at two timepoints, yielding 56 ESCS sessions. Annotated sessions lasted 27.1 minutes on average (SD = 4.6), resulting in a total of 25.3 hours of coded interaction. The two arms were closely matched on baseline characteristics as they did not statistically differ on sex, age, and nonverbal cognitive ability (Table 5). Across the sample, scores on nonverbal cognitive ability also matched the parent trial (Freitag et al., 2025).

Table 5

Sample Descriptives by Intervention Group.

Characteristic	A-FFIP (n = 14)	EIAU (n = 14)	Overall (N = 28)	Test
Sex, n male (%)	11 (78.6)	10 (71.4)	21 (75.0)	Fisher's exact $p = 1.000$
Baseline age in months, M (SD)	47.2 (11.8)	44.6 (9.0)	45.9 (10.4)	$t(24.4) = 0.67$, $p = .511$
range	33–64	26–60	26–64	-
Cognitive ability, M (SD)	59.0 (20.9)	61.4 (19.1)	60.2 (19.7)	$t(25.8) = -0.32$, $p = .748$
range	32–97	38–108	32–108	-

Note. Baseline characteristics of the 28 children included in the gaze-annotation sub-sample. Cognitive ability is the baseline nonverbal developmental quotient (cognitive age / chronological age × 100), harmonized across the Bayley-III and WPPSI-III as in the parent trial. Group comparisons are Welch two-sample t-tests for continuous variables and Fisher's exact test for sex, as the smallest expected cell count was 3.5. M, mean; SD, standard deviation; A-FFIP, Frankfurt Early Intervention Program; EIAU, early intervention as usual.

Measures

Demographics. Child age at baseline (in months), sex (male/female), and treatment-group assignment were obtained from the trial's enrollment records.

Cognitive ability. Nonverbal cognitive ability at baseline was assessed by trained, blinded testers using one of two age-appropriate instruments: the Bayley Scales of Infant and Toddler Development-Third Edition (Bayley-III; Reuner & Rosenkranz, 2014) for children with a developmental age at or below 42 months, and the Wechsler Preschool and Primary Scale of Intelligence–Third Edition (WPPSI-III; Petermann et al., 2014) for children above this developmental age. No single German-normed instrument spans the full eligible developmental range, which motivated this two-instrument approach (Kitzerow et al., 2020). For the Baley-III, nonverbal ability was expressed as a ratio developmental quotient (nonverbal cognitive age divided by chronological age, × 100). For the WPPSI-III, the nonverbal index was used. These were combined into a single harmonized nonverbal developmental quotient, following the convention of the clinical trial (Freitag et al., 2025), and entered the model as the cognitive-ability covariate.

Early Social Communication Scales. The ESCS (Mundy et al., 2003) is a semi-structured observational protocol that elicits prelinguistic joint-attention and social-interaction behaviors through a standardized sequence of child-tester episodes in early development. The tester sits opposite the child and presents a fixed series of objects (e.g., wind-up toys, a ball, a picture book, a hat) just out of the child's reach, together with visual targets placed around the room (e.g., posters), to evoke initiating and responding behaviors. A detailed description of all tasks conducted is provided in Supplementary Table S6. Rather than applying the conventional ESCS event scoring, we used the recorded sessions as a standardized elicitation context for video annotation. Sessions were recorded under uniform conditions at the Frankfurt site, with the examination room, play materials, and video equipment held constant across both timepoints, and served as the input to the coding scheme. Anonymized example frames of the setup are shown in Supplementary Figure S2.

Coding scheme and video annotation

Each ESCS administration was recorded from two camera angles with a Canon Legria HF G40 camcorder at 25 frames per second. The two streams were synchronized and played back in parallel during annotation, so that the gaze of both the child and the interaction partner could be resolved even when one view was momentarily occluded. Annotation was carried out in Mangold INTERACT (Mangold, 2018), which exported each coded segment with onset and offset timestamps.

The coding scheme was developed for the dyadic ESCS interaction and extends the atomic state taxonomy of Fan and colleagues (2019), who decompose JA behavior into a set of elementary states: single, mutual, avert, refer, follow, and share. For each state, onset and offsets were defined. Each state was annotated also encoding together the

interactant it involved, distinguishing the child (C), the tester (T), and, where present, an accompanying caregiver (P), so that the same elementary act was coded as a distinct state depending on who performed it. This cross-classification captured behaviors such as mutual eye contact, shared attention to a common stimulus, child and tester referring, child and tester gaze-following, and gaze aversion by either partner. Referring is the one state not defined by gaze alone. In the ESCS it may be realized through pointing, showing and gazing, and it was coded whenever one partner directed the other's attention to a target by any of these means. Supporting gestures such as pointing or vocalizing were annotated alongside the state they accompanied. The complete annotation label set and definitions are given in Supplementary Table S7.

Annotation was anchored on the child. The child's focus of attention defined the state, and the partners' gaze was coded in relation to it, so that partner movements which did not alter the child's focus did not break an ongoing state. A sustained period of shared attention, for example, was treated as a single continuous episode even when the tester briefly glanced back to the child's face to monitor attention, as long as the child remained oriented to the shared target. Within this child-centered frame, states were assigned based on the social meaning of the gaze behavior in context rather than the gaze direction of any single person. Shared attention, for instance, required that the child was attending to a stimulus that the partner had established or was sharing as a common focus, rather than the child simply looking at an object alone, and child gaze-following was coded only when the child reoriented in response to a preceding partner cue. In this way the sequential embedding of each behavior, and not momentary gaze direction alone, determined the assigned state.

Annotation was performed by four trained raters (one biology student, three psychology students), blind to group assignment, participants' cognitive ability, and other clinical scores. Each completed a standardized protocol covering the coding manual (Table S7), the Fan et al. (Fan et al., 2019) framework, and the operation of Mangold INTERACT. Training proceeded in stages: observation, joint coding with an experienced rater, and independently coded practice videos with feedback, with regular consensus meetings to resolve ambiguous sequences. Inter-rater reliability was assessed after training on two independently double-coded sessions using global rater agreement, the proportion of matching codes across all compared annotations, which is robust to the skewed category frequencies typical of these data (McHugh, 2012). Agreement was computed for all six rater pairs and ranged from 0.76 to 0.87, above the 0.70 threshold adopted as acceptable.

Pre-processing of annotated atomic interaction states

For each session, the annotation produced a chronologically ordered set of atomic interaction states with onset and offset times. The raw labelling data was not yet a mutually exclusive, continuous-time state sequence the model requires. Because the model treats the time between successive onsets as the duration spent in a state and estimates transition rates from those durations and the transitions between them, two

distortions were avoided. Deleting any behavior not of interest to the analysis would make the preceding state appear to persist across the deleted interval, artificially lengthening its sojourn and biasing its exit rate downward. Compressing the timeline to close such gaps would instead splice together two states that were never adjacent in the video, creating a transition that did not occur. Thus, a pre-processing pipeline was designed to preserve the true temporal structure of each session, so that both the durations spent in states and the transitions between them reflect what was observed rather than artefacts of data handling. The full pipeline is specified step by step in Supplementary Methods 1 while key decisions are summarized here.

Contiguous annotations of the same behavior were first merged into single intervals, and behaviors occurring at the same recorded onset were ordered by their interactional precedence and separated by a single frame so that the sequence was strictly ordered in time. Spans in which no interaction state could be coded, for example when a relevant person left the frame or gaze could not be determined, were not deleted but retained as an explicit "Break" state spanning the elapsed interval, which preserves real time without asserting that the surrounding states continued. Child off-target fixation and the comparatively rare caregiver-directed behaviors, neither of which forms part of the focal joint attention transitions in this study, were likewise retained rather than removed, combined into a single "Other" state so that they too held their place on the timeline and did not generate spurious direct transitions between focal states. After mapping to the ten-state ontology (see Table 2) and merging consecutive states reintroduced by the pipeline again, each session was a gap-free sequence of distinct states with their durations.

A further decision concerned brief interruptions of an otherwise direct focal transition. The two child-directed focal sequences that index the onset of social engagement (Mutual-CT → Refer-C, Refer-T → Follow-C) were frequently interrupted by a short intervening state (e.g., Avert-C) before the focal destination is reached. Treating such an interruption as a genuine excursion would fragment a single coordinated act into several states and obscure the transition of interest, whereas ignoring it would discard temporal information about how the act unfolded. These brief intervening spans, termed bridges, were therefore absorbed into the focal transition under explicit limits on their duration and number (Supplementary Table S9), so that longer or more complex excursions were left intact as genuine interruptions.

The direction in which a bridge's time was credited was made content-aware rather than uniform. For an initiating sequence, any intervening behavior is part of the child's own build-up to the initiation, and its time was credited forward to the initiating state as latency to initiation. For a responding sequence the credit depended on content. When the child had briefly looked away between the tester's referring and arriving at the target itself, the time was credited forward to the response. When the child was instead looking at the referent alone, the tester's bid was still operative and the child was still processing

it, so the time was credited back to the tester's bid as latency to respond. Mixed interruptions were divided between the two states in proportion to their content. In every case the total duration of the session was preserved exactly, with no gap or overlap introduced between the states on either side.

The pipeline was applied under a single primary specification for all analyses reported here. As the duration and count limits placed on bridge absorption are necessarily judgement calls, the same pipeline was re-run across a ladder of eight specifications ranging from no bridging at all to progressively more permissive limits, and the robustness of the substantive findings across these specifications is reported in Supplementary Methods 2 and the accompanying sensitivity tables.

Participant covariates were then added to each session and the two continuous covariates, baseline age and nonverbal cognitive ability, were standardized to the analytic sample's own mean and standard deviation. The resulting sequences were then converted into the sojourn table the model is fitted to, in which each state's duration is calculated from its onset to the onset of the next state. Occupancies crossing a session-phase boundary were split, and segments with no observed destination (i.e., session end, phase-splits) were right-censored.

Model building and estimation

We modelled each session's sequence of atomic interaction states as a continuous-time semi-Markov process over the ten states, in which the time spent in a state and the choice of the next state are governed by separate but linked components. A semi-Markov process admits two parameterizations (Asanjarani et al., 2022). One is built from transition intensity functions, with a separate hazard for each origin-to-destination pair, and is the form adopted in the Bayesian hierarchical duration modelling of behavioral observation on which our work builds (Dagne & Snyder, 2009). The other combines a sojourn-time distribution for each transition with an embedded Markov chain over destinations. The two are formally interchangeable but differ in what they make accessible to inference, and the intensity-based form has the practical advantage that its likelihood separates by transition. Under either, the probability of each destination shifts with elapsed time whenever the hazards, or equivalently the sojourn distributions, differ across destinations, as Asanjarani and colleagues show analytically and illustrate on observed data. That shift, however, is a consequence of the competing transition-specific components rather than a parameter of the model, so it carries no estimate and no interval of its own.

We therefore adopt a restricted form of the sojourn-based parameterization. Each state has one sojourn distribution, whose Weibull shape is estimated at the level of the origin state rather than the individual transition and whose scale is set by the exit rate, and a separate categorical model determines which state follows. Holding the sojourn independent of the destination holds the next-state probabilities constant with respect to elapsed duration. We take that as the default and reintroduce the dependence through a

single term, described below, so that it enters as an estimated coefficient. Inverting the parameterization forgoes the likelihood separation, but it disentangles two quantities that a per-transition hazard combines, and it allows the five interaction domains to be imposed directly on the embedded-chain rates.

States, transitions, and domains. The process transitions among ten states. Following the convention of the multistate literature, r denotes the state currently occupied and s the state entered next, and $\mathcal{S}(r)$ denotes the set of states that can be entered directly from r (Jackson, 2011). Six of these states participate in the nine focal transitions of interest. Tester gaze following and tester aversion enter no focal transition but remain available destinations, and are retained on that basis, since exit rates sum over every transition out of a state and next-state probabilities are normalized against the same set. Omitting genuine alternatives would inflate the focal next-state probabilities. Other and Break are retained for the same reason and additionally preserve the temporal structure of the session. Each focal transition $r \to s$ is assigned to one of five interaction domains, indexed d, which are initiating, responding, disengagement, reengagement, and synchrony. The domain is the level at which within-session dynamics, individual variation, and several modelling terms are expressed. Non-focal transitions carry a baseline rate alone, and none of the structured effects below are estimated for them. We write n for a single sojourn, that is, one uninterrupted occupancy of a state. Each sojourn therefore has exactly one origin state, so that $r = r(n)$, while the destination s is realized only when the sojourn ends. A sojourn also identifies the child $i(n)$, the phase $m(n)$ in which it occurs, the ESCS task in progress, and whether the preceding state was one of the two joint-engagement states. Timepoint enters as an element of the covariate vector rather than as a separate index. Because occupancies that cross a phase boundary are split at that boundary (see Estimation), every sojourn falls within a single phase. The index n therefore denotes the occasion rather than the states: the pair $r \to s$ identifies which transition, and hence which parameters, are meant, while n identifies which of that transition's occurrences is being evaluated. Quantities that vary across occurrences carry n in the equations below, so that one transition takes a different value at each occurrence according to the child, phase, task, and history of that occurrence.

Intensity, sojourn, and destination. The transition intensity q_{rs} is the instantaneous rate of moving from r to s (Jackson, 2011), and $q_{rs}(n)$ denotes the value it takes for sojourn n. For each focal transition we modelled it on the log scale (equation 1), combining a transition-specific baseline log-rate μ_{rs}, the effect $\varphi_{d,m(n)}$ of session phase $m(n)$ on domain d, the person-level covariates x_n for the child and timepoint of that sojourn weighted by coefficients β_{rs} estimated separately for each of the nine focal transitions, the task effects τ_d acting through w_n, the deviation-coded indicator vector for the ESCS activity in progress during that sojourn, a history effect γ_d acting through h_n, an indicator equal to one when the immediately preceding state was one of the two joint-engagement states, Mutual-CT or Share-CT, and a subject-by-domain random intercept $\eta_{i(n),d}$ for the child to whom the sojourn belongs. Non-focal transitions retain μ_{rs} alone.

$$\log q_{rs}(n) = \mu_{rs} + \varphi_{d,m(n)} + x_n^\top \beta_{rs} + w_n^\top \tau_d + \gamma_d\, h_n + \eta_{i(n),d} \qquad (1)$$

The exit rate $q_r(n)$ is the total rate of leaving state r regardless of destination, obtained by summing the intensities of the competing transitions out of r for that sojourn (equation 2). It is the negative of the diagonal entry q_{rr} in the standard intensity-matrix convention (Jackson, 2011), and it follows the competing-risks logic in which the total exit hazard is the sum of the cause-specific rates (Putter et al., 2007). The sojourn u_n is Weibull with a state-specific shape α_r and a scale determined entirely by the exit rate rather than estimated separately (equation 3). The exit rate therefore fixes the overall level of the hazard of leaving the state, while the shape governs how that hazard changes as the sojourn continues. A shape below one denotes a hazard that declines with elapsed time, describing a state that becomes more persistent the longer it lasts. A shape of one recovers the constant-hazard exponential case that a continuous-time Markov model imposes on every state, in which the hazard equals the exit rate at every elapsed duration (Aalen et al., 2008).

$$q_r(n) = \sum_{s \in \mathcal{S}(r)} q_{rs}(n) \qquad (2)$$

$$u_n \sim \text{Weibull}\big(\alpha_r,\, q_r(n)^{-1/\alpha_r}\big) \qquad (3)$$

Given that the state is left, the destination is drawn from a categorical distribution over the competing transitions. The rate entering that distribution carries one further term, the effect δ_d of $\tilde{u}$, the standardized log of the duration the current sojourn has already reached, and is therefore written as a function of that duration (equation 4). Standardization is applied within origin state, so that $\tilde{u}$ expresses elapsed duration relative to the mean and standard deviation of log sojourn time in the state being left, and δ_d is accordingly an effect per standard deviation of within-state log duration. A value of zero therefore marks the state's mean log duration. Under this scaling, $q_{rs}(n)$ is the intensity at that mean, and $q_{rs}(n, \tilde{u})$ the intensity when the elapsed duration lies $\tilde{u}$ standard deviations from it. In the likelihood, $\tilde{u}$ takes the value $\tilde{u}_n$ that sojourn n reached. Normalizing these rates within the origin state gives the next-state probabilities $p_{rs}(n, \tilde{u})$ (equation 5).

$$\log q_{rs}(n, \tilde{u}) = \log q_{rs}(n) + \delta_d\, \tilde{u} \qquad (4)$$

$$p_{rs}(n, \tilde{u}) = \frac{q_{rs}(n, \tilde{u})}{\sum_{s' \in \mathcal{S}(r)} q_{rs'}(n, \tilde{u})} \qquad (5)$$

Without δ_d these would be constants, and the argument $\tilde{u}$ marks the single dependence we reintroduce. It enters the destination distribution only and not the exit rate of equation 2. Separating the sojourn distribution from the next-state probabilities is established practice, with the transition intensity written as the product of an exit rate and a next-state probability (Jackson, 2025; Titman & Sharples, 2010). In that literature the destination is reached with a probability explicitly assumed independent of the time

already spent in the state, so the dependence is neither estimated nor tested, which is defensible as their data are intermittently observed and the sojourn is interval-censored rather than known. In contrast, our annotation records exact onsets and offsets, so the sojourn is observed and the dependence can be estimated. It has been estimated before in discrete time within the hidden semi-Markov family, where a sojourn distribution attached to each state is combined with next-state probabilities indexed by the number of time steps already spent in the current one (Marhasev et al., 2009; S.-Z. Yu, 2010). Those models infer a latent state sequence from emissions at discrete time steps, whereas the interaction states modelled here are annotated directly and evolve in continuous time. Accordingly, our dependence is represented by a single coefficient on standardized log-elapsed duration rather than by a duration-indexed transition matrix, which keeps it interpretable on the odds scale and comparable across the five domains.

Covariates and hierarchy. The person-level covariates are the only terms in the model estimated at the level of the individual transition. Each of the nine focal transitions carries its own coefficient vector β_{rs}, giving 54 covariate-by-transition coefficients, whereas the phase, task, history, and elapsed-duration effects and the random intercepts are all estimated across the five interaction domains. The asymmetry is deliberate. Within-session dynamics are defined theoretically at the domain level, whereas a person-level feature such as cognitive ability may plausibly act on one transition within a domain and not on its partner, so the covariates are given the finer resolution. The vector $\mathbf{x}$ holds an indicator for follow-up versus baseline, an indicator for intervention group, their interaction, nonverbal cognitive ability, age at enrollment, and sex. The two continuous covariates were standardized to the analytic sample, so their coefficients in β_{rs} are expressed per standard deviation. The phase effect $\varphi_{d,m}$ divides each session into three equal intervals of elapsed time and allows the focal intensities to drift across the early, middle, and late thirds of an administration. The subject-by-domain random intercepts were given a non-centered parameterization (equation 6), in which $z_{i,d}$ is a standard normal deviate and σ_d is the between-child standard deviation for domain d, so that the degree of individual heterogeneity is estimated separately for each kind of social dynamic. This places the model within the transition-specific random-effects lineage of hierarchical multistate models (Dagne & Snyder, 2009; Hougaard, 2000), with the hierarchy organized by interaction domain rather than by individual transition. The intercepts are estimated independently across domains instead of a full covariance structure, which is a simplification relative to models that estimate correlations among competing risks from the same origin state (Dagne & Snyder, 2009).

$$\eta_{i,d} = z_{i,d}\,\sigma_d, \qquad z_{i,d} \sim \mathcal{N}(0,1) \qquad (6)$$

Prior choices. Priors were weakly informative and set on the scale of the linear predictor. The baseline log-rates μ_{rs} were given a normal prior with mean −1 and standard deviation 1.5, which places the central 95% of prior mass for a single transition intensity between 0.02 and 7 per second. The transition-specific covariate coefficients β_{rs} were given a

normal prior with standard deviation 0.5. On the hazard-ratio scale this permits a one standard deviation change in a continuous covariate, or the presence of a binary one, to multiply a focal transition intensity by between 0.38 and 2.67 with 95% prior probability, and places 17% of prior mass on effects exceeding twofold in either direction. This prior bears most directly on the reported person-level effects, because those are informed by contrasts between 28 children rather than by the full set of sojourns, so it constrains them more than it constrains the baseline rates. The domain-level task effects $\boldsymbol{\tau}_d$ were given a tighter normal prior with standard deviation 0.35, reflecting the smaller number of tasks and the sparsity of some task-by-transition combinations, and the phase effects $\varphi_{d,m}$ a standard deviation of 0.5. The history and elapsed-duration effects γ_d and δ_d were given standard deviations of 0.3, corresponding to 95% prior intervals from 0.56 to 1.80 on the ratio scale. The between-child standard deviations σ_d were given an exponential prior with rate 2, whose median implies about a twofold range in baseline intensity across children within a domain while permitting considerably more. The Weibull shapes α_r were centred on an empirical per-state value $\hat{\alpha}_r$ estimated from the observed sojourns and allowed to vary around it on the log scale through a standard normal deviate ε_r scaled by 0.3 (equation 7), which regularizes the shapes towards data-driven values without fixing them.

$$\alpha_r = \exp(\log\hat{\alpha}_r + 0.3\,\varepsilon_r), \qquad \varepsilon_r \sim \mathcal{N}(0,1) \qquad (7)$$

Estimation. The model was fitted in a fully Bayesian framework using Hamiltonian Monte Carlo in Stan through the cmdstanr interface (cmdstanr 0.9.0, CmdStan 2.38.0; R version 4.4.2). The likelihood was evaluated over all 25,850 sojourn segments, with sojourns that crossed a phase boundary or reached the end of a session treated as right-censored at the boundary so that no transition was attributed where none was observed. Of these, 112 were censored at a phase boundary (two per session) and 56 at the end of a session (one per session), leaving the 25,682 completed sojourns reported in Table 2. We ran four chains with 1,500 warm-up and 3,000 sampling iterations each, retaining 12,000 posterior draws, with the target acceptance probability raised to 0.92 and the maximum tree depth set to 12. Convergence was assessed by the potential scale-reduction factor, effective sample sizes, divergent transitions, and the estimated Bayesian fraction of missing information. All parameters met standard convergence criteria, with a maximum potential scale-reduction factor of 1.002, no divergent transitions, and bulk effective sample sizes in the thousands, with full diagnostics in the Supplementary Results 1. Posterior summaries are reported with 95% equal-tailed credible intervals. Model parameters, hazard ratios, and odds ratios derived from them are reported as posterior means. The sample-average next-state probabilities and model-implied sojourn durations are summarized by the posterior median. An effect is described as credible when the interval excludes the value of no effect..

Distinguishing intensities, exit rates, and probabilities. As the model separates the rate of leaving a state, the choice of destination, and the duration of the sojourn, we use

a consistent vocabulary throughout. The destination-specific instantaneous rate q_{rs} is the transition intensity, and covariate, phase, history, and task effects on it are reported as hazard ratios. The aggregate rate of leaving a state, q_r, is the exit rate, and together with the Weibull shape it determines the sojourn distribution. Normalizing the intensities within the origin state gives the next-state probabilities p_{rs}, used for the average structure and the network representations. The effect of elapsed duration on the destination is reported as an odds ratio on a focal relative to a non-focal destination. These are kept distinct because a covariate can change an intensity without changing a probability, when a competing transition from the same state responds more strongly, and because the exit rate and the destination distribution are separately parameterized.

Nested models. To establish what each layer of the model contributes, we fitted a sequence of four nested models to the same data. A baseline model retained the baseline rates, the Weibull shapes, and the subject-by-domain random effects, but no within-session dynamics and no covariates. A dynamics model added the phase, history, sojourn-duration, and task effects. The full model added the participant and time covariates. All results are reported from the full model. A fourth, supplementary model removed the subject random effects from the full model, to show that the covariates do not absorb all the individual heterogeneity. The four models were compared by approximate leave-one-out cross-validation using Pareto-smoothed importance sampling (Vehtari et al., 2017) and by a per-domain Bayesian R^2 (Gelman et al., 2019), with posterior predictive checks of state occupancy, sojourn-duration distributions, and transition counts. The nested comparison serves to attribute explanatory contribution across the layers and is reported with the model-evaluation results and in the Supplementary Results 1.

Average network structure and covariate contrasts. To obtain interpretable sample-level quantities from the fitted model, the structured effects were not directly extracted from the estimated rates for two reasons. First, the destination probabilities are competing intensities that have been normalized within the origin state (equation 5), so a covariate that changes one intensity changes every probability in that row. A hazard ratio on an intensity therefore does not map linearly onto a change in any single next-state probability, and a credible effect on an intensity can coincide with little or no change, or even an opposing change, in the corresponding probability when a competing transition from the same state responds more strongly. Effects that are to be interpreted at the level of state changes are therefore reported as next-state probabilities rather than as intensities. Second, the model is hierarchical and these quantities are non-linear functions of the parameters, so that the average quantity over the distribution of subject random effects is not equal to that quantity evaluated at the average random effect. Reading the structure off the baseline rates, or off any single covariate profile with the random effect set to zero, therefore returns the structure of an artificial average individual rather than the sample-average structure, a form of non-collapsibility familiar from non-linear and survival models (Mood, 2010; Pang et al., 2016). We therefore used the

observed-value approach (Hanmer & Kalkan, 2013), which estimates the average effect across individuals rather than the effect for an artificial average individual (i.e., covariate values set to means for continuous variables and proportions for categorical variables).

For the average transition structure, we reconstructed, within each posterior draw, the transition intensity matrix for every observed session at that session's actual covariates and phase and the corresponding child's posterior random effect. The resulting quantities were averaged across sessions within the draw before summarizing across draws with credible intervals. This yields a sample-average structure rather than the structure of a reference session and follows the established practice of reporting model-implied transition probabilities at meaningful covariate configurations (de Wreede et al., 2011). For the cognitive-ability contrasts, we used the same per-draw construction but set the nonverbal cognitive ability estimate to a low and a high value while leaving every other covariate, and each child's random effect, at its observed value, so that the contrast reflects the average effect of cognitive ability across the sample as it is rather than at a single synthetic profile (Hanmer & Kalkan, 2013). From the resulting matrices per draw, we derived the focal next-state probabilities and the model-implied median sojourn durations and summarized the high-versus-low differences and ratios with credible intervals. The contrast width is itself a choice. Thus, we computed it at a gradient of widths, from the interquartile range through one standard deviation to the extremes of the observed range, and report the standard-deviation contrast in the main text with the wider contrasts in Figure 3 and supplementary Tables S2-S3.

Code and data availability. The pre-processing pipeline, the model-ready sojourn tables it produces for all eight bridge specifications, the model-fitting code for the full model, the fitted results object, and the post-fit analysis of the cognitive ability contrasts are shared on Open Science Framework (OSF) The raw video recordings cannot be shared due to participant identifiability.

Acknowledgements

This work was supported by the LOEWE Center DYNAMIC, funded by the LOEWE program of the Hessian Ministry of Higher Education, Research, Science and the Arts (grant number LOEWE1/16/519/03/09.001(0009)/98). The underlying A-FFIP trial was funded by the German Research Foundation (DFG, project number FR 2069/8-1).

References

Aalen, O. O., Borgan, Ø., & Gjessing, H. K. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer. https://doi.org/10.1007/978-0-387-68560-1

Alie, D., Mahoor, M. H., Mattson, W. I., Anderson, D. R., & Messinger, D. S. (2011). *Analysis of eye gaze pattern of infants at risk of autism spectrum disorder using Markov models*. 282–289. https://doi.org/10.1109/WACV.2011.5711515

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, *11*(2), 91–115. https://doi.org/10.1191/0962280202SM276ra

Asanjarani, A., Liquet, B., & Nazarathy, Y. (2022). Estimation of semi-Markov multi-state models: A comparison of the sojourn times and transition intensities approaches. *The International Journal of Biostatistics*, *18*(1), 243–262. https://doi.org/10.1515/ijb-2020-0083

Betancourt, M. (2016). *Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1604.00695

Billeci, L., Narzisi, A., Campatelli, G., Crifaci, G., Calderoni, S., Gagliano, A., Calzone, C., Colombi, C., Pioggia, G., Muratori, F., ALERT group, Raso, R., Ruta, L., Rossi, I., Ballarani, A., Fulceri, F., Darini, A., Maroscia, E., Lattarulo, C., … Comminiello, V. (2016). Disentangling the initiation from the response in joint attention: An eye-tracking study in toddlers with autism spectrum disorders. *Translational Psychiatry*, *6*(5), e808–e808. https://doi.org/10.1038/tp.2016.75

Bölte, S., Rühl, D., Schmötzer, G., & Poustka, F. (2006). *Diagnostisches Interview für Autismus – Revidiert (ADI-R): Deutsche Fassung des Autism Diagnostic Interview–Revised von M. Rutter, A. Le Couteur und C. Lord*. Huber.

Bottema-Beutel, K. (2016). Associations between joint attention and language in autism spectrum disorder and typical development: A systematic review and meta-regression analysis. *Autism Research*, *9*(10), 1021–1035. https://doi.org/10.1002/aur.1624

Bruschetta, R., Famà, F. I., Spadaro, L., Leonardi, E., Carrozza, C., Aiello, S., Campisi, A., Mastrogiuseppe, M., Campisi, S., Chakrabarti, B., Ruta, L., Pioggia, G., De Gaetano, A., Borri, A., & Tartarisco, G. (2025). Modeling gaze behavior with continuous-time Markov chains to investigate social attention dynamics in autism. *Scientific Reports*, *15*(1), 39692. https://doi.org/10.1038/s41598-025-23366-4

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), 1–143. https://doi.org/10.2307/1166214

Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Cox, A., & Drew, A. (2000). Testing joint attention, imitation, and play as infancy precursors to language and theory of

mind. *Cognitive Development*, *15*(4), 481–498. https://doi.org/10.1016/S0885-2014(01)00037-5

Cook, R. J., & Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. Chapman & Hall/CRC. https://doi.org/10.1201/9781315119731

Dagne, G. A., & Snyder, J. (2009). Bayesian hierarchical duration model for repeated events: An application to behavioral observations. *Journal of Applied Statistics*, *36*(11), 1267–1279. https://doi.org/10.1080/02664760802587032

de Wreede, L. C., Fiocco, M., & Putter, H. (2011). mstate: An R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, *38*(7), 1–30. https://doi.org/10.18637/jss.v038.i07

Del Bianco, T., Mason, L., Charman, T., Tillmann, J., Loth, E., Hayward, H., Shic, F., Buitelaar, J., Johnson, M. H., & Jones, E. J. H. (2021). Temporal profiles of social attention are different across development in autistic and neurotypical people. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(8), 813–824. https://doi.org/10.1016/j.bpsc.2020.09.004

Erskine, H. E., Norman, R. E., Ferrari, A. J., Chan, G. C. K., Copeland, W. E., Whiteford, H. A., & Scott, J. G. (2016). Long-term outcomes of attention-deficit/hyperactivity disorder and conduct disorder: A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(10), 841–850. https://doi.org/10.1016/j.jaac.2016.06.016

Fairchild, G., Hawes, D. J., Frick, P. J., Copeland, W. E., Odgers, C. L., Franke, B., Freitag, C. M., & De Brito, S. A. (2019). Conduct disorder. *Nature Reviews Disease Primers*, *5*(1), 43. https://doi.org/10.1038/s41572-019-0095-y

Falck-Ytter, T., Carlström, C., & Johansson, M. (2015). Eye contact modulates cognitive processing differently in children with autism. *Child Development*, *86*(1), 37–47. https://doi.org/10.1111/cdev.12273

Falck-Ytter, T., Kleberg, J. L., Portugal, A. M., & Thorup, E. (2023). Social attention: Developmental foundations and relevance for autism spectrum disorder. *Biological Psychiatry*, *94*(1), 8–17. https://doi.org/10.1016/j.biopsych.2022.09.035

Fan, L., Wang, W., Huang, S., Tang, X., & Zhu, S.-C. (2019). *Understanding human gaze communication by spatio-temporal graph reasoning*. 5724–5733. https://doi.org/10.1109/ICCV.2019.00582

Freitag, C. M., Kitzerow-Cleven, J., Polzer, L., Schwarz, T., Bast, N., Kim, Z., Schnettler, L., Raji, N., Bracht, M., Jensen, K., & Kieser, M. (2025). Complex, low-intensity, individualised naturalistic developmental behavioural intervention in toddlers and pre-schoolers with autism spectrum disorder: The multicentre, observer-blind, parallel-group randomised-

controlled A-FFIP trial. *Journal of Child Psychology and Psychiatry*, *66*(10), 1500–1513. https://doi.org/10.1111/jcpp.14162

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Georgescu, A. L., Koeroglu, S., Hamilton, A. F. D. C., Vogeley, K., Falter-Wagner, C. M., & Tschacher, W. (2020). Reduced nonverbal interpersonal synchrony in autism spectrum disorder independent of partner diagnosis: A motion energy study. *Molecular Autism*, *11*(1), 11. https://doi.org/10.1186/s13229-019-0305-1

Hanmer, M. J., & Kalkan, K. O. (2013). Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models. *American Journal of Political Science*, *57*(1), 263–277. https://doi.org/10.1111/j.1540-5907.2012.00602.x

Hedger, N., & Chakrabarti, B. (2021). Autistic differences in the temporal dynamics of social attention. *Autism*, *25*(6), 1615–1626. https://doi.org/10.1177/1362361321998573

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer. https://doi.org/10.1007/978-1-4612-1304-8

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, *38*(8), 1–29. https://doi.org/10.18637/jss.v038.i08

Jackson, C. H. (2025). Stable and practical semi-Markov modelling of intermittently-observed data. *arXiv*. (arXiv:2508.20949). https://arxiv.org/abs/2508.20949

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, *42*(4), 855–865. https://doi.org/10.2307/2530699

Kim, S. H., Thurm, A., Shumway, S., & Lord, C. (2013). Multisite study of new autism diagnostic interview-revised (ADI-R) algorithms for toddlers and young preschoolers. *Journal of Autism and Developmental Disorders*, *43*(7), 1527–1538. https://doi.org/10.1007/s10803-012-1696-4

Kitzerow, J., Hackbusch, M., Jensen, K., Kieser, M., Noterdaeme, M., Fröhlich, U., Taurines, R., Geißler, J., Wolff, N., Roessner, V., Bast, N., Teufel, K., Kim, Z., & Freitag, C. M. (2020). Study protocol of the multi-centre, randomised controlled trial of the Frankfurt Early Intervention Programme A-FFIP versus early intervention as usual for toddlers and preschool children with Autism Spectrum Disorder (A-FFIP study). *Trials*, *21*(1), 217. https://doi.org/10.1186/s13063-019-3881-7

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., & Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A

dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Lai, M.-C., Kassee, C., Besney, R., Bonato, S., Hull, L., Mandy, W., Szatmari, P., & Ameis, S. H. (2019). Prevalence of co-occurring mental health diagnoses in the autism population: A systematic review and meta-analysis. *The Lancet Psychiatry*, *6*(10), 819–829. https://doi.org/10.1016/S2215-0366(19)30289-5

Lasch, C., Carlson, S. M., & Elison, J. T. (2023). Responding to joint attention as a developmental catalyst: Longitudinal associations with language and social responsiveness. *Infancy*, *28*(2), 339–366. https://doi.org/10.1111/infa.12515

Lord, C. (2015). *ADOS-2: Diagnostische Beobachtungsskala für autistische Störungen-2: Manual*. Huber.

Mangold, P. (2018). Discover the invisible through tool-supported scientific observation. In H. Böttger, K. Jensen, & T. Jensen (Eds.), *Mindful Evolution: Conference Proceedings*. Klinkhardt.

Marhasev, E., Hadad, M., Kaminka, G. A., & Feintuch, U. (2009). The use of hidden semi-Markov models in clinical diagnosis maze tasks. *Intelligent Data Analysis*, *13*(6), 943–967. https://doi.org/10.3233/IDA-2009-0402

Masedu, F., Vagnetti, R., Pino, M. C., Valenti, M., & Mazza, M. (2022). Comparison of visual fixation trajectories in toddlers with autism spectrum disorder and typical development: A Markov chain model. *Brain Sciences*, *12*(1), 10. https://doi.org/10.3390/brainsci12010010

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Mundy, P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *European Journal of Neuroscience*, *47*(6), 497–514. https://doi.org/10.1111/ejn.13720

Mundy, P., Delgado, C., Block, J., Venezia, M., Hogan, A., & Seibert, J. (2003). *Early Social Communication Scales (ESCS)*. University of Miami.

Mundy, P., Sigman, M., & Kasari, C. (1994). Joint attention, developmental level, and symptom presentation in autism. *Development and Psychopathology*, *6*(3), 389–401. https://doi.org/10.1017/S0954579400006003

Mundy, P., Sullivan, L., & Mastergeorge, A. M. (2009). A parallel and distributed-processing model of joint attention, social cognition and autism. *Autism Research*, *2*(1), 2–21. https://doi.org/10.1002/aur.61

Pang, M., Kaufman, J. S., & Platt, R. W. (2016). Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Statistical Methods in Medical Research*, *25*(5), 1925–1937. https://doi.org/10.1177/0962280213505804

Petermann, F., Ricken, G., Fritz, A., Schuck, K.-D., & Preuss, U. (Eds.). (2014). *WPPSI-III: Wechsler Preschool and Primary Scale of Intelligence – Third Edition. German adaptation after D. Wechsler* (3rd ed.). Pearson Assessment.

Portugal, A. M., Viktorsson, C., Taylor, M. J., Mason, L., Tammimies, K., Ronald, A., & Falck-Ytter, T. (2024). Infants' looking preferences for social versus non-social objects reflect genetic variation. *Nature Human Behaviour*, *8*(1), 115–124. https://doi.org/10.1038/s41562-023-01764-w

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, *26*(11), 2389–2430. https://doi.org/10.1002/sim.2712

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Reuner, G., & Rosenkranz, J. (Eds.). (2014). *Bayley Scales of Infant and Toddler Development – Third Edition (German version)*. Pearson Assessment.

Ryan, F., Bati, A., Lee, S., Bolya, D., Hoffman, J., & Rehg, J. M. (2025). *Gaze-LLE: Gaze target estimation via large-scale learned encoders*. 28874–28884.

Sandbank, M., Bottema-Beutel, K., LaPoint, S. C., Feldman, J. I., Barrett, D. J., Caldwell, N., Dale, S., Dunham, K., Crank, J., Albarran, S., & Woynaroski, T. (2023). Autism intervention meta-analysis of early childhood studies (Project AIM): Updated systematic review and secondary analysis. *BMJ*, *383*, e076733. https://doi.org/10.1136/bmj-2023-076733

Sano, M., Yoshimura, Y., Hirosawa, T., Hasegawa, C., An, K.-M., Tanaka, S., Naitou, N., & Kikuchi, M. (2021). Joint attention and intelligence in children with autism spectrum disorder without severe intellectual disability. *Autism Research*, *14*(12), 2603–2612. https://doi.org/10.1002/aur.2600

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*, *253*(5489), 265–266. https://doi.org/10.1038/253265a0

Sodian, B., & Kristen-Antonow, S. (2015). Declarative joint attention as a foundation of theory of mind. *Developmental Psychology*, *51*(9), 1190–1200. https://doi.org/10.1037/dev0000039

Tafasca, S., Gupta, A., & Odobez, J.-M. (2023). *ChildPlay: A new benchmark for understanding children's gaze behaviour*. 20935–20946.

Tafasca, S., Gupta, A., & Odobez, J.-M. (2024). *Sharingan: A transformer architecture for multi-person gaze following*. 2008–2017.

Thorup, E., Nyström, P., Gredebäck, G., Bölte, S., & Falck-Ytter, T. (2016). Altered gaze following during live interaction in infants at risk for autism: An eye tracking study. *Molecular Autism*, *7*(1), 12. https://doi.org/10.1186/s13229-016-0069-9

Thurman, A. J., & Dimachkie Nunnally, A. (2022). Joint attention performance in preschool-aged boys with autism or fragile X syndrome. *Frontiers in Psychology*, *13*, 918181. https://doi.org/10.3389/fpsyg.2022.918181

Tiede, G., & Walton, K. M. (2019). Meta-analysis of naturalistic developmental behavioral interventions for young children with autism spectrum disorder. *Autism*, *23*(8), 2080–2095. https://doi.org/10.1177/1362361319836371

Titman, A. C., & Sharples, L. D. (2010). Semi-Markov models with phase-type sojourn distributions. *Biometrics*, *66*(3), 742–752. https://doi.org/10.1111/j.1541-0420.2009.01339.x

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691. https://doi.org/10.1017/S0140525X05000129

Tönsing, D., Schiller, B., Vehlen, A., Nickel, K., Tebartz van Elst, L., Domes, G., & Heinrichs, M. (2025). Altered interactive dynamics of gaze behavior during face-to-face interaction in autistic individuals: A dual eye-tracking study. *Molecular Autism*, *16*(1), 12. https://doi.org/10.1186/s13229-025-00645-5

Tronick, E. Z. (1989). Emotions and emotional communication in infants. *American Psychologist*, *44*(2), 112–119. https://doi.org/10.1037/0003-066X.44.2.112

Vaughan Van Hecke, A., Mundy, P., Block, J. J., Delgado, C. E. F., Parlade, M. V., Pomares, Y. B., & Hobson, J. A. (2012). Infant responding to joint attention, executive processes, and self-regulation in preschool children. *Infant Behavior and Development*, *35*(2), 303–311. https://doi.org/10.1016/j.infbeh.2011.12.001

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved R^ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Wu, Y., Jarvis, E. D., & Sarkar, A. (2024). Bayesian semiparametric Markov renewal mixed models for vocalization syntax. *Biostatistics*, *25*(3), 648–665. https://doi.org/10.1093/biostatistics/kxac050

Wu, Y., & Sarkar, A. (2024). BMRMM: An R package for Bayesian Markov (renewal) mixed models. *arXiv*. (arXiv:2409.10835). https://arxiv.org/abs/2409.10835

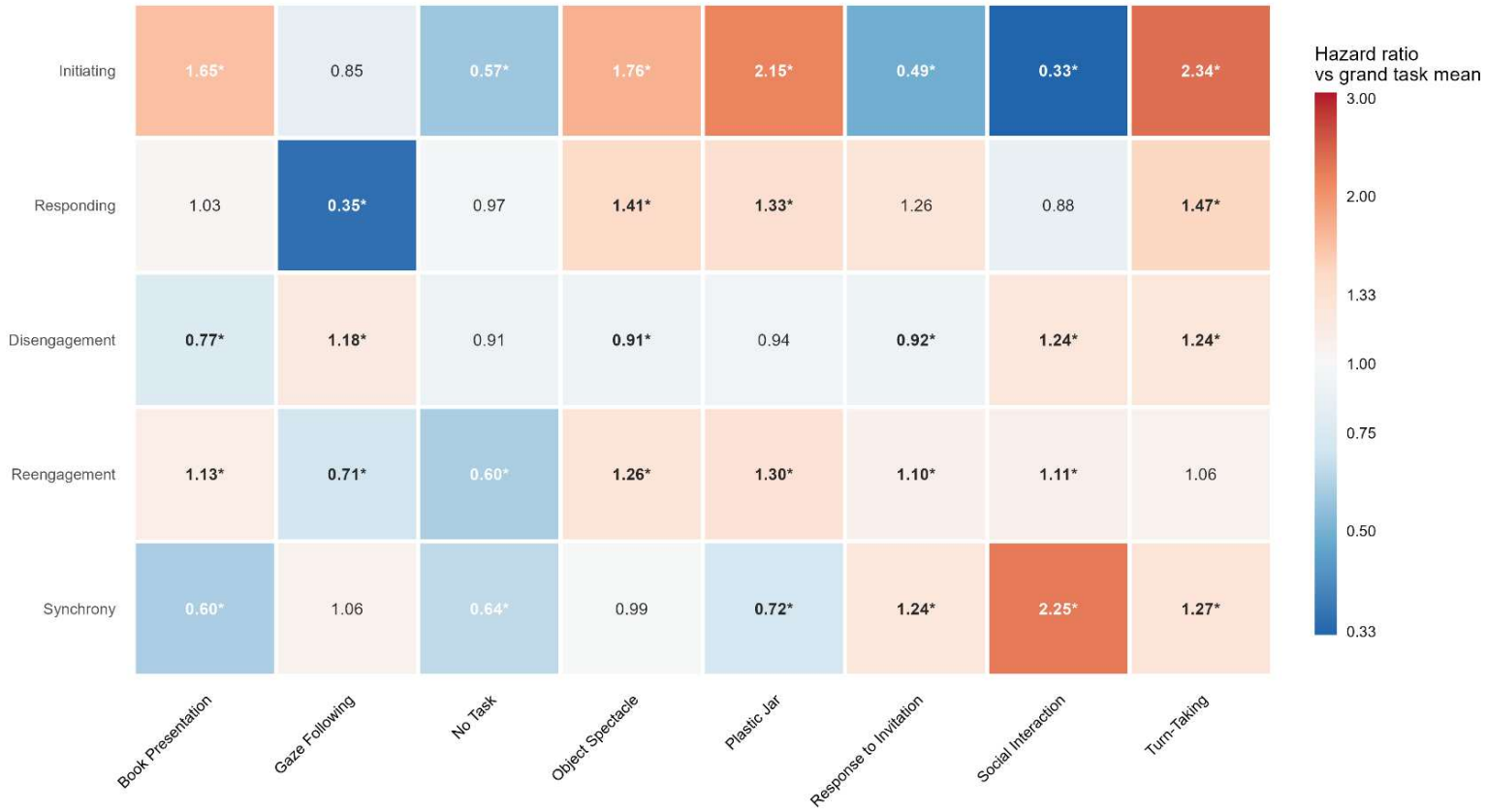
Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE*, *8*(11), e79659. https://doi.org/10.1371/journal.pone.0079659

Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, *174*(2), 215–243. https://doi.org/10.1016/j.artint.2009.11.011

Supplementary Materials

Figure S1

Contextual dynamic of social interaction as Hazard ratios intensity changes for each predefined social interaction domain (rows) between ESCS tasks (columns).



Note. Task effects on focal transition intensities by interaction domain, from the full hierarchical semi-Markov model (12,000 posterior draws; N = 28 children, 56 sessions). An overview of task descriptions is displayed in Table S6. Rows are the five interaction domains, columns the eight ESCS task categories. Tasks were entered with sum-to-zero deviation coding, so each cell is the hazard ratio (HR) of a domain's transition intensity for that task relative to the grand mean across tasks; values above 1 (red) indicate elevated and below 1 (blue) suppressed intensity relative to the average task. Asterisks (*) denote credible effects (95% CrI excluding 1). HR = hazard ratio; CrI = credible interval; ESCS = Early Social Communication Scales.

Table S1

Full Person-Level Covariate Effects on Focal Transition Intensities.

Covariate	Focal transition	Domain	HR	95% CrI
Baseline age (per 1 SD)	Mutual-CT → Refer-C	Initiating	0.94	[0.56, 1.46]
	Follow-C → Share-CT	Responding	1.15	[0.84, 1.55]
	Refer-T → Follow-C	Responding	1.15	[0.86, 1.51]
	Mutual-CT → Avert-C	Disengagement	1.03	[0.83, 1.27]

Covariate	Focal transition	Domain	HR	95% CrI
	Share-CT → Avert-C	Disengagement	0.97	[0.79, 1.18]
	Avert-C → Mutual-CT	Reengagement	**1.47***	**[1.28, 1.68]**
	Avert-C → Share-CT	Reengagement	**1.20***	**[1.05, 1.35]**
	Mutual-CT → Share-CT	Synchrony	0.98	[0.72, 1.27]
	Share-CT → Mutual-CT	Synchrony	1.19	[0.90, 1.53]
Sex (male vs female)	Mutual-CT → Refer-C	Initiating	0.88	[0.40, 1.72]
	Follow-C → Share-CT	Responding	0.90	[0.53, 1.43]
	Refer-T → Follow-C	Responding	**0.57***	**[0.34, 0.89]**
	Mutual-CT → Avert-C	Disengagement	0.92	[0.62, 1.34]
	Share-CT → Avert-C	Disengagement	1.14	[0.77, 1.62]
	Avert-C → Mutual-CT	Reengagement	0.96	[0.72, 1.24]
	Avert-C → Share-CT	Reengagement	**1.34***	**[1.02, 1.70]**
	Mutual-CT → Share-CT	Synchrony	0.86	[0.50, 1.36]
	Share-CT → Mutual-CT	Synchrony	1.18	[0.71, 1.83]
Nonverbal Cognitive Ability (per 1 SD)	Mutual-CT → Refer-C	Initiating	0.93	[0.58, 1.42]
	Follow-C → Share-CT	Responding	1.05	[0.75, 1.42]
	Refer-T → Follow-C	Responding	**1.89***	**[1.45, 2.43]**
	Mutual-CT → Avert-C	Disengagement	**0.80***	**[0.66, 0.97]**
	Share-CT → Avert-C	Disengagement	**0.82***	**[0.68, 0.98]**
	Avert-C → Mutual-CT	Reengagement	**1.73***	**[1.53, 1.94]**
	Avert-C → Share-CT	Reengagement	**1.20***	**[1.07, 1.34]**
	Mutual-CT → Share-CT	Synchrony	1.29	[0.99, 1.65]
	Share-CT → Mutual-CT	Synchrony	1.25	[0.97, 1.59]
Intervention (EIAU vs A-FFIP, at T2)	Mutual-CT → Refer-C	Initiating	0.84	[0.40, 1.56]
	Follow-C → Share-CT	Responding	0.80	[0.47, 1.28]
	Refer-T → Follow-C	Responding	**2.12***	**[1.33, 3.20]**
	Mutual-CT → Avert-C	Disengagement	1.04	[0.73, 1.42]
	Share-CT → Avert-C	Disengagement	0.93	[0.67, 1.25]
	Avert-C → Mutual-CT	Reengagement	1.10	[0.86, 1.39]
	Avert-C → Share-CT	Reengagement	1.16	[0.94, 1.41]
	Mutual-CT → Share-CT	Synchrony	**1.62***	**[1.02, 2.42]**
	Share-CT → Mutual-CT	Synchrony	1.25	[0.79, 1.83]

Covariate	Focal transition	Domain	HR	95% CrI
Time point (T5 vs T2, in A-FFIP)	Mutual-CT → Refer-C	Initiating	1.12	[0.78, 1.55]
	Follow-C → Share-CT	Responding	1.36	[0.99, 1.83]
	Refer-T → Follow-C	Responding	1.08	[0.80, 1.43]
	Mutual-CT → Avert-C	Disengagement	**1.49***	**[1.28, 1.72]**
	Share-CT → Avert-C	Disengagement	1.02	[0.94, 1.11]
	Avert-C → Mutual-CT	Reengagement	**1.40***	**[1.20, 1.61]**
	Avert-C → Share-CT	Reengagement	1.10	[1.00, 1.21]
	Mutual-CT → Share-CT	Synchrony	**1.36***	**[1.07, 1.70]**
	Share-CT → Mutual-CT	Synchrony	0.94	[0.78, 1.14]
Group × Time interaction	Mutual-CT → Refer-C	Initiating	**1.84***	**[1.14, 2.83]**
	Follow-C → Share-CT	Responding	1.02	[0.68, 1.46]
	Refer-T → Follow-C	Responding	1.17	[0.80, 1.66]
	Mutual-CT → Avert-C	Disengagement	**0.75***	**[0.60, 0.92]**
	Share-CT → Avert-C	Disengagement	1.07	[0.95, 1.20]
	Avert-C → Mutual-CT	Reengagement	1.03	[0.83, 1.27]
	Avert-C → Share-CT	Reengagement	0.91	[0.79, 1.04]
	Mutual-CT → Share-CT	Synchrony	0.94	[0.68, 1.25]
	Share-CT → Mutual-CT	Synchrony	1.07	[0.82, 1.37]

Note. All 54 person-level covariate effects on the nine focal transitions, from the full hierarchical semi-Markov model (the 16 effects with 95% credible intervals excluding 1 are reported in Table 4 of the main text). Effects are hazard ratios (HR) on focal transition intensities, reported as posterior means with 95% equal-tailed credible intervals (CrI). Continuous covariates (baseline age, nonverbal cognitive ability) are scaled per one standard deviation; categorical covariates are indicators (sex: male = 1; intervention: EIAU = 1; time point: T5 = 1). Because the model includes a group × time-point interaction, the intervention and time-point main effects are interpreted within their reference category (A-FFIP and T2, respectively). Bold values with an asterisk (*) denote effects whose 95% CrI excludes 1. HR = hazard ratio; CrI = credible interval.

Table S2

Effect of Nonverbal Cognitive Ability on Focal Transition Next-State Probabilities Across Four In-Sample Contrast Widths.

Transition	Domain	Contrast	Raw cog. ability (low/high)	P (low cog. ability)	P (high cog. ability)	Difference [95% CrI]	Ratio [95% CrI]
Mutual-CT → Refer-C	Initiating	IQR (Q1–Q3)	46.0 / 72.8	0.067	0.066	−0.002 [−0.040, 0.034]	0.97 [0.58, 1.68]
		±1 SD	40.5 / 79.9	0.068	0.064	−0.003 [−0.058, 0.051]	0.95 [0.45, 2.14]
		P10–P90	38.3 / 83.9	0.068	0.064	−0.004 [−0.067, 0.060]	0.94 [0.39, 2.40]
		Min–Max	32.0 / 108.0	0.067	0.058	−0.009 [−0.099, 0.110]	0.86 [0.19, 3.99]
Refer-T → Follow-C	Responding	IQR (Q1–Q3)	46.0 / 72.8	0.239	0.402	**0.163 [0.098, 0.223]***	**1.68 [1.36, 2.08]***
		±1 SD	40.5 / 79.9	0.211	0.451	**0.239 [0.143, 0.325]***	**2.13 [1.56, 2.88]***
		P10–P90	38.3 / 83.9	0.201	0.478	**0.278 [0.167, 0.376]***	**2.39 [1.67, 3.36]***
		Min–Max	32.0 / 108.0	0.173	0.644	**0.471 [0.286, 0.621]***	**3.73 [2.25, 5.89]***
Follow-C → Share-CT	Responding	IQR (Q1–Q3)	46.0 / 72.8	0.842	0.848	0.006 [−0.047, 0.063]	1.01 [0.95, 1.08]
		±1 SD	40.5 / 79.9	0.841	0.850	0.009 [−0.070, 0.092]	1.01 [0.92, 1.12]
		P10–P90	38.3 / 83.9	0.841	0.851	0.011 [−0.082, 0.106]	1.01 [0.91, 1.14]
		Min–Max	32.0 / 108.0	0.839	0.858	0.018 [−0.147, 0.162]	1.02 [0.83, 1.21]
Mutual-CT → Avert-C	Disengagement	IQR (Q1–Q3)	46.0 / 72.8	0.565	0.479	**−0.085 [−0.149, −0.019]***	**0.85 [0.75, 0.96]***
		±1 SD	40.5 / 79.9	0.581	0.456	**−0.125 [−0.218, −0.028]***	**0.78 [0.65, 0.95]***
		P10–P90	38.3 / 83.9	0.588	0.442	**−0.146 [−0.251, −0.033]***	**0.75 [0.60, 0.94]***
		Min–Max	32.0 / 108.0	0.605	0.358	**−0.247 [−0.411, −0.063]***	**0.59 [0.38, 0.88]***
Share-CT → Avert-C	Disengagement	IQR (Q1–Q3)	46.0 / 72.8	0.630	0.549	**−0.081 [−0.140, −0.020]***	**0.87 [0.79, 0.97]***
		±1 SD	40.5 / 79.9	0.645	0.526	**−0.119 [−0.205, −0.030]***	**0.82 [0.70, 0.95]***

Transition	Domain	Contrast	Raw cog. ability (low/high)	P (low cog. ability)	P (high cog. ability)	Difference [95% CrI]	Ratio [95% CrI]
		P10–P90	38.3 / 83.9	0.651	0.513	**−0.138 [−0.237, −0.035]***	**0.79 [0.66, 0.94]***
		Min–Max	32.0 / 108.0	0.668	0.433	**−0.236 [−0.396, −0.061]***	**0.65 [0.45, 0.90]***
Avert-C → Mutual-CT	Reengagement	IQR (Q1–Q3)	46.0 / 72.8	0.160	0.251	**0.091 [0.076, 0.106]***	**1.57 [1.46, 1.69]***
		±1 SD	40.5 / 79.9	0.145	0.280	**0.135 [0.113, 0.157]***	**1.93 [1.74, 2.15]***
		P10–P90	38.3 / 83.9	0.139	0.297	**0.158 [0.132, 0.184]***	**2.14 [1.90, 2.42]***
		Min–Max	32.0 / 108.0	0.123	0.408	**0.285 [0.237, 0.334]***	**3.31 [2.77, 3.97]***
Avert-C → Share-CT	Reengagement	IQR (Q1–Q3)	46.0 / 72.8	0.511	0.497	−0.014 [−0.040, 0.013]	0.97 [0.92, 1.03]
		±1 SD	40.5 / 79.9	0.510	0.488	−0.022 [−0.060, 0.016]	0.96 [0.89, 1.03]
		P10–P90	38.3 / 83.9	0.510	0.482	−0.028 [−0.071, 0.017]	0.95 [0.87, 1.03]
		Min–Max	32.0 / 108.0	0.507	0.431	**−0.075 [−0.140, −0.009]***	**0.85 [0.74, 0.98]***
Mutual-CT → Share-CT	Synchrony	IQR (Q1–Q3)	46.0 / 72.8	0.091	0.142	**0.050 [0.015, 0.087]***	**1.55 [1.13, 2.12]***
		±1 SD	40.5 / 79.9	0.083	0.158	**0.075 [0.021, 0.130]***	**1.90 [1.19, 2.99]***
		P10–P90	38.3 / 83.9	0.080	0.167	**0.088 [0.025, 0.154]***	**2.10 [1.23, 3.53]***
		Min–Max	32.0 / 108.0	0.071	0.232	**0.161 [0.041, 0.304]***	**3.27 [1.39, 7.13]***
Share-CT → Mutual-CT	Synchrony	IQR (Q1–Q3)	46.0 / 72.8	0.087	0.129	**0.042 [0.009, 0.076]***	**1.48 [1.09, 2.00]***
		±1 SD	40.5 / 79.9	0.079	0.142	**0.062 [0.014, 0.113]***	**1.78 [1.14, 2.75]***
		P10–P90	38.3 / 83.9	0.077	0.149	**0.073 [0.016, 0.134]***	**1.95 [1.16, 3.21]***
		Min–Max	32.0 / 108.0	0.069	0.203	**0.133 [0.026, 0.265]***	**2.92 [1.27, 6.23]***

Note. Effects of nonverbal cognitive ability on focal transition next-state probabilities, estimated from the full hierarchical semi-Markov model on a sample-average (observed-value) basis, across four in-sample contrast widths. P (low cognitive ability) and P (high cognitive ability) are posterior median next-state probabilities at the low and high contrast endpoints; differences and ratios are formed within each posterior draw and summarized with 95% equal-tailed credible intervals (CrI). Contrast

widths span the observed cognitive ability range and are given as observed developmental quotient values (low / high) in the Raw cog. ability column (sample M = 60.2, SD = 19.7): IQR (25th–75th percentile), ±1 SD, P10–P90 (10th–90th percentile), and Min–Max (observed minimum–maximum). Bold values with an asterisk (*) denote credible effects (95% CrI excludes the null: 0 for differences, 1 for ratios). CrI = credible interval; P = next-state probability.

Table S3

Effect of Nonverbal Cognitive Ability on Model-Implied Median Sojourn Durations Across Four In-Sample Contrast Widths.

State	Contrast	Raw cog. ability (low/high)	Median (low cog. ability)	Median (high cog. ability)	Difference [95% CrI]	Ratio [95% CrI]
Mutual-CT	IQR (Q1–Q3)	46.0 / 72.8	0.73	0.82	0.09 [−0.01, 0.18]	1.12 [0.98, 1.27]
	±1 SD	40.5 / 79.9	0.71	0.83	0.12 [−0.02, 0.26]	1.17 [0.98, 1.41]
	P10–P90	38.3 / 83.9	0.70	0.84	0.14 [−0.02, 0.30]	1.20 [0.97, 1.48]
	Min–Max	32.0 / 108.0	0.68	0.88	0.20 [−0.07, 0.46]	1.29 [0.92, 1.79]
Share-CT	IQR (Q1–Q3)	46.0 / 72.8	3.45	4.08	0.63 [−0.11, 1.35]	1.18 [0.97, 1.43]
	±1 SD	40.5 / 79.9	3.31	4.23	0.92 [−0.17, 1.97]	1.28 [0.96, 1.69]
	P10–P90	38.3 / 83.9	3.25	4.31	1.06 [−0.21, 2.28]	1.32 [0.95, 1.82]
	Min–Max	32.0 / 108.0	3.10	4.70	1.60 [−0.46, 3.70]	1.52 [0.88, 2.47]
Refer-C	IQR (Q1–Q3)	46.0 / 72.8	0.61	0.61	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	±1 SD	40.5 / 79.9	0.61	0.61	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	P10–P90	38.3 / 83.9	0.61	0.61	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	Min–Max	32.0 / 108.0	0.61	0.61	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
Refer-T	IQR (Q1–Q3)	46.0 / 72.8	0.79	0.57	**−0.22 [−0.30, −0.13]***	**0.72 [0.64, 0.82]***
	±1 SD	40.5 / 79.9	0.83	0.51	**−0.32 [−0.44, −0.19]***	**0.62 [0.51, 0.75]***
	P10–P90	38.3 / 83.9	0.85	0.48	**−0.37 [−0.51, −0.22]***	**0.56 [0.45, 0.72]***
	Min–Max	32.0 / 108.0	0.89	0.29	**−0.60 [−0.79, −0.37]***	**0.32 [0.19, 0.54]***
Follow-C	IQR (Q1–Q3)	46.0 / 72.8	0.33	0.32	−0.01 [−0.12, 0.10]	0.96 [0.70, 1.35]

State	Contrast	Raw cog. ability (low/high)	Median (low cog. ability)	Median (high cog. ability)	Difference [95% CrI]	Ratio [95% CrI]
	±1 SD	40.5 / 79.9	0.33	0.31	−0.02 [−0.18, 0.14]	0.94 [0.59, 1.55]
	P10–P90	38.3 / 83.9	0.33	0.31	−0.02 [−0.21, 0.16]	0.93 [0.54, 1.66]
	Min–Max	32.0 / 108.0	0.33	0.30	−0.04 [−0.32, 0.30]	0.89 [0.35, 2.28]
Follow-T	IQR (Q1–Q3)	46.0 / 72.8	0.13	0.13	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	±1 SD	40.5 / 79.9	0.13	0.13	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	P10–P90	38.3 / 83.9	0.13	0.13	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	Min–Max	32.0 / 108.0	0.13	0.13	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
Avert-C	IQR (Q1–Q3)	46.0 / 72.8	2.27	1.63	**−0.63 [−0.88, −0.39]***	**0.72 [0.64, 0.82]***
	±1 SD	40.5 / 79.9	2.40	1.48	**−0.92 [−1.28, −0.57]***	**0.61 [0.51, 0.74]***
	P10–P90	38.3 / 83.9	2.46	1.39	**−1.07 [−1.47, −0.66]***	**0.57 [0.46, 0.70]***
	Min–Max	32.0 / 108.0	2.61	0.93	**−1.68 [−2.24, −1.10]***	**0.36 [0.24, 0.52]***
Avert-T	IQR (Q1–Q3)	46.0 / 72.8	0.05	0.05	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	±1 SD	40.5 / 79.9	0.05	0.05	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	P10–P90	38.3 / 83.9	0.05	0.05	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]
	Min–Max	32.0 / 108.0	0.05	0.05	0.00 [0.00, 0.00]	1.00 [1.00, 1.00]

Note. Model-implied median sojourn durations (in seconds) at low versus high nonverbal cognitive ability for each interaction state, estimated from the full hierarchical semi-Markov model on a sample-average (observed-value) basis, across four in-sample contrast widths. Estimates are posterior medians; differences and ratios are formed within each posterior draw and summarized with 95% equal-tailed credible intervals (CrI). Contrast widths span the observed cognitive ability range and are given as observed developmental quotient values (low / high) in the Raw cog. ability column (sample M = 60.2, SD = 19.7): IQR (25th–75th percentile), ±1 SD, P10–P90 (10th–90th percentile), and Min–Max (observed minimum–maximum). Bold values with an asterisk (*) denote credible effects (95% CrI excludes the null: 0 for differences, 1 for ratios). CrI = credible interval.

Table S4

Nested model comparison.

Model	Description	ELPD	ΔELPD vs M0 (SE)	Bayesian R^2 by domain (95% CrI)					Credible covariate effects
				Init	Resp	Diseng	Reeng	Synch	
M0 (baseline)	Baseline + random effects	−78,020	—	0.42 (0.30, 0.56)	0.36 (0.29, 0.44)	0.20 (0.18, 0.22)	0.11 (0.10, 0.13)	0.34 (0.31, 0.37)	—
M1	M0 + temporal/contextual variables	−77,274	+745.2 (43.1)	0.43 (0.31, 0.57)	0.38 (0.32, 0.44)	0.20 (0.19, 0.22]	0.12 (0.11, 0.13)	0.29 (0.26, 0.32)	—
M2 (full model)	M1 + covariates	−77,144	+875.5 (47.3)	0.47 (0.36, 0.61)	0.42 (0.37, 0.47)	0.20 (0.18, 0.22)	0.15 (0.14, 0.17)	0.30 (0.27, 0.33)	16
M_NoRE	M2 without random effects	−77,776	+244.2 (62.0)	0.13 (0.08, 0.20)	0.40 (0.36, 0.44)	0.14 (0.13, 0.16)	0.14 (0.12, 0.15)	0.24 (0.21, 0.27)	30

Note. *ELPD*, expected log predictive density from PSIS-LOO. SE, standard error of the ELPD difference relative to M0. R^2, posterior mean of the Bayesian R^2 per domain. R^2 sub-columns are abbreviated as Init for initiating, Resp for responding, Diseng for disengagement, Reeng for reengagement, and Synch for synchrony. The final column gives the number, out of 54 covariate × focal-transition coefficients, whose 95% credible intervals exclude 1. The column is not applicable to M0 or M1, which omit person-level covariates (denoted —). All four models share the same nine focal transitions, ten Weibull state-level sojourn distributions, and Bayesian estimation via Stan. M_NoRE differs from M2 only in the omission of subject-by-domain random effects.

Table S5

Convergence and Sampling Diagnostics for All Fitted Models.

Model	R̂ max	R̂ > 1.01	ESS bulk (min)	ESS bulk (med)	ESS tail (min)	ESS tail (med)	Divergences	Max-treedepth hits	E-BFMI (min)
M0 (baseline)	1.0037	0	876	14,043	1,685	8,941	0	0	0.84
M1	1.0022	0	1,716	20,310	3,195	9,085	0	0	0.83
M2 (full)	1.0020	0	3,309	16,786	5,428	9,312	0	0	0.80
M_NoRE	1.0024	0	5,907	14,900	4,635	9,284	0	0	1.01

Note. Convergence and sampling diagnostics for the four models compared in the main text, each fitted with the same Hamiltonian Monte Carlo configuration. R̂, rank-normalized split potential scale reduction factor (maximum across parameters, and the count of parameters exceeding 1.01). ESS, effective sample size (minimum and median across parameters), reported separately for bulk and tail. Divergences, number of divergent transitions; Max-treedepth hits, number of iterations saturating the maximum tree depth. E-BFMI, minimum chain-level energy Bayesian fraction of missing information. M0, baseline with subject-by-domain random effects; M1, M0 plus within-session predictors; M2, full model adding person-level covariates; M_NoRE, M2 without random effects.

Supplementary Results 1. Convergence, predictive accuracy, and posterior predictive checks

All four models converged under Hamiltonian Monte Carlo sampling. In the full model the maximum rank-normalized split $\hat{R}$ was 1.0020, indicating close agreement across chains (Vehtari et al., 2021). The minimum bulk effective sample size was 3,309, with a median of 16,786. No divergent transitions or maximum-treedepth saturations occurred. The minimum chain-level energy Bayesian fraction of missing information was 0.80, indicating adequate exploration of the posterior energy distribution (Betancourt, 2016). Diagnostics for the remaining models met the same thresholds and are reported in Supplementary Table S5.

Predictive accuracy increased across the nested model sequence (Supplementary Table S4). Pareto-smoothed importance sampling leave-one-out cross-validation estimates the expected log predictive density for held-out observations, with higher values indicating better expected predictive performance (Vehtari, Gelman & Gabry, 2017). Fewer than 2% of observations had Pareto-k values above 0.7 in any model, supporting the reliability of the approximation. Removing the subject-by-domain random effects from the full model reduced the cumulative improvement to 244.2 (SE = 62.0).

Posterior predictive checks showed that the full model reproduced state occupancy and mean sojourn duration more closely than median sojourn duration. State occupancy fell within the 95% posterior predictive band for all ten states, and marginal sojourn means were recovered within approximately 10% for every state. Sojourn medians were over-predicted by approximately 10 to 30% for most states. This discrepancy was larger for states with Weibull shapes below 1, including Share-CT (+26%, $\alpha = 0.74$), Break (+24%, $\alpha = 0.86$), and Refer-T (+20%, $\alpha = 0.71$). The most likely source is that a single shape per state must accommodate sojourns generated under different exit rates, since phase, task, recent engagement, and between-child variation all shift the rate while the shape is held fixed. The marginal sojourn distribution for a state is therefore a scale mixture of Weibull distributions sharing one shape, which is more right-skewed than any single Weibull, so fitting one shape to it recovers the mean while placing the median too high. The overprediction of Avert-T (29%, $\alpha = 0.85$) may further be explained by most sojourns being a single frame, which no continuous distribution can reproduce. Because the shape was estimated at the state level and did not vary with covariates, the resulting multiplicative bias is shared across covariate scenarios within a state and does not affect the within-state sojourn ratios reported in the main text, although the absolute model-implied medians retain the upward bias disclosed here. Focal transition counts fell within the 95% posterior predictive band for five of the nine

focal transitions, with deviations on the remaining four no larger than 15% and showing no systematic direction.

Figure S2

Example frames of the ESCS recording setup.



Note. Single frames from the two camera angles during the gaze-following task within the Early Social Communication Scales (ESCS), in which the tester points toward a visual target (poster) placed on the wall.

The images display an instance of shared attention (Share-CT) after the child has followed the tester's referential bid. Faces are blurred for anonymization.

Table S6

Description of the eight ESCS task categories.

Task	Description
Book Presentation	The tester places an open picture book with large, distinct pictures on the table within the child's reach and asks what the child sees, allowing approximately 20s of free examination. The tester then points to six pictures in turn, holding each point for about 3s roughly two inches from the page without touching it, saying the child's name while pointing to a picture on the left side of the book and then pointing to a picture on the right side before turning the page. Presented once per session.
Gaze Following	The tester first draws the child's attention to his/her own face, then turns his/her whole torso toward one of four posters on the walls and points at it with the elbow held against her side, while saying the child's name three times with increasing emphasis. Each pointing episode is maintained for at least 6s, and the four targets are addressed in the order left, left-behind, right, right-behind. Administered twice per session, with four pointing trials each time.
No Task	Intervals in which no ESCS task was in progress. These comprise the period before the first task began, the period after the last task ended, and the transitions in which the tester had put away one set of materials and had not yet retrieved the next. Interaction states occurring in these intervals were retained in the sequence and assigned to this category rather than to an adjacent task.
Object Spectacle	The tester activates a mechanical toy on the table in front of, but out of reach of, the child, and remains silent but attentive while the toy is active, for at least 6s. Each toy is presented three times, and up to five times if the child shows sustained interest. Once the toy stops, it is placed within the child's reach for approximately 10s of play. Six episodes per session, drawn from three wind-up toys and three hand-operated toys.
Plastic Jar	The tester shows the child a transparent jar with two novel wind-up toys sealed inside, unscrews the lid and pours the toys onto the table, then returns them to the jar and seals the lid so that the child cannot open it. The jar is handed to the child, who holds it for approximately 10s before it is requested back. The tester then opens the jar, winds one toy and places it on the table, and the sequence is repeated with the second toy. Presented once per session.
Response to Invitation	The tester places a hat, a comb, or a pair of glasses directly in front of the child, who is allowed to play with it for approximately 15s. If the child uses the object conventionally, that is, hat on head, glasses on face, or comb to hair, the tester leans forward, tilts his/her head, and asks whether he/she may play, repeating the invitation three times at 2s intervals or until the child moves the object toward her. If the child does not use the object conventionally, the tester first demonstrates by placing it on the child. Three episodes per session, one for each object.

Task	Description
Social Interaction	The tester clears all toys from the table, announces a game or a song, and sings for approximately 10s. He/she then runs his/her fingers across the table with an accompanying sound and touches or tickles the child, withdraws his/her hands to his/her own side of the table, and pauses for approximately 5s to allow the child to bid for a repetition. Administered twice per session, with three such trials each time.
Turn-Taking	The tester places a ball or a car within the child's easy reach and holds both hands apart on the table in a posture ready to receive it, remaining so for approximately 10s. When the child returns the object, the tester rolls it back, and the exchange continues until the child stops returning it or twelve turns have been taken. Two episodes per session, one with the ball and one with the car.

Note. Task procedures follow the ESCS manual (Mundy et al., 2003) and the protocol administered at the study site. Categories are listed alphabetically, matching the column order of Figure S1. For the tasks that use materials, a task began when the tester retrieved the material from the storage box and ended when she returned it or retrieved the material for the next task. For the tasks that use no materials (Social Interaction, Gaze Following), a task began when the tester placed both hands on the table and began to sing or gesture and ended when she placed both hands on the table for the last time. Each interaction state was assigned the task active at its onset, and any interval not falling within a task segment was assigned to No Task. Task was entered in the model with sum-to-zero deviation coding across the eight categories. ESCS = Early Social Communication Scales.

Table S7

Full gaze annotation label set and operational definitions.

Code	Description
Single-C *single gaze, child*	Child focuses visual attention on a single point or object, without social-communicative intent.
Mutual-CT *mutual gaze, child-tester*	Child and tester look into each other's eyes at the same time.
Mutual-CP *mutual gaze, child-caregiver*	Child and caregiver look into each other's eyes at the same time.
Avert-CT *gaze aversion, child (from tester)*	Child shifts gaze away from the tester following mutual or shared gaze; captures the move away from face or stimulus.
Avert-CP *gaze aversion, child (from caregiver)*	Child shifts gaze away from the caregiver following mutual or shared gaze.
Avert-T *gaze aversion, tester*	Tester shifts gaze away from the child, interrupting dyadic or shared orientation.
Avert-P *gaze aversion, caregiver*	Caregiver shifts gaze away from the child, interrupting dyadic or shared orientation.

Code	Description
Refer-CT *referential gaze, child-tester*	Child directs gaze to induce the tester's attention toward a target; the child is the initiating party.
Refer-CP *referential gaze, child - caregiver*	Child directs gaze to induce the caregiver's attention toward a target; the child is the initiating party.
Refer-T *referential gaze, tester*	Tester directs gaze to induce the child's attention toward a target; the tester is the initiating party.
Refer-P *referential gaze, caregiver*	Caregiver directs gaze to induce the child's attention toward a target; the caregiver is the initiating party.
Follow-CT *gaze following, child (tester cue)*	Child perceives the tester's gaze and follows to the target stimulus the tester is attending to.
Follow-CP *gaze following, child (caregiver cue)*	Child perceives the caregiver's gaze and follows to the target stimulus the caregiver is attending to.
Follow-T *gaze following, tester*	Tester perceives the child's gaze and follows to the target stimulus the child is attending to.
Follow-P *gaze following, caregiver*	Caregiver perceives the child's gaze and follows to the target stimulus the child is attending to.
Share-CT *shared gaze, child-tester*	Child and tester attend to the same stimulus; also coded when only the child gazes at the stimulus but attention was induced by the tester.
Share-CP *shared gaze, child-caregiver*	Child and caregiver attend to the same stimulus; also coded when only the child gazes at the stimulus but attention was induced by the caregiver.
Supporting-CT *supporting gesture, child-tester*	Child uses gesture, pointing, vocalization, or language to direct the tester's attention toward a target
Supporting-CP *supporting gesture, child-caregiver*	Child uses gesture, pointing, vocalization, or language to direct the caregiver's attention toward a target.
Supporting-T *supporting gesture, tester*	Tester uses gesture, pointing, language, or sound to direct the child's attention toward a target; often co-occurs with referential gaze.
Supporting-P *supporting gesture, caregiver*	Caregiver uses gesture, pointing, language, or sound to direct the child's attention toward a target.

Note. Codes follow the atomic gaze-act taxonomy of Fan et al. (2019), adapted for the dyadic ESCS context. C = child; T = tester; P = caregiver; CT = child–tester; CP = child–caregiver. This is the raw annotation scheme applied during video coding. In pre-processing, these labels are mapped to the ten-state model ontology (see

Methods, Pre-processing): caregiver-related and residual non-focal behaviors are combined into the Other state, uncodable gaps into the Break state, supporting gestures are not modelled as separate states, and the child–caregiver and child–tester variants are consolidated.

Supplementary Methods 1. State sequence construction and full pre-processing

This section specifies the pipeline that converts the raw state annotations into the model-ready event-time sequences. The rationale for preserving the temporal structure of each session is given in the main text and is not repeated here. The pipeline was applied to each session independently. Each step is listed in Table S8 in the order applied.

Table S8

Overview of all implemented pre-processing steps.

Step	Operation	Description
1	**Merge same-behavior intervals**	Contiguous annotations of the same behavior are combined into single intervals.
2	**Merge child aversion and off-target fixation**	A child off-target fixation following a child gaze aversion within 0.5s is absorbed into the aversion, so that the full duration of the aversion is captured rather than just the initial shift.
3	**Resolve nesting**	Overlapping annotations are resolved by a fixed rule set that either splits the outer interval around the inner one or drops the inner interval, removing temporal overlap.
4	**Order simultaneous onsets**	Behaviors sharing an exact onset are ordered by their interactional precedence so that the sequence is unambiguous.
5	**Separate tied onsets**	Remaining tied onsets are separated by one frame (0.04s) so that the sequence is strictly ordered in time.
6	**Collapse single-frame aversion**	A single-frame child aversion lying between two engaged states collapses into the preceding engaged state
7	**Insert Break states**	Any gap of 0.5s or longer in which no behavior is coded is filled with an explicit Break state spanning the gap, preserving elapsed time without asserting state continuity.
8	**Map behaviors to states**	Each behavior is mapped to one of the ten model states. Child off-target fixation, caregiver-directed behaviors, and residual non-focal codes map to Other, with the originating code retained as an origin tag.
9	**Merge consecutive identical states**	Reintroduced/remaining adjacent occurrences of the same state are merged, yielding a gap-free sequence of distinct states with their durations.
10	**Absorb bridges**	Brief intervening states interrupting a focal initiating or responding transition are absorbed into the transition under

Step	Operation	Description
		explicit duration and count limits, with content-aware credit direction. Longer or more complex interruptions are left intact.
11	**Assign tasks**	Each state is assigned the ESCS task active at its original onset.
12	**Add session and timing indices**	Session and timing indices are added, producing the gap-free model-ready event-time sequence for each session.
13	**Form sojourns**	Each state's duration is computed as the interval from its own onset to the onset of the next state, so that sojourns capture the session without gap or overlap.
14	**Split at phase boundaries and introduce censoring**	Each session is divided into three equal intervals of elapsed time. A sojourn crossing a boundary is split at it. A segment interrupted by a phase boundary or by the end of the session has no observed destination and is marked as right-censored. This produces the model-ready sojourn table.

Note. Off-target fixation denotes a child fixation directed neither at a shared referent nor at a social partner (coded as single gaze and mapped to the "Other" state). The duration and state caps in step 10 define the primary specification and are varied in the sensitivity analysis below. Times refer to the recorded frame rate of 25 frames per second (0.04 s per frame).

Supplementary Methods 2. Sensitivity analysis of the bridging decision

Bridge specifications

We defined a bridge as a brief intervening state that interrupts one of the two child-directed focal sequences before its destination is reached, absorbed into the transition under limits on the total duration and number of intervening states and with content-aware credit direction, as described in the main text. Because these limits are discretionary, eight specifications were defined, from no bridging (v0) to progressively more permissive limits on the duration, number, and content of intervening states (v7). The primary specification used throughout the main text, here labelled v4, lies at the midpoint of this ladder.

Table S9 reports, for each specification, the number of bridges absorbed and the resulting count of focal initiating and responding transitions, with the gain over the no-bridge baseline in parentheses. The two content kinds recover the two sequences asymmetrically. Permitting only child aversion as bridge content (v1, v2) recovers initiating transitions but almost no responding transitions, whereas permitting only off-target fixation (v6) recovers responding transitions but almost no initiating transitions. This asymmetry is the empirical basis for the content-aware credit direction: a child looking away interrupts an initiation, whereas a child inspecting the referent alone is still engaged in responding to the tester's bid. The primary specification admits both content kinds at a 1.0 s duration limit and a two-state limit, which together recover the majority of brief interruptions while excluding the long, multi-state excursions that the most permissive setting (v7) absorbs.

Robustness of the fitted effects across specifications

Each specification was carried through the full estimation pipeline, and the credible effects reported in the main text were compared across the eight resulting fits. All 16 credible covariate effects remained credible in all eight specifications, with no sign disagreement in any specification and a hazard-ratio range across specifications of at most 1.08 (that is, the point estimate varied by no more than eight per cent between the least and most permissive setting, and by less than four per cent for most effects). All 12 credible treatment-by-time contrasts were likewise credible and sign-consistent across all eight specifications. The substantive conclusions are therefore not artefacts of the bridge-absorption settings.

Convergence across specifications

All eight fits met the same convergence standards (Table S10). Across specifications the maximum R-hat was 1.0021, no parameter exceeded an R-hat of 1.01, and there were no divergent transitions and no maximum-treedepth saturations. The minimum bulk effective sample size across all specifications was 3,085 and the minimum estimated Bayesian fraction of missing information was 0.83.

Table S9

Bridge specifications and recovered focal transitions.

Specification	Bridge content	Duration cap	State cap	Bridges absorbed	Initiating recovered	Responding recovered	Sojourns
v0	None	—	—	0	0 (167)	0 (787)	25,848
v1	Aversion only	0.04 s	1	28	+28 (195)	+0 (787)	25,820
v2	Aversion only	0.5 s	2	91	+81 (248)	+10 (797)	25,757
v3	Aversion + off-target	0.5 s	2	130	+82 (249)	+48 (835)	25,719
v4 (primary)	Aversion + off-target	1.0 s	2	**167**	+97 (264)	+70 (857)	25,682
v5	Aversion + off-target	2.0 s	2	189	+111 (278)	+78 (865)	25,660
v6	Off-target only	1.0 s	2	55	+2 (169)	+53 (840)	25,794
v7	Aversion + off-target	none	2	217	+125 (292)	+92 (879)	25,630

Note. Bridge content lists the kinds of intervening state eligible for absorption. The duration cap is the maximum total duration of intervening material and the state cap the maximum number of intervening states for a bridge to be absorbed. Bridges absorbed is the number of absorbed bridges. Initiating recovered and Responding recovered give the gain in focal transitions of each kind over the no-bridge baseline (v0), with the absolute count in parentheses. Sojourns is the resulting number of completed sojourns across all 56 sessions. Caregiver-related and mixed-origin material is never eligible as bridge content. Off-target fixation denotes child single gaze.

Table S10

Convergence diagnostics for the full model under each bridge specification.

Specification	Max R-hat	Divergences	Max-treedepth hits	Min bulk ESS	Min E-BFMI
v0	1.0018	0	0	3,654	0.853
v1	1.0019	0	0	3,463	0.834
v2	1.0021	0	0	3,521	0.831
v3	1.0019	0	0	3,486	0.843
v4 (primary)	1.0020	0	0	3,506	0.877
v5	1.0017	0	0	3,286	0.845
v6	1.0021	0	0	3,320	0.859
v7	1.0018	0	0	3,085	0.864

Note. Diagnostics are for the full model fitted to each specification. Max R-hat is the largest potential scale-reduction factor across parameters. Min bulk ESS is the smallest bulk effective sample size across parameters. Min E-BFMI is the smallest estimated Bayesian fraction of missing information across chains. Divergences and max-treedepth hits are summed across the four chains.