UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Human Action Scaffolds Visual Non-Adjacent Dependency Learning Beyond Surface Form

Permalink

https://escholarship.org/uc/item/2bf5112h

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Darvishi, Neshat
Jiang, Landi
Mintz, Toby

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Human Action Scaffolds Visual Non-Adjacent Dependency Learning Beyond Surface Form

Neshat Darvishi (neshatda@usc.edu)[1], Landi Jiang (lcjiang@usc.edu)[1] &
Toben H. Mintz (tmintz@usc.edu)[1,2]
[1]Department of Psychology, SGM 501, 3620 S. McClintock Avenue, Los Angeles, CA 90089-1061, USA
[2]Department of Linguistics, GFS 301, 3601 Watt Way, Los Angeles, CA 90089-1693, USA

Abstract

Adult learners can acquire non-adjacent dependencies (NADs) in visual sequences, yet learning is far stronger for human actions than for dynamic objects (Lu & Mintz, 2023). We asked what makes human action stimuli special by separating two candidate contributors: appearance-based structure (a visible body with form and texture) versus biological motion cues (joint trajectories). In Experiment 1, participants viewed full-body avatar actions for 6 minutes and were tested on matched point-light displays (PLDs); they rated NAD-consistent triplets as more familiar than positional foils, showing that knowledge learned from rich action displays transfers to sparse motion-only tests. Experiments 2 and 3 then asked whether motion cues alone can support learning when used during encoding. With PLDs throughout, learning was not detectable after 6 minutes but emerged with extended exposure, suggesting that the human-action advantage stems from stronger support during encoding, not simply from perceiving the stimuli as actions.

Keywords: Non-adjacent dependencies; Biological motion; Point-light displays

Introduction

Humans are constantly immersed in structured streams of events. Sounds, sights, and actions are rarely random; instead they contain regularities that can be learned and used to form expectations. A long-standing idea in cognitive science is that people can learn patterns incidentally, without conscious awareness, through exposure alone (Cleeremans, Destrebecqz, & Boyer, 1998; Seger, 1994). This capacity is commonly referred to as statistical learning, and it is thought to provide a foundational mechanism for discovering structure in the environment across domains and across development.

A large body of evidence supports the idea that statistical learning operates broadly across sensory modalities and stimulus types, especially when regularities are expressed in adjacent relationships. Throughout development, infants and adults reliably track adjacent transitional probabilities in speech streams (Aslin, Saffran, & Newport, 1998; Romberg & Saffran, 2013), and similar sensitivity has been documented in non-linguistic auditory sequences such as music tones (Saffran, Johnson, Aslin, & Newport, 1999). Parallel findings in the visual domain show that learners extract adjacent regularities from sequences of shapes (Fiser & Aslin, 2001, 2002), and such structure also appears in more complex event streams, including human action sequences (Baldwin, Andersson, Saffran, & Meyer, 2008). Yet the structure of real-world events is not always expressed through adjacent transitions. Many meaningful regularities span intervening material, such that one event predicts another across temporal gaps. Learning these non-adjacent dependencies (NADs) is substantially more challenging, as it requires maintaining representations across intervening elements, selecting relevant positions, and binding them while resisting interference as the stream unfolds (Conway, 2020; Perruchet & Pacton, 2006), making NAD learning particularly sensitive to memory and processing constraints.

Consistent with these demands, research across domains shows that NAD learning often benefits from additional cues that enhance salience and structure. For example, inserting pauses that delineate triplets can improve learners' ability to track non-adjacent relationships (Peña, Bonatti, Nespor, & Mehler, 2002). Beyond segmentation, studies show that dependencies become easier to acquire when dependent elements share perceptual features, making them stand out against the intervening material (Creel, Newport, & Aslin, 2004; Gebhart, Newport, & Aslin, 2009; Onnis, Monaghan, Richmond, & Chater, 2005; Weyers & Mueller, 2022). A complementary finding shows that increasing variability in intervening elements can also promote NAD learning by discouraging memorization of whole sequences and encouraging learners to focus on stable relationships between the edge elements (Gómez, 2002).

Together, these findings suggest that NAD learning depends not only on exposure, but also on how input is organized and whether the stream provides scaffolding that supports segmentation, attention, and memory. The present work builds on this framework by examining how visual format shapes the efficiency of NAD learning in dynamic visual sequences.

Encoding Support Across Visual Formats

Visual experience is inherently sequential: objects transform, bodies move, and scenes unfold to create temporal regularities. This makes vision a compelling domain for studying non-adjacent dependency (NAD) learning. As in language and audition, NAD learning in vision often requires additional support, and outcomes can vary dramatically even when statistical structure and procedures are matched. That is, learning efficiency is shaped by the visual format itself.

A clear demonstration that stimulus format matters comes from work directly comparing NAD learning across different kinds of visual sequences while holding the dependency

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

structure and overall procedure constant. In a series of studies, Li and Mintz (2015) and Lu and Mintz (2023) used triplet-based NAD sequences across visual instantiations, including human action and object transformations. Across these comparisons, learners showed more reliable learning when the dependencies were embedded in sequences of human actions than when the same structure was embedded in sequences of object transformations, with object-based learning typically requiring longer exposure to reach comparable performance (Li & Mintz, 2015; Lu & Mintz, 2023). These findings align with broader evidence suggesting that temporally coherent human movement provides strong support for learning relations across distance (Endress & Wood, 2011). Lu and Mintz (2023) also reported that, even within human stimuli, NAD learning was stronger for dynamic actions than for static postures, consistent with the idea that continuity of motion can facilitate binding across intervening elements. Together, these results point to an advantage for human movement in visual sequences.

Lu and Mintz (2023) argue that the efficiency advantage for human action sequences may reflect how biological actions are encoded and organized during learning. One possibility is that biological actions support richer event representations than visually arbitrary transformations due to their interpretable structure and predictable internal relations. In this view, human movement may provide a natural scaffold for sequence processing, allowing observers to chunk the stream into coherent events and maintain continuity across intervening elements, thereby supporting the binding of non-adjacent relationships (Lu & Mintz, 2023). A complementary account is that observing human action recruits perceptual and predictive systems, including mechanisms linked to motor simulation, which can enhance encoding and temporal integration of sequential input (Reid, Kaduk, & Lunn, 2019; Salo, Ferrari, & Fox, 2019; Wilson & Knoblich, 2005). On both accounts, human actions carry structured information that may help learners preserve and connect distant elements.

This observation raises a central question for the present study: what properties of biological motion enable efficient NAD learning? One possibility is that the advantage comes from the structure of the movement itself—the coordinated motion of an articulated body, with lawful temporal dynamics and predictable relationships among joints as actions unfold. Biological motion is highly informative in the sense that even when most visual detail is removed, observers rapidly recover the presence of an agent from motion alone, suggesting that the visual system is strongly tuned to the kinematic organization of human movement (Cutting, 1981; Johansson, 1973; Troje, 2013; Troje & Westhoff, 2006). A second possibility, however, is that efficient learning depends on appearance-based cues available in full human avatars, such as body form, surface detail, and a stable visual identity, which may support continuity, memory, and sustained tracking across the sequence. Distinguishing between these possibilities matters because it clarifies whether the human action advantage in visual NAD learning—and visual sequence processing more broadly—reflects sensitivity to biological movement structure, depends on richer visual scaffolding, or is strongest when both sources of information are jointly available.

One way to separate these sources of information is to use point-light displays (PLDs)—a minimal biological motion format in which actions are represented only by a small set of moving points placed at major joints. In Johansson's (1973) classic demonstrations, these sparse displays were sufficient for observers to perceive a coherent human figure and recognize actions despite the absence of texture, contour, and detailed body form. Thus, PLDs remove much of the appearance-based information present in full avatars while preserving the joint trajectories and temporal coordination that define biological motion (Johansson, 1973; Troje, 2013; Troje & Westhoff, 2006). Consistent with this, neuroimaging studies show that point-light hand actions engage regions similar to fully visible actions, indicating that biologically meaningful structure remains accessible under strong visual reduction (Ziccarelli, Errante, & Fogassi, 2022, 2024). For the present question, PLDs therefore provide an informative testbed: if robust visual NAD learning depends primarily on the visual richness of the human body, learning should be reduced when actions are rendered as point-light motion; if it depends on movement structure, efficient learning should persist even when only motion dynamics are available. Comparing full avatars and PLDs thus isolates which properties of human movement support rapid NAD learning.

To test what makes human action sequences especially supportive for visual NAD learning, we ran three experiments that isolate the roles of visual richness and biological motion. Experiment 1 tests whether NAD knowledge learned from visually rich avatar actions involves representations that are tied to surface properties of the stimuli or transfer to a more abstract action format in point-light displays. Experiments 2–3 then test whether PLDs alone support learning at encoding. Because point-light displays remove surface-form information and may limit encoding—especially for simpler actions (e.g., arm raises, leg lifts)—we include a more complex set of whole-body movements (e.g., plié, jumping jacks) to test whether increased action complexity facilitates learning. Following (Lu & Mintz, 2023), Experiment 2 uses a 6-minute exposure, and Experiment 3 extends training to 9 minutes. If rich appearance-based scaffolding is critical, transfer should be limited and PLD-only learning should be weak; if kinematic structure is sufficient, learning should generalize to PLDs and emerge with increased exposure.

Experiment 1: Cross-Stimulus NAD Generalization

Building on evidence that visual NADs can be learned from human avatar actions within six minutes (Lu & Mintz, 2023), Experiment 1 tested whether learning acquired from visually rich human action sequences generalizes to a minimal biological-motion format. We adopted the same full-avatar

stimuli used by Lu and Mintz (2023) and matched their exposure duration during training. These avatar actions provide rich appearance-based information—including color, texture, and connected body contours—across the stream. We then created a corresponding set of PLDs depicting the same actions, which remove surface form and texture while preserving joint-based motion dynamics, from which we assembled our test stimuli. This transfer design asks what kind of representations result after full avatar training. It asks whether the resulting NAD is tied to the surface properties of the trained avatar clips (i.e., a stimulus-specific representation) or whether participants acquire a more abstract, representation of the kinematic organization of biological action. If the format of the representation is action dynamics, learning should transfer to PLDs; if they rich appearance-based cues, evidence of NAD learning should be absent with PLD testing.

Methods

Participants We recruited 110 undergraduate students from the Psychology Department subject pool at the University of Southern California. Detailed demographic information (e.g., age, gender, language background, cultural background) was not collected. Participants received course credit as compensation. Thirty-eight participants were excluded for failing the attention check (see Procedure), for a final sample of 72.

Materials and Procedure The visual stimulus set consisted of 15 short video clips, each depicting a simple biological action (e.g., raising an arm or lifting a leg) performed by a fully visible avatar figure borrowed from Lu and Mintz (2023) during training and by the point-light display (PLD) figure during testing. To create PLD versions of these same actions, we generated corresponding PLD clips in Autodesk Maya (Autodesk, 2023) using the same actor performance as the avatar stimuli. A virtual stick figure composed of 13 points was modeled to correspond to the actor's body: one point marked the head, and 12 points were positioned at major joints (shoulders, elbows, wrists, hips, knees, and ankles). The point-light markers were aligned to the actor's motion so that each dot precisely tracked the underlying joint trajectory over time. Each clip began in a neutral stance, progressed smoothly to the movement's maximum displacement and returned to the starting position, creating a clear perceptual boundary between successive actions. The neutral position and the target pose of each action clip—the midpoint of the action clip—for the full avatar and PLD are depicted in Figure 1 and Figure 2, respectively.

The experiment followed a two-phase design consisting of a training phase and a testing phase.

Training stimuli. During training, participants viewed a continuous stream of short video clips depicting dynamic action sequences. Each clip lasted 625 ms, and action triplets were formed by sequencing three clips. A 125 ms pause separated triplets during the training phase. Training sequences were structured around three non-adjacent dependencies: a_1Xb_1, a_2Xb_2, and a_3Xb_3, where the dependent

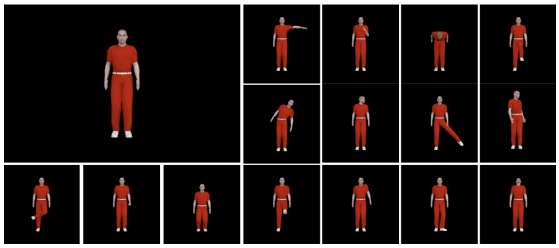

Figure 1: The large image shows the human avatar in its neutral posture, while the smaller images depict the maximum displacement frames for each of the 15 distinct movements.

elements occurred in the first and third positions of the triplet, separated by an intervening item. The stimulus set consisted of 15 distinct clips: three served as a-elements (a_1–a_3), three as b-elements (b_1–b_3), and nine as interveners (X_1–X_9). The intervening X-elements varied across repetitions of an NAD but were constrained by dependency: X_1–X_3 occurred in the $a_1\_b_1$ frame, X_4–X_6 in $a_2\_b_2$, and X_7–X_9 in $a_3\_b_3$. This restriction allowed us to create test stimuli with the necessary characteristics (see below). Thus, each dependency was observed across multiple interveners.

All participants viewed clips drawn from the same stimulus pool, but assignment of specific clips to the a, b, and X roles was randomized across individuals, yielding 45 unique stimulus sets. Participants were exposed to repeated presentations of the resulting 9 NAD triplets, with each triplet repeated 20 times, for a total of 180 triplets (6 minutes of exposure). Triplets were presented in a pseudo-randomized order, with no more than two identical triplets appearing consecutively.

Testing stimuli. The testing phase assessed whether participants learned the trained non-adjacent dependencies using a familiarity-rating task. The full test set contained 48 items: 18 *NAD-consistent triplets*, 18 *positional triplets*, and 12 *catch trials*. The two critical test conditions—NAD-consistent and positional—contrasted triplets that either preserved or violated the trained a_b dependency structure. NAD-consistent triplets preserved the trained non-adjacent pairings ($a_1\_b_1$, $a_2\_b_2$, $a_3\_b_3$) but introduced new intervening X-elements, producing triplets that were novel in surface composition (e.g.,

Figure 2: The large image shows the PLDs in neutral posture, while the smaller images depict the maximum displacement frames for each of the 15 distinct movements recorded from a human actor.

$a_1X_4b_1$, $a_2X_9b_2$). Positional triplets retained the same positional frame (an a-element in the first position and a b-element in the third) but recombined elements into a_b pairings that had never co-occurred during training (e.g., $a_1X_9b_2$). Importantly, in both test types adjacent transitional probabilities between successive items were zero, meaning that all triplets were equally novel in their specific action-to-action transitions. This was possible because X items were restricted to a single NAD during training. The 12 catch trials consisted of repeated-item sequences that violated the training format (e.g., $a_1a_1b_2$ or $a_5b_4b_4$) and served as an attention check. Participants who rated more than 4 of the 12 catch trials with a familiarity score of 3 or higher were excluded from the final analyses.

Each triplet was presented twice consecutively, separated by a 125 ms pause, after which participants rated its familiarity on a 5-point Likert scale (1 = definitely had not seen to 5 = definitely had seen). Test items were shown in pseudo-random order, with no more than two items of the same type appearing consecutively. Both test triplet types were novel, but they used familiar actions in the same positions as during training. Therefore, we expected participants to rate all test triplets as somewhat familiar overall. However, if participants learned the non-adjacent dependencies, they should give higher familiarity ratings to the NAD-consistent triplets because only these items preserve the same a_b pairings encountered during exposure, whereas the positional foils recombine the edge elements into untrained pairings.

Testing environment. The experiments were programmed in PsychoPy (Peirce et al., 2019) and deployed online via Pavlovia (Open Science Tools, 2020). Recruitment and consent were managed through Qualtrics (Qualtrics, 2022). Participants completed the experiment on their own laptops or desktop computers; mobile devices were not permitted. Each session lasted about 30 minutes.

Results and Discussion

Participants who did not meet the attention-check criterion were excluded, leaving a final sample of 72 participants. We fitted an ordinal logistic regression model using the ordinal package in R (Christensen, 2019) to investigate the relationship between participants' familiarity ratings and test triplet type. The model included fixed effects of triplet type (NAD-consistent vs. positional) and test trial number, along with participant-level random effects, with random slopes included when the model fit reliably. To complement the frequentist analysis, we also fit Bayesian ordinal regression models in brms (v2.20.4) (Bürkner, 2018), using bayestestR (v0.13.2) (Makowski, Ben-Shachar, & Lüdecke, 2019) to summarize posterior evidence. For each effect, we evaluated whether the 89% highest density interval (HDI) excluded 0, and we report the proportion of the posterior falling within the region of practical equivalence (ROPE: [-0.18, 0.18]; Kruschke, 2015).

Participants' average familiarity ratings were 3.32 ± 1.24 for NAD-consistent triplets and 3.15 ± 1.28 for positional triplets. Ratings for NAD-consistent items were reliably higher than ratings for positional foils ($\beta = 0.15$, $z = 2.14$, $p = .02$). Bayesian analysis converged and the posterior for stimulus type was positive (89% HDI [0.06, 0.58]), suggesting reliable sensitivity to non-adjacent triplets (left column of Figure 3).

These results suggest that learning acquired from full-avatar action sequences can generalize to a more visually reduced biological-motion format at test. After 6 minutes of exposure to richly detailed avatar actions, participants still showed sensitivity to the trained non-adjacent structure when the same actions were rendered as point-light displays. This pattern is consistent with the idea that the human-action advantage in visual NAD learning reflects sensitivity to action structure that remains available under substantial visual reduction. Experiment 2 builds on this finding by testing whether NAD learning persists when training and test are both carried out using only point-light motion cues.

Experiment 2: Action-Based PLD

Experiment 1 showed that NAD knowledge acquired from richly detailed human-avatar actions can generalize to a point-light version of the same movements at test. Experiment 2 extends this logic by asking whether appearance-based scaffolding is necessary during encoding, or whether learners can acquire visual NADs when surface form and texture are removed and only the spatiotemporal structure of human motion is available. Prior work has demonstrated robust NAD learning from fully visible human action sequences (Endress & Wood, 2011; Li & Mintz, 2015; Lu & Mintz, 2023), but those displays combine action dynamics with a stable body form and rich surface detail. Here, we trained and tested participants using point-light displays, which eliminate body shape and texture while preserving coordinated joint trajectories over time, allowing us to test whether 6 minutes of exposure to visually reduced yet dynamically intact human action supports learning of non-adjacent dependencies.

Methods

Participants We recruited 120 undergraduate participants from the same subject pool; none had participated in Experiment 1. Applying the same exclusion criterion used in Experiment 1, 60 participants did not meet the attention check criterion, resulting in a final sample of 60. Participants received course credit as compensation.

Materials and Procedure Stimuli consisted of 15 short point-light display video clips depicting a new set of more complex whole-body actions (e.g., plié, tuck jump). PLDs were created in Autodesk Maya (Autodesk, 2023) from reference videos of a live human actor performing the same discrete movements. A virtual point-light figure was constructed with 13 points (one at the head and 12 at major joints). Each marker tracked the actor's joint trajectory over time, producing a sparse animation that preserved biological kinematics. Each clip began and ended in a neutral posture, making transitions between actions easier to parse.

Experiment 2 used the same two-phase design and task structure as Experiment 1. The only differences were the stimulus set and visual format in which participants were trained and tested on were PLDs depicting a new set of more complex whole-body actions, rather than the simpler avatar actions used previously. At the end of testing, participants were also asked whether, at any point during the study, they recognized the PLDs as depicting human motion.

Results and Discussion

As in Experiment 1, we fitted an ordinal logistic regression model to evaluate whether familiarity ratings distinguished NAD-consistent from positional triplets, while accounting for test trial number. Participants' average familiarity ratings were similar for NAD-consistent triplets (3.19 ± 1.23) and positional triplets (3.13 ± 1.24). The middle column of Figure 3 depicts the distribution of participants' mean rating differences (NAD versus positional). The model indicated no evidence of higher ratings for NAD-consistent compared to positional triplets ($\beta = -0.05$, $p = .18$). There was a small effect of test trial number ($\beta = -0.007$, $p = .05$), reflecting a modest downward drift in ratings across the test phase. Bayesian analysis confirmed the null result of test trial type, with 89.66% of the posterior estimates falling within the ROPE [−0.18,0.18] (89% HDI [−0.03,0.22]), indicating an absence of NAD learning.

Unlike Experiment 1, where training with full-avatar actions supported transfer to PLD testing within 6 minutes of exposure, training and testing entirely in PLDs did not yield detectable sensitivity to the non-adjacent structure after the same exposure duration. Importantly, all participants who passed the attention check reported recognizing the stimuli as depictions of human motion, consistent with evidence that PLDs engage action-related neural systems similar to fully visible movements (Ziccarelli et al., 2022). This pattern suggests that, although PLDs preserve biological kinematics, the reduced visual format may make it harder to form or stabilize the representations needed to bind distant elements over a short exposure window. In other words, minimal biological motion cues may be sufficient for perceiving action, but not necessarily sufficient for rapid extraction of NADs under the current training duration. Motivated by the idea that additional exposure can support consolidation of weakly encoded structure (Lu & Mintz, 2023), Experiment 3 increased training time to test whether longer input is enough to reveal NAD learning in the same point-light format.

Experiment 3: Action-Based PLD with Extended Exposure Time

Experiment 3 followed up on the null effect in Experiment 2 by asking whether additional exposure changes what learners can extract from point-light action sequences. Motivated by evidence that longer training can reveal NAD sensitivity in more challenging visual formats (Lu & Mintz, 2023), we tested whether extending PLD exposure during encoding yields detectable NAD learning.

Methods

Participants Eighty-three new undergraduate participants who had not taken part in Experiments 1 and 2 were recruited from the same subject pool. Fifty-two participants remained in the final sample after passing the attention check criterion. Participants received course credit as compensation.

Materials and Procedure The procedure and materials were identical to Experiment 2 in terms of visual format and overall task structure. The key change was that the training stream was extended from 180 to 270 triplets, corresponding to 10 additional repetitions of each of the 9 NAD triplets (i.e., 30 repetitions per triplet). To help maintain attention during the longer exposure period, participants viewed a 1-minute drawing-break video after the first 6 minutes of training. The test phase was identical to the previous experiments.

Results and Discussion

As in Experiments 1–2, we fitted an ordinal logistic regression model. Participants' average rating was 3.33 ± 1.24 for NAD-consistent triplets and 3.13 ± 1.29 for positional triplets, showing a clearer separation than in Experiment 2. The model revealed a reliable effect of triplet type ($\beta = -0.15$, SE = 0.07, $z = -2.27$, $p = .02$) and a modest decrease in ratings over test trials ($\beta \approx -0.15$, SE ≈ 0.004, $z \approx -3.25$, $p < .001$). The Bayesian ordinal model converged on the same conclusion, the posterior for triplet type was credibly positive (89% HDI [0.08, 0.54]), consistent with higher familiarity ratings for NAD-consistent than positional triplets.

These results provide evidence that adults can learn NADs of visual action sequences from point-light displays, but that longer exposure is required compared to learning NADs with richer stimuli (i.e., animated human avatars). As in Experiment 2, all participants whose data were analyzed reported recognizing the stimuli as depicting human motion. After 9 minutes of PLD training, participants distinguished NAD-consistent from positional triplets, indicating that sparse bi-

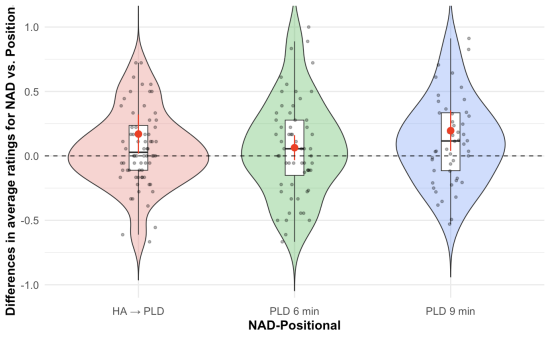

Figure 3: The distribution of participants' mean rating differences (NAD triplets – positional triplets) across Experiments 1–3. Black dots indicate individual participants.

ological motion cues can support extraction of non-adjacent structure once enough evidence accumulates. The effect, although modest, matched the magnitude reported for extended object transformation conditions in Lu and Mintz (2023) and remained smaller than that observed with full human avatar, suggesting that removing surface-form cues increases the amount of exposure needed for learning. Nevertheless, the presence of learning here—contrasted with the null result after 6 minutes in Experiment 2—supports the idea that kinematic information in biological motion can serve as a sufficient basis for NAD learning, but that it may require longer training to stabilize the dependency representation in the absence of richer visual scaffolding.

General Discussion

Across three experiments, we asked which properties of human movement support efficient visual NAD learning by manipulating available information at encoding. The results show a consistent pattern: learners can acquire and use NAD knowledge when encoding is sufficiently supported, either through richer visual structure (full-body human actions) or through longer exposure when the input is visually reduced (point-light displays). Specifically, NAD knowledge acquired from brief exposure to fully visible avatar actions generalized to sparse point-light depictions at test (Experiment 1), whereas learning was not detectable after the same brief exposure when both training and test were conducted entirely in the point-light format (Experiment 2). Extending exposure in the point-light format yielded reliable sensitivity to the NAD triplets (Experiment 3). These findings align with prior work showing that NAD learning depends on supportive conditions that reduce interference and facilitate maintaining relationships across time (Conway, 2020; Gómez, 2002; Peña et al., 2002; Perruchet & Pacton, 2006). Together, these results isolate two sources of encoding support — visual richness and exposure duration — and clarify the conditions under which the human-action advantage emerges.

Experiment 1 is informative because it moves beyond demonstrating strong learning from human actions—a result already established in visual NAD learning (Endress & Wood, 2011; Li & Mintz, 2015; Lu & Mintz, 2023)—and instead tests what kind of information participants carry forward. Transfer from avatar training to point-light testing suggests that participants did not rely exclusively on surface-form cues at retrieval (e.g., texture, body shape, stable appearance), but could still use the learned dependency when the same actions were rendered as sparse joint trajectories. This pattern is consistent with the idea that human action sequences provide supportive structure for encoding through familiar, coherent event dynamics and predictable motion organization, making it easier to maintain and link elements across intervening items (Endress & Wood, 2011; Lu & Mintz, 2023). It is also compatible with accounts in which action perception recruits predictive and motor-resonance mechanisms that can enrich sequential encoding of observed movement (Reid et al., 2019; Salo et al., 2019; Urgesi, Candidi, Imaizumi, & Aglioti, 2006; Wilson & Knoblich, 2005). However, transfer in Experiment 1 does not by itself show that biological kinematics are sufficient at encoding. Experiments 2 and 3 therefore used point-light displays as a stricter test, because they preserve the spatiotemporal organization of human movement while removing the additional perceptual scaffolding provided by a fully visible body. Classic work shows that observers can nonetheless recover animacy and action from these sparse joint trajectories (Cutting, 1981; Johansson, 1973; Troje, 2013; Troje & Westhoff, 2006). Yet perceiving biological motion is not the same as learning higher-order temporal relations across distance. In Experiment 2, six minutes of PLD-only exposure produced no detectable advantage for NAD-consistent triplets over positional triplets, whereas extending exposure in Experiment 3 yielded reliable sensitivity. This exposure-dependent shift fits an encoding-strength account such that, when encoding is both visually reduced and brief, learners may not have sufficient support to form a stable representation of the non-adjacent structure. Extending exposure can compensate for this visual information reduction by providing additional repetitions for the dependency to stabilize. This tradeoff mirrors Lu and Mintz's (2023) broader finding that stimulus formats providing weaker support at encoding (e.g., object transformations) tend to require longer training to yield reliable NAD sensitivity compared to full-body human-action sequences. Effects were modest but consistent across experiments and expected given the similarity of test item types and the demands of maintaining non-adjacent structure across intervening elements (Lu & Mintz, 2023).

The present findings extend beyond the laboratory. Many naturalistic visual sequences contain meaningful long-distance relations, such as the predictable link between initiation and completion of a goal-directed action or recurring motifs embedded within dance phrases. Research shows that novice observers implicitly acquire sequential structure from unfamiliar human movement through passive exposure (Opacic, Stevens, & McKenzie, 2009), and that kinematic cues alone support event segmentation even without conceptual goal knowledge (Monroy, Gerson, & Hunnius, 2023), suggesting that the kinematic structure inherent in biological motion may be a key substrate for extracting long-distance regularities in naturalistic action perception. Yet these findings also raise an open question: which component of biological motion provides the critical support for abstraction? One possibility is that learners primarily benefit from temporally coherent kinematic information as it unfolds over time; another is that learning depends on preserving a stable, body-like spatial organization of moving parts; and a third is that learning is strongest when these cues jointly engage perceptual-motor mechanisms for action perception and prediction. Future work can separate these candidates by manipulating whether displays preserve temporal dynamics while disrupting spatial coherence, and by examining whether behavioral learning tracks activity in neural systems associated with biological-motion processing under these reduced conditions.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

Autodesk. (2023). *Autodesk maya (version 2023) [computer software]*. Computer software.

Baldwin, D. A., Andersson, A., Saffran, J. R., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382–1407.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. doi: 10.32614/RJ-2018-017

Christensen, R. H. B. (2019). ordinal—regression models for ordinal data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ordinal` (R package version 2019.12-10)

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, *2*(10), 406–416.

Conway, C. M. (2020). How does the brain learn environmental structure? ten core principles for understanding the neurocognitive mechanisms of statistical learning. *Neuroscience & Biobehavioral Reviews*, *112*, 279–299.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1119–1130.

Cutting, J. E. (1981). Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(1), 71–87. doi: 10.1037/0096-1523.7.1.71

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology*, *63*(3), 141–171. doi: 10.1016/j.cogpsych.2011.07.001

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458–467.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review*, *16*(3), 486–490.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436. doi: 10.1111/1467-9280.00476

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211. doi: 10.3758/BF03212378

Kruschke, J. K. (2015). *Doing bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

Li, J., & Mintz, T. H. (2015). Constraints on learning nonadjacent dependencies (nads) of visual stimuli. In *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 1304–1309). Cognitive Science Society. Retrieved from `https://escholarship.org/uc/item/275k23m`

Lu, H. S., & Mintz, T. H. (2023). Dynamic motion and human agents facilitate visual nonadjacent dependency learning. *Cognitive Science*, *47*(4), e13344. doi: 10.1111/cogs.13344

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence, and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. doi: 10.21105/joss.01541

Monroy, M., Gerson, S., & Hunnius, S. (2023). Finding structure in modern dance. *Cognitive Science*, *47*(11), e13375. doi: 10.1111/cogs.13375

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*(2), 225–237.

Opacic, T., Stevens, C., & McKenzie, S. (2009). Unspoken knowledge: Implicit learning of structured human dance movement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1570–1577. doi: 10.1037/a0017244

Open Science Tools. (2020). *Pavlovia (version 2020) [computer software]*. Computer software.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. doi: 10.3758/s13428-018-01193-y

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604–607.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238.

Qualtrics. (2022). *Qualtrics (version xm) [computer software]*. Computer software. Provo, UT.

Reid, V. M., Kaduk, K., & Lunn, J. (2019). Links between action perception and action production in 10-week-old infants. *Neuropsychologia*, *126*, 69–74. doi: 10.1016/j.neuropsychologia.2018.01.009

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science*, *37*(7), 1290–1320.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Salo, V. C., Ferrari, P. F., & Fox, N. A. (2019). The role of the motor system in action understanding and communication: Evidence from human infants and non-human primates. *Developmental Psychobiology*, *61*(3), 390–401. doi: 10.1002/dev.21835

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*,

115(2), 163–196.

Troje, N. F. (2013). What is biological motion? Definition, stimuli, and paradigms. In H. Hecht, R. S. Johnson, & M. Shiffrar (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention* (pp. 13–36). MIT Press.

Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a "life detector"? *Current Biology*, *16*(8), 821–824. doi: 10.1016/j.cub.2006.03.022

Urgesi, C., Candidi, M., Imaizumi, S., & Aglioti, S. M. (2006). Representation of body identity and body actions in extrastriate body area and ventral premotor cortex. *Nature Neuroscience*, *9*(8), 1102–1109. doi: 10.1038/nn1759

Weyers, I., & Mueller, J. L. (2022). A special role of syllables, but not vowels or consonants, for nonadjacent dependency learning. *Journal of Cognitive Neuroscience*, *34*(8), 1467–1487.

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, *131*(3), 460–473. doi: 10.1037/0033-2909.131.3.460

Ziccarelli, S., Errante, A., & Fogassi, L. (2022). Decoding point-light displays and fully visible hand grasping actions within the action observation network. *Human Brain Mapping*, *43*(14), 4293–4309. doi: 10.1002/hbm.25914

Ziccarelli, S., Errante, A., & Fogassi, L. (2024). Direction and velocity kinematic features of point-light display grasping actions are differentially coded within the action observation network. *NeuroImage*, *303*, 120939. doi: 10.1016/j.neuroimage.2024.120939