SCHA-REACT: STAGED CAUSAL HISTORY-ASSISTED REFINEMENT FOR ONLINE REACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online reaction generation requires a policy to produce causal responses from bounded recent interaction context. Yet temporal proximity is not equivalent to interaction relevance: completed interactions outside the current window may remain informative, while directly enlarging the inference context introduces redundant history and additional runtime dependencies. We introduce SCHA-React, a staged causal-history-assisted refinement framework that augments a bounded reaction policy with selected historical evidence during training. Specifically, SCHA-React selects causally completed historical events, compresses them into compact event representations, and adaptively fuses them with recent interaction features during refinement. The shared policy uses conditional Flow Matching for next-latent prediction and an Online Motion Decoder for executable motion generation. We instantiate SCHA-React on DuoBox with 2,525 multimodal-model-assisted, manually verified annotations, including 1,310 admissible guidance events. Across the Reactive, Two-character, and 1,800-frame closed-loop evaluation settings, SCHA-React achieves the lowest reported per-frame, per-transition, and per-clip FIDs among evaluated methods. At inference, the historical pathway is disabled; the policy uses bounded recent motion and maintained latent buffers. Overall, the results support causal historical evidence as a training-time refinement resource that improves the bounded policy without adding inference-time inputs.

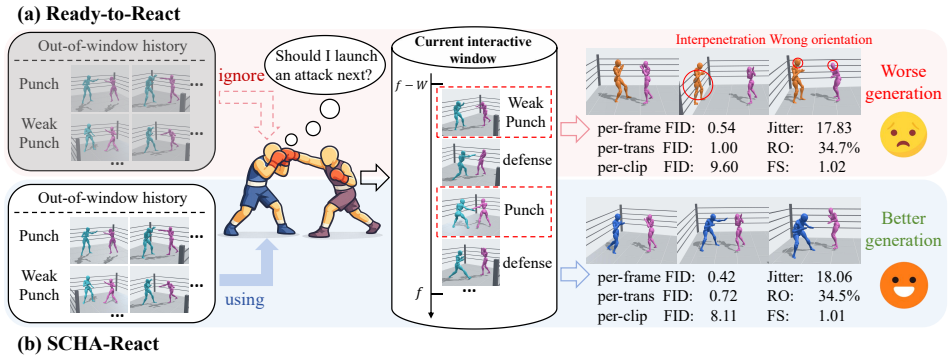


Figure 1: **Motivation.** (a) A bounded recent-context policy may exclude earlier completed interactions relevant to subsequent responses. (b) SCHA-React uses selected causally completed interactions as historical evidence during policy refinement. The resulting bounded policy benefits from longer-range interaction experience while retaining the bounded recent-context inference interface.

1 INTRODUCTION

Online reaction generation continuously synthesizes a character's response to an interacting counterpart as the interaction unfolds. Because each response must be generated causally from currently available observations, an online policy cannot access future counterpart motion and typically relies on bounded recent context for repeated prediction (Cen et al., 2025). This bounded interface is practical but restrictive: once an interaction leaves the context window, it can no longer directly

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

influence the policy input even if it remains relevant to a later response. In fast-changing interactions such as boxing, earlier attacks, defenses, or counters may still provide useful evidence after leaving the local window. This raises a central question: can response-relevant causal history improve an online reaction policy during training while inference remains restricted to bounded recent context?

Existing temporal mechanisms mainly control how much past information remains available through recent windows, autoregressive histories, recurrent states, or closed-loop updates (Chopin et al., 2023; Cen et al., 2025; Ye et al., 2026; Geng et al., 2026). Such mechanisms address temporal availability, but availability is not equivalent to interaction relevance: recency largely determines what a bounded policy retains, although an earlier completed interaction may remain more informative than routine recent motion. We refer to this discrepancy as the *recency–relevance mismatch*. Rather than enlarging the inference window, we use selected, causally completed interaction events as training-time historical evidence. Motivated by sports-anticipation studies associating effective responses with both unfolding kinematics and context accumulated through prior exposure (Mann et al., 2014; Thomas et al., 2022; Gredin et al., 2023), this design separates information used for policy refinement from information required online: a general recent-context policy is learned first and then refined with selected causal history while preserving its bounded inference interface.

Accordingly, we introduce SCHA-React, a Staged Causal History-Assisted Refinement framework for online reaction generation. Causal Key-Event Selection identifies completed historical interactions satisfying the current causal boundary, Historical Event Representation compresses each selected interaction into a compact event token, and Event-Guided Context Fusion adaptively combines these tokens with recent interaction features during refinement. The shared policy uses conditional Flow Matching for next-latent prediction and an Online Motion Decoder for executable motion generation. The historical pathway is active only during refinement; at inference, event selection, historical representation, and event-guided fusion are bypassed, while the refined policy operates on bounded recent motion context and role-specific latent buffers.

Our contributions are threefold. (1) We formulate the *recency–relevance mismatch* in bounded online reaction generation and study whether causal historical evidence can improve policy learning without expanding the inference context. (2) We introduce SCHA-React, which selects causally completed interactions, represents them as compact event-level evidence, and adaptively integrates them during staged policy refinement while preserving the bounded inference interface. (3) We instantiate SCHA-React on DuoBox with 2,525 multimodal-model-assisted, manually verified annotations, including 1,310 admissible guidance events, and evaluate it across Reactive, Two-character, and 1,800-frame closed-loop generation with targeted ablations of event selection, representation, fusion, and historical content.

2 Related Work

2.1 Human–Human Interaction Generation

Human–human interaction generation models coordinated motion between two or more participants under text, role, contact, or interaction constraints, with work on joint control, variable participants, discrete generation, and model composition (Tanaka & Fujiwara, 2023; Liang et al., 2024; Xu et al., 2024a; Wang et al., 2025; Tanke et al., 2025; Maluleke et al., 2025; Chen et al., 2024a; Ruiz-Ponce et al., 2024; Wang et al., 2024; Fan et al., 2024; Javed et al., 2025; Wu et al., 2026; Ruiz-Ponce et al., 2025). Unlike these joint or globally conditioned settings, reaction generation predicts one participant from another's observed motion under asymmetric, causal conditioning.

2.2 Reactive Motion Generation

Reactive motion generation predicts a response conditioned on an observed participant using motion, semantic or intention-aware conditions, and action–reaction objectives (Chopin et al., 2023; Ghosh et al., 2024; Xu et al., 2024b; Tan et al., 2025; Lin et al., 2026; Gupta et al., 2026; Liu et al., 2025); recent work also targets causal or low-latency use (Cen et al., 2025; Ye et al., 2026; Geng et al., 2026; Ji et al., 2025). Ready-to-React is closest to our setting, using DuoBox, bounded causal Agent–Opponent context, a shared policy, a VQ representation, and online decoding. SCHA-React preserves this interface but uses selected, causally completed interactions as training-time evidence for policy refinement.

2.3 Temporal Modeling in Online Motion Generation

Online and long-horizon motion generators manage past information through bounded windows, generated histories, recurrent states, and blockwise or autoregressive updates (Cen et al., 2025; Ye et al., 2026; Geng et al., 2026; Tanke et al., 2025; Maluleke et al., 2025). Related single-character work addresses long-sequence generation or sampling efficiency (Barquero et al., 2024; Dai et al., 2024; Zhang et al., 2024). These approaches modify how context is retained or processed; we instead ask which completed interactions outside the recent window help refine a bounded reaction policy.

Motion representation and historical-context selection are separate axes. Latent representations provide compact generation targets (van den Oord et al., 2017; Zhang et al., 2023; Li et al., 2024; Cen et al., 2025), while diffusion and flow formulations provide conditional generative objectives for motion synthesis and autoregressive prediction (Lipman et al., 2023; Liu et al., 2023; Hu et al., 2023; Canales Cuba et al., 2025; Jiang et al., 2025; Geng et al., 2026). SCHA-React uses conditional Flow Matching, while its contribution is selected, causally completed events as training-time evidence rather than an explicit opponent-tendency variable.

3 Method

SCHA-React is trained in three stages to learn a bounded online boxing reaction policy. Stage I pre-trains a VQ-VAE motion representation. Stage II learns a complete recent-context reaction policy for next-latent prediction and executable motion decoding. Stage III initializes from this policy and refines its shared modules using selected, causally completed Opponent events. At inference, the refined shared policy operates on bounded recent interaction context and policy-maintained latent buffers, while the historical-event pathway is bypassed.

3.1 Problem Formulation

We follow the causal Agent–Opponent interface of Ready-to-React (Cen et al., 2025). Let A and O denote the Agent and Opponent, respectively. At an invocation beginning at frame f, let $\mathcal{B}_f^A = A_{[f-W_f,f)}$ and $\mathcal{B}_f^O = O_{[f-W_f,f)}$ denote the available bounded motion buffers, where $W_f \le W$. Let $\mathcal{Z}_f^A = [\boldsymbol{z}_{f,1}^A, \dots, \boldsymbol{z}_{f,T_c}^A]$, where $T = W/U$ is the number of quantized latent positions in a complete training window and $T_c \le T$ is the current number of continuous endpoints retained online, denote the bounded latent buffer of the currently generated Agent. The policy consumes these motion buffers and this recurrent latent state, and returns both the next U-frame Agent block and its continuous Flow-Matching endpoint:

$$\left(\widetilde{A}_{[f,f+U)}, \widehat{\boldsymbol{z}}_{\text{next}}^A\right) = \mathcal{P}_\theta\left(\mathcal{B}_f^A, \mathcal{B}_f^O, \mathcal{Z}_f^A\right). \tag{1}$$

Here, f denotes the first generated frame, W the bounded context length, U the generated block length, and $\tau = f - W_f$ the current-window boundary. The two motion buffers are externally observable bounded interaction context, whereas $\mathcal{Z}_f^A$ is bounded recurrent state maintained across policy invocations. After generation, $\widehat{\boldsymbol{z}}_{\text{next}}^A$ is appended directly to form $\mathcal{Z}_{f+U}^A$, with older endpoints discarded once the buffer is full; generated motion is not re-encoded to recover this state. During training, a complete W-frame window contains T quantized ground-truth embeddings and yields $T-1$ adjacent latent targets; the next Agent block and latent are targets only. During rollout, the buffer instead contains the policy's own continuous Flow-Matching endpoints. All inputs to Equation 1 are available before f, so the policy remains causal. Historical events provide additional conditioning only during Stage III refinement and are not runtime arguments of the official inference policy. In *Reactive* generation, O is causally observed and only an Agent latent buffer $\mathcal{Z}_f^A$ is maintained. In *Two-character* generation, the parameter-shared policy is invoked with exchanged roles, with a separate latent buffer for each generated role. The same boundary τ determines whether a historical event is causally eligible for Stage III refinement.

3.2 General Reaction Policy

Figure 2 summarizes the overall framework. The recent-interaction pathway constitutes the complete inference backbone: it combines bounded Agent–Opponent motion context with the current role's

latent buffer, predicts the next continuous latent, and decodes the corresponding motion block. Stage III augments this shared policy with the historical pathway described in Section 3.3.

Recent Interaction Representation. During training, the frozen VQ tokenizer maps each ground-truth U-frame Agent block to a quantized embedding $\boldsymbol{z}_j^A$, while the reaction policy predicts continuous latent vectors rather than code indices. A W-frame training window contains $T = W/U$ quantized latent positions and yields $T-1$ adjacent supervision pairs $(\boldsymbol{z}_j^A, \boldsymbol{z}_{j+1}^A)$ for $j = 1, \ldots, T-1$. This training-window count is distinct from the current online-buffer occupancy $T_c \leq T$. During rollout, each generated Flow-Matching endpoint is appended directly to $\mathcal{Z}_f^A$ without re-encoding the generated motion.

At latent position j, $\boldsymbol{o}_j$ denotes the first Opponent pose in the aligned U-frame block, expressed in the contemporaneous Agent root frame. Linear projections of $\boldsymbol{z}_j^A$ and $\boldsymbol{o}_j$ form the recent interaction token:

$$\boldsymbol{x}_j = E_A(\boldsymbol{z}_j^A) + E_O(\boldsymbol{o}_j), \qquad \boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T-1}], \qquad \boldsymbol{c}_j = \mathrm{CausalEnc}_\phi(\boldsymbol{x}_{\leq j}). \tag{2}$$

The causal state $\boldsymbol{c}_j$ therefore depends only on tokens available before the target $\boldsymbol{z}_{j+1}^A$. Here, j, f, and t index latent blocks, motion frames, and continuous Flow time, respectively. This ordering allows the same policy to consume quantized ground-truth latents during training and its own continuous Flow endpoints during rollout. The frozen VQ decoder is distinct from the learned Online Motion Decoder. Appendix A.1 specifies the tokenizer, while Appendix A.2 provides the alignment, causal masks, and architecture details.

Flow-Matching Next-Latent Predictor. Given a causal conditioning state $\boldsymbol{c}_j^\star$, we train the shared latent predictor along a linear conditional Flow-Matching path from Gaussian source noise to the next Agent latent:

$$\boldsymbol{z}(t) = (1-t)\boldsymbol{\xi} + t\boldsymbol{z}_{j+1}^A, \qquad \boldsymbol{u}_t = \boldsymbol{z}_{j+1}^A - \boldsymbol{\xi}, \qquad \widehat{\boldsymbol{u}}_t = \boldsymbol{v}_\theta(\boldsymbol{z}(t), t, \boldsymbol{c}_j^\star). \tag{3}$$

Here, $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The stage-dependent definition of $\boldsymbol{c}_j^\star$ is given in Equation 7. During Stage II and inference, the predictor is conditioned on the recent-context state $\boldsymbol{c}_j$.

The ground-truth target $\boldsymbol{z}_{j+1}^A$ is used only to construct the training path and target velocity. At inference, integration begins from sampled $\boldsymbol{\xi}$ and never observes the target latent. Integrating the learned velocity field from $t=0$ to $t=1$ yields $\widehat{\boldsymbol{z}}_{j+1}^A$, which is passed to the Online Motion Decoder and appended to the bounded latent buffer for the next policy invocation. Appendix C.2.1 specifies the solver settings.

Online Motion Decoder. The Online Motion Decoder combines the preceding Agent motion block, the relative-root block computed from the current Agent–Opponent states, and the previous and predicted latent endpoints to reconstruct the next U-frame motion and auxiliary relative-root blocks:

$$(\widetilde{A}_{[f,f+U)}, \widetilde{R}^{A|O}_{[f,f+U)}) = \mathrm{OnlineDec}_\psi\Big(A_{[f-U,f)}, R^{A|O}_{[f-U,f)}, \boldsymbol{z}^A_{\mathrm{prev}}, \boldsymbol{z}^A_{\mathrm{next}}\Big). \tag{4}$$

During training, $\boldsymbol{z}^A_{\mathrm{prev}}$ is the ground-truth previous latent and $\boldsymbol{z}^A_{\mathrm{next}} = \overline{\boldsymbol{z}}^A_{j+1}$ is the stop-gradient analytic endpoint estimate. During rollout, $\boldsymbol{z}^A_{\mathrm{prev}}$ is the latest entry of $\mathcal{Z}_f^A$ and $\boldsymbol{z}^A_{\mathrm{next}} = \widehat{\boldsymbol{z}}^A_{\mathrm{next}}$ is the newly generated continuous endpoint. The Online Motion Decoder is distinct from the frozen VQ decoder and does not interpret these vectors as discrete code indices. Because the training endpoint is detached, motion- and root-reconstruction losses update only the Online Motion Decoder, whereas the Flow-Matching loss trains the conditional latent generator and its conditioning modules. Appendix A.3 provides the analytic endpoint estimate and complete decoder interface.

3.3 Causal History-Assisted Refinement

Stage III augments the recent-context policy with selected historical evidence. The refinement pathway first identifies causally eligible completed interactions, compresses each selected event into a compact representation, and then fuses the resulting event tokens with the recent interaction sequence. This changes the policy conditioning during refinement without changing the next-latent target or the Online Motion Decoder interface.

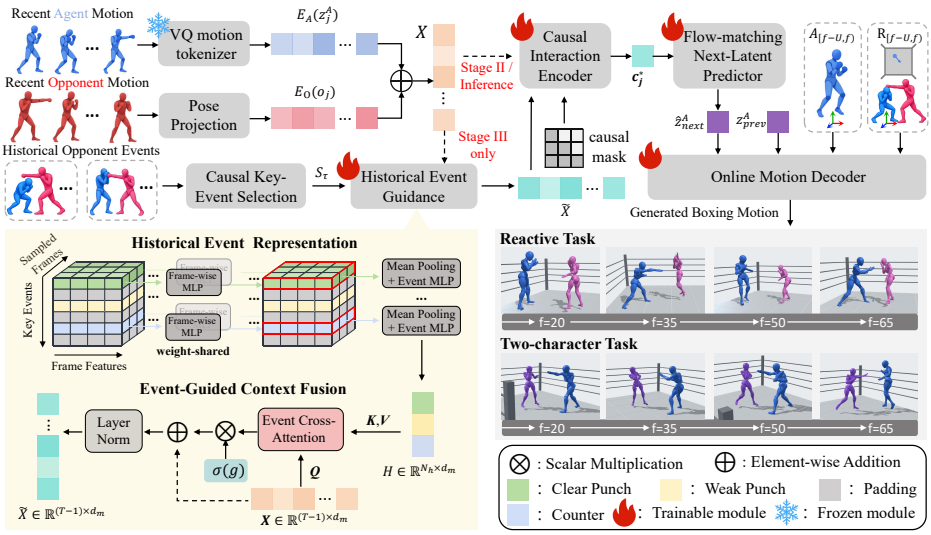


Figure 2: **SCHA-React framework.** General Reaction Learning first establishes a bounded recent-context reaction policy. Causal History-Assisted Refinement then encodes selected prior interactions and fuses the resulting event representations with recent interaction features to refine the shared policy. Online generation uses the refined policy with only bounded recent motion context and role-specific latent buffers; historical-event inputs are bypassed.

Offline Historical Event Annotation. Existing DuoBox sequences are rendered offline and processed with a structured multimodal annotation pipeline, after which all annotations are manually verified or corrected. Each event record stores its temporal interval, initiating role, event category, and key score. The resulting archive contains 2,525 atomic events from 117 annotated interaction sequences, including 1,310 events in the admissible clear-punch, weak-punch, and counter categories. These statistics describe the complete annotation archive rather than the events selected for an individual training window. The annotation model is used only to construct the offline archive and is not invoked during policy training or inference. Appendix B describes the annotation pipeline, prompts, review rules, and coverage statistics.

Causal Key-Event Selection. For a current window beginning at τ, the candidate pool contains only same-sequence events that are initiated by the current Opponent, belong to an admissible category, and end no later than τ. These constraints exclude both overlap with the current window and access to future events. A score–recency rule ranks the remaining causal events and retains at most N_h:

$$\mathcal{S}_\tau = \mathrm{TopK}_{N_h}(\mathcal{E}_\tau; \text{score–recency}). \tag{5}$$

The selected set $\mathcal{S}_\tau$ provides historical evidence during Stage III. Selection is precomputed from the training archive; causal filtering is applied before ranking so that neither score nor recency can admit an ineligible event. The retained events are restored to chronological order before representation. Appendix B specifies the ranking weights, tie handling, and complete annotation rubric.

Historical Event Representation. For each selected event $e_i \in \mathcal{S}_\tau$, we sample at most N_f frames. Each sampled frame contains the Opponent pose in the Agent root frame and the relative-root state $R^{A|O}$, which records the planar position and facing of the Agent in the Opponent root frame together with its distance to the fixed ring reference center. A shared Frame-wise MLP embeds the sampled frames, Masked Mean Pooling aggregates valid frame features, and a shared Event MLP produces the event representation $\boldsymbol{h}_i$. The same two MLPs are reused across all events, and padded frame slots are excluded from pooling. This compresses each variable-duration interaction into a fixed-width event token while preserving event identity until fusion.

The resulting $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{N_h}]$ forms an Opponent-centric event set rather than a historical Agent skeleton trajectory. Appendix C provides the frame-sampling procedure, padding masks, and empty-event handling for variable-size event sets.

Event-Guided Context Fusion. Event-level cross-attention fuses the historical event set $\boldsymbol{H}$ with the recent interaction tokens $\boldsymbol{X}$:

$$\begin{aligned} &\boldsymbol{Q} = \boldsymbol{X}, \quad \boldsymbol{K} = \boldsymbol{V} = \boldsymbol{H}, \quad \boldsymbol{G} = \mathrm{MHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}), \\ &\widetilde{\boldsymbol{X}} = \mathrm{LN}(\boldsymbol{X} + \sigma(g)\boldsymbol{G}), \qquad \widetilde{\boldsymbol{C}} = \mathrm{CausalEnc}_{\phi}(\widetilde{\boldsymbol{X}}). \end{aligned} \tag{6}$$

Here, $\boldsymbol{G}$ denotes the retrieved event context and $\sigma(g)$ is a learned scalar gate controlling its residual contribution. The causal encoder maps the fused sequence to event-conditioned states $\widetilde{\boldsymbol{C}}$, whose j-th row is $\widetilde{\boldsymbol{c}}_j$. Compressing each historical interaction before attention allows fusion to operate over event tokens rather than extending the recent sequence with every historical frame. The event representation and fusion pathway is active only during Stage III refinement.

3.4 Training Objectives and Staged Optimization

The recent-context and event-conditioned pathways share the same latent prediction and motion-decoding modules. Their difference is the causal conditioning state supplied to the Flow-Matching predictor:

$$\boldsymbol{c}_j^{\star} = \begin{cases} \boldsymbol{c}_j, & \text{Stage II and inference,} \\ \widetilde{\boldsymbol{c}}_j, & \text{Stage III.} \end{cases} \tag{7}$$

At inference, $\boldsymbol{c}_j$ is computed from the current bounded motion context and $\mathcal{Z}_f^A$; the Stage III event set is not retained as rollout state.

Stage I: VQ-VAE motion representation pre-training. We pre-train the VQ-VAE on raw motion blocks and freeze its encoder, codebook, and VQ decoder before reaction-policy learning. Appendix A.1 provides the reconstruction and commitment objectives.

Stage II: General Reaction Learning. Using bounded recent Agent–Opponent windows, we jointly train the Causal Interaction Encoder, Flow-Matching Next-Latent Predictor, and Online Motion Decoder. Each valid adjacent target is conditioned on $\boldsymbol{c}_j$. This stage establishes the complete inference pathway and provides the checkpoint used to initialize Stage III.

Stage III: Causal History-Assisted Refinement. Starting from the Stage-II checkpoint, we introduce Historical Event Representation and Event-Guided Context Fusion using event sets precomputed by Causal Key-Event Selection. The recent windows, adjacent latent targets, Flow-Matching paths, decoder inputs, and motion/root supervision remain unchanged from Stage II. Stage III instead conditions each valid adjacent target on $\widetilde{\boldsymbol{c}}_j$.

Over the reported 40-epoch (40×1000-update) schedule, both the shared policy and newly introduced event modules are updated, while the VQ tokenizer remains frozen. The staged procedure therefore first learns general reaction dynamics from dense recent-context windows and then uses sparser historical event evidence to refine the shared policy. After Stage III, inference retains the updated shared parameters and returns to the recent-context condition $\boldsymbol{c}_j$ in Equation 7.

For every valid adjacent target, Stages II and III minimize the Flow-Matching, motion-reconstruction, and root-reconstruction losses:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\|\boldsymbol{v}_\theta(\boldsymbol{z}(t), t, \boldsymbol{c}^{\star}) - \boldsymbol{u}_t\|_2^2\right], \quad \mathcal{L}_{\mathrm{mot}} = \mathbb{E}[\|\widetilde{A} - A^{\mathrm{gt}}\|_2^2], \quad \mathcal{L}_{\mathrm{root}} = \mathbb{E}[\|\widetilde{R}^{A|O} - R_{\mathrm{gt}}^{A|O}\|_2^2], \tag{8}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{mot}}\mathcal{L}_{\mathrm{mot}} + \lambda_{\mathrm{root}}\mathcal{L}_{\mathrm{root}}. \tag{9}$$

$\mathcal{L}_{\mathrm{FM}}$ supervises the target velocity, while $\mathcal{L}_{\mathrm{mot}}$ and $\mathcal{L}_{\mathrm{root}}$ supervise the decoded motion and relative-root blocks. We set $\lambda_{\mathrm{mot}} = \lambda_{\mathrm{root}} = 1$ in all reported experiments. The expectations are taken over sampled valid adjacent targets, Flow times, and Gaussian source noise, with padded latent positions excluded. Appendix C.2.1 provides the masking, normalization, gradient, optimization, and sampling details.

3.5 Online Two-Character Interaction Generation

At inference, the refined shared policy uses only the bounded recent-context pathway defined above. At each generation step, the parameter-shared policy produces a U-frame motion block and appends its continuous latent endpoint to the corresponding bounded motion and latent buffers.

In *Reactive* generation, causally observed Opponent motion advances the Opponent motion buffer while Equation 1 generates the next Agent block and updates only $\mathcal{Z}_f^A$; no Opponent latent buffer is used. In *Two-character* generation, two role-swapped invocations separately use $\mathcal{Z}_f^A$ and $\mathcal{Z}_f^O$. Both read the same pre-update motion and latent buffers, after which their generated blocks and endpoints are committed synchronously. Thus, neither character conditions on the other's newly generated block within the same update. Historical event selection, representation, and fusion are bypassed throughout online generation. Appendix D.1.4 provides the complete rollout algorithm; Appendix D.2 gives its complexity analysis.

4 EXPERIMENTS

4.1 DATASETS AND EXPERIMENTAL SETTING

Dataset. We evaluate on DuoBox (63.4 minutes; approximately 457K frames at 120 FPS) using the established 80/20 interaction split (Cen et al., 2025). Following the benchmark protocol, motion is downsampled to 30 FPS and role swapping is applied during training. Stage III additionally uses the verified historical event archive described in Section 3.3; Appendix B details its construction. In Reactive generation, the Opponent is observed and the Agent is generated. In Two-character generation, the shared policy recursively generates both roles from a four-frame prefix, so generated motion becomes subsequent context. We further evaluate 1,800-frame rollouts to assess long-horizon closed-loop behavior. Unless stated otherwise, evaluations use bounded recent motion and role-specific latent buffers without historical-event input; this is our "recent-context-only" setting.

Evaluation metrics. Following Ready-to-React (Cen et al., 2025), per-frame, per-transition, and per-clip FID compare the distributions of frame-level poses, adjacent-state transitions, and complete motion clips, respectively. Together, they measure distributional fidelity at complementary temporal scales. Jitter measures temporal instability; Root Orientation (RO) is the fraction of frames whose facing difference exceeds $45°$; and Foot Sliding (FS) is the average foot displacement while the foot is within 5 cm of the ground. Lower FID is better, whereas Jitter, RO, and FS are interpreted by their absolute deviation from the corresponding Real-data statistics. In particular, a very low FS may also result from near-stationary motion and should not be interpreted in isolation. Appendix D.3 provides the metric implementation details.

4.2 COMPARISON WITH EXISTING METHODS

Baselines. Ready-to-React (Cen et al., 2025) is the closest matched baseline, sharing DuoBox, the bounded causal interface, VQ representation, parameter-shared policy, Online Motion Decoder, and generation protocol. InterFormer, CAMDM, and T2MGPT-online are recent-context baselines adapted to DuoBox; MAGNet, Intention Guidance, and ReGenNet are contextual comparisons under the same metrics but different interfaces or training procedures. Appendix F details their provenance, adaptation, and comparison scope.

Distributional fidelity. SCHA-React achieves the lowest reported per-frame, per-transition, and per-clip FIDs in both Reactive and Two-character settings among evaluated methods (Table 1). Relative to the closest matched baseline, Ready-to-React, per-clip FID decreases from 9.60 to 8.11 in Reactive generation and from 25.28 to 21.28 in Two-character generation. The corresponding frame- and transition-level improvements show that the gain is not confined to full-clip statistics, but also appears in frame-level pose distributions and adjacent-state dynamics.

Closed-loop generation. Reactive generation conditions on observed Opponent motion, whereas Two-character generation recursively synthesizes both roles and feeds the generated states back into subsequent predictions. This recursive feedback allows local errors in pose, transition, or relative configuration to become part of future conditioning. Under this setting, SCHA-React reduces Ready-to-React's per-frame, per-transition, and per-clip FIDs from 1.39/2.11/25.28 to 1.15/1.58/21.28, respectively. The improvement across all three temporal scales remains when both characters are generated recursively; Section 4.3 evaluates the same policy over 1,800-frame closed-loop rollouts.

Physical statistics. The FID gains are not uniform across physical statistics. In Reactive generation, SCHA-React has the smallest non-Real absolute deviation from the Real FS (1.01 versus 0.97), while

Table 1: **Comparison with existing methods.** Results for Reactive and Two-character generation. Lower FID is better; Jitter, RO, and FS are ranked by absolute deviation from the Real-data statistics. Bold and underlined values denote the best and second-best displayed non-Real results, respectively; ties share the same style. A dash denotes an unavailable or numerically divergent measurement. The final three SCHA-React rows are staged-training variants analyzed in Section 4.4.

Methods	Reactive						Two-character					
	FID↓ Per-frame	FID↓ Per-tr.	FID↓ Per-clip	Jitter→	RO→	FS→	FID↓ Per-frame	FID↓ Per-tr.	FID↓ Per-clip	Jitter→	RO→	FS→
Real	–	–	–	21.33	24.7%	0.97	–	–	–	21.33	24.7%	0.97
InterFormer (Chopin et al., 2023)	0.72	1.99	15.06	6.71	40.3%	1.08	2.50	4.59	47.19	6.12	31.5%	1.05
CAMDM (Chen et al., 2024b)	1.61	4.04	28.50	53.62	40.1%	1.96	4.16	7.49	70.42	53.99	22.6%	2.01
T2MGPT-online (Zhang et al., 2023)	1.72	1.57	30.16	78.29	50.4%	2.65	8.25	3.35	122.46	78.33	46.2%	2.63
Ready-to-React (Cen et al., 2025)	0.54	1.00	9.60	17.83	34.7%	1.02	1.39	2.11	25.28	16.84	**24.1%**	**0.97**
MAGNet (Maluleke et al., 2025)	4.11	12.85	80.46	**19.90**	76.9%	0.55	9.93	31.76	129.60	56.23	35.3%	3.97
Intention Guidance (Zhang et al., 2025)	2.28	21.37	37.24	155.76	52.6%	6.49	12.38	39.94	175.26	165.29	53.7%	7.27
ReGenNet (Xu et al., 2024b)	1.18	4.07	34.13	131.19	**33.7%**	2.30	5.35	8.74	105.45	103.52	7.3%	1.61
SCHA-React (w/o Stage III)	0.45	0.88	8.95	17.30	35.4%	1.03	1.23	1.93	24.60	16.88	23.9%	**0.97**
SCHA-React (w/o Stage II)	0.54	1.13	11.63	17.26	36.7%	1.02	1.55	2.58	30.10	16.63	22.1%	0.98
SCHA-React	**0.42**	**0.72**	**8.11**	18.06	34.5%	**1.01**	**1.15**	**1.58**	**21.28**	**17.29**	22.3%	**0.97**

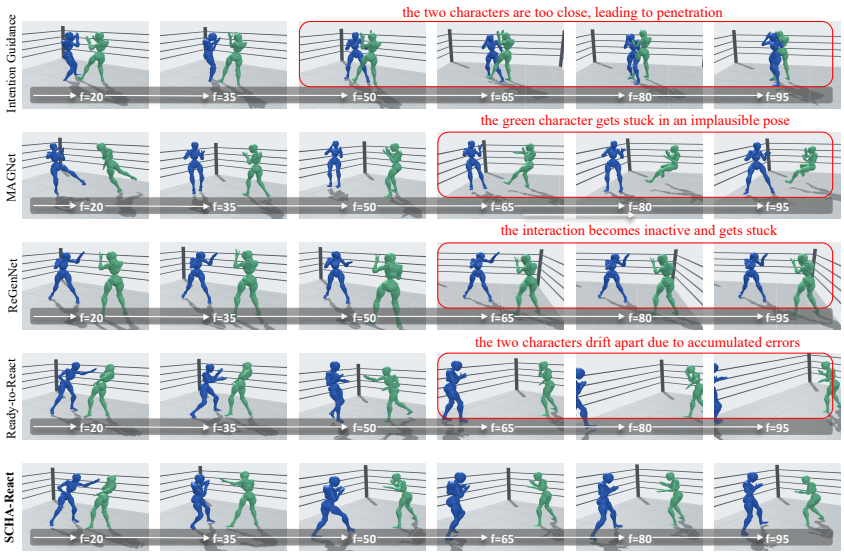


Figure 3: **Two-character closed-loop comparison.** Matched rollouts show penetration (Intention Guidance), stagnation (MAGNet), interaction loss (ReGenNet), and separation (Ready-to-React), whereas SCHA-React maintains the interaction. Such recursive failures re-enter later predictions; Appendix I provides additional examples.

MAGNet has the smallest Jitter deviation and ReGenNet the smallest RO deviation. In Two-character generation, SCHA-React has the smallest Jitter deviation and matches the Real FS of 0.97, tied with Ready-to-React, whereas Ready-to-React has the smallest RO deviation. Thus, SCHA-React's main advantage is consistent distributional fidelity rather than uniform dominance across Jitter, RO, and FS.

4.3 Extended Closed-Loop Robustness

Longer closed-loop rollouts expose error propagation as recursively generated states become future conditioning inputs (Table 2). Starting from a four-frame prefix, SCHA-React achieves the lowest reported per-frame, per-transition, and per-clip FIDs among evaluated methods. Relative to Ready-to-React, these values decrease from 2.39/2.38/36.76 to 1.54/1.95/25.69, including a 30.1% per-clip reduction. The simultaneous reductions show that the gain is not confined to clip-level statistics. RO is close to Real (24.9% versus 24.7%), although the smallest deviations vary by physical statistic.

Historical-event retrieval is disabled throughout: the refined policy uses only bounded recent motion context and role-specific latent buffers.

Table 2: **Long-horizon two-character closed-loop generation.** Results for 1,800-frame recursive rollouts from a four-frame prefix; conventions follow Table 1.

Methods	FID↓			Jitter→	RO→	FS→
	Per-frame	Per-transition	Per-clip			
Real	–	–	–	21.33	24.7%	0.97
InterFormer	5.63	5.94	87.70	4.78	44.6%	0.90
CAMDM	7.56	13.47	109.65	25.24	13.7%	0.86
T2MGPT-online	21.27	4.46	273.09	72.48	70.4%	2.34
Ready-to-React	2.39	2.38	36.76	**19.20**	31.3%	0.83
MAGNet	23.96	13.31	288.05	27.74	63.5%	3.34
Intention Guidance	46.06	29.27	499.71	140.98	62.1%	6.58
ReGenNet	4.83	15.07	127.12	60.89	12.1%	**0.99**
SCHA-React	**1.54**	**1.95**	**25.69**	17.25	**24.9%**	0.81

4.4 Ablation Studies

Effect of staged reaction learning. The final rows of Table 1 isolate the stages. Continued Stage II gives Reactive/Two-character per-clip FIDs of 8.95/24.60; historical guidance without it gives 11.63/30.10, whereas Stage III after it reaches 8.11/21.28. This ordering supports dense recent-context learning as the stronger base for sparse historical refinement.

Effect of historical evidence selection. Replacing random past motion with completed key events lowers all six FIDs, including Two-character per-clip FID from 24.70 to 23.63 (Table 3). Recency-only and quality-only selection give 23.55 and 24.30, while Full combines score–recency selection with complete representation and fusion to reach 21.28. This supports both event identity and selection cues.

Effect of historical evidence representation and fusion. Event mean pooling gives Two-character transition/clip FIDs of 1.98/24.87, versus 2.19/27.21 for frame-level attention. Removing or fixing the learned gate raises per-clip FID to 28.10 or 30.42, versus 21.28 for Full. Compact event representation with adaptive gating thus gives the strongest tested recursive Two-character FID profile.

Table 3: **Core ablation of Causal History-Assisted Refinement.** Recent-context-only inference; conventions follow Table 1.

Variants	Reactive						Two-character					
	FID↓ Per-frame	FID↓ Per-tr.	FID↓ Per-clip	Jitter→	RO→	FS→	FID↓ Per-frame	FID↓ Per-tr.	FID↓ Per-clip	Jitter→	RO→	FS→
Real	–	–	–	21.33	24.7%	0.97	–	–	–	21.33	24.7%	0.97
Historical evidence selection												
random past motion	0.49	0.90	9.63	16.99	34.8%	1.01	1.33	2.00	24.70	16.38	21.8%	0.98
random key events	0.46	0.88	9.31	17.13	35.7%	1.04	1.27	1.81	23.63	16.57	23.4%	0.99
recency-only selection	0.45	0.87	9.11	17.19	34.7%	1.06	1.26	1.89	23.55	16.47	22.8%	0.98
quality-only selection	0.46	0.88	9.27	17.27	34.5%	1.04	1.30	1.86	24.30	16.62	22.4%	0.99
Representation and fusion												
use global projection	0.48	0.88	9.40	17.37	34.5%	1.04	1.33	1.89	24.73	16.83	23.3%	1.01
use event mean pooling	0.48	0.81	8.99	17.54	35.8%	1.03	1.37	1.98	24.87	16.73	**23.6%**	1.02
use frame-level attention	0.51	1.13	10.72	16.52	34.3%	1.01	1.50	2.19	27.21	16.17	18.6%	0.99
use Event Cross-Attention without gating	0.54	1.11	10.94	16.55	33.0%	**1.00**	1.61	2.51	28.10	15.75	18.5%	0.98
use Event Cross-Attention with fixed gating	0.58	1.15	10.97	16.39	**32.8%**	**1.00**	1.75	2.67	30.42	15.59	18.0%	**0.97**
SCHA-React (Full)	**0.42**	**0.72**	**8.11**	**18.06**	34.5%	1.01	**1.15**	**1.58**	**21.28**	**17.29**	22.3%	**0.97**

Additional module and solver analyses appear in Appendices G and H.1.

5 Conclusion

We address the *recency–relevance mismatch* with SCHA-React, which refines a bounded recent-context policy using completed historical interactions only during training. It attains the lowest reported FIDs among evaluated methods across Reactive, Two-character, and 1,800-frame tests; ablations support staged training, event selection, and learned fusion, while physical-statistic deviations from Real remain metric-dependent. Inference uses only bounded recent motion and role-specific latent buffers; evaluation is limited to pre-annotated DuoBox events.

AI USE STATEMENT

The multimodal model in Appendix B produced initial event annotations, all of which were manually reviewed and corrected. OpenAI Codex assisted with translation, organization, language editing, LaTeX, literature and reference checks, and presentation of existing results. The authors verified all assisted content against source materials, implementation records, tables, and references. Generative AI produced no reported measurement and was not used in policy training or official inference; the authors remain responsible for the final work.

REFERENCES

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 457–469, 2024.

Manolo Canales Cuba, Vinícius do Carmo Melício, and João Paulo Gois. FlowMotion: Target-predictive conditional flow matching for jitter-reduced text-driven human motion generation. *Computers & Graphics*, 132:104374, 2025. doi: 10.1016/j.cag.2025.104374.

Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Ready-to-React: Online reaction policy for two-character interaction generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mm0cqJ2O3f.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024a. doi: 10.52202/079017-0759.

Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers*, 2024b. doi: 10.1145/3641519.3657440.

Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 25:8842–8854, 2023. doi: 10.1109/TMM.2023.3242152.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. MotionLCM: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pp. 390–408, 2024. doi: 10.1007/978-3-031-72640-8_22.

Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. FreeMotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pp. 93–109, 2024. doi: 10.1007/978-3-031-73242-3_6.

Zichen Geng, Zeeshan Hayder, Wei Liu, Hesheng Wang, and Ajmal Saeed Mian. ARMFlow: AutoRegressive MeanFlow for online 3D human reaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30718–30728, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Geng_ARMFlow_AutoRegressive_MeanFlow_for_Online_3D_Human_Reaction_Generation_CVPR_2026_paper.html.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. ReMoS: 3D motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision*, pp. 418–437, 2024. doi: 10.1007/978-3-031-72764-1_24.

N. Viktor Gredin, David P. Broadbent, Joseph L. Thomas, and A. Mark Williams. The role of action tendencies in expert anticipation. *Asian Journal of Sport and Exercise Psychology*, 3(1):30–38, 2023. doi: 10.1016/j.ajsep.2023.02.001.

Prerit Gupta, Shourya Verma, Ananth Grama, and Aniket Bera. Unified multi-modal interactive and reactive 3D motion generation via rectified flow. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a783b0bba6aaa9dec34aca6500d853cd-Abstract-Conference.html.

Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, Yunlu Chen, Basura Fernando, Yuki M. Asano, Efstratios Gavves, Pascal Mettes, Björn Ommer, and Cees G. M. Snoek. Motion flow matching for human motion synthesis and editing. *CoRR*, abs/2312.08895, 2023. doi: 10.48550/arXiv.2312.08895.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. InterMask: 3d human interaction generation via collaborative masked modeling. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZAyuwJYN8N.

Kaiyang Ji, Ye Shi, Zichen Jin, Kangyi Chen, Lan Xu, Yuexin Ma, Jingyi Yu, and Jingya Wang. Towards immersive Human-X interaction: A real-time framework for physically plausible motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10173–10183, October 2025. doi: 10.1109/ICCV51701.2025.00948.

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

Siyao Li, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower GPT with off-policy reinforcement learning for dance accompaniment. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GW4j4n2cjH.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. doi: 10.1007/s11263-024-02042-6.

Jingzhong Lin, Xinru Li, Yuanyuan Qi, Bohao Zhang, Wenxiang Liu, Kecheng Tang, Wenxuan Huang, Xiangfeng Xu, Bangyan Li, Changbo Wang, and Gaoqi He. ReactDance: Hierarchical representation for high-fidelity and coherent long-form reactive dance generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FvMyAMbbX0.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yunze Liu, Changxi Chen, and Li Yi. Interactive Humanoid: Online full body human motion reaction synthesis with social affordance forecasting and canonicalization. In *International Conference on 3D Vision*, pp. 1440–1450. IEEE, 2025. doi: 10.1109/3DV66043.2025.00136.

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

David L. Mann, Teuntje Schaefers, and Rouwen Cañal-Bruland. Action preferences and the anticipation of action outcomes. *Acta Psychologica*, 152:1–9, 2014. doi: 10.1016/j.actpsy.2014.07.004.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2IN: Leveraging individual information to generate human INteractions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1941–1951, 2024.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. MixerMDM: Learnable composition of human motion diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12380–12390, 2025.

Wenhui Tan, Boyuan Li, Chuhao Jin, Wenbing Huang, Xiting Wang, and Ruihua Song. Think Then React: Towards unconstrained action-to-reaction motion generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UxzKcIZedp.

Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15999–16009, 2023.

Julian Tanke, Takashi Shibuya, Kengo Uchida, Koichi Saito, and Yuki Mitsufuji. Dyadic Mamba: Long-term dyadic human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2868–2877, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/HuMoGen/html/Tanke_Dyadic_Mamba_Long-term_Dyadic_Human_Motion_Synthesis_CVPRW_2025_paper.html.

Joseph L. Thomas, David P. Broadbent, N. Viktor Gredin, Bradley J. Fawver, and A. Mark Williams. Skill-based differences in the detection and utilization of opponent action preferences following increasing exposure and changes in tendencies. *Journal of Sport and Exercise Psychology*, 44(5): 370–381, 2022. doi: 10.1123/jsep.2021-0244.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. TIMotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7169–7178, 2025.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. InterControl: Zero-shot human interaction generation by controlling every joint. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3345.

Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, and Lingjie Liu. Text2Interact: High-fidelity and diverse text-to-two-person interaction generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a99f732df9b668284b449da0214a3286-Abstract-Conference.html.

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-X: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22260–22271, 2024a.

Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. ReGenNet: Towards human action-reaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1759–1769, 2024b.

Yaoqin Ye, Yiteng Xu, Qin Sun, Xinge Zhu, Yujing Sun, and Yuexin Ma. ReMoGen: Real-time human interaction-to-reaction generation via modular learning from diverse data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16475–16485, June 2026.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14730–14740, 2023.

Xiaotang Zhang, Ziyi Chang, Qianhui Men, and Hubert P. H. Shum. Real-time and controllable reactive motion synthesis via intention guidance. *Computer Graphics Forum*, 44(6), 2025. doi: 10.1111/cgf.70222.

Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion Mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pp. 265–282, 2024. doi: 10.1007/978-3-031-73232-4_15.