DyaChain: Strategizing Concurrent, Holistic and Dyadic Motion Generation via Progressive Interaction Chain

Anonymous authors
Paper under double-blind review

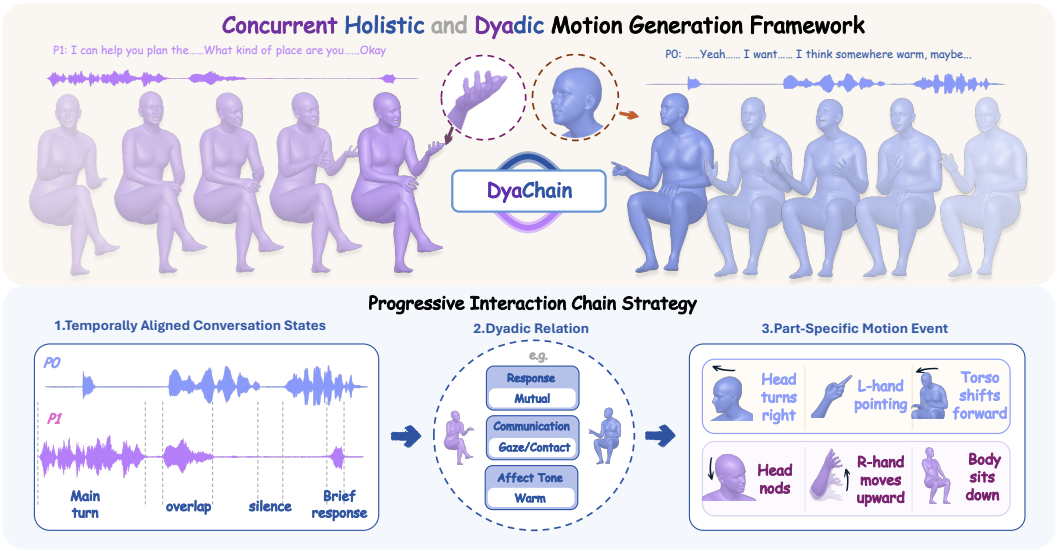


Figure 1: Our **DyaChain** framework strategizes a **Progressive Interaction Chain**, providing hierarchical guidance to **concurrently** generate **holistic** motions for both interlocutors.

Abstract

Generating conversational motion is fundamental to creating lifelike virtual humans for immersive communication and human–computer interaction. While existing methods enable reactive individual movement synthesis, they overlook the **concurrent** and **holistic** behaviors of both speakers in an interactive manner. Recent datasets also rarely provide aligned annotations at both the interlocutor and interaction levels. To address these limitations, we introduce **HOWDY**, a dyadic conversation dataset that augments synchronized speech and holistic 3D motion from both participants with fine-grained body-part motion descriptions and hierarchical interaction annotations. Additionally, we propose **DyaChain**, a holistic motion generation framework to concurrently model dyadic speakers by strategizing a **Progressive Interaction Chain** approach. To ensure spatiotemporal coherence between the two concurrently evolving motion streams while preserving asynchronous action with response dynamics, the chain progresses through three levels: conversational states, dyadic relations, and motion events. Here, we first derive temporally aligned conversational states to delineate speaking and response periods, and then reason dyadic relations between interlocutors. Conditioned on these cues, we design a Temporal-Relational Interactor (**TRI**) module to capture cross-individual body-part dependencies and asynchronous responses using causal counterpart attention and learnable response delays. Moreover, we retrieve static posture descriptions from the HOWDY and then infer part-aligned dynamic motion events, motivating a newly proposed Part-Aligned Semantic Routing (**PSR**) module. PSR decomposes holistic motion into semantically coherent body regions and injects static and dynamic motion semantics into the corresponding motion representations to guide region-specific dynamics. Extensive experiments demonstrate that **DyaChain** outperforms SOTA methods for dyadic motion generation.

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.

1 INTRODUCTION

Given the speech of two interlocutors, dyadic conversational motion generation aims to jointly synthesize their facial expressions, hand gestures, and body movements. These non-verbal behaviors complement verbal communication and are essential to natural and expressive human interaction. Dyadic conversational motion generation has broad applications in avatar animation Lee et al. (2026), immersive communication, human-computer interaction Son & Rubo (2025), and embodied AI Cai et al. (2024). Existing methods primarily focus on reactive individual motion synthesis in conversational settings Li et al. (2026); Ki et al. (2026).

Most conversational motion generation approaches usually separately model the individual movement, conditioned on dialogue audios or contextual cues Mughal et al. (2026); Peng et al. (2026). A few recent studies further generate motions for both interlocutors, but remain limited to gestures or partial behavioral modalities, leaving the concurrent and holistic behaviors insufficiently modeled Qi et al. (2025); Zhao et al. (2026); Shan et al. (2026). This limitation results in drawbacks for coordination among different body parts (*e.g.*, facial expressions, gestures, and body movements) *w.r.t.* the motion dynamics of both interlocutors concurrently. Moreover, recent datasets rarely provide paired holistic body parameters with concrete semantic-aligned motion annotations at both the interlocutor and interaction levels, slowing the path to learning coordinated dyadic behaviors.

Dyadic motion synthesis relies on coordinating the concurrently evolving holistic behaviors of both interlocutors. Achieving this goal presents two main challenges: ❶ Building a high-quality dyadic dataset that pairs holistic (*e.g.*, face, gestures) motion with fine-grained body-part aware descriptions and hierarchical interaction annotations. ❷ Modeling holistic movements for both interlocutors while maintaining coherent cross-individual interactions and preserving asynchronous response dynamics.

To address the data limitations, we construct a dyadic conversation dataset with synchronized speech and holistic 3D motion for both participants, dubbed **HOWDY**. In particular, each part of the body, gestures, and head/ face belonging to both interlocutors are annotated with spatially aligned static configurations from dynamic motion events, providing fine-grained semantics for individual activities. Meanwhile, these annotations are synchronized with conversational states and dyadic relations, connecting individual motion details to the temporal and relational context of the interaction. In this manner, our HOWDY dataset provides complementary guidance for jointly modeling individual behaviors and coordinated dyadic interactions from the perspective of spatiotemporal consistency.

Building upon HOWDY, we propose **DyaChain**, a framework for concurrently generating fine-grained holistic motions for both interlocutors from conversational speech, as depicted in Figure 1. The key insight of our framework is to explicitly structure the interactions by strategizing a **Progressive Interaction Chain**. Considering two motion streams should maintain spatiotemporal coherence while preserving asynchronous responses, we formulate this chain of progress from conversational states through dyadic relations to motion events. These states provide temporal context through large language model reasoning about dyadic relations, which characterize how the interlocutors engage with and respond to one another. Based on temporal and relational cues, we infer part-aligned motion events by exploring "**`What movements does each participant perform with specific body regions?`**", to provide concrete dynamics guidance for joint generation.

Specifically, to capture cross-individual body-part interactions under the guidance of conversational states and dyadic relations, we present a Temporal-Relational Interactor (**TRI**). Such interactions involve dependencies across body regions of the two interlocutors. Here, TRI employs causal counterpart attention within a temporal window to associate body-part representations of each participant *w.r.t.* the counterpart's preceding motion context. Considering that responses of the interlocutor may lag behind the counterpart, we incorporate learnable response delays to adjust attention weights across temporal offsets. Then, the contribution of the attended counterpart is modulated by these temporal and relational cues to each participant's motion representations. TRI thus facilitates the modeling of coordinated dyadic interactions with asynchronous body-part responses.

While TRI captures interactive dependencies among different body parts across individuals, semantic-aligned spatial guidance also encourages the specific body region displaying expressive movements. To this end, we propose a Part-Aligned Semantic Routing (**PSR**) module. Specifically, PSR draws on static and dynamic motion semantics as complementary conditions for body-part motion modeling. Here, we retrieve static motion descriptions from the HOWDY dataset and infer part-aligned dynamic

motion events to characterize evolving body-part movements, conditioning on the aforementioned temporal and relational context. To ensure the semantic coherence between the obtained complementary information and corresponding body regions, PSR first decomposes the holistic human body into several concrete parts. Then, we inject static and dynamic motion semantics into the motion representations of related body parts to guide region-specific dynamics. Extensive experiments demonstrate that **DyaChain** outperforms SOTA methods for dyadic motion generation.

Overall, our contributions are summarized as follows:

- We introduce **HOWDY**, a dyadic conversation dataset that pairs synchronized speech and holistic 3D motion with fine-grained body-part descriptions and hierarchical interaction annotations to support concurrent holistic motion generation.
- We propose **DyaChain**, a framework for concurrent holistic motion generation that enables hierarchical interaction modeling by strategizing a **Progressive Interaction Chain**.
- We design a Temporal-Relational Interactor (**TRI**) to model cross-individual body-part dependencies and asynchronous responses, ensuring coordinated dyadic interactions.
- We propose a Part-Aligned Semantic Routing (**PSR**) module to align static and dynamic motion semantics with corresponding body regions, ensuring semantically coherent motion dynamics.

2 Related Work

Co-Speech Motion Generation. Early methods primarily focused on synthesizing co-speech gestures for individual speakers Ginosar et al. (2019). Subsequent studies incorporate textual content, speaker identity, and emotion to guide gesture generation Yoon et al. (2020); Liu et al. (2022a). Further efforts explore hierarchical speech-motion associations and diffusion-based generation to improve semantic alignment and motion diversity Liu et al. (2022b); Zhu et al. (2023). However, gestures capture only part of a speaker's non-verbal behavior. To capture holistic behavior, TalkSHOW Yi et al. (2023) combines separate facial modeling with compositional body-hand representations, while EMAGE Liu et al. (2024) integrates regional motion priors through masked gesture modeling. However, these methods overlook the interactive nature of human conversations. To account for this context, several studies use both participants' speech to guide a target character's movements Ng et al. (2024); Mughal et al. (2024). Recent methods further explore reactive individual motion synthesis, modeling the target character's responses to an interlocutor Mughal et al. (2026); Peng et al. (2026). Despite these advances, generation remains centered on a single participant, leaving both interlocutors' concurrent and holistic behaviors and their mutual interactions without joint modeling.

Co-Speech Motion Dataset. High-quality datasets are essential for advancing co-speech motion generation. Early resources obtain speech-aligned gesture sequences through motion capture Ferstl & McDonnell (2018) or pose estimation from videos Yoon et al. (2020). BEAT, SHOW, and BEAT2 broaden motion coverage to include the body, hands, and face Liu et al. (2022a); Yi et al. (2023); Liu et al. (2024). These datasets support holistic individual motion modeling but lack paired motion sequences for studying interpersonal interactions. In conversational settings, Talking With Hands captures synchronized body and finger motions with speech from both participants Lee et al. (2019). DnD Group Gesture extends the scope to multi-party conversations, providing body and hand motions, speech, and gesture-type annotations Mughal et al. (2024). HoCo and GES-Inter further provide paired holistic motion representations aligned with conversational speech Sun et al. (2025); Qi et al. (2025). InterAct incorporates relationship and emotion labels Ho et al. (2025), while Seamless Interaction provides behavioral descriptions and internal-state annotations for selected moments of interest Agrawal et al. (2025). However, existing datasets rarely provide paired holistic motions together with aligned annotations of individual body-part semantics and hierarchical interaction context. HOWDY addresses this gap with fine-grained static and dynamic body-part descriptions and hierarchical interaction annotations aligned with both participants' holistic motion sequences.

3 HOWDY Dataset

Data Acquisition and Alignment. Following prior video-based reconstruction pipelines Yi et al. (2023); Qi et al. (2025); Agrawal et al. (2025), we construct HOWDY from in-the-wild dyadic

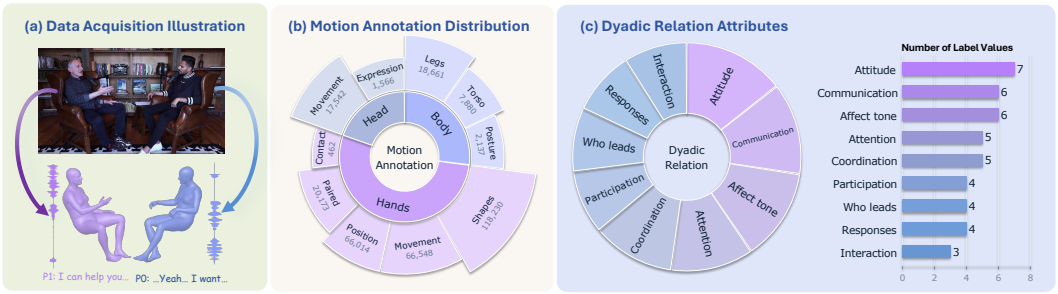


Figure 2: **Overview of HOWDY data and annotations.** (a) Recovery of interactive motion, speech, and transcripts from dyadic conversation videos. (b) Distribution of motion annotations across body regions. (c) Dyadic relation attributes and the number of label values for each attribute.

talk-show and interview videos (Figure 2(a)). We estimate body and hand motions with SAM 3D Body Yang et al. (2026) and facial expression parameters with EMOCA Daněček et al. (2022), consolidating these estimates into a unified SMPL-X-based full-body parametric mesh representation Pavlakos et al. (2019). For speech processing, we perform human voice separation Bredin et al. (2020), associate speech segments with the corresponding participants, and obtain transcripts with word-level timestamps through ASR Bain et al. (2023). Finally, we align the motion sequences, speech, and transcripts of both participants along a shared timeline, providing synchronized multimodal data for subsequent annotation.

Part-level Motion Annotation. To capture fine-grained motion semantics in dyadic conversations, we organize the behavior of each participant into descriptions of the body, hands, and head/face. Figure 2(b) summarizes these annotation categories and their distribution. We employ MotionScript Yazdian et al. (2025) to detect salient motion patterns in the recovered joint and facial motions and localize their temporal intervals. After an initial review with GPT-5.6OpenAI (2026), human annotators verify the motion descriptions and event boundaries to ensure annotation accuracy. Specifically, static descriptions characterize relatively stable configurations, such as torso orientation and hand shape, while dynamic events describe movements such as hand gestures and head nods, together with their temporal intervals. The resulting annotations associate each motion description with a specific participant, body region, and, for dynamic events, time interval, providing complementary semantic cues for stable configurations and evolving movements.

Hierarchical Interaction Annotation. To situate individual movements within their conversational context, we organize interaction annotations into conversational states and dyadic relations. We use participant-aligned transcripts with word-level timestamps obtained through automatic speech recognition Bain et al. (2023) to annotate frame-wise conversational states. The state indicators include main speaking turns, brief responses, turn acquisition and loss, and silence by one or both participants. We use Qwen3.6-27BTeam (2026a) to annotate predefined dyadic relation attributes from videos and transcripts. These attributes characterize the participation, responsiveness, affective tone, and behavioral coordination of the interlocutors. Figure 2(c) summarizes the attributes and the number of label values for each. Together with the part-level descriptions, these annotations link individual motion details to the temporal and relational context of dyadic interaction.

4 Proposed Method

4.1 Problem Formulation

Given the dialogue speech signals $\mathcal{A} = \{A_0, A_1\}$ and paired transcripts $\mathcal{C} = \{C_0, C_1\}$ of two interlocutors, the goal of our **DyaChain** framework is to concurrently generate corresponding holistic motion sequences $\mathcal{M} = \{M_0, M_1\}$, including facial expressions, gestures, and body movements. As illustrated in Figure 3, DyaChain first derives interaction conditions through a **Progressive Interaction Chain** strategy. Then, we synthesize the motions of both interlocutors guided by the speech signals within the obtained conditions. The overall process is expressed as:

$$\mathcal{M} = \mathrm{DyaChain}(\mathcal{A}, \mathcal{C}). \tag{1}$$

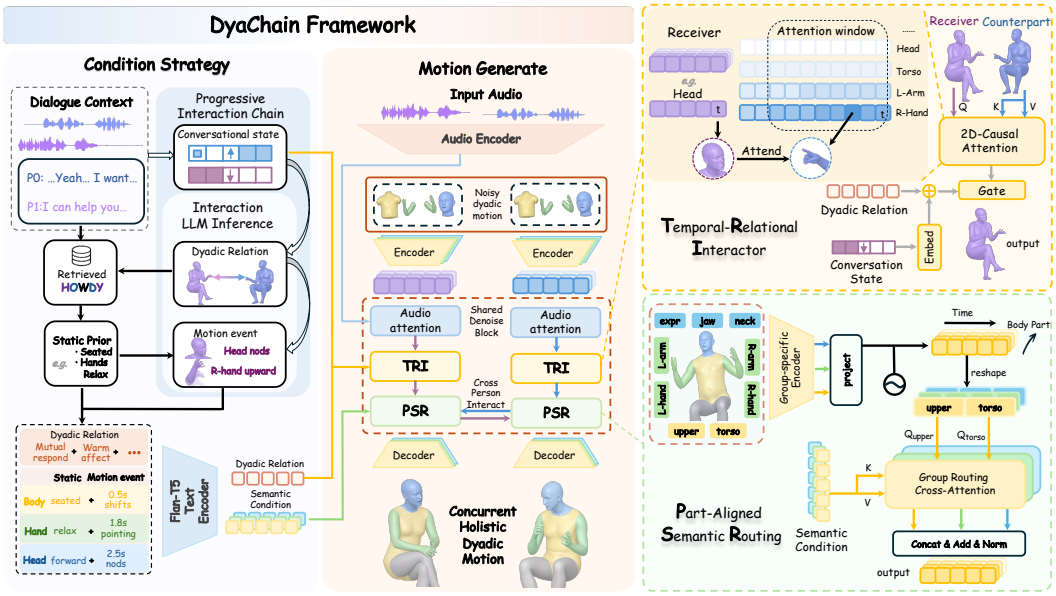


Figure 3: We propose DyaChain for concurrent holistic dyadic motion generation. We derive hierarchical interaction conditions by strategizing a Progressive Interaction Chain. During motion generation, we employ TRI to capture cross-person dependencies and PSR to route motion semantics to corresponding body parts.

Here, each pose is formulated as J joints with the 6D rotation representation, accompanied by FLAME facial expression coefficients. The selected joints cover the upper body, detailed fingers, and head involved in conversational behavior.

4.2 Progressive Interaction Chain

Considering that the behaviors of interlocutors are usually coherent with the evolving dialogue context, we design a Progressive Interaction Chain strategy to provide interactive dynamics guidance from the given context. Specifically, we progress this interaction chain in three folds: conversational states, dyadic relations, and motion events. Conversational states delineate speaking and response periods, establishing temporal context for interaction modeling. Within this context, dyadic relations characterize the response pattern between one interlocutor *w.r.t.* the other. Building on these temporal and relational cues, motion events describe the specific body-part movements of each participant.

Conversational States. We derive temporally aligned conversational states from transcripts with word-level timestamps to delineate speaking and response periods as feature vectors, formulated as:

$$\mathcal{S} = \{s_{i,t} \mid i \in \{0,1\},\ t = 1, \ldots, T\}\,, \tag{2}$$

where $s_{i,t} \in \mathbb{R}^6$ denotes six states we pre-defined in the HOWDY dataset (*e.g.*, *speaking, brief response, lose in turn, etc.*). i means the interlocutor ID and T is the motion sequence length.

Dyadic Relations. Once we obtain the conversational state indicators, we leverage the annotated attributes of dyadic relations in the dataset to model the response pattern by LLM fine-tuning. Given the speech transcripts with accurate timestamps, we adopt the LoRA Hu et al. (2021) technique on Qwen3-4B Yang et al. (2025) to predict corresponding relations. The predicted relation results are then converted into relational descriptions, providing semantic cues (*e.g.*, a tense atmosphere, mutual attention, and a competitive stance) for motion-event inference.

Motion Events. To further organize the temporal and relational cues into concrete actions for dyadic interaction, we define the motion event as a specific body-part movement (*e.g.*, head nodding or wrist raising) and its temporal interval. As illustrated in Figure 3, we first retrieve static posture descriptions for both interlocutors from a single HOWDY clip matched by dialogue content and predicted relations. These descriptions provide posture priors grounded in observed dyadic interactions, retaining the body configurations of both participants as context for subsequent motion-event inference. The same

fine-tuned LLM takes the static descriptions, dialogue transcripts, and predicted relations as input to infer appropriate motion events for both interlocutors. The static posture descriptions and inferred dynamic motion events are then encoded as complementary semantic conditions for PSR to guide the generation of corresponding body-part motions.

4.3 TEMPORAL-RELATIONAL INTERACTOR

To ensure spatiotemporal coherence while preserving asynchronous response dynamics, we propose a Temporal-Relational Interactor (TRI) to capture cross-individual motion dependencies, as shown in Figure 3. Concretely, we model dynamic dependencies between the motion of each participant and that of the dialogue partner across body regions at the current and preceding frames. Each interlocutor's motion features are represented as $\mathbf{H}_i \in \mathbb{R}^{T\times G\times D}$, where G is the number of body-part groups and D is the feature dimension. Taking one interlocutor as an example, we use the current body-part features as queries and the counterpart features across all body parts within a causal temporal window as keys and values. A learnable temporal bias for modeling response delay adjusts attention across relative time offsets to account for asynchronous responses. The aggregation is expressed as:

$$\mathbf{Z}_t = \mathrm{Softmax}\left(\frac{(\mathbf{H}_t\mathbf{W}_Q)(\mathbf{H}_{\mathcal{W}_t}\mathbf{W}_K)'}{\sqrt{d}} + \mathbf{B}_t\right)(\mathbf{H}_{\mathcal{W}_t}\mathbf{W}_V), \tag{3}$$

where $\mathbf{H}_t$ and $\mathbf{H}_{\mathcal{W}_t}$ denote the current body-part features and stacked counterpart features within causal window, respectively. $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable projection matrices, and d is the feature dimension of an attention head. Empirically, the causal window is defined as $\mathcal{W}_t = \{\max(1, t-L), \ldots, t\}$, where L denotes the maximum number of preceding frames included in the window. $\mathbf{B}_t$ denotes the bias matrix of each counterpart frame from the current one, temporally. In this fashion, the model promotes temporally coordinated actions and responses in dyadic interaction.

Considering that the responsiveness of each interlocutor to the counterpart varies over time with their "**`dyadic relations`**" and "**`conversational states`**", we fuse relation embeddings with frame-wise state encoding to predict an adaptive gate. The gate regulates the contribution of the aggregated counterpart features through residual learning:

$$\widehat{\mathbf{H}}_t = \mathbf{H}_t + \alpha_t\overline{\mathbf{Z}}_t, \tag{4}$$

where $\widehat{\mathbf{H}}_t$ denotes the updated representation, $\overline{\mathbf{Z}}_t$ is the combined multi-head output and α_t is the corresponding interaction gate. The upgrades of both interlocutors follow a synchronized manner.

4.4 PART-ALIGNED SEMANTIC ROUTING

Considering the distinct motion patterns of different body regions during interaction, we propose a Part-Aligned Semantic Routing (PSR) module to ensure the generated body-part display semantic consistency with the given paired motion descriptions. As shown in Figure 3, group-specific encoders extract corresponding body-part motion features. To encode body-part identities and temporal positions in the motion representation, we incorporate body-part semantic embeddings (*e.g.*, features obtained by encoding "head" with a text encoder) and temporal positional encodings, respectively. The result is expressed as:

$$\mathbf{F} = \{\mathbf{f}_{t,g} \mid t = 1, \ldots, T,\ g = 1, \ldots, G\} \in \mathbb{R}^{T\times G\times D}, \tag{5}$$

where $\mathbf{f}_{t,g}$ denotes the D-dimensional feature of body-part group g at frame t, T is the sequence length, and G is the number of body-part groups. The static and dynamic semantic conditions are then injected into the corresponding body-part representations.

To associate each body region with relevant motion semantics, we introduce Group Routing Cross-Attention (GRCA). We utilize the motion features of each group as the query (Q), while the key (K) and value (V) are derived from the corresponding semantic features. Groups within the same semantic category share text inputs but use group-specific projections for Q, K, and V to extract part-specific information. The semantic aggregation is formulated as:

$$\mathbf{Z}_g = \mathrm{GRCA}_g(\mathbf{F}_g, \mathbf{E}), \tag{6}$$

where $\mathbf{F}_g$ denotes the motion feature sequence of group g, and $\mathbf{E}$ represents the corresponding text features from its assigned body, hand, or head/face category. GRCA_g denotes group routing cross-attention with projections specific to group g. In this fashion, the routed semantic features enable fine-grained control over body-part dynamics during holistic motion generation.

4.5 Training Objective

During the training phase, our denoiser concurrently reconstructs the clean motions of both interlocutors from noisy sequences, conditioned on their speech and interaction cues. Similar to prior work Qi et al. (2025); Chen et al. (2024), we supervise 6D joint rotations and FLAME expression coefficients with squared-error reconstruction losses $\mathcal{L}_{\text{body}}$ and $\mathcal{L}_{\text{expr}}$, respectively. We further incorporate velocity losses $\mathcal{L}_{\text{body-vel}}$ and $\mathcal{L}_{\text{expr-vel}}$ Tevet et al. (2022)to provide supervision on motion smoothness. The overall objective is formulated as:

$$\mathcal{L} = \lambda_b \mathcal{L}_{\text{body}} + \lambda_e \mathcal{L}_{\text{expr}} + \lambda_{bv} \mathcal{L}_{\text{body-vel}} + \lambda_{ev} \mathcal{L}_{\text{expr-vel}}, \tag{7}$$

where λ_b, λ_e, λ_{ev} and λ_{bv} are trade-off coefficients.

5 Experiments

5.1 Dataset and experimental setting

HOWDY Dataset. Since existing conversational motion datasets rarely provide aligned body-part descriptions together with hierarchical interaction annotations, we evaluate our approach on HOWDY. We construct HOWDY from 1,193 videos through extensive data processing and annotation. The dataset is split into training, validation, and test sets with an 80%/10%/10% ratio, yielding a total of 42,031 dyadic clips. More details are in the appendix.

Implementation Details. For interaction-chain inference, we use Qwen3-4B Yang et al. (2025) fine-tuned for 2 epochs using LoRA Hu et al. (2021) with a rank of 32. Audio features combine wav2vec 2.0 Baevski et al. (2020) features and mel-spectrograms, while a frozen FLAN-T5-base Chung et al. (2024) encodes the text. We train the motion generator for 300 epochs using AdamW with an initial learning rate of 10^{-5} and a batch size of 32. The loss weights are set to $\{\lambda_b, \lambda_e, \lambda_{bv}, \lambda_{ev}\} = \{10, 0.4, 1, 0.1\}$, with velocity weights gradually increased from zero. Training uses 1,000 diffusion timesteps with a cosine noise schedule, and inference uses DDIM Song et al. (2020a) sampling with 50 denoising steps. All experiments are conducted on 8 NVIDIA RTX 4090 GPUs. Our DyaChain concurrently synthesizes 90-frame sequences of body movements, hand gestures, and facial expressions for both interlocutors at 15 FPS. Each interlocutor's output is represented as $\mathbb{R}^{90\times326}$, comprising 46 joint rotations in the 6D representation Zhou et al. (2019) and 50-dimensional FLAME Li et al. (2017) expression coefficients per frame.

Evaluation Metrics. To evaluate the realism, rhythmic consistency, and diversity of the generated holistic motions, we adopt the following metrics: ❶**FD:** Fréchet Distance (FD) is calculated as the distribution distance between generated and real motions for gestures (FGD) Yoon et al. (2020), facial expressions (FED), and holistic motions (FHD). Pairwise FD (P-FD) and Single-person FD (S-FD) assess dyadic and individual quality, respectively. ❷**MSE:** Mean Squared Error (MSE) evaluates facial reconstruction accuracy between vertices of meshes. ❸**BC:** Beat Consistency (BC) Liu et al. (2022a; 2024)measures the rhythmic alignment between generated movements and the input speech. ❹**Div:** Diversity (Div)Qi et al. (2025); Zhu et al. (2023)measures motion variability based on the feature-space distance of extracted motion features. More details are in the appendix.

5.2 Quantitative Results

Comparisons with SOTA Methods. To evaluate the effectiveness of our framework, we retrain various state-of-the-art (SOTA) methods on our HOWDY dataset, covering single-person co-speech gesture generation(*i.e.*, SemGes (Liu et al., 2025a), GestureLSM (Liu et al., 2025b), GlobalDiff Zhang et al. (2026)) and conversational motion generation (*i.e.*, ConvoFusion (Mughal et al., 2024), Co^{3}Gesture (Qi et al., 2025), DyaDiT (Peng et al., 2026)). For fair comparisons, all competing methods are implemented using their official code and retain their original audio encoders. The remaining motion generation components are trained from scratch on HOWDY, with the output layers adapted to predict the holistic motion representation for both interlocutors.

As reported in Table 1, we compare methods in terms of gesture, facial expression, and holistic motion quality. Our DyaChain achieves substantial improvements over competing methods across all

Table 1: Comparison with state-of-the-art methods on gesture and face generation. ↓ indicates lower is better, ↑ indicates higher is better, and → indicates closer to GT is better. The dotted line separates methods adapted from single-person co-speech gesture generation and conversational ones.

	Gesture				Face				Holistic	
Methods	**P-FGD↓**	**S-FGD↓**	**BC↑**	**Div→**	**P-FED↓**	**S-FED↓**	**MSE↓**	**Div→**	**P-FHD↓**	**S-FHD↓**
GT	–	–	–	143.942	–	–	–	786.462	–	–
SemGes Liu et al. (2025a) *ICCV*	3.936	3.178	0.589	107.845	8.239	3.253	10.663	583.897	14.596	7.536
GestureLSM Liu et al. (2025b) *ICCV*	1.443	2.890	0.605	133.454	5.171	0.921	10.913	713.333	10.024	5.109
GlobalDiff Zhang et al. (2026) *AAAI*	1.893	1.293	0.598	137.489	5.228	1.594	11.793	1035.739	17.647	8.497
ConvoFusion Mughal et al. (2024) *CVPR*	1.198	1.231	0.595	134.336	6.693	1.118	9.227	772.931	11.825	4.079
Co^{3}Gesture Qi et al. (2025) *ICLR*	1.158	1.524	0.603	138.976	7.879	3.672	9.516	707.410	10.055	5.198
DyaDiT Peng et al. (2026) *CVPR*	1.207	1.738	0.592	151.374	8.104	2.773	9.080	819.509	11.321	5.352
DyaChain (Ours)	**0.801**	**1.080**	**0.609**	**140.316**	**4.017**	**0.850**	**8.978**	**784.649**	**7.447**	**2.919**

evaluation metrics, with particularly strong gains in pairwise Fréchet distance (P-FD), highlighting its superiority in motion quality and interaction. This is attributed to our progressive interaction chain, which guides TRI in modeling coherent interactions between interlocutors. Besides, in terms of FHD, our method outperforms other counterparts. This demonstrates that PSR effectively aligns semantic cues with body-part dynamics, enhancing the quality of holistic motion generation.

Ablation Study. To further evaluate the effectiveness of DyaChain, we conduct extensive ablation studies on different components and experimental settings.

Effects of the PSR and TRI: To verify the effectiveness of PSR and TRI, we conduct detailed experiments as reported in Table 2. Specifically, we use global text to condition all motion groups, replacing the original part-aligned semantic routing. Subsequently, we replace TRI with full-sequence cross-attention to model interactions between interlocutors. We also remove the gating based on relations and states from TRI while retaining its temporal window and relative temporal bias. Replacing part-aligned routing with global text conditioning increases both P-FED and P-FHD, despite retaining textual guidance. This suggests that directing relevant semantic cues to individual body parts enables the model to capture their distinct motion patterns. The degradation becomes more pronounced when text conditioning is removed entirely, further highlighting the importance of semantic guidance. Moreover, compared with full-sequence cross-attention, TRI's temporal modeling lowers P-FGD and P-FHD, suggesting that explicitly modeling temporal dependencies benefits dyadic motion generation. Introducing gating further improves both metrics, indicating that conversational states and dyadic relations effectively regulate cross-person information exchange.

Table 2: Ablation study of PSR, TRI, and their internal mechanisms on our HOWDY dataset.

Methods	P-FGD ↓	Div →	P-FED ↓	P-FHD ↓
w/o PSR	1.149	138.702	6.168	10.920
w/o Routing	1.026	133.171	4.423	9.663
w/o TRI	1.084	136.989	4.219	9.307
w/o Gate	0.926	136.702	4.071	8.698
DyaChain (Full)	**0.801**	**140.316**	**4.017**	**7.447**

Effects of the Progressive Interaction Chain: We further evaluate the effectiveness of our progressive interaction chain under different inference settings and use GT annotations as an oracle reference. Specifically, we evaluate three variants: replacing retrieved static descriptions with randomly sampled pairs from the same training set, using motion events from the retrieved clip, and directly inferring events without predicted dyadic relations. As shown in Table 3, the full chain outperforms all three variants. Random static descriptions cause the largest degradation, highlighting the importance of matching posture priors to conversational context. The decline under full retrieval suggests that events from similar dialogues may not match the current speaking turns and response timing, while direct event inference further demonstrates the value of relation

Table 3: Ablation study of the Progressive Interaction Chain on our HOWDY dataset.

Methods	P-FGD ↓	Div →	P-FED ↓	P-FHD ↓
w/ GT text	0.688	142.339	3.074	5.551
w/ Random text	1.404	132.229	5.235	10.255
w/ Full retrieval	1.029	136.713	4.324	8.360
w/ Direct event inference	0.945	134.326	4.182	7.926
DyaChain (Full)	**0.801**	**140.316**	**4.017**	**7.447**

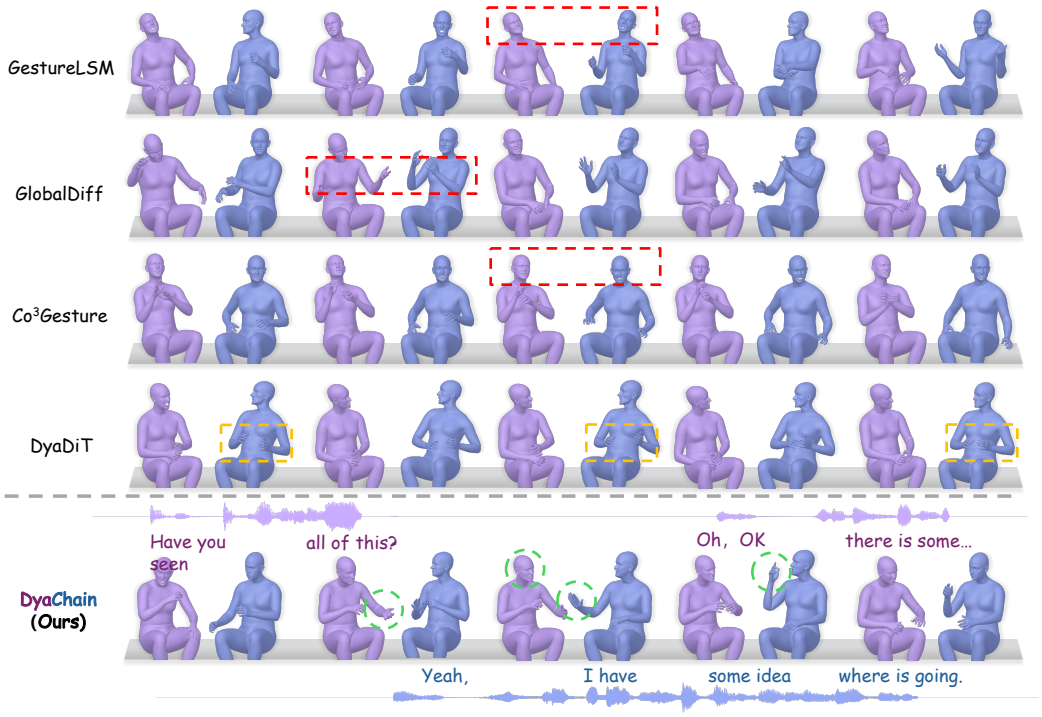


Figure 4: Visualization of our generated concurrent holistic motion against various **SOTA** methods.

guidance. The full chain also produces diversity closer to the ground truth, supporting both motion realism and variation. More ablations are in the Appendix.

5.3 Qualitative Results

To qualitatively evaluate the realism and interaction coherence of the generated holistic motions, we compare DyaChain with representative single-person and conversational motion generation methods on HOWDY. Figure 4 shows key frames sampled from motion sequences generated by different methods for the same conversation. Specifically, GestureLSM and GlobalDiff produce plausible individual movements, but the two interlocutors appear to move independently rather than engage in a coordinated exchange (as illustrated by the red boxes). Co^{3}Gesture generates more coherent interactions between the interlocutors. DyaDiT also captures their mutual attention, yet its hand movements remain limited in amplitude (as depicted in the yellow boxes), restricting the expressiveness of the interaction. In contrast, our DyaChain generates expressive and coherent holistic motions. The green circles highlight an open-hand gesture accompanying "**`all of this`**" and the partner's raised index finger during "**`some idea`**". The questioner also turns toward the respondent, integrating speech-aligned gestures and attentive listening into a coordinated exchange. For more visualizations, please refer to our demo page and appendix.

6 Conclusion

In this paper, we propose DyaChain, a framework for concurrently generating holistic motions for both interlocutors from conversational speech. Meanwhile, we introduce HOWDY, a dataset containing 42,031 dyadic clips with aligned body-part descriptions and hierarchical interaction annotations. Specifically, we establish a Progressive Interaction Chain that provides hierarchical interaction guidance for holistic motion generation. To preserve interaction coherence while aligning body-part movements with semantic descriptions, we design the Temporal-Relational Interactor and Part-Aligned Semantic Routing. Extensive experiments conducted on HOWDY show the superiority of our model. In the future, we will explore complex conversations involving multiple participants across diverse environments.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets and to clean and reformat datasets. We did not use generative AI tools for other required tasks; relevant details are provided in the main paper and appendix. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 7124–7128. IEEE, 2020.

Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, et al. Digital life project: Autonomous 3d characters with social intelligence. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 582–592. IEEE, 2024.

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7352–7361. IEEE, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20311–20322, 2022.

Aaron Ferguson, Ahmed AA Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, et al. Mhr: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025.

Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th international conference on intelligent virtual agents*, pp. 93–98, 2018.

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. Springer, 2024.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3492–3501. IEEE, 2019.

Leo Ho, Yinghao Huang, Dafei Qin, Mingyi Shi, Wangpok Tse, Wei Liu, Junichi Yamagishi, and Taku Komura. Interact: A large-scale dataset of dynamic, expressive and interactive activities between two people in daily scenarios. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(4):1–27, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, and Sung Ju Hwang. Avatar forcing: Real-time interactive head avatar generation for natural conversation. *arXiv preprint arXiv:2601.00664*, 2026.

Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 763–772, 2019.

Seungeun Lee, SeungJun Moon, Hah Min Lew, Ji-Su Kang, and Gyeong-Moon Park. Audioavatar: Personalized audio-driven whole-body talking avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3998–4010, 2026.

Shiying Li, Xingqun Qi, Bingkun Yang, Weile Chen, Zezhao Tian, Muyi Sun, Qifeng Liu, Man Zhang, and Zhenan Sun. Vividlistener: Expressive and controllable listener dynamics modeling for multi-modal responsive interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6396–6404, 2026.

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*, volume 2024, pp. 810–829, 2024.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

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.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pp. 612–630. Springer, 2022a.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1144–1154. IEEE, 2024.

Lanmiao Liu, Esam Ghaleb, Asli Ozyurek, and Zerrin Yumak. Semges: Semantics-aware co-speech gesture generation using semantic coherence and relevance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13963–13973, 2025a.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10929–10939. IEEE, 2025b.

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10452–10462. IEEE, 2022b.

M Hamza Mughal, Rishabh Dabral, Vera Demberg, and Christian Theobalt. Miburi: Towards expressive interactive gesture synthesis. *arXiv preprint arXiv:2603.03282*, 2026.

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1388–1398. IEEE, 2024.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20363–20373. IEEE, 2022.

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1001–1010. IEEE, 2024.

OpenAI. GPT-5.6 Luna. `https://developers.openai.com/api/docs/models/gpt-5.6-luna`, 2026. Large language model. Accessed: 2026-09-19.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Yichen Peng, Jyun-Ting Song, Siyeol Jung, Ruofan Liu, Haiyang Liu, Xuangeng Chu, Ruicong Liu, Erwin Wu, Hideki Koike, and Kris Kitani. Dyadit: A multi-modal diffusion transformer for socially favorable dyadic gesture generation. *arXiv preprint arXiv:2602.23165*, 2026.

Xingqun Qi, Yatian Wang, Hengyuan Zhang, Jiahao Pan, Wei Xue, Shanghang Zhang, Wenhan Luo, Qifeng Liu, and Yike Guo. Co^{3}gesture: Towards coherent concurrent co-speech 3d gesture generation with interactive diffusion. In *International Conference on Learning Representations*, volume 2025, pp. 31814–31832, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2in: Leveraging individual information to generate human interactions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1941–1951. IEEE, 2024.

Mengyi Shan, Shouchieh Chang, Ziqian Bai, Shichen Liu, Yinda Zhang, Luchuan Song, Rohit Pandey, Sean Fanello, and Zeng Huang. Talking together: Synthesizing co-located 3d conversations from audio. *arXiv preprint arXiv:2603.08674*, 2026.

Gayoung Son and Marius Rubo. Social virtual reality elicits natural interaction behavior with self-similar and generic avatars. *International Journal of Human-Computer Studies*, 199:103488, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020b.

Mingze Sun, Chao Xu, Xinyu Jiang, Yang Liu, Baigui Sun, and Ruqi Huang. Beyond talking–generating holistic 3d human dyadic motion for communication. *International Journal of Computer Vision*, 133(5):2910–2926, 2025.

Qwen Team. Qwen3. 6-27b: Flagship-level coding in a 27b dense model, 2026a.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026b. URL `https://qwen.ai/blog?id=qwen3.5`.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xitong Yang, Devansh Kukreja, Don Pinkus, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jia-Wei Liu, Nicolás Ugrinovic, Anushka Sagar, et al. Sam 3d body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7209–7219, 2026.

Payam Jome Yazdian, Rachel Lagasse, Hamid Mohammadi, Eric Liu, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 21574–21581. IEEE, 2025.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 469–480. IEEE, 2023.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

Xiangyue Zhang, Jianfang Li, Jianqiang Ren, and Jiaxu Zhang. Mitigating error accumulation in co-speech motion generation via global rotation diffusion and multi-level constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12834–12842, 2026.

Weiyu Zhao, Chenyang Wang, Liangxiao Hu, Zonglin Li, Wei Yu, and Shengping Zhang. Dialogen: Towards dialog gesture generation via identity-decoupled style guidance in interactive diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13253–13261, 2026.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746. IEEE, 2019.

Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10544–10553. IEEE, 2023.

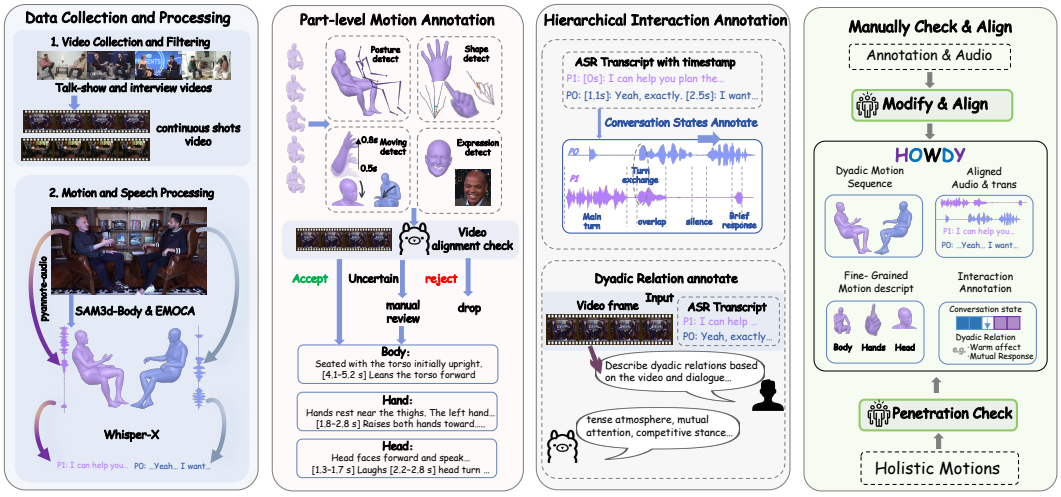


Figure 5: **Overview of the HOWDY dataset construction pipeline.** We process conversational videos to obtain high-quality holistic motions and corresponding annotations.

A Appendix

A.1 Details of the HOWDY Dataset Construction

In this section, we describe the construction of HOWDY in detail. We first introduce the data collection and processing pipeline, followed by the annotation protocol and dataset statistics.

A.1.1 Data Collection and Processing

Video Collection and Filtering. HOWDY is constructed from 1,193 talk-show and interview videos retained after screening. We select English-language videos featuring real people, with a minimum resolution of 480p. The selected footage contains two visible interlocutors with clearly observable upper-body movements, while non-conversational scenes, severe occlusions, and speech from off-screen speakers are excluded. Since the source videos may contain camera cuts, we segment them into continuous shots before pose estimation. Specifically, we use PySceneDetect to locate shot boundaries, allowing motion recovery to be performed within each shot without spanning camera transitions. This step yields 10,044 continuous shots for subsequent motion and speech processing.

Motion and Speech Processing. Within each continuous shot, we first use YOLOv8 to detect the body and face regions of both interlocutors. We recover body and hand motions at the original video frame rate using SAM 3D Body Yang et al. (2026), which represents the human body using the Momentum Human Rig (MHR) Ferguson et al. (2025). The recovered MHR representations are then converted to SMPL-X Pavlakos et al. (2019) using the official conversion tool. To maintain consistent participant identities throughout each shot, we associate the reconstructed poses with the two interlocutors based on their root-joint positions. For facial motion, we apply EMOCA Daněček et al. (2022) to the detected face regions to estimate facial expression coefficients and jaw poses. These estimates replace the corresponding expression and jaw parameters in the converted SMPL-X sequences, combining body and hand reconstruction with facial motion recovery in a unified representation. We preserve each participant's estimated body shape and the reconstructed relative positions of the two interlocutors. For speech processing, we use pyannote-audio Bredin et al. (2020) for human voice separation and manually associate each speaker with the corresponding interlocutor in the video. We then use Whisper large-v3 Radford et al. (2023) for speech recognition and WhisperX Bain et al. (2023) for word-level temporal alignment. The audio is stored at 16,000 Hz and aligned with the transcripts and recovered motions on a shared timeline.

Quality Control. We combine automatic outlier detection with manual inspection to assess the quality of the recovered motions. The automatic checks flag anomalous motion parameters, followed by manual screening for body or hand reconstruction failures, pronounced jitter, and unnatural poses. Identity switches between the two interlocutors are corrected to maintain consistent identities within

each shot. Sequences with extensive motion errors are discarded, while brief anomalies, such as short periods of jitter, are manually corrected. We also exclude segments with unclear audio, failed transcriptions, or speech from off-screen speakers.

A.1.2 Annotation Protocol

Part-level Motion Annotation. For each six-second clip, we independently derive body, hand, and head/face annotations for both interlocutors, primarily from the reconstructed motions and supplemented by video semantics. Building on MotionScript Yazdian et al. (2025), we adapt its geometric rules to extract dominant states and temporal events from SMPL-X joint positions and rotations. Sitting and standing are distinguished by knee angles and thigh orientations, while hand locations are determined in a body-local coordinate system and normalized by torso length and shoulder width. Hand shapes are characterized by finger extension, flexion relative to the relaxed hand pose, and finger spread, including differences across fingers. Static descriptions summarize prevailing postures, hand configurations, and head orientations through frame-wise statistics, while motion events capture local changes and their time intervals. Hand trajectories and head orientations are analyzed relative to the torso to distinguish local movements from overall body motion. Qwen3.5 Team (2026b) provides supplementary facial semantics from sampled video frames, while body and hand motion labels remain grounded in geometric analysis. We use GPT-5.6-luna to review candidate events against timestamped frames from the original videos, checking the visible motion, participant identity, and temporal extent. Uncertain cases are further reviewed by human annotators, and rejected candidates are excluded from the final annotations. After merging adjacent intervals and resolving conflicting labels, we use fixed templates to combine the states and events into structured descriptions for each body part.

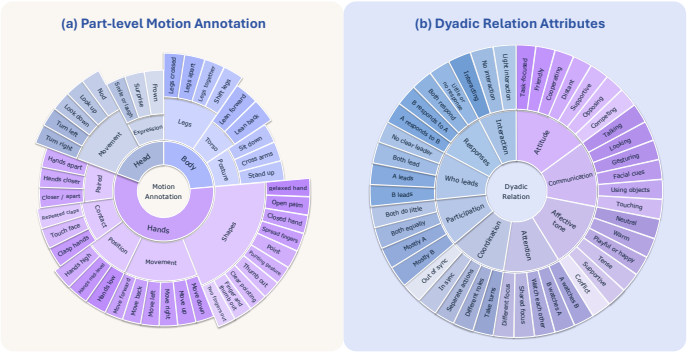


Figure 6: Sample illustrations of the HOWDY dataset and the multiple types of annotations.

Hierarchical Interaction Annotation. For each six-second clip, we annotate conversational states at the frame level and dyadic relations at the clip level. Conversational states are derived from participant-aligned transcripts and WhisperX word-level alignments using predefined rules. Brief responses must match a feedback lexicon, contain at most three words, last at most 1.2 s, and have a mean word-alignment score of at least 0.55. They must also overlap the counterpart's speech or occur within 0.45 s of its boundary. Turn acquisition and loss are represented by triangular soft signals over five frames centered on a main-speaker change at 15 fps; brief responses alone do not trigger a turn change. The speech roles are mutually exclusive for each participant at each frame, whereas turn-transition signals can overlap with these roles. For dyadic relations, we provide Qwen3.6-27B-FP8 with video frames sampled at 2 fps and participant-aligned transcripts containing speech intervals. The model assigns labels to nine predefined relation attributes using fixed label sets and attribute-specific decision criteria. The model combines verbal cues with visible gaze, facial expressions, gestures, and coordinated actions to assess how the two interlocutors engage with one another. For example, participation balance is judged from speaking time, turn-taking, and visible nonverbal involvement, while initiative concerns who initiates or drives the interaction within the clip. Responsiveness includes both spoken replies and observable nonverbal reactions, such as nods, smiles, and gaze shifts. The generated relation labels are subsequently reviewed by human annotators. Examples of the resulting motion and interaction annotations are shown in Figure 6.

A.1.3 Dataset Statistics and Splits

We first partition the continuous shots into training, validation, and test sets, and then segment the processed sequences within each set into six-second clips with a three-second stride. All clips derived from the same continuous shot are assigned to the same split, preventing data leakage caused by

overlapping temporal segments. The resulting dataset contains 42,031 dyadic clips, with 33,566 for training, 4,202 for validation, and 4,263 for testing, corresponding to an 80%/10%/10% split.

A.2 More Details of Experiments

A.2.1 Evaluation Metrics

To evaluate the realism, rhythmic consistency, and diversity of the generated holistic motions, we adopt the following metrics:

- **FD:** Fréchet Distance (FD) is calculated as the distribution distance between generated and real motions for gestures (FGD) Yoon et al. (2020), facial expressions (FED), and holistic motions (FHD). FGD and FHD use pre-trained autoencoder features, while FED uses FLAME coefficients. Pairwise FD (P-FD) and Single-person FD (S-FD) assess dyadic and individual quality, respectively.
- **MSE:** Mean Squared Error (MSE) evaluates facial reconstruction accuracy by measuring the mean squared difference between corresponding vertex coordinates of meshes reconstructed from generated and ground-truth facial parameters. The MSE reported in the table is scaled by 10^{-7}.
- **BC:** Beat Consistency (BC) Liu et al. (2022a; 2024) measures the rhythmic alignment between generated movements and the input speech. We report the average score of the two interlocutors.
- **Div:** Following previous work Qi et al. (2025); Zhu et al. (2023), Diversity (Div) is measured as the average feature distance between randomly sampled pairs of generated sequences, using motion embeddings extracted by the FD autoencoder.

A.2.2 Progressive Interaction Chain

Joint Fine-tuning. We fine-tune Qwen3-4B Yang et al. (2025) using LoRA Hu et al. (2021), with rank 32 and scaling parameter $\alpha = 64$. Relation prediction and motion-event inference share the same model and LoRA parameters, with distinct task prompts. Each training clip provides two examples: one maps the conversation to the annotated nine-field relation JSON, while the other predicts their motion events from the dialogue, predicted relations, and paired static descriptions. The event output contains five channels for each interlocutor: body motion, hand movement, hand shape, facial expression, and head motion, with temporal intervals. The two tasks contribute 33,566 examples each, yielding 67,132 examples mixed at a 1:1 ratio. We compute the generation loss only on response tokens and train for two epochs, using the final checkpoint at step 4,196.

Training Conditions. To expose event inference to relation-prediction errors, we construct its relation conditions through two-fold cross-fitting: each auxiliary model is trained on one fold and predicts relations for the other. Static conditions come from the current clip's annotations for approximately half of the event-training examples and from a retrieved training clip for the remainder. For examples with retrieved conditions, each static-description category for each interlocutor is independently marked as unavailable with probability 0.1. In both cases, the supervision consists of the current clip's annotated events.

Static Description Retrieval. We use a frozen `all-mpnet-base-v2` encoder Song et al. (2020b) to encode each interlocutor's transcript separately. The semantic score averages the cosine similarities between corresponding interlocutors, excluding silent query participants. We retrieve the top 50 training clips and rerank them by combining semantic similarity, speaking-pattern compatibility, and dyadic relation consistency. Speaking-pattern compatibility accounts for speech-structure buckets based on speaking durations, speaker dominance, and segment counts, together with differences in the two participants' speaking durations and segment counts. Relation consistency compares the query's predicted relations with the candidate's annotated relations, using exact matches for single-valued attributes and Jaccard similarity for multi-valued attributes, averaged over valid fields. Unknown fields are omitted from this average. We select the highest-scoring clip and retrieve both interlocutors' static descriptions together, without penalizing source-clip reuse.

Sequential Inference. The fine-tuned model is called twice: first to predict relations, and then to infer both interlocutors' timed events from the dialogue, predicted relations, and retrieved static descriptions. Relation prediction uses greedy decoding with a maximum of 384 new tokens. Event inference uses sampling with temperature 1.0, top-$p = 0.95$, and a maximum of 768 new tokens; input prompts are preserved without truncation. We apply format and temporal consistency checks to

Table 4: Motion groups and their corresponding text conditions.

Condition	Motion group	SMPL-X joints / parameters	Dim.
BODY	Torso	Spine: 3, 6, 9	18
	Shoulder girdle	Clavicles: 13, 14	12
HAND	Left arm	Shoulder, elbow, wrist: 16, 18, 20	18
	Right arm	Shoulder, elbow, wrist: 17, 19, 21	18
	Left fingers	Joints 25–39	90
	Right fingers	Joints 40–54	90
FACE	Head, neck, eyes	Joints 12, 15, 23, 24	24
	Jaw	Joint 22	6
	Expression	Complete expression vector	50
Total (10 motion tokens)			**326**

Notes. Joint indices are zero-based, and each rotation joint contributes six dimensions. The expression representation uses two motion tokens to encode all 50 coefficients.

the generated outputs. We combine the retrieved static descriptions with the predicted events to form part-level motion text, and retain the current predicted relations as relational text conditions for the motion generator.

A.2.3 IMPLEMENTATION DETAILS

We use an eight-layer Transformer with a hidden dimension of 1,280, comprising ten 128-dimensional motion tokens. The backbone uses 16-head attention, while TRI and PSR use four and two heads, respectively. The feed-forward dimension is 2,560, with GELU activation and dropout 0.1. Audio conditioning combines frame-aligned log-Mel features, Wav2Vec2 convolutional features, and final hidden states, projected to 128, 384, and 768 dimensions, respectively, and concatenated. Each six-second sample contains 90 motion frames at 15 fps. During training, motion descriptions and dyadic relations are taken from dataset annotations; text conditions are dropped with probability 0.2 for classifier-free guidance, while audio conditioning is always retained. Table 4 summarizes the motion groups, input dimensions, and corresponding text conditions.

A.3 USER STUDY

To assess the perceptual quality of generated dyadic motions, we conducted a user study with 20 participants using 20 randomly selected six-second conversations from the HOWDY test set. We compared DyaChain with six baseline methods through pairwise comparisons. In each comparison, participants viewed one dyadic motion video generated by DyaChain and one baseline from the same conversation. The videos were displayed side by side with a shared audio track, allowing participants to compare the motions under the same speech conditions.

Specifically, participants evaluated each pair along four dimensions: ❶**motion naturalness**, assessing whether facial expressions, hand gestures, and body movements appear plausible; ❷**motion smoothness**, assessing temporal continuity and the absence of jitter or abrupt transitions; ❸**interaction coherence**, assessing how well the two interlocutors coordinate their behaviors during speaking, listening, and turn-taking; and ❹**speech–motion appropriateness**, assessing how well the motions match the rhythm, affect, and content of the conversation. Participants selected one preferred result independently for each dimension, allowing different aspects of the same motion pair to be evaluated separately.

As shown in Figure 7, participants consistently preferred DyaChain over all six baselines across the four dimensions, with preference rates ranging from 60.6% to 80.9%. The preferences for naturalness and smoothness indicate that DyaChain produces more plausible and temporally continuous holistic motions. Meanwhile, the higher preference rates for interaction coherence and speech–motion appropriateness suggest that its generated motions better convey coordinated conversational exchanges

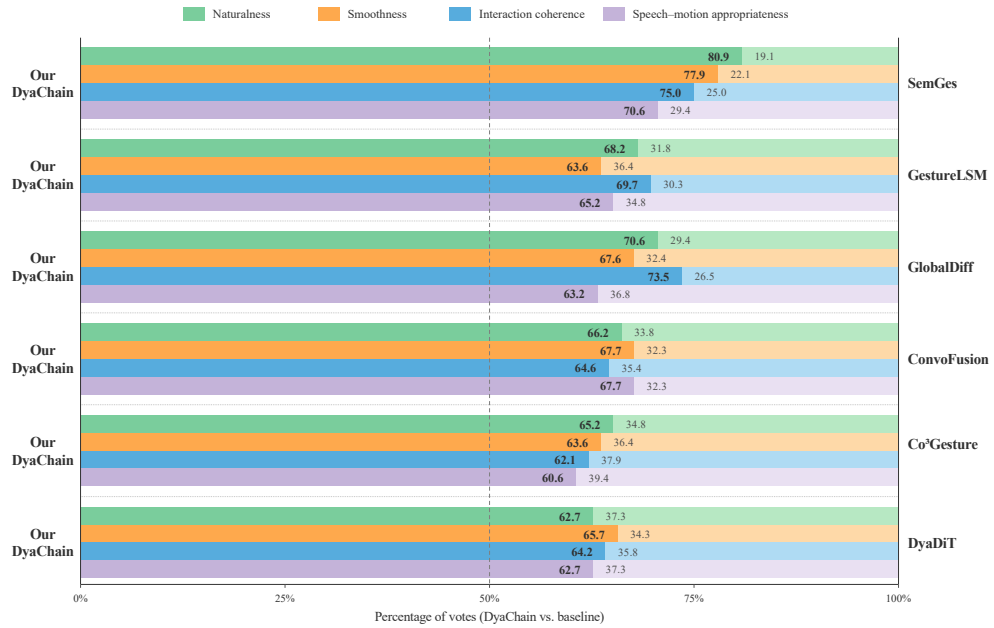


Figure 7: User study results showing the percentage of votes preferring DyaChain over each baseline in pairwise comparisons. Results are reported separately for motion naturalness, motion smoothness, interaction coherence, and speech–motion appropriateness.

and align with the accompanying speech. These results demonstrate improvements in both the perceived quality of holistic motions and the coherence of dyadic interactions.

A.4 Additional Experiments

A.4.1 Effects of the TRI Window Size

To examine the effect of temporal context, we vary TRI's history window among 8, 16, 24, and 32 frames while retaining its other components. As shown in Table 5, the 8-frame window gives poorer results, suggesting that limited access to the counterpart's preceding motions restricts interaction modeling. The default 24-frame window achieves the lowest P-FGD and P-FED, whereas further extending the window provides no additional improvement in these metrics. Although 16 frames yield slightly lower P-FHD and 32 frames yield higher diversity, the default setting offers a favorable balance across gesture, facial, and holistic motion quality.

Table 5: Effects of the TRI history window on HOWDY. L denotes the number of preceding frames; the current frame is also included.

Window	P-FGD ↓	Div →	P-FED ↓	P-FHD ↓
$L = 8$	0.994	137.766	4.137	8.463
$L = 16$	0.865	139.429	4.081	**7.393**
$L = 24$ **(Full)**	**0.801**	140.316	**4.017**	7.447
$L = 32$	0.920	**140.519**	4.068	7.454

A.4.2 Effects of Relation and State Conditions

To evaluate relation and state guidance, we independently remove TRI's gate, relation condition, or state condition while retaining the default temporal window and PSR. The latter two variants retain the gate with the remaining condition. As shown in Table 6, removing either condition worsens all three distributional distances, while removing the gate causes the largest degradation in P-FHD and diversity. These results suggest

Table 6: Ablation study of TRI gating and its relation and state conditions on HOWDY.

Methods	P-FGD ↓	Div →	P-FED ↓	P-FHD ↓
w/o Gate	0.926	136.702	4.071	8.698
w/o Relation	0.934	139.731	4.022	7.754
w/o State	0.864	139.553	4.035	7.558
DyaChain (Full)	**0.801**	**140.316**	**4.017**	**7.447**

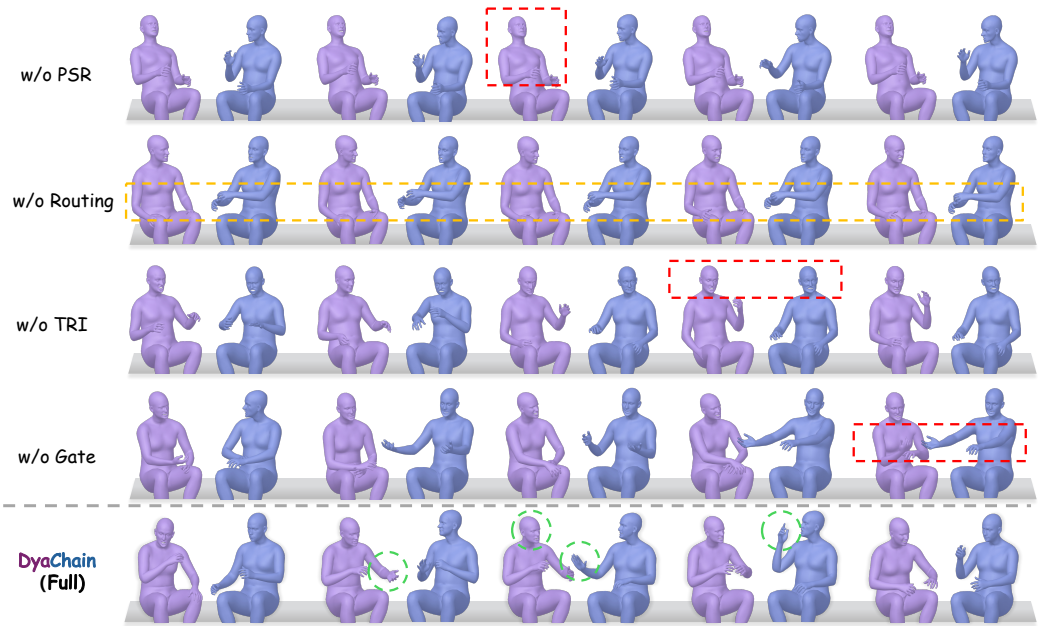


Figure 8: Visualization of our generated concurrent holistic motion in the ablation study.

that clip-level relations and frame-level states provide complementary cues for regulating cross-person information exchange, helping preserve both the realism and diversity of holistic motions.

A.5 Additional Qualitative Results

We provide additional visual comparisons between DyaChain and its ablated variants in Figure 8. The full model generates more realistic holistic motions with more coherent interactions between interlocutors, demonstrating the effectiveness of its key components. More detailed sequence visualizations are available on our demo page.

A.6 Additional Related Work

Human Interaction Modeling. A central challenge in interactive human motion modeling is capturing the mutual influence between participants' movements. Early work combines speaker audio and motion through cross-modal attention and autoregressively predicts listener facial reactions in a discrete motion space Ng et al. (2022). To capture richer body and hand interactions, ReMoS Ghosh et al. (2024) uses spatiotemporal cross-attention to model dependencies between participants' joints for reactive motion synthesis. Duolando Li et al. (2024) autoregressively predicts follower motion and relative translation conditioned on the leader's movements to coordinate the dancers' poses and positions. However, these methods generate responses to given partner behavior without jointly synthesizing both participants' motions. Joint generation methods couple both participants' motion streams through cooperative denoising and mutual attention Liang et al. (2024); Qi et al. (2025). Beyond Talking Sun et al. (2025) explores chain-like autoregressive coupling between speaker and listener motion tokens for holistic dyadic generation. To explicitly guide individual behaviors, in2IN Ruiz-Ponce et al. (2024) incorporates individual action descriptions and overall interaction descriptions into self-attention and cross-person attention, respectively. Dynamic speaking/listening role embeddings further modulate generation in conversational facial animation Shan et al. (2026). However, modeling dyadic conversational motion with explicit guidance from evolving interaction context remains underexplored. In this work, we propose DyaChain to model dyadic conversational motion by strategizing a Progressive Interaction Chain from conversational states through dyadic relations to motion events.