
Salsa as a Nonverbal Embodied Language— The CoMPAS3D Dataset and Benchmarks

Bermet Burkanova,* Payam Jome Yazdian*
 Chuxuan Zhang, Trinity Evans, Paige Tuttosí, Angelica Lim

School of Computing Science

Simon Fraser University

Burnaby, BC, Canada

{bba60, pjomeyaz, cza152, trinitye, ptuttosi, angelica}@sfu.ca

Abstract

Imagine a humanoid that can safely and creatively dance with a human, adapting to its partner’s proficiency, using haptic signaling as a primary form of communication. While today’s AI systems excel at text or voice-based interaction with large language models, human communication extends far beyond text—it includes embodied movement, timing, and physical coordination. Modeling coupled interaction between two agents poses a formidable challenge: it is continuous, bidirectionally reactive, and shaped by individual variation. We present CoMPAS3D, the largest and most diverse motion capture dataset of improvised salsa dancing, designed as a challenging testbed for interactive, expressive humanoid AI. The dataset includes 3 hours of leader-follower salsa dances performed by 18 dancers spanning beginner, intermediate, and professional skill levels. For the first time, we provide fine-grained salsa expert annotations, covering over 2,800 move segments, including move types, combinations, execution errors and stylistic elements. We draw analogies between partner dance communication and natural language, evaluating CoMPAS3D on two benchmark tasks for synthetic humans that parallel key problems in spoken language and dialogue processing: leader or follower generation with proficiency levels (speaker or listener synthesis), and duet (conversation) generation. Towards a long-term goal of partner dance with humans, we release the dataset, annotations, and code, along with a multitask SalsaAgent model capable of performing all benchmark tasks, alongside additional baselines to encourage research in socially interactive embodied AI and creative, expressive humanoid motion generation.

1 Introduction

Salsa is “arguably the world’s most popular partnered social dance form” [3], practiced globally across a wide range of skill levels and cultural contexts. Recent work in formal linguistics has increasingly suggested that structured, rule-governed communication extends beyond spoken and signed language to include modalities such as gesture, facial expression, music, and dance. As Patel-Grosz et al. [23] note, “formal linguistics may come to encompass aspects of human communication (such as gestures and facial expressions) that were traditionally left outside its purview, as well as non-linguistic systems such as animal communication, visual narratives, music and dance.” Building on this broader framing—and longstanding views of dance as nonverbal communication [12]—we propose that salsa duet improvisation can be usefully analyzed as an embodied language, complete with vocabulary, grammar, conversational dynamics, fluency levels, stylistic expression, and dialectical variation.

*Equal contribution

Given its global reach, improvisational structure, and established evaluation criteria, salsa offers an ideal starting point for embodied interaction benchmarks—serving a similar role as English in early spoken language-based model development.

Existing embodied interaction datasets typically focus on isolated or acted actions, such as “hug” or “handshake” [41], rather than capturing the continuous, adaptive flow of embodied dialogue. Dance motion datasets, in particular, often label entire sequences simply as “jive” or “samba,” offering little detail about the compositional structure of interaction. To the best of our knowledge, no motion capture partner dance resources provide frame-level annotations of moves, errors, or stylistic variation. In addition, most focus exclusively on professional performers, limiting their generalizability. In contrast, spoken language research benefits from diverse, spontaneous, and richly transcribed corpora like the Switchboard dataset [5]. This gap in embodied data matters: humans rely on in-person interactions [38]—including gesture, touch, synchrony, and shared attention—to feel truly connected. Embodied AI systems that overlook these cues risk appearing unresponsive or socially inappropriate in human-facing roles.

We present CoMPAS3D, a large-scale motion capture dataset of improvised salsa dancing that captures the richness of nonverbal social interaction. It includes leader-follower improvisation across three skill levels, with frame-level annotations for moves, errors, and stylistic elements. By framing salsa as an embodied language, CoMPAS3D opens new directions for modeling not just individual actions, but dynamic, context-sensitive, multimodal dialogue.

Our contributions are as follows:

- We introduce CoMPAS3D, the largest and most diverse motion capture dataset of improvised salsa dance available for machine learning applications, with over 3 hours of motion capture data, as a benchmark for nonverbal, embodied communication
- We provide fine-grained annotations spanning three levels of dancer proficiency (beginner, intermediate, professional), including move labels, styling variations, and execution errors—generated through over 120 hours of expert salsa annotation effort.
- We provide results on benchmark tasks for modeling embodied dialogue with 3D virtual humans, including solo dance generation and duet dance generation
- We release the dataset, annotations, and baseline code publicly to support research in embodied AI, nonverbal interaction, and socially interactive systems.

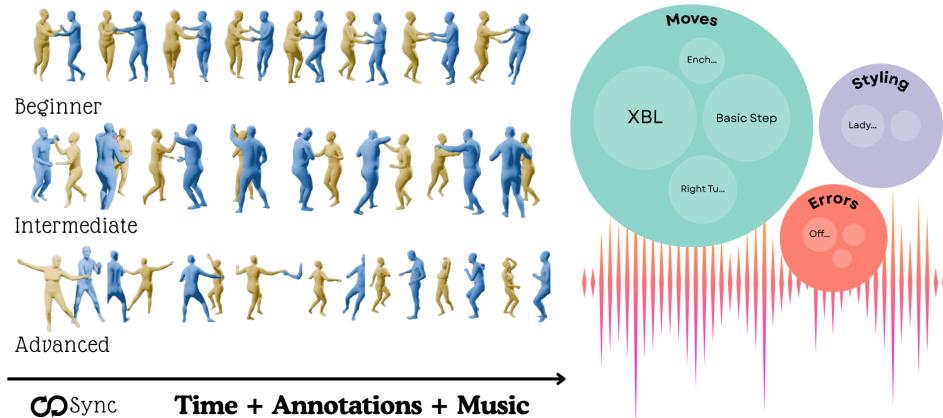


Figure 1: CoMPAS3D comprises 3 hours of improvised salsa dance with beginner (top), intermediate (middle) and professional (bottom) pairs with synchronised music and fine-grained annotations.

Table 1: Comparison of publicly available dance datasets capturing human-human interaction (HHI). \bar{T}/s represents the average duration per sequence in seconds. CoMPAS3D uniquely combines improvisation, multiple proficiency levels, long sequence durations, and fine-grained annotations.

Dataset	# Participants	Audio	Experience	\bar{T}/s	T	Improvised	Mocap	Annotated	Representation
DuetDance	Unknown	✓	Unknown	Varies	1.09h	✗	✗	✗	3D Skeletons
ReMoCap	9	✗	Pro	Varies	2.04h	✗	✗	✗	3D Skeletons
ExPI	4	✗	Pro	10.4	0.33h	✗	✓	✓	3D Meshes
InterHuman	30	✗	Pro	10	6.56h	✗	✓	✓	SMPL
InterDance	Unknown	✓	Pro	142.7	3.93h	✓	✓	✗	SMPL-X
DD100	10	✓	Pro	70.2	1.95h	✓	✓	✗	SMPL-X
CoMPAS3D	18	✓	Beg, Int., Pro	150	3.0h	✓	✓	✓	SMPL-X

2 Related Work

In this section, we review related work on human-human motion datasets, social interaction modeling, and dance datasets, highlighting the need for naturalistic, skill-diverse, and richly annotated resources such as CoMPAS3D.

Human-Human Motion Datasets. Several datasets have captured aspects of human-human interaction, typically focusing on short-term, scripted, or task-specific actions. Datasets such as NTU RGB+D 120 [17], SBU [13], and Inter-X [41] offer labeled interactions for action recognition, primarily covering isolated, repetitive activities like handshaking and hugging. Other datasets, including CHI3D [8], ShakeFive2 [36], and Hi4D [43], record close-proximity social interactions with annotated contact events, but remain limited to short, scripted encounters under controlled settings. Additionally, resources such as MuCo3DHP [19], MI-Motion [25], and the MultiHuman dataset focus on multi-person poses and static interactions, without capturing continuous improvisational dynamics. While these datasets provide valuable snapshots of social signals, they do not model the sustained, unscripted, and fluent interactions characteristic of embodied conversations. CoMPAS3D addresses this gap by capturing long-term, improvised duet dances across multiple skill levels, enabling the study of naturalistic nonverbal communication over extended timeframes.

Dance and Movement Datasets. Professional close-contact sports and couple dances represent a promising source of long-term physical interaction, offering structured movements with well-defined labels and scoring criteria. Several dance-specific datasets have been introduced, focusing primarily on choreographed performances by professional dancers. ExPI [11] captures Lindy Hop dancing actions with 3D body poses and shapes, while DuetDance [15] extracts 3D skeletons from YouTube videos of couple dances. ReMoCap [10] presents multi-view captures of Lindy Hop and Ninjutsu with 3D skeletons and RGB videos. InterHuman [26] includes sequences of martial arts and dance, alongside daily activities. More recently, InterDance [16] offers 3.93 hours of optical motion capture from professional duets across 15 genres, and DD100 [32] collects 117 minutes of music-synchronized SMPL-X data from five professional dance pairs. A closely related study is the work by Senecal et al. [29, 30, 28], which introduced a salsa dance motion capture dataset containing clips from beginner, intermediate, and advanced dancers. However, the dataset is not publicly available for machine learning research, and no detailed annotations were provided.

While these datasets offer valuable resources, they differ from real-world embodied communication in several key ways: they often rely on choreographed (i.e. acted) rather than spontaneous performances, capture only professional dancers rather than a diversity of skill levels, and lack fine-grained annotations of moves, errors, or styling. In addition, extracting accurate 3D skeletons from video is especially difficult in close-contact dances like salsa, where frequent occlusions and continuous physical contact lead to pose estimation errors. This limitation makes motion capture essential for precision, and also explains why existing duet datasets remain relatively small in scale.

As shown in Table 1, CoMPAS3D is distinguished by its improvised duet recordings, inclusion of beginner, intermediate, and professional dancers, longer average sequence durations. Notably, it is the first spontaneous dance dataset with frame-level annotations for moves, styling, and execution errors, important for benchmarking the legibility and correctness of generated motions.

Summary. Existing datasets for human-human interaction largely focus on short-term, scripted performances by professional participants, limiting their applicability for modeling naturalistic,

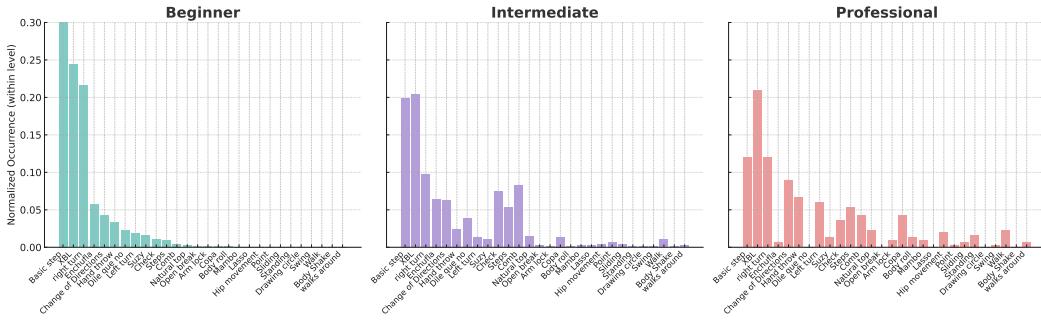


Figure 2: Distribution over the 30 move classes (sorted by beginner move frequency) in CoMPAS3D for beginner, intermediate and pro pairs. Beginners tend to primarily use the “basic step”, which professionals use less. Instead, pros use a wider variety of moves such as left turns and copa.

embodied communication. CoMPAS3D addresses these gaps by providing a large-scale, improvised duet dance dataset with multi-level proficiency variation and fine-grained annotations, offering a new foundation for studying nonverbal interaction, fluency, and style in embodied AI systems.

3 The CoMPAS3D Dataset

To support the study of improvised, naturalistic nonverbal communication in physical interactions, we introduce **CoMPAS3D** (Complex Multi-Level Person-Interaction Annotated Salsa Dataset)², a large-scale motion capture dataset of salsa duet dances. CoMPAS3D, *compas* meaning rhythm in Spanish, consists of over 3.0 hours of improvised leader-follower interactions performed by 18 participants spanning beginner, intermediate, and professional skill levels. Each recording captures long-duration sequences of continuous social improvisation, annotated at the frame level for move types, stylistic variations, and execution errors. The dataset includes synchronized audio recordings, high-fidelity 3D motion data and SMPL-X parametric body model fits [24], enabling detailed analysis and modeling of embodied conversational dynamics across skill levels.

Participants. CoMPAS3D includes 18 participants, forming 9 dancing pairs. Participants were recruited from a university salsa club, community dance groups, and professional dance schools. To capture variation in fluency and style, dancers self-reported their experience level as beginner, intermediate, or professional, based on years of training and social dance experience. This diversity enables the study of movement improvisation and fluency across a wide proficiency spectrum. This study was approved by the Research Ethics Board of Simon Fraser University in Canada. Each participant was compensated \$100 for 1 hour of study participation time and provided informed consent for their anonymized motion capture data release prior to data collection.

Collection Setup. Recordings were conducted in a controlled studio environment using a Vicon motion capture system equipped with 20 cameras operating at 120 frames per second. Each dancer wore 53 markers following the Vicon “FrontWaist” marker set. Improvisation sessions used four salsa music tracks (90–105 beats per minute) chosen to vary in mood and tempo. Each pair performed two improvised takes per song, each lasting approximately 2.5 minutes, resulting in a total of 72 sequences.

Data Representation. We release the dataset to facilitate a wide range of machine learning and animation applications. Each sequence includes 55-joint SMPL-X [24] human body joint trajectories and fitted parameters (.npz), as well as visualizations with synchronized music tracks (.mp4). We also provide ELAN annotation files (.txt) aligned frame-by-frame with the motion data.

Annotation. Approximately half of the recorded sequences (2803 segments) were annotated manually by an expert salsa dancer with 15 years of salsa dance experience and competition judging experience. Salsa moves are performed in 8-beat cycles, where the leader typically provides a signal in the early part of the cycle, and the follower completes the move by the end of the 8th beat. Therefore, each

²<https://huggingface.co/datasets/Rosie-Lab/compas3d>

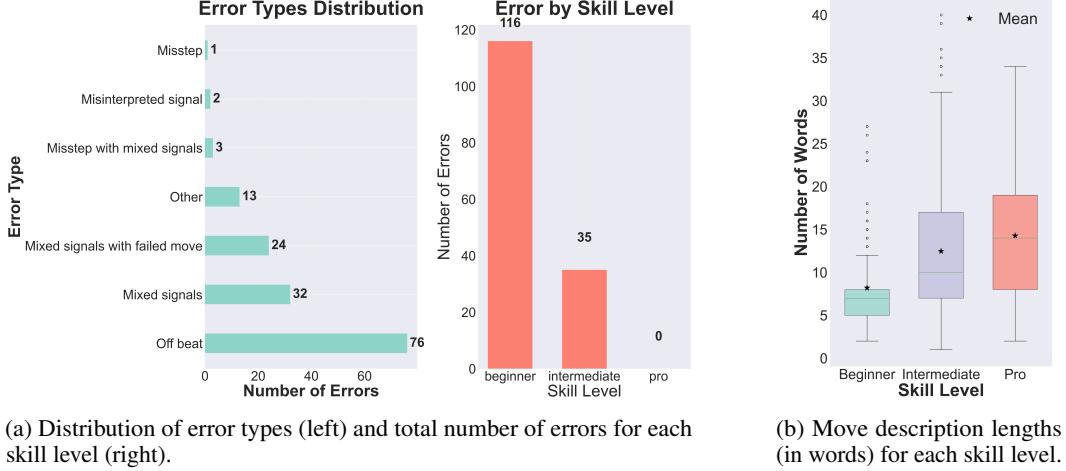


Figure 3: Analysis of errors and move description annotations across proficiency levels.

sequence was split into 8-beat segments and annotated. Each annotation contains a primary move category selected from among 30 move categories; these move categories are listed and explained in the Appendix. Annotations also include common execution errors (e.g., off-beat errors, mixed signals), and presence of styling (e.g., arm styling, hip accents, annotated as “lady styling” or “man styling”). In addition to each broad move category, a detailed description of the move, including hand holds and secondary combinations, is provided for each segment. This detailed annotation effort using the ELAN software [1] required over 120 hours. Half the sequences remain unannotated, offering a clean set for future evaluation.

Analysis. Analysis of the annotations reveal distinctions between the populations of dancers in our dataset. In Fig. 2, we compare the move distributions between beginner, intermediate and professional dancers. We notice that professionals employ a wider variety of moves and use fewer “basic steps”. In Fig. 3a, we notice that a common error is “off beat” suggesting that multimodal information including music is important in detecting errors. Another error is unclear signals from the leader resulting, in some cases, in a failed move. In Fig. 3b, we notice that professional descriptions have a longer average length (in words), suggesting a higher number of secondary moves and variations within the 8-beat segment. An analysis of the styling annotations show that professionals execute 54.5 styling moves per performance. This stands in sharp contrast with intermediate dancers, who perform 12.9 styling moves per performance, and beginners, who incorporate 5.1 styling moves per performance. This tenfold increase in styling density from beginner to pro level provides evidence for the inclusion of expressive styling elements as the dancer achieves mastery.

4 Salsa Dance as an Embodied Language

Salsa is more than just dance — it is a dynamic, structured form of nonverbal, embodied interaction that shares many characteristics with language, including recognizable vocabulary, grammar, conversational dynamics, fluency, and personal style. In this section, we elaborate on this analogy to consider how we can apply established tasks and tools from language modeling, such as fluency metrics, sequence modeling, and dialogue structure, to the domain of embodied interaction.

Lexicon and Grammar. The lexical vocabulary of salsa consists of standardized moves such as the Cross Body Lead, Copa, and Enchufa, each functioning as a meaningful unit of action. These moves are combined following an implicit grammar, where certain transitions are natural (e.g., a Cross Body Lead followed by a follower turn) and others are less conventional [12]. This compositional structure and hierarchical grouping [23] enables dancers to build coherent movement sequences dynamically during improvisation, much like speakers form sentences by combining words and syntactic rules.

Fluency Levels. Dance fluency, like linguistic fluency, varies with experience. In second language fluency [35], factors to measure fluency include word choice, speech rate, silent pause ratio, and so on. Linguists also distinguish *seemingly* fluent speakers (who speak without hesitation) with

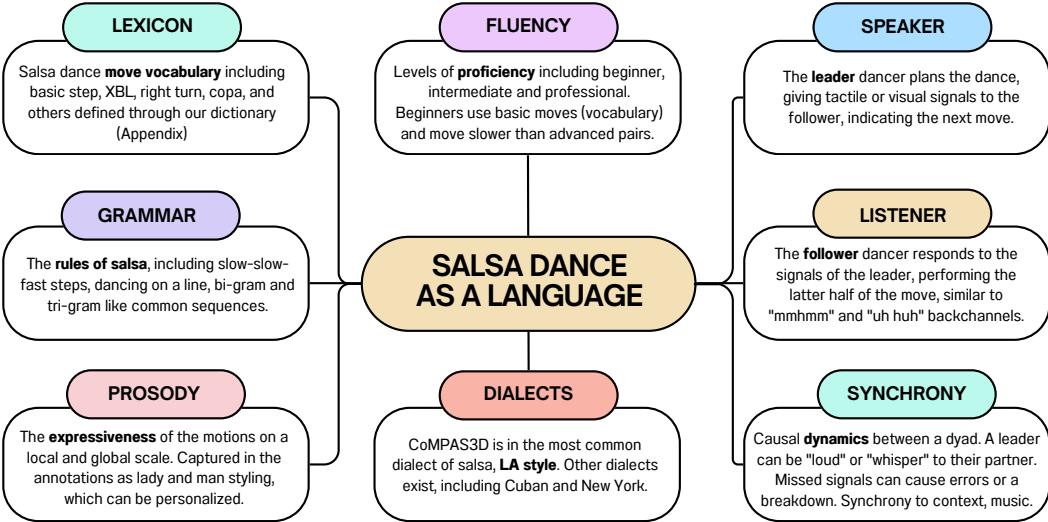


Figure 4: Salsa dance can be usefully analyzed as a form of embodied language.

those with extensive vocabulary knowledge and accurate grammar. The annotations in CoMPAS3D aim to provide insight into avoiding fluent-looking motion generation that would appear inaccurate or illegible to trained dancers. Our analyses of the dataset (Fig. 2-3 and Appendix) show that beginners tend to use a smaller set of basic moves, make more timing mistakes, with little or no styling. Intermediate dancers employ richer combinations and begin to incorporate stylistic accents. Professionals utilize a more advanced set of move vocabularies, exhibit individualized styling, and move at higher speeds (more moves per second).

Personal Expression and Style. Beyond move selection, dancers communicate affect and intent through movement quality, analogous to prosody and accent in speech. In our dataset, we annotate the stylistic moments by both the leader and follower which “accessorize” their dance. Such embodied variations can be highly individual, with subtle variations in space, time and body motion to convey emotion and attitude nonverbally.

Dialects and Variations. As in spoken language, salsa dance features dialectical variations based on region and tradition, including LA-style salsa, New York (Mambo) style, and Cuban Casino. Each dialect differs in timing, movement structure, and styling, paralleling linguistic accents and regional syntax. CoMPAS3D specifically captures LA-style salsa—the most commonly practiced global variant. These dialectical distinctions further support framing salsa as a living, structured communicative system with culturally grounded variations, similar to spoken language dialects.

Speaker and Listener Roles. In salsa duets, the leader assumes the role of the speaker, initiating moves through physical cues and timing, while the follower acts as the listener, interpreting and responding in real time [7]. In this analogy, the salsa dance follower provides responses akin to listener backchanneling such as “mmhm” [37]. A desirable follower can be a “light lead” [18], indicating that only the slightest signaling will produce the desired response. An interesting challenge is that communication between the leader and follower almost completely haptic, signaled by subtle pushes and pulls.

Synchrony and Conversational Dynamics. Improvised salsa dancing exhibits conversational properties, including back-and-forth exchanges. These dynamics are bidirectional: research in impaired backchanneling suggests that poor active listening can have a negative effect on the speaker’s narrative quality [4], which may suggest that dance follower generation tasks [32] should not be unidirectional. As dancers describe: “A rough lead feels like shouting,” while “a soft lead feels like whispering.” Partners negotiate movement in both directions, adapting to each other’s timing, style, and intended complexity, akin to conversational repair and accommodation in spoken dialogue.

Evaluability and Structured Judgement. Unlike many forms of nonverbal interactive behavior such as facial expressions, salsa performances can be systematically evaluated by trained judges using established criteria. Competitions and grading systems assess elements such as timing, partner

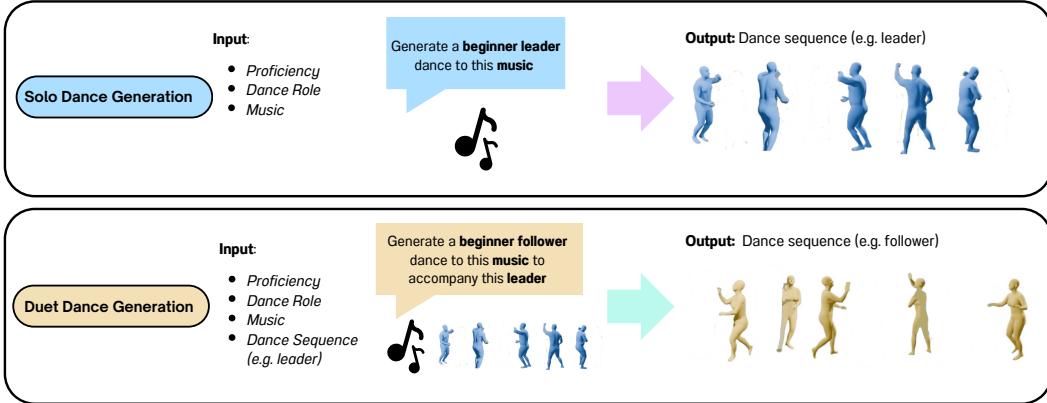


Figure 5: Proposed benchmark tasks for the CoMPAS3D dataset: solo dance generation and duet generation. The duet task is conditioned on music, proficiency and the partner’s dance sequence.

connection, technical execution, and expressive styling. This provides an objective basis for measuring dance generation quality, making salsa particularly well-suited as a benchmark for embodied interaction modeling.

We provide the above linguistic analogy to inspire an array of future tasks on CoMPAS3D, including, but not limited to, move classification, transcription, short and long-term proficiency adaptation, error detection, personalization and dialect translation. In this paper, we focus on two tasks as a starting point for long-form embodied dance interaction benchmarking.

5 Benchmark Tasks

To promote research on embodied nonverbal communication, we offer two main benchmark tasks on CoMPAS3D, involving solo and duet dance generation. Figure 5 summarizes the inputs and outputs for each task. We briefly describe our metrics and tasks below:

Solo Generation. Similar to monologue generation conditioned on a topic, this task generates a leader or follower’s motion sequence based on the accompanying music and the proficiency level. Evaluation metrics include Fréchet Inception Distance (both FID kinematic [21] and graphical [20], denoted by FID_k and FID_g respectively), diversity (Div) of the motion, and Beat-Align Score (BAS) [33] which measure how aligned the generated motions to the music rhythm.

Duet Dance Generation. Analogous to listener modeling or dialogue response generation, this task predicts the follower’s motion based on the leader’s motion and the shared musical context. In this paper, following [32], the entirety of the leader’s motion is used to generate the follower’s motion. We compute the metrics for single-person motion generation evaluation. In addition, for duet dance generation evaluation, we adopted the FID and Div measurement computed with cross-distance, and Beat Echo Degree (BED) proposed by [32] to evaluate the consistency of dynamic rhythms of two dancers. In future work, instead of full leader motions as input, shorter leader motions can be input into the follower response, and vice-versa, towards real-time interactions between independent SalsaAgents. This is to provide room for bi-directional adaptation, as the responses of the follower can affect the movements of the leader (e.g. consider a professional dancing with a beginner).

6 Benchmark Experiments

We present baseline results for key benchmark tasks defined in Section 5. Due to space constraints, architecture and training details are provided in supplementary materials.

6.1 SalsaAgent: A Unified Multitask Model for Humanoid Salsa Interaction

To demonstrate the utility of CoMPAS3D across multiple embodied tasks, we introduce SalsaAgent, a unified model trained to perform both benchmark tasks: solo dance generation (as either a leader

Table 2: Dance generation task results on CoMPAS3D. Metrics grouped by Solo, Interactive, and Rhythmic dimensions. **Bold** indicates best, and underline second best.

Task	Method	Solo (S)				Interactive (I)			Rhythmic (R)
		FID _{k↓}	FID _{g↓}	Div _{k↑}	Div _{g↑}	FID _{cd↓}	Div _{cd↑}	BED↑	BAS↑
Solo Gen.	MotionLLM	>1e ⁶	>1e ¹⁷	68.20	>1e ⁸	n/a	n/a	n/a	0.24
	SalsaAgent (Ours)	153.31	90.80	17.29	12.54	n/a	n/a	n/a	0.24
Duet Gen.	Groundtruth	0.00	0.00	12.00	8.32	0.00	14.38	0.50	0.22
	Duolando (PT)	3051.38	148.03	22.86	6.65	229.06	<u>9.54</u>	0.22	0.23
	Duolando (FT)	464.84	75.32	<u>19.19</u>	<u>7.69</u>	<u>57.64</u>	11.86	<u>0.28</u>	0.23
	SalsaAgent (Ours)	80.62	<u>107.24</u>	12.46	13.10	20.16	9.08	0.37	0.23

or follower, with a proficiency of beginner, intermediate or pro), and duet dance generation. Unlike prior baselines, which are typically specialized for a single task, SalsaAgent is pretrained on our motion token vocabulary and fine-tuned in a multitask setting across tasks using shared motion representations.

SalsaAgent is built upon the MotionLLM [39] text-to-motion backbone. The transformer architecture is conditioned on task-specific prompts and multimodal inputs (i.e., music, leader motion, target proficiency). It is trained in a two-stage pipeline: (1) motion token pretraining using masked modeling on CoMPAS3D sequences, and (2) task-specific supervised fine-tuning using our CoMPAS3D detailed move annotations as well as MotionScript fine-grained motion descriptions [42]. This multitask approach aligns with the goal of creating a dancing agent able to flexibly take on multiple roles and interact with varied partners.

6.2 Experiments

We compare SalsaAgent to task-specific baselines across benchmark tasks described in Fig. 5.

Solo Generation. We evaluate our salsa leader motion generation against each of the sequences in the test set. The baseline MotionLLM [39] is a general model that receives a text prompt (see Appendix) to produce humanoid motion at our 3 different proficiency levels. We prompt it to produce leader salsa dance and it is able to produce motions that resemble rhythmic latin dance. We observe that the FID scores and graphical diversity are very high, indicating a large divergence from the groundtruth beginner, intermediate or professional salsa dances. Kinetic diversity is also relatively high. Our SalsaAgent receives a text prompt, as well as music tokens in addition, and produces a baseline for our dataset, as shown in Table 2.

Duet Generation. In the offline follower generation task, we benchmark the state-of-the-art Duolando [32] model which was trained on latin dance data *Duolando (PT)*, as well as a model fully retrained on our dataset *Duolando (FT)*. The task involves predicting a follower’s motion sequence given the leader’s groundtruth motion, as well as music and proficiency. We evaluate our salsa follower motion generation against each of the sequences in the test set. As indicated in Table 2, we find that our SalsaAgent is able to produce motions considerably closer to the groundtruth in terms of kinetic FID, contact distance FID and higher BED, indicating that the two dancers are well synchronized. Videos are provided in the supplemental material.

7 Limitations

CoMPAS3D focuses on a single dance genre (salsa) and future work can increase the current number of dancing pairs (9 pairs, 18 participants) and annotation coverage. Although the dataset includes beginner, intermediate, and professional skill levels, the diversity of salsa variants, dance styles, and partner matching can further be increased, e.g. New York or Cuban salsa styles, tango or swing, or matching a professional with a beginner. A current technical limitation is the absence of significance testing for differences in experimental metrics, consistent with prior work [32].

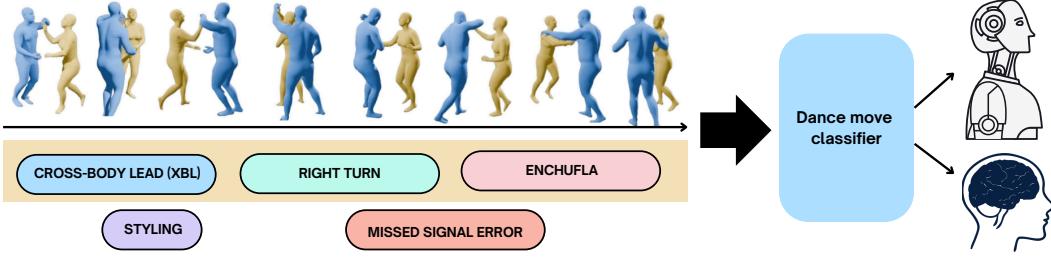


Figure 6: Example of potential application for move classification using CoMPAS3D annotations, to train both humans and generative models to produce legible, fluent motions.

8 Applications and Broader Impact

CoMPAS3D opens up multiple avenues for a wide range of future applications and uses.

Applications. CoMPAS3D provides a new testbed for developing embodied AI agents capable of social physical interaction. As a primary application, similar to natural language learning systems, salsa dance training systems (e.g. virtual or augmented reality generated partner) can adapt to learner fluency levels, especially when a fluent leader or follower is not available for practice. Other potential applications include developing a move classifier using our annotated labels (Fig. 6). This task may remain challenging due to lack of fine-grained data on salsa dance, as well as heavy label imbalance which is common in naturalistic data; a simple unsupervised vector quantization (VQ-VAE) with nearest neighbor retrieval in embedding space resulted in approximately 52% accuracy over the 30 leader and 30 follower classes in the test set, suggesting an open direction for exploration.

Broader Impacts. CoMPAS3D provides a foundation for learning fluent nonverbal communication patterns, toward socially interactive embodied agents capable of improvising and responding [2] in physical conversations. That is, AI systems trained on CoMPAS3D could help bridge gaps between verbal and physical communication, particularly for domains where speech is non-primary (e.g., face and gesture generation for interactive agents, accessibility technologies). In addition, it may spur development of computational models of synchrony that can be used in studying human-human interactions, including interpersonal synchrony breakdowns [9]. Similar to autonomous cars interacting with pedestrians, a potential negative impact is safety if the dataset is used to develop humanoid robots that dance with real people—developers should carefully develop and test their robots to avoid contact accidents: for instance, using virtual or augmented reality with digital agents as explored here, physics-based simulators, and/or dancing between 2 robots.

9 Future Work and Conclusion

Future work includes taking advantage of salsa’s formal judging standards to use as evaluation criteria, refining language metrics such as BLEU score [22] for motion, and using timestamped, localized contact annotations for exploration of physical touch and haptic signaling between partners. In addition, future work can consider other tasks such as automatic segmentation and transcription, especially to support the creation of an automatic score for intelligibility, similar to the use of Automatic Speech Recognition for speech synthesis tasks (Fig. 6). Ultimately, an extremely challenging task involves training individual humanoid robot SalsaAgents that can safely and creatively dance with each other and with real humans, using only haptic signaling as nonverbal communication. Achieving this goal, with salsa judging as a metric, would indicate a large step forward in interactive motion planning.

In conclusion, we introduce CoMPAS3D, a large-scale, richly annotated dataset capturing improvised salsa duet dances across diverse proficiency levels. Our benchmark tasks—including classification, solo and duet dancer generation—lay the groundwork for modeling nonverbal, embodied social intelligence. We invite the research community to build upon CoMPAS3D, extend it with tasks given our analogy with natural language, towards advancing socially interactive AI, embodied modeling, and nonverbal human-AI collaboration.

References

- [1] Pierre-Emmanuel Aguera, Karim Jerbi, Anne Caclin, and Olivier Bertrand. Elan: a software package for analysis and visualization of meg, eeg, and lfp signals. *Computational intelligence and neuroscience*, 2011(1):158970, 2011.
- [2] Sames Al Moubayed, Malek Baklouti, Mohamed Chetouani, Thierry Dutoit, Ammar Mahdhaoui, J-C Martin, Stanislav Ondas, Catherine Pelachaud, Jérôme Urbain, and Mehmet Yilmaz. Generating robot/agent backchannels during a storytelling experiment. In *2009 IEEE International Conference on Robotics and Automation*, pages 3749–3754. IEEE, 2009.
- [3] Juliet McMains and. Salsa steps toward intercultural education. *Journal of Dance Education*, 16(1):27–30, 2016.
- [4] Janet B Bavelas, Linda Coates, and Trudy Johnson. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941, 2000.
- [5] Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44:387–419, 2010.
- [6] Canada Salsa and Bachata Congress. Rules, definitions and judging criteria 2024. <https://www.canadasalsacongress.com/rules>, 2024. Accessed: 2025-05-06.
- [7] Hanne De Jaeger and Ezequiel A Di Paolo. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4):485–507, 2007.
- [8] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *arXiv preprint arXiv:2308.01854*, 2023.
- [9] Alexandra Livia Georgescu, Sevim Koeroglu, A F de C Hamilton, Kai Vogeley, Christine M Falter-Wagner, and Wolfgang Tschacher. Reduced nonverbal interpersonal synchrony in autism spectrum disorder independent of partner diagnosis: a motion energy study. *Molecular autism*, 11:1–14, 2020.
- [10] 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*, pages 418–437. Springer, 2024.
- [11] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13053–13064, 2022.
- [12] Judith Lynne Hanna. *To dance is human: A theory of nonverbal communication*. University of Chicago Press, 1987.
- [13] Xiaowei Hu, Zhenghao Xing, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. Unveiling deep shadows: A survey on image and video shadow detection, removal, and generation in the era of deep learning. *arXiv preprint arXiv:2409.02108*, 2024.
- [14] Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.
- [15] Jogendra Nath Kundu, Himanshu Buckhash, Priyanka Mandikal, Anirudh Jamkhandi, Venkatesh Babu Radhakrishnan, et al. Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2724–2733, 2020.
- [16] Ronghui Li, Youliang Zhang, Yachao Zhang, Yuxiang Zhang, Mingyang Su, Jie Guo, Ziwei Liu, Yebin Liu, and Xiu Li. Interdance: Reactive 3d dance generation with realistic duet interactions. *arXiv preprint arXiv:2412.16982*, 2024.

- [17] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.
- [18] Janice L Mahinka. *The Musicality of Salsa Dancers: An Ethnographic Study*. City University of New York, 2018.
- [19] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 international conference on 3D vision (3DV)*, pages 120–130. IEEE, 2018.
- [20] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. In *ACM SIGGRAPH 2005 Papers*, pages 677–685. 2005.
- [21] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. Fmdistance: A fast and effective distance function for motion capture data. *Eurographics (Short Papers)*, 7(10), 2008.
- [22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- [23] Pritty Patel-Grosz, Salvador Mascarenhas, Emmanuel Chemla, and Philippe Schlenker. Super linguistics: an introduction. *Linguistics and Philosophy*, 46(4):627–692, 2023.
- [24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.
- [25] Xiaogang Peng, Xiao Zhou, Yikai Luo, Hao Wen, Yu Ding, and Zizhao Wu. The mi-motion dataset and benchmark for 3d multi-person motion prediction. *arXiv preprint arXiv:2306.13566*, 2023.
- [26] 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*, pages 1941–1951, 2024.
- [27] Salsa is Good. Salsa dancing dictionary. https://www.salsaisgood.com/dictionary/Salsa_dictionary.htm, n.d. Accessed: 2025-04-07.
- [28] Simon Senecal, Niels A Nijdam, Andreas Aristidou, and Nadia Magnenat-Thalmann. Salsa dance learning evaluation and motion analysis in gamified virtual reality environment. *Multimedia Tools and Applications*, 79(33):24621–24643, 2020.
- [29] Simon Senecal, Niels A Nijdam, and Nadia Magnenat Thalmann. Motion analysis and classification of salsa dance using music-related motion features. In *Proceedings of the 11th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–10, 2018.
- [30] Simon Senecal, Niels Alexander Nijdam, and Nadia Magnenat Thalmann. Classification of salsa dance level using music and interaction based motion features. In *VISIGRAPP (1: GRAPP)*, pages 100–109, 2019.
- [31] Rebecca Simpson-Litke and Chris Stover. Theorizing fundamental music/dance interactions in salsa. *Music Theory Spectrum*, 41(1):74–103, 2019.
- [32] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. *arXiv preprint arXiv:2403.18811*, 2024.

- [33] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022.
- [34] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- [35] Ron I Thomson. Fluency. *The handbook of English pronunciation*, pages 209–226, 2015.
- [36] Coert Van Gemeren, Ronald Poppe, and Remco C Veltkamp. Spatio-temporal detection of fine-grained dyadic human interactions. In *Human Behavior Understanding: 7th International Workshop, HBU 2016, Amsterdam, The Netherlands, October 16, 2016, Proceedings* 7, pages 116–133. Springer, 2016.
- [37] Sheida White. Backchannels across cultures: A study of americans and japanese1. *Language in society*, 18(1):59–76, 1989.
- [38] Simon N Williams, Christopher J Armitage, Tova Tampe, and Kimberly Dienes. Public perceptions and experiences of social distancing and social isolation during the covid-19 pandemic: a uk-based focus group study. *BMJ Open*, 10(7), 2020.
- [39] Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. Mg-motionllm: A unified framework for motion comprehension and generation across multiple granularities. *arXiv preprint arXiv:2504.02478*, 2025.
- [40] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. *arXiv preprint arXiv:2405.17013*, 2024.
- [41] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22260–22271, 2024.
- [42] Payam Jome Yazdian, Eric Liu, Rachel Lagasse, Hamid Mohammadi, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv preprint arXiv:2312.12634*, 2023.
- [43] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023.
- [44] 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*, pages 14730–14740, 2023.

Appendix - Supplementary Material for *Salsa as a Nonverbal Embodied Language—The CoMPAS3D Dataset and Benchmarks*

A CoMPAS3D: Additional Dataset Details

A.1 Training and Testing Split

The CoMPAS3D dataset is comprised of 72 salsa duet dances of 2.5min each. Each of the 9 pairs performed two takes for each of 4 songs, resulting in 8 takes each. The details on each pair, their annotations, and the test set is in Table 3.

Pair	Proficiency	Public Annotations	Test Set
Pair 1	Beginner	100%	Song1_Take1
Pair 2	Intermediate	100%	Song1_Take2
Pair 3	Beginner	100%	Song2_Take1
Pair 4	Intermediate	100%	Song2_Take2
Pair 5	Professional	50%	Song3_Take1
Pair 6	Intermediate	n/a	Song3_Take2
Pair 7	Professional	n/a	Song4_Take1
Pair 8	Beginner	n/a	Song4_Take2
Pair 9	Professional	n/a	Song1_Take1

Table 3: Pair proficiency levels, annotations and corresponding sequences held out for testing.

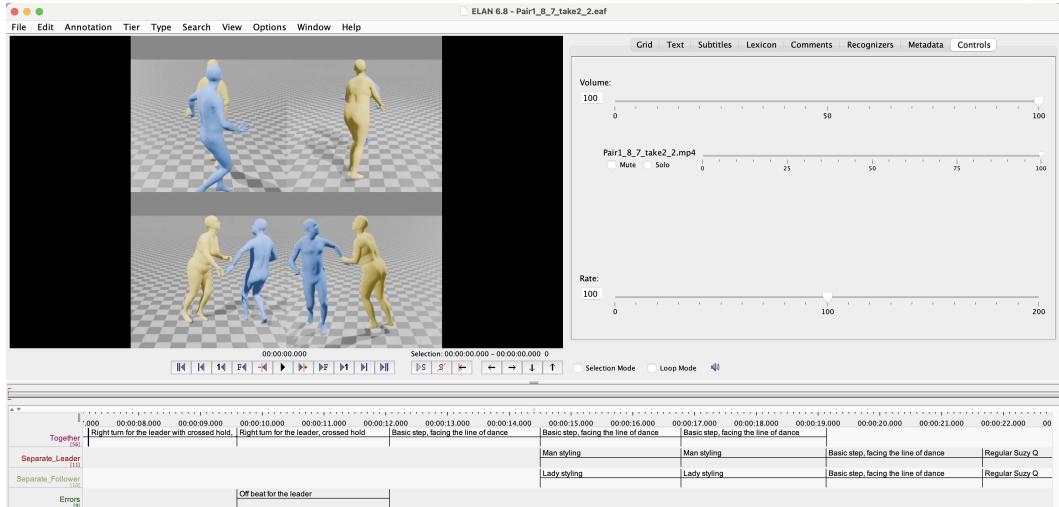


Figure 7: ELAN annotation tool used for segmenting and labeling dance moves in animated SMPL-X representation files. The annotation includes four tracks: Together – when dancers execute the move as a pair; Separate_Leader – when the leader dances solo or adds "Man Styling" to the base move; Separate_Follower – when the follower dances solo or incorporates "Lady Styling"; and Errors – for marking mistakes.

Move, Styling, or Error Name	Category	Detailed Description
Arm lock	Move	A locking arm movement often used to create tension or highlight transitions.
Basic step	Move	Fundamental salsa step with variations including side, cross-back, and back basic steps.
Body shake	Move	A rapid shaking movement emphasizing torso dynamics.
Body roll	Move	A fluid, wave-like motion passing through the body.
Change of Directions	Move	Transition step involving directional changes, including position swaps.
Check	Move	A checking step used to halt or redirect movement.
Comb	Move	A styling-influenced move where the hand is combed over the head.
Copa	Move	A pivoting movement redirecting the follower after a forward step.
Dile que no	Move	A foundational salsa move, translating to "tell her no."
Hand throw	Move	A dramatic throwing motion of one or both hands.
Right turn	Move	A clockwise rotational turn performed by the dancer.
Drawing circle	Move	Circular motion with hands or body to accentuate movement.
Enchufa	Move	Salsa turn pattern where partners switch places.
Walks around	Move	Continuous walking around a partner, often in a circular path.
Suzy Q	Move	Classic salsa footwork emphasizing rhythm and flair.
Hip movement	Move	Emphasized hip motion often synchronized with the rhythm.
Kicks	Move	Kicking action integrated within footwork patterns.
Lasso	Move	Overhead arm motion resembling lassoing.
Natural top	Move	Continuous circular motion performed with a partner.
Left turn	Move	A counterclockwise turn executed by the dancer.
Mambo	Move	Latin dance step characterized by forward and backward movements.
Open break	Move	A breaking step where partners create distance.
Point	Move	Pointing gesture typically with feet or hands.
Sliding	Move	Smooth gliding motion across the floor.
Standing	Move	Stationary stance often used for resets or transitions.
Steps	Move	General term for footwork elements.
Swing	Move	Rhythmic swinging motion involving torso or arms.
Walk	Move	Basic locomotion step in any direction.
XBL (Cross Body Lead)	Move	Core salsa move where the follower is led across the leader.
Indescribable	Move	Complex or ambiguous movements not fitting other categories.
Markers Swap issue	Move	Technical artifact caused by marker misalignment.
Lady styling	Styling	Feminine aesthetic enhancements involving hands, hips, and posture.
Man styling	Styling	Masculine aesthetic embellishments emphasizing strength and rhythm.
Misinterpreted signal	Error	Occurs when the follower misunderstands the leader's cue.
Misstep	Error	Incorrect foot placement deviating from the intended movement.
Mixed signals	Error	Conflicting cues from the leader resulting in follower confusion.
Off beat	Error	Deviation from the musical rhythm during execution.

Table 4: Comprehensive Overview of Move, Styling, and Error Annotations in the CoMPAS3D Dataset. This table categorizes the various elements annotated during the dataset creation process, specifying whether each element pertains to a dance move, styling, or error classification. Detailed descriptions provide insight into the contextual significance of each annotation.

A.2 Annotation Tool

We utilized the ELAN annotation tool (Figure 7) to facilitate precise temporal and semantic labeling of the captured dance sequences. SMPL-X representations were manually synchronized with the musical tracks using the witness camera audiovisual footage, generating video files imported into ELAN. We created four annotation tracks: paired move labels, individual dancer move and styling annotations, and error classification.

A.3 Segmentation

Frame-accurate segmentation was achieved through rhythmic alignment based on the clave pattern, a fundamental rhythmic structure in salsa [31]. The clave pattern, characterized by alternating bars of three and two beats (2-3 or 3-2), provides the dance's temporal framework. Segmentation involved marking the start and end frames of each 8-count dance sequence, typically corresponding to a complete dance move, based solely on the musical rhythm.

Table 5: Songs used in the CoMPAS3D dataset with artist names and tempos.

Song	Artist	Title	Tempo (BPM)
Song 1	Tito Rojas	<i>Lo que te queda</i>	90
Song 2	Louie Ramirez, Ray de La Paz	<i>Lluvia</i>	105
Song 3	Leoni Torres	<i>Idilio</i>	95
Song 4	Johnny Ventura	<i>Dilema</i>	93

A.4 Annotation

Moves. Dance move annotations were derived from expert knowledge and standardized salsa terminology [27]. Each segmented sequence was labeled with base moves and their variations, compiled from a 20-entry dictionary (Table 4). This dictionary, based on external resources and expert additions, defined moves with base names and descriptive add-ons. For instance, a sequence could be labeled ‘cross body lead’ followed by ‘follower’s right turn with normal open hold’, specifying the base move, follower action, and hand hold. Move complexity included simultaneous or sequential execution of multiple base moves within an 8-count cycle. To derive the primary move class from a detailed annotation, the move class, e.g. used for the classification task, was determined (Table 4) using the first four words of the detailed annotation.

Styling. Styling annotations captured ‘man styling’ and ‘lady styling’, which are aesthetic embellishments of base moves through hand, foot, hip, head, shoulder, or full-body accessorization. These were classified into ‘no styling’ (standard execution), ‘lady styling’ (feminine embellishments), and ‘man styling’ (masculine embellishments). These stylings, including balance, posture, locomotion, timing, body isolation, and partner connection, were annotated to analyze role-specific stylistic variations.

Errors. Five error classes were defined: ‘no error’, ‘misinterpreted signal’ (leader cue misunderstanding), ‘misstep’ (incorrect foot placement), ‘mixed signals’ (conflicting cues), and ‘off beat’ (deviation from musical rhythm). For example, a ‘Mixed signals and failed move’ occurred during a ‘cross body lead with left (inside) crossed hold and hand change’ at 00:01:56.510 - 00:01:58.860 for the second pair, second song, first take (Pair2_8_7_take2_1), where leader hesitation and an ambiguous hand movement led to follower confusion and a subsequent ‘copa’ move. These error annotations aim to support analysis of skill levels, non-verbal communication, and identifying undesirable dance patterns.

A.5 Music

To capture a diverse range of couple dance dynamics, we selected 4 popular musical pieces with varying beats per minute (BPM), tempi, and musical moods (Table 5). The music is copyrighted, with all rights remaining with the original performers. The release of the music in our dataset within .mp4 video files was reviewed by the Simon Fraser University copyright office and deemed Fair Dealing (Canadian version of Fair Use). The memorandum with rationale is included in the Supplementary Materials.

B Salsa Agent Model Details

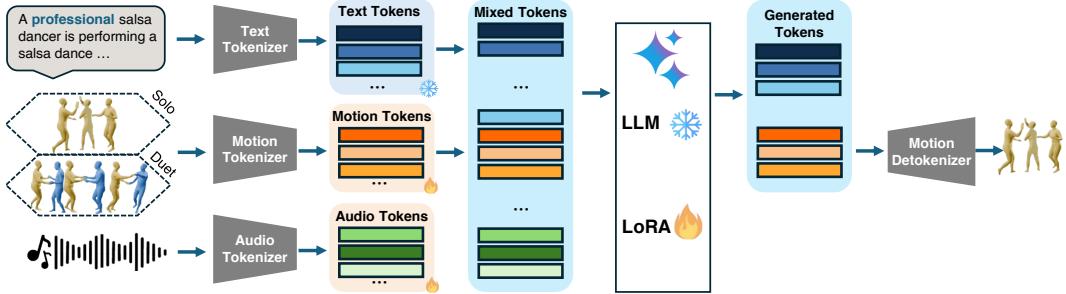


Figure 8: SalsaAgent framework integrates text, audio, and motion to generate salsa dance motions based on input requirements. The generated motion tokens are concatenated and decode to raw motion space.

To enable a unified understanding and generation of multimodal human behavior, we introduce a framework that integrates text, motion, fine-grained motion scripts, and music. Our approach includes a discrete motion representation module that converts raw motion sequences into compact motion tokens using a VQ-VAE-based tokenizer [44]. In parallel, we apply a wave-based tokenizer [14] to transform audio signals into discrete audio tokens. Textual data is processed using a pre-trained tokenizer, while fine-grained motion scripts serve as an intermediate semantic layer bridging high-level language and low-level motion. As the backbone causal language model, we utilize Gemma2-2b-it [34], a lightweight and open-source LLM developed by Google, chosen for its accessibility and ability to run efficiently on a single consumer-grade GPU.

These modality-specific tokenizers output a unified sequence of synchronized discrete tokens, which are then fed into an instruction-tuned language model trained to reason over and generate temporally aligned multimodal content. To enable effective multimodal alignment and generalization across different modalities, we adopt a two-stage training strategy. The first stage focuses on granularity-aware pretraining, learning synergy between different modalities. The second stage involves task-specific instruction tuning, where the model is guided by curated multimodal prompts for dance motion understanding and generation tasks. Furthermore, we used **Low-Rank Adaptation (LoRA)** method, following the hyperparameter settings reported in the MotionLLM [40]. This approach allowed for efficient adaptation by updating only a small subset of parameters while keeping the pre-trained backbone mostly frozen to preserve the generalizable knowledge of the language model.

In the following section, we explain MotionScript briefly followed by stage I, stage II training as well as prompt examples.

Stage I: Pretraining. In the first stage, we pretrained the model using a large collection of aligned audio, follower/leader dancer motion and coarse/fine-grained textual description. Motion data was sampled using a fixed-size sliding window with a duration of **5 seconds** and a stride of **1 second**. Pretraining was conducted for **5 epochs** with a **batch size of 4**, enabling the model to learn general associations between body movements and linguistic descriptions.

Stage II: Task-specific Fine-Tuning. After pretraining, we fine-tuned the model on task-specific datasets such as follower generation or skill-level generation. During this stage, we trained the model for **100 epochs** with an increased **batch size of 16** and frozen embedding for additional special tokens. The fine-tuning objective was to enhance task-specific performance while preserving general motion, audio, and text alignment learned in the pretraining phase.

MotionScript: Fine-grained Motion Captioning Within each 5-second motion window, our MotionScript [42] module automatically produced temporally grounded and linguistically structured descriptions at **0.5-second intervals**. This granularity provides language models with a textual interpretation of compact motion token sequences. Each caption includes explicit start and end timestamps and provides a structured language description of joint positions, movements, and

orientations, proximity, and so on. An illustrative example of MotionScript output is shown in Table 6.

Table 6: Example output from MotionScript for a 5-second motion segment.

Time (s)	Fine-grained Structured Description (MotionScript)
0.0–0.5	[His right hand spreads significantly apart from the left hand]
0.5–1.0	[He shifts slightly to the left]
1.0–1.5	[He is turning counter clockwise]
1.5–2.0	[Both arms begin to raise from below the neck to above]
:	:

B.1 Example Prompts

In this section, we provide illustrative examples of the text prompts used during the fine-tuning and evaluation phases. These prompts are designed to be interpretable by a Large Language Model (LLM) operating on multimodal input, including text, audio, and motion data.

The language model is guided using structured, task-specific prompts. For instance, in the Leader-to-Follower task, the prompt includes the leader’s motion and corresponding audio, and the model is expected to generate the follower’s motion in response.

Each prompt is carefully designed to provide clarity, maintain task relevance, and support effective multimodal alignment. We also introduce special tokens such as [`<LeaderMotion>`] and [`</LeaderMotion>`] that clearly delineate modality-specific inputs. In the following, we illustrate examples of prompts used in various tasks.

Task Name	Input	Output
caption to motion	Components: Coarse Caption, Implicit Role (via output tag). Example Prompt: ### Instruction: Generate a motion sequence based on the description. ### Input: A beginner salsa dancer practices simple steps with careful timing.	Response: <code><LeaderMotion></code> Component: Motion Tokens. Example Output: <code><Motion_1><Motion_5><Motion_10></code> ... <code></Motion_10></Motion_5></code> <code></Motion_1></LeaderMotion></code>
caption to motionscript	Components: Coarse Caption, Implicit Role (via output tag). Example Prompt: ### Instruction: Generate a detailed motion script based on the description. ### Input: A beginner salsa dancer practices simple steps with careful timing.	Response: <code><LeaderScript></code> Component: Motion Script. Example Output: 0.0s-0.5s: Move your right leg forward. <code><SEP></code> 0.5s-1.0s: Left knee bends <code><SEP></code> ... <code></LeaderScript></code>

Table 7: Tasks utilizing text (caption) as the primary input modality.

B.2 Training, Hyperparameters and Hardware Details

We list our training hyperparameters in Table 11. During training, we set aside 10% of the training data for validation. To minimize padding overhead arising from varying sequence lengths, each pretraining batch is drawn from a single randomly chosen task rather than mixing tasks within a batch. All experiments were performed on a single NVIDIA A6000 GPU. Supporting a broad variety of tasks makes pretraining relatively slow: completing five epochs over the full training set requires approximately 12 hours. Fine-tuning on each individual task is more efficient, requiring about 10 hours per task for 100 epochs.

Task Name	Input	Output
caption script to motion	<p>Components: Coarse Caption, Motion Script, Role.</p> <p>Example Prompt:</p> <p>### Instruction: Generate a motion sequence based on the description and motion script.</p> <p>### Input: An intermediate salsa dancer combines footwork and turns with growing confidence.</p> <p>### Script: <LeaderScript> 0.0s-0.5s: Step forward. <SEP> 0.5s-1.0s: Turn left. <SEP> ... </LeaderScript></p>	<p>Response: <LeaderMotion> Component: Motion Tokens.</p> <p>Example Output:</p> <pre><Motion_3><Motion_8><Motion_12> ... </Motion_12></Motion_8> </Motion_3></LeaderMotion></pre>
caption script audio to motion	<p>Components: Coarse Caption, Motion Script, Audio Tokens, Role.</p> <p>Example Prompt:</p> <p>### Instruction: Generate a motion sequence based on description, motion script, and music.</p> <p>### Input: A professional salsa dancer dazzles with sharp, synchronized, and rhythmic movements.</p> <p>### Script: <FollowerScript> 0.0s-0.5s: Move right. <SEP> 0.5s-1.0s: Turn. <SEP> ... </FollowerScript></p> <p>### Audio: <Audio_120><Audio_121><Audio_122> ...</p>	<p>Response: <FollowerMotion> Component: Motion Tokens.</p> <p>Example Output:</p> <pre><Motion_15><Motion_20><Motion_25> ... </Motion_25></Motion_20> </Motion_15></FollowerMotion></pre>
caption audio to motionscript	<p>Components: Coarse Caption, Audio Tokens, Role.</p> <p>Example Prompt:</p> <p>### Instruction: Generate a detailed motion script based on the description and music.</p> <p>### Input: A beginner salsa dancer moves cautiously to the rhythm.</p> <p>### Audio: <Audio_5><Audio_6><Audio_7> ...</p>	<p>Response: <LeaderScript> Component: Motion Script.</p> <p>Example Output:</p> <pre>0.0s-0.5s: Step left. <SEP> 0.5s-1.0s: Shift weight. <SEP> 1.0s-1.5s: Turn clockwise. <SEP> ... </LeaderScript></pre>

Table 8: Tasks utilizing text (caption) plus script and/or audio as input.

Task Name	Input	Output
motionscript to motion	<p>Components: Coarse Caption, Motion Script, Role.</p> <p>Optional: Audio Tokens (50% chance).</p> <p>Example Prompt:</p> <p>### Instruction: Generate a motion sequence from the provided motion script.</p> <p>### Input: A seasoned salsa dancer performs an intricate routine with confidence.</p> <p>### Script: <LeaderScript> 0.0s-0.5s: Step forward. <SEP> 0.5s-1.0s: Turn right. <SEP> ... </LeaderScript></p> <p>### Audio: <Audio_102><Audio_29> <Audio_419> ...</p>	<p>Response: <LeaderMotion> Component: Motion Tokens.</p> <p>Example Output:</p> <pre><Motion_4><Motion_9><Motion_14> ... </Motion_14></Motion_9> </Motion_4></Motion_41> </LeaderMotion></pre>
motion to motionscript	<p>Components: Coarse Caption, Motion Tokens, Role.</p> <p>Example Prompt:</p> <p>### Instruction: Describe the following motion sequence in a detailed motion script.</p> <p>### Input: A professional salsa dancer flows through complex moves with ease.</p> <p><FollowerMotion> <Motion_2><Motion_7><Motion_13> ... </Motion_13></Motion_247></p>	<p>Response: <FollowerScript> Component: Motion Script.</p> <p>Example Output:</p> <pre>0.0s-0.5s: move to the left. <SEP> 0.5s-1.0s: Turning clockwise. <SEP> 1.0s-1.5s: Moving backward. <SEP> ... </FollowerScript></pre>

Table 9: Tasks converting between motion scripts and motion tokens.

Task Name	Input	Output
leader to follower	<p>Components: Coarse Caption, Leader Motion Tokens, Audio Tokens.</p> <p>Example Prompt:</p> <pre>### Instruction: Given leader motion, predict follower motion. ### Input: A mid-level salsa dancer executes a balanced and expressive routine. <LeaderMotion> <Motion_10><Motion_15><Motion_20> ... </Motion_20></Motion_15></Motion_10> ### Audio: <Audio_102><Audio_29> <Audio_419> ...</pre>	<p>Response: <FollowerMotion></p> <p>Component: Motion Tokens.</p> <p>Example Output:</p> <pre><Motion_11><Motion_16><Motion_21> ... </Motion_21></Motion_16> </Motion_11></FollowerMotion></pre>
follower to leader	<p>Components: Coarse Caption, Follower Motion Tokens, Audio Tokens.</p> <p>Example Prompt:</p> <pre>### Instruction: Given follower motion, predict leader motion. ### Input: A salsa duo at intermediate level blends technical movements with smoother transitions. <FollowerMotion> <Motion_5><Motion_8><Motion_12> ... </Motion_12></Motion_8></Motion_5> ### Audio: <Audio_98><Audio_243> <Audio_70> ...</pre>	<p>Response: <LeaderMotion></p> <p>Component: Motion Tokens.</p> <p>Example Output:</p> <pre><Motion_6><Motion_9><Motion_13> ... </Motion_13></Motion_9> </Motion_6></LeaderMotion></pre>
motion completion	<p>Components: Coarse Caption, Partial Motion Tokens (Leader or Follower), Audio Tokens.</p> <p>Example Prompt:</p> <pre>### Instruction: Given a partial motion sequence, complete the motion. ### Input: An expert salsa dancer dazzles with swift, precise, and rhythmic movements. <LeaderMotion> <Motion_2><Motion_4><Motion_7> ... (first 30% tokens) </LeaderMotion> ### Audio: <Audio_22><Audio_343> <Audio_501> ...</pre>	<p>Response: <LeaderMotion></p> <p>Component: Remaining Motion Tokens.</p> <p>Example Output:</p> <pre><Motion_8><Motion_12><Motion_16> ... </Motion_16></Motion_12> </Motion_8></LeaderMotion></pre>

Table 10: Tasks converting or predicting between leader and follower motions, including motion completion.

Table 11: Hyperparameters of our models used in the main experiments.

Component	Hyperparameter	Value
LLM Backbone	Model	Gemma2-2b-it
Adapter	Type	LoRA
Adapter	Rank	64
Adapter	Dropout	0.1
Training	Batch Size Stage I	6
Training	Batch Size Stage II	16
Training	Learning Rate Stage I	2e-5
Training	Learning Rate Stage II	1e-5
Training	# Epochs Stage I	5
Training	# Epochs Stage II	100
Tokenization	Motion Token Length	512
Tokenization	Text Token Length	256000
Tokenization	Audio Token Length	4096
Tokenization	Special Tokens	14
Optimizer	Type	AdamW

B.3 Visualizations

Solo dance generation. Example videos of solo dance generation can be viewed in the Supplementary Video “solo-dance-examples.mp4”. In this video, we notice that MotionLLM (baseline) generated samples tend to start off well, but degrade in quality over time (e.g. 0:20-0:30) eventually walking or stopping completely. As for Salsa Agent, we can observe a progression in speed and complexity as levels progress from the beginner, intermediate to professional. However, fine-grained details such as hip movements are less prominent in Salsa Agent samples compared to the originally captured motion files, likely due to the VQ-VAE tokenization process.

Duet dance generation. Example videos of duet dance generation can be viewed in Supplementary Video “duet-dance-examples.mp4”. We notice relatively good synchrony for Salsa Agent, reflecting the high beat echo degree (BED) score in our quantitative metrics. Future work should include a human study with expert judges to score this task [6].

C Classification

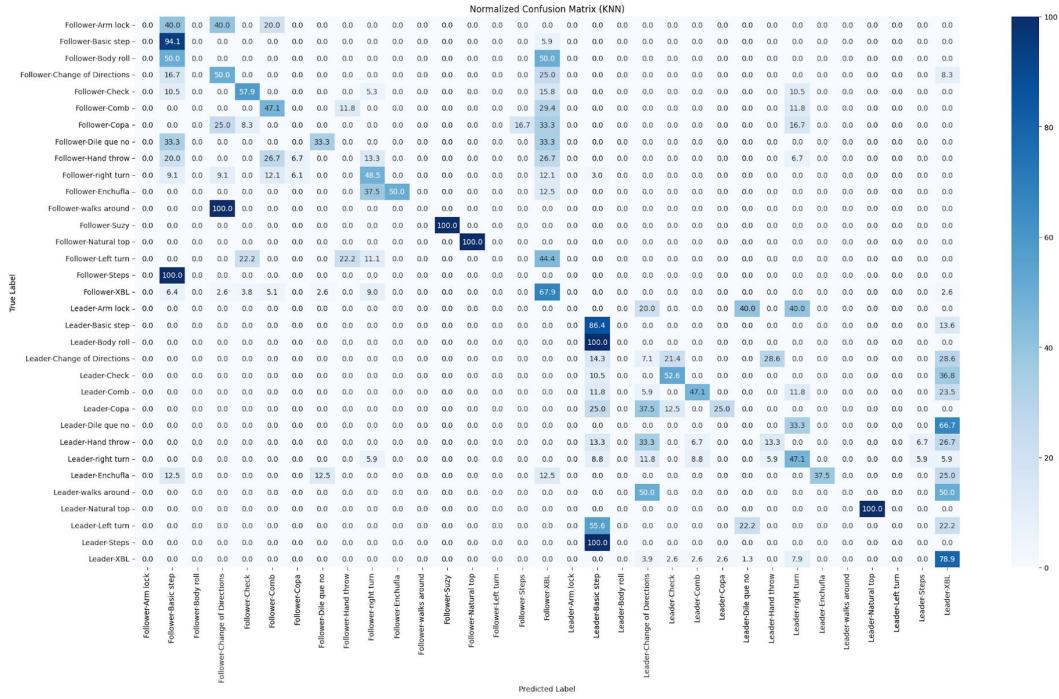


Figure 9: KNN confusion matrix results on the test set for $k = 5$ KNN on the VQ-VAE embedding space. The model over-classifies for the over-represented classes of XBL and basic step.

We provide initial explorations on the suggested use case of classification, via basic machine learning methods on the embedding space, following encoding from a VQ-VAE. Here, we provide further results and model details.

Tokenization. The data was segmented to approximately 3 second segments to align exactly with the annotation times. The data is first passed through our DAE and VQ-VAE that cluster the data in an unsupervised manner. The DAE and VQ-VAE were trained for 5 and 60 epochs respectively on a single NVIDIA GeForce RTX 4070. Training time for the DAE was approximately 4 minutes and 4 hours for the VQ-VAE. Classification was run on the resulting embedding space. A validation set of 3 takes (Pair1_Song2_Take2, Pair2_Song3 Pair5_Song1_Take1) was used for hyperparameter tuning.

C.1 K Nearest Neighbours

We conducted K Nearest Neighbours (KNN) classification on both $k = 1$ and $k = 5$. In both cases, cosine distances were used. We observed an accuracy of 52% for both $k = 1$ and $k = 5$ and a weighted F1-score of 0.51 ($k = 5$) and 0.48 ($k = 1$). The resulting confusion matrix for $k = 5$ can be seen in Fig. 9. Training time was negligible and ran on a CPU.

C.2 XG-Boost

We conducted XG-Boost classification with the following hyper parameters: objective=soft probability, number of estimators=500, max depth=3, learning rate=0.01, subsample=0.5, colsample bytree=0.5, alpha=1.0, lambda=5.0, tree method=histogram. We observed an accuracy of 46% and a weighted F1-score of 0.38. The resulting confusion matrix can be seen in Fig. 10. Training time was approximately 7 minutes and ran on a CPU.

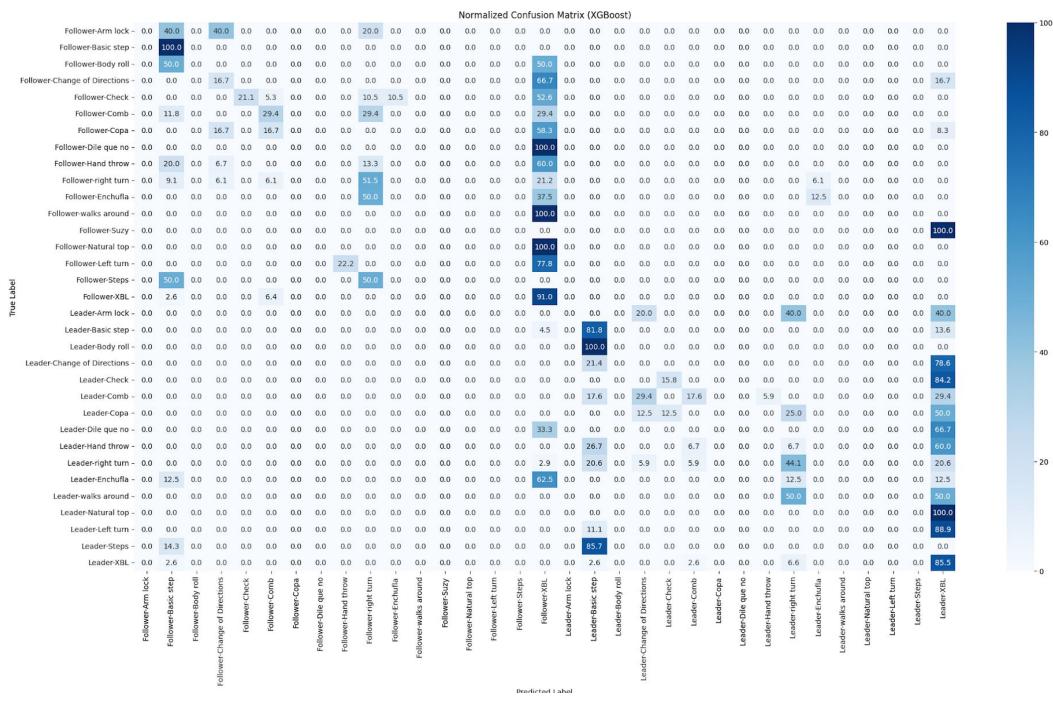


Figure 10: **XG-Boost** confusion matrix results on the test set, conducted on the VQ-VAE embedding space. The model strongly over-classifies for the over-represented classes of XBL and basic step.

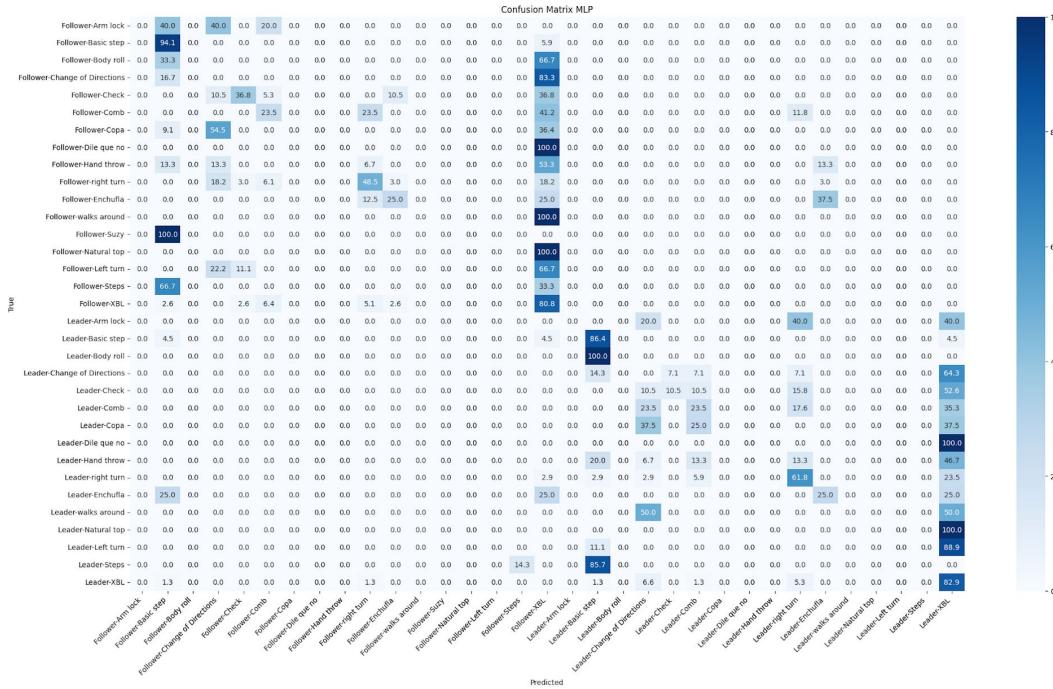


Figure 11: **Multi-layer Perceptron** confusion matrix results on the test set conducted on the VQ-VAE embedding space. The model strongly over-classifies for the over-represented classes of XBL and basic step despite weight balancing in the loss function.

C.3 Multi-Layer Perceptron

We trained a simple multi-layer perceptron (MLP). The MLP contained two linear layers of (128 dimensions) with ReLu activation and a dropout of 0.3 between the two hidden layers. The model used an Adam optimizer with a learning rate of 0.0005, a batch size of 5 and was trained for 5 epochs. We used a class-weighted cross-entropy loss as an attempt to control for class imbalance. We observed an accuracy of 46% and a weighted F1-score of 0.49. The resulting confusion matrix can be seen in Fig. 11. Training time was approximately 6 minutes and ran on a single NVIDIA GeForce RTX 4070.

D Acknowledgments

We would like to thank Giorgio Becherini and Dr. Michael Black for their assistance in MOSH conversion to SMPL-X format. We also thank Ahmet Tasel, Jim Su and ScanlineVFX for their help in learning the motion capture process, discussions and training. This work would also not be possible without support from the Rajan Family and NSERC grant RGPIN-2019-06908.