Tokenizing Nonverbal Communication in Salsa Dance

Anonymous Authors[1]

Figure 1. CoMPAS3D is an improvised salsa dance mocap dataset with a diverse set of individuals at varied experience levels.

Abstract

Partner dance offers a compelling testbed for studying tokenization in multimodal, bidirectional communication. In salsa, a lifted hand may signal a turn; musical accents may shape both dancers' motion. These interactions are continuous and improvisational, and hinge on discrete, interpretable cues—gestures, beats, and movement segments—that can be modeled as tokens. In this paper, we introduce a language model and tokenization framework for social dance, treating salsa as a form of embodied dialogue grounded in motion, music, and role-based interaction. To support this, we present CoMPAS3D, a large-scale motion capture dataset of improvised salsa dancing, capturing over 3 hours of leader-follower interaction across three skill levels. The dataset includes frame-level annotations of moves, styling, and execution errors, created through over 120 hours of expert effort. We use tokens as a foundation for generative and classification tasks, including follower motion prediction and move recognition, demonstrating the utility of token-based models for interactive, expressive virtual agents.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Salsa is "arguably the world's most popular partnered social dance form" (McMains, 2016), 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. (Patel-Grosz et al., 2023) 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 (Hanna, 1987)—we suggest that salsa duet improvisation may be usefully analyzed as an embodied language, complete with vocabulary, grammar, conversational dynamics, fluency levels, stylistic expression, and contextual variation. Given its global reach, improvisational structure, and established evaluation criteria, salsa offers framework for studying for embodied interaction—serving a similar role as English in early spoken language-based model development.

1.1. Salsa Dance as an Embodied Language

Salsa is more than just dance — it is a dynamic, structured form of nonverbal, embodied communication that shares many characteristics with language, including recognizable vocabulary, implicit grammar, conversational dynamics, fluency, and personal style. We first introduce this analogy to consider how we can apply established paradigms from tokenization and language modeling to this multimodal task.

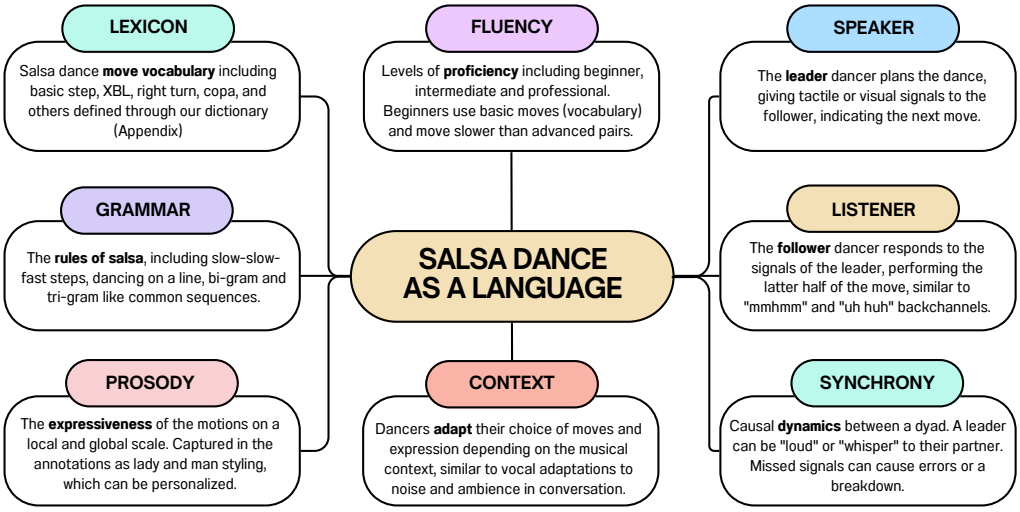

Figure 1. Salsa dance can be usefully analyzed as a form of embodied language.

Lexicon and Grammar. The lexical vocabulary of salsa consists of standardized moves such as the Cross Body Lead, Copa, and Enchufla, each functioning as a meaningful unit of action (see Appendix). 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 (Hanna, 1987). This compositional structure and hierarchical grouping (Patel-Grosz et al., 2023) enables dancers to build coherent movement sequences dynamically during improvisation, much like speakers form sentences by combining words and syntactic rules.

Prosody. 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. As in spoken language, salsa dance also features dialectical variations based on region and tradition, including LA-style salsa, New York (Mambo) style, and Cuban Casino.

Fluency Levels. Dance fluency, like linguistic fluency, varies with experience. In second language fluency (Thomson, 2015), factors to measure fluency include word choice, speech rate, silent pause ratio, and so on. In dance, beginners tend to use a smaller set of basic moves, make more timing mistakes, with little or no styling (Fig. 2), whereas professionals utilize a more advanced set of move vocabularies, exhibit individualized styling, and move at higher speeds. Linguists also distinguish *seemingly* fluent speakers (who speak without hesitation) with those with extensive vocabulary knowledge and accurate grammar (Thomson, 2015). In speech synthesis, automatic speech recognition systems are used to assess this notion of intelligibility (Taylor & Richmond, 2021); similarly, it is important to guard against fluent-looking generative dance motion that would appear inaccurate or illegible to trained dancers.

Context. It is known that speakers adapt their voices to the context. For instance, in noisy environments (Hazan & Baker, 2011), both speaker and listener may use Lombard voice to be heard (Lombard, 1911). Speakers will also adapt to the person they are speaking to (Burnham et al., 2002) (Lam & Kitamura, 2012). Due to this, it is important to model information from the context in which an interaction is taking place. In the case of salsa dance, the ambient music acts as a shared context, modifying both leader and follower actions. The quality of the follower's responses also can result in leader adaptations.

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 (De Jaegher & Di Paolo, 2007). In this analogy, the salsa dance follower provides responses akin to listener backchanneling such as "mmhm" (White, 1989). A desirable follower can be a "light lead" (Mahinka, 2018), 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. Interaction dynamics are bidirectional: research in impaired backchanneling suggests that poor active listening can have a negative effect on the speaker's narrative quality (Bavelas et al., 2000), which may suggest that dance follower generation

tasks (Siyao et al., 2024) 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.

1.2. Multimodal Salsa Dance Motion Capture

In this paper, 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, against a backdrop of musical context. By framing salsa as an embodied language, CoMPAS3D opens new directions for modeling not just individual actions, but dynamic, context-sensitive, multimodal dialogue.

- 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 use tokens as a foundational multimodal representation, and provide results on dance generation and move classification, towards modeling embodied dialogue with 3D virtual humans.

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 (Liu et al., 2019), SBU (Hu et al., 2024), and Inter-X (Xu et al., 2024) offer labeled interactions for action recognition, primarily covering isolated, repetitive activities like handshaking and hugging. Other datasets, including CHI3D (Fieraru et al., 2023), ShakeFive2 (Van Gemeren et al., 2016), and Hi4D (Yin et al., 2023), record close-proximity social interactions with annotated contact events, but remain limited to short, scripted encounters under controlled settings. Additionally, resources such as MuCo3DHP (Mehta et al., 2018), MI-Motion (Peng et al., 2023), 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 (Guo et al., 2022) captures Lindy Hop dancing actions with 3D body poses and shapes, while DuetDance (Kundu et al., 2020) extracts 3D skeletons from YouTube videos of couple dances. ReMoCap (Ghosh et al., 2024) presents multi-view captures of Lindy Hop and Ninjutsu with 3D skeletons and RGB videos. InterHuman (Ruiz-Ponce et al., 2024) includes sequences of martial arts and dance, alongside daily activities. More recently, InterDance (Li et al., 2024) offers 3.93 hours of optical motion capture from professional duets across 15 genres, and DD100 (Siyao et al., 2024) 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. (Senecal et al., 2018; 2019; 2020), 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: 1) they often rely on choreographed (i.e. acted) rather than spontaneous performances, 2) capture only professional dancers rather than a diversity of skill levels, and 3) 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 future benchmarking the legibility and correctness of generated motions.

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

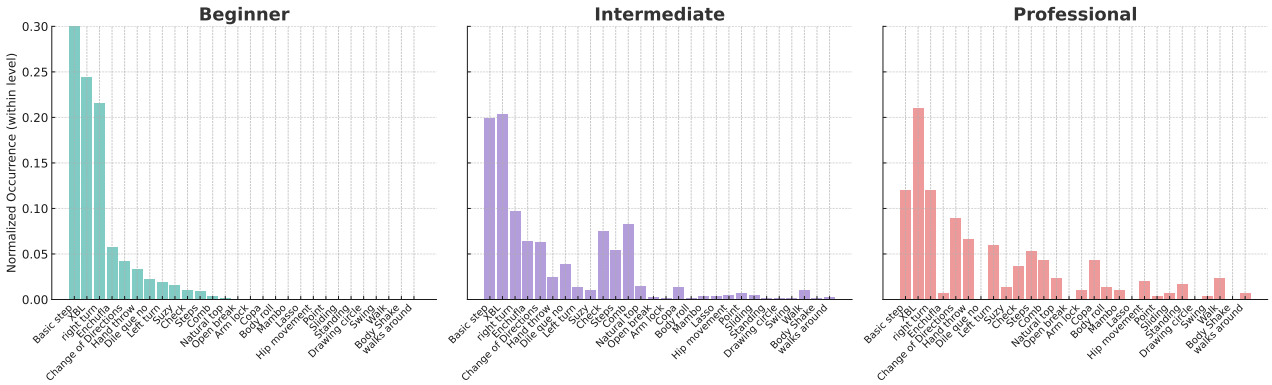

Figure 2. Distribution over the 30 move classes (sorted by beginner move frequency) in CoMPAS3D for beginner, intermediate and pro pairs. Beginners tend to primarily use the "basic step", which professionals use less. Instead, pros use a wider variety of moves such as left turns and copa.

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)[1], 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 (Pavlakos et al., 2019), 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 university ethics board. 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. The dancers

[1]https://huggingface.co/datasets/Rosie-Lab/compas3d

used LA-style salsa—the most commonly practiced global variant.

Data Representation. We release the dataset to facilitate a wide range of machine learning and animation applications. Each sequence includes 55-joint SMPL-X (Pavlakos et al., 2019) 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 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 (Aguera et al., 2011) required over 120 hours. Half the sequences remain unannotated, offering a clean set for future evaluation.

Analysis. Analysis of the annotations reveal distinctions between 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". An analysis of the styling annotations show that professionals execute 54.5 styling moves per performance, followed by intermediates with 12.9 styling moves per performance, and beginners, who incorporate 5.1 styling moves per performance. The most 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.

4. Dance Generation

Recently, large language models (LLMs) are increasingly being used for motion generation. MotionGPT (Zhang et al., 2024) introduced a unified framework for the generation of text-to-motion and motion-to-text using tokenized representations. MG-MotionLLM (Wu et al., 2025) further incorporated hierarchical tasks to understand coarse and fine motion. At the same time, music-conditioned solo and duet dance generation (Siyao et al., 2024) has seen advancements. To the best of our knowledge, the application of motion-generating LLMs to music-guided dance generation has yet to be explored. In the following section, we describe generative tasks on CoMPAS3D (Figure 3) and an LLM-based approach for salsa dance generation by leveraging multimodal tokenization.

Solo Generation Task. Similar to monologue generation conditioned on a topic, this task generates a leader or follower's motion sequence based on the accompanying music (audio) and the proficiency level (text). Evaluation metrics include Fréchet Inception Distance (both FID kinematic (Onuma et al., 2008) and graphical (Müller et al., 2005), denoted by FID_k and FID_g respectively), diversity (Div) of the motion, and Beat-Align Score (BAS) (Siyao et al., 2022) which measure how aligned the generated motions to the music rhythm.

Duet Dance Generation Task. Analogous to listener modeling or dialogue response generation, this task predicts the follower's proficiency (text) and motion (skeleton pose sequence) based on the leader's motion (skeleton pose sequence) and the shared musical context (audio). In this paper, following (Siyao et al., 2024), 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 (Siyao et al., 2024) to evaluate the consistency of dynamic rhythms of two dancers.

Towards Real Time Generation. In future work, instead of full leader motions as input, shorter leader motions (tokens) can be input into the follower response, and vice-versa, towards real-time interactions between independent SalsaAgents. This would 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).

4.1. SalsaAgent: A Unified Multitask Model for Humanoid Salsa Interaction

We introduce SalsaAgent, a unified LLM-based model trained to perform multiple tasks: solo dance generation (as either a leader 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 (Wu et al., 2025)

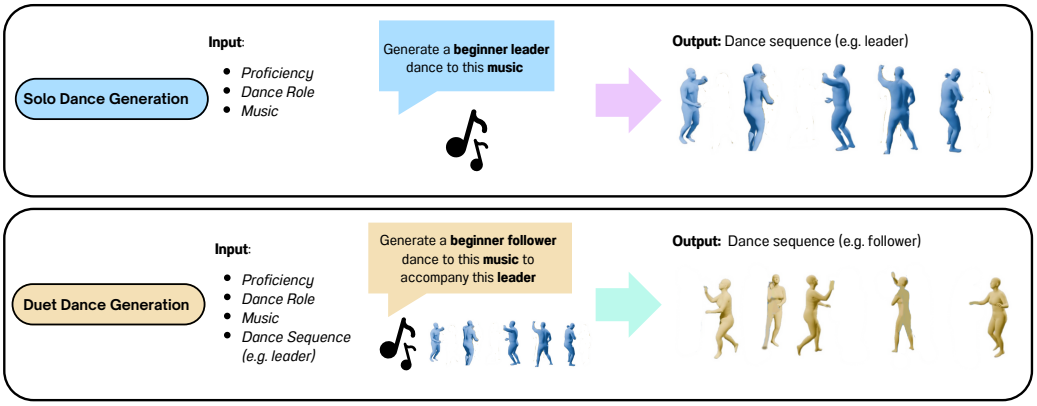

Figure 3. Generative tasks: Solo dance generation and duet generation. The duet task is conditioned on music, proficiency and the partner's dance sequence.

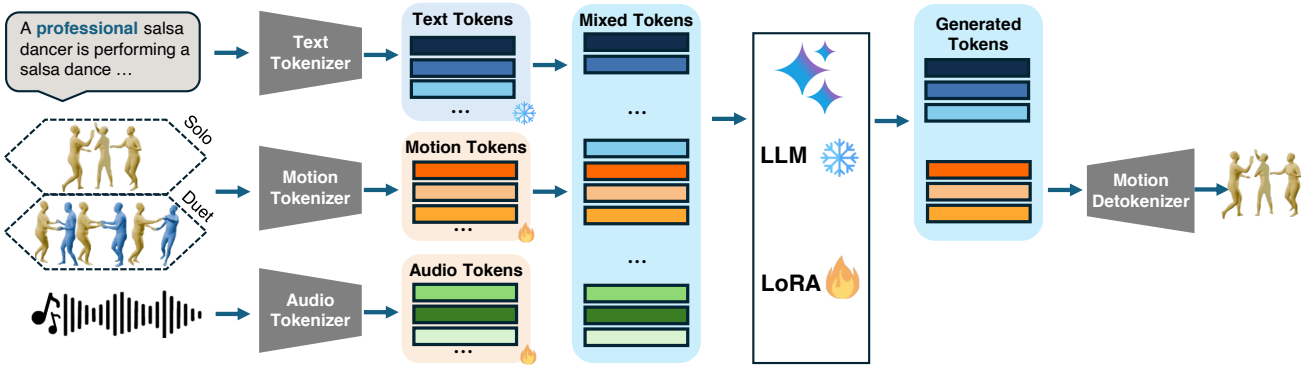

Figure 4. SalsaAgent framework integrates text, audio, and motion to generate salsa dance motions based on input requirements. The generated motion tokens are concatenated and decode to raw motion space.

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 (Yazdian et al., 2023). This multitask approach aligns with the goal of creating a dancing agent able to flexibly take on multiple roles and interact with varied partners.

4.2. Multimodal Tokenization

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 (Zhang et al., 2023). In parallel, we apply a wave-based tokenizer (Ji et al., 2024) 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 language model, we utilize Gemma2-2b-it (Team et al., 2024), 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.

We adopt a two-stage training strategy to enable effective multimodal alignment and generalization across different modalities. The first stage focuses on granularity-aware pretraining, learning correspondences 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 genera-

Table 2. 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]
⋮	⋮

Table 3. Tokenization lengths used in our models.

Tokenization	Length
Motion Token Length	512
Text Token Length	256
Audio Token Length	4096
Special Tokens	14

tion tasks. Furthermore, we used **Low-Rank Adaptation (LoRA)** method, following the hyperparameter settings reported in the MotionLLM (Wu et al., 2024).

Stage I: Pretraining. In the first stage, we pretrained the model using our dataset's 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** (Yazdian et al., 2023) 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 2.

4.3. Text Prompts

In Table 4 and in the Appendix, 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. 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.

5. Experiments

We compare SalsaAgent to task-specific baselines across the tasks described in Fig. 3.

Solo Generation. We evaluate our salsa leader motion generation against each of the sequences in the test set. The baseline MotionLLM (Wu et al., 2025) 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 rythmic 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 5.

Duet Generation. In the offline follower generation task, we benchmark the state-of-the-art Duolando (Siyao et al., 2024) 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 5, we find that our

Table 4. Example prompt for leader-to-follower task

Task Name	Input	Output

`leader to follower`	Components: Coarse Caption, Leader Motion Tokens, Audio Tokens.
	Example Prompt:
	`### 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> ...`
		Response: `<FollowerMotion>`
		Component: Motion Tokens.
		Example Output:
		`<Motion_11><Motion_16><Motion_21>`
		`... </Motion_21></Motion_16>`
		`</Motion_11></FollowerMotion>`

Table 5. 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)
		$FID_k\downarrow$	$FID_g\downarrow$	$Div_k\uparrow$	$Div_g\uparrow$	$FID_{cd}\downarrow$	$Div_{cd}\uparrow$	BED↑	BAS↑
Solo Gen.	MotionLLM	$>1e^6$	$>1e^{17}$	**68.20**	$>\mathbf{1}e^8$	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	9.54	0.22	0.23
	Duolando (FT)	464.84	**75.32**	19.19	7.69	57.64	**11.86**	0.28	0.23
	SalsaAgent (Ours)	**80.62**	107.24	12.46	**13.10**	**20.16**	9.08	**0.37**	0.23

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.

5.1. Visualizations

Supplementary videos can be viewed at this anonymous link: https://osf.io/6dfpy/files/osfstorage?view_only=34d4b26152344a89b7c11ed912abd6b3.

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 (Canada Salsa and Bachata Congress, 2024).

6. Discussion on Tokens and Semantics

How should we tokenize multimodal data in human embodied communication? Typical tokenization approaches in human motion generation pay little attention to semantics. For instance, our dance generation models used a fixed token length (5s) and stride (1s). It is unclear whether this is the ideal representation for a multi-purpose, LLM-based SalsaAgent. To illustrate, consider that many of the tokens in the codebook are devoid of meaning: a token may contain half of a basic step and half of a right turn, for example. That is, tokenization did not consider annotations for segmentation or as part of the optimization process.

We further illustrate the issue using the results of a simple experiment on classification (Fig. 5). Our first attempts used the same VQVAE used for dance generation for a classification (move recognition) task. However, this approach did not yield convergence using trainable classifier methods,

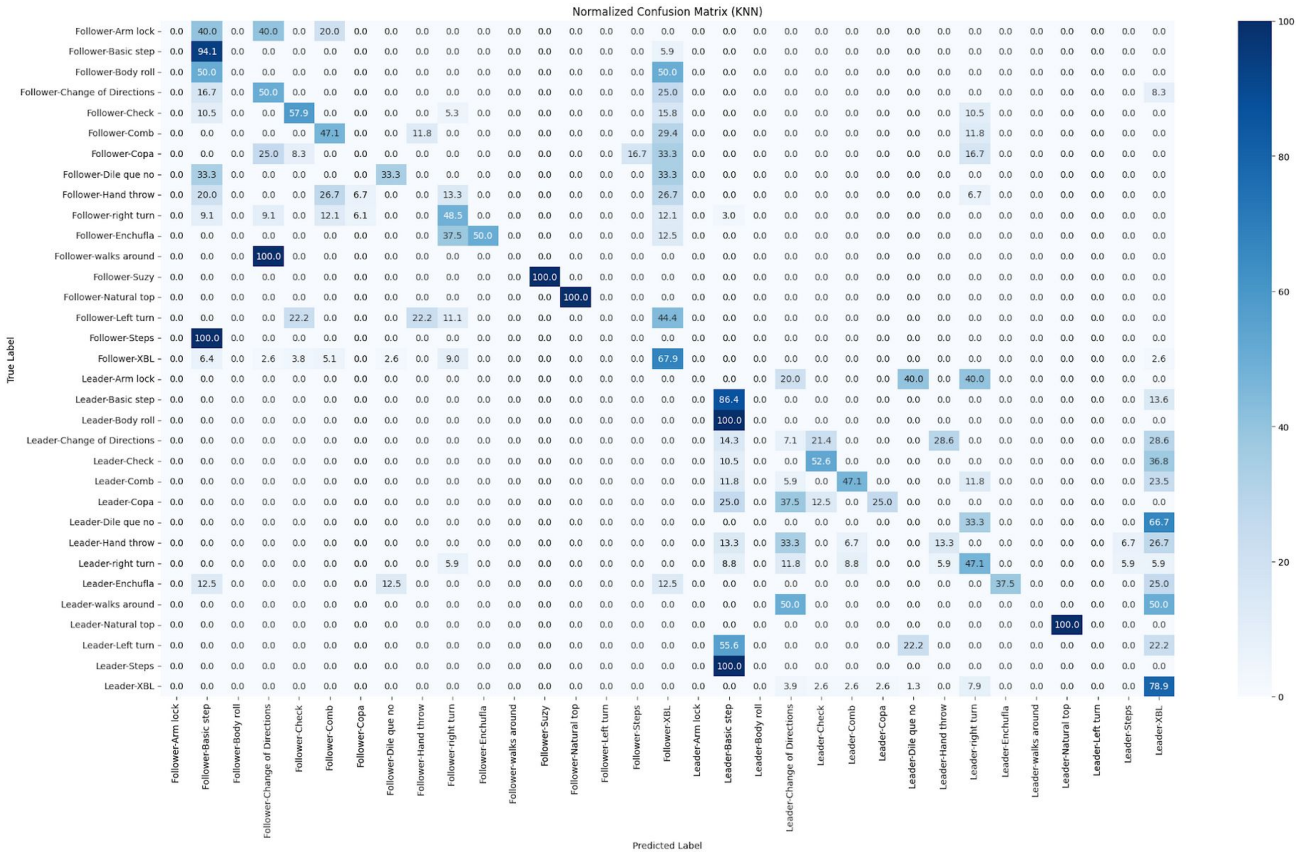

Figure 5. **KNN** confusion matrix results on the test set for k = 5 KNN on the VQ-VAE embedding space aligned with annotation start times.

and very poor results for the KNN approach.

Upon reflection, this result can be easily explained. The tokenization process was fixed at a fixed 5s length with arbitrary start times. Yet, semantic labels such as “XBL" had a precise start time, and a sometimes variable length for each move. When we aligned the tokens with move annotation times and lengths (3s ave. move length), we achieved reasonable results of approximately 52% accuracy over the 30 leader and 30 follower classes in the test set (Fig. 5). Future work can consider how to create tokenizers that consider not only static-length tokenization, but variable length segmentation and semantics. Similar to language, motions can be sliced into small “character"-like chunks, with semantics that are understandable by humans on the “word" level. Metrics for generative motion could then consider, similar to automatic speech recognition, using classification on the semantic level to ensure intelligibilty of the resulting moves.

7. 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 (Papineni et al., 2002) for motion, and using timestamped, localized contact annotations for exploration of physical touch and haptic signaling between partners. In addition, future work can tackle automatic segmentation and transcription, especially to support the creation of an automatic score for intelligibility. 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.

In conclusion, we introduced CoMPAS3D, a large-scale, richly annotated dataset capturing improvised salsa duet dances across diverse proficiency levels. We also proposed a SalsaAgent that can perform solo and duet dance generation using a multimodal tokenization framework. 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

Aguera, P.-E., Jerbi, K., Caclin, A., and Bertrand, O. Elan: a software package for analysis and visualization of meg, eeg, and lfp signals. *Computational intelligence and neuroscience*, 2011(1):158970, 2011.

Bavelas, J. B., Coates, L., and Johnson, T. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941, 2000.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. What's new, pussycat? on talking to babies and animals. *Science*, 296:1435, 2002.

Canada Salsa and Bachata Congress. Rules, definitions and judging criteria 2024. https://www.canadasalsacongress.com/rules, 2024. Accessed: 2025-05-06.

De Jaegher, H. and Di Paolo, E. A. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4):485–507, 2007.

Fieraru, M., Zanfir, M., Oneata, E., Popa, A.-I., Olaru, V., and Sminchisescu, C. Reconstructing three-dimensional models of interacting humans. *arXiv preprint arXiv:2308.01854*, 2023.

Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., and Slusallek, P. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision*, pp. 418–437. Springer, 2024.

Guo, W., Bie, X., Alameda-Pineda, X., and Moreno-Noguer, F. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13053–13064, 2022.

Hanna, J. L. *To dance is human: A theory of nonverbal communication*. University of Chicago Press, 1987.

Hazan, V. and Baker, R. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acoust. Soc*, 130:2139–52, 2011.

Hu, X., Xing, Z., Wang, T., Fu, C.-W., and Heng, P.-A. 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.

Ji, S., Jiang, Z., Wang, W., Chen, Y., Fang, M., Zuo, J., Yang, Q., Cheng, X., Wang, Z., Li, R., et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Kundu, J. N., Buckchash, H., Mandikal, P., Jamkhandi, A., Radhakrishnan, V. B., 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*, pp. 2724–2733, 2020.

Lam, C. and Kitamura, C. Mommy, speak clearly: induced hearing loss shapes vowel hyperarticulation. *Dev. Sci.*, 15 (2):212–21, 2012.

Li, R., Zhang, Y., Zhang, Y., Zhang, Y., Su, M., Guo, J., Liu, Z., Liu, Y., and Li, X. Interdance: Reactive 3d dance generation with realistic duet interactions. *arXiv preprint arXiv:2412.16982*, 2024.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C. 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.

Lombard, E. Le signe de l'élévation de la voix. *Ana. d. Mal. de L'Oreillexdu du larynx [etc]*, 37:101–119, 1911.

Mahinka, J. L. *The Musicality of Salsa Dancers: An Ethnographic Study*. City University of New York, 2018.

McMains, J. Salsa steps toward intercultural education. *Journal of Dance Education*, 16(1):27–30, 2016. doi: 10.1080/15290824.2015.1048865. URL https://doi.org/10.1080/15290824.2015.1048865.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., and Theobalt, C. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 international conference on 3D vision (3DV)*, pp. 120–130. IEEE, 2018.

Müller, M., Röder, T., and Clausen, M. Efficient content-based retrieval of motion capture data. In *ACM SIGGRAPH 2005 Papers*, pp. 677–685. 2005.

Onuma, K., Faloutsos, C., and Hodgins, J. K. Fmdistance: A fast and effective distance function for motion capture data. *Eurographics (Short Papers)*, 7(10), 2008.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Patel-Grosz, P., Mascarenhas, S., Chemla, E., and Schlenker, P. Super linguistics: an introduction. *Linguistics and Philosophy*, 46(4):627–692, 2023.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Peng, X., Zhou, X., Luo, Y., Wen, H., Ding, Y., and Wu, Z. The mi-motion dataset and benchmark for 3d multi-person motion prediction. *arXiv preprint arXiv:2306.13566*, 2023.

Ruiz-Ponce, P., Barquero, G., Palmero, C., Escalera, S., and García-Rodríguez, J. in2in: Leveraging individual information to generate human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1941–1951, 2024.

Salsa is Good. Salsa dancing dictionary. https://www.salsaisgood.com/dictionary/Salsa_dictionary.htm, n.d. Accessed: 2025-04-07.

Senecal, S., Nijdam, N. A., and Thalmann, N. M. 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*, pp. 1–10, 2018.

Senecal, S., Nijdam, N. A., and Magnenat-Thalmann, N. Classification of salsa dance level using music and interaction based motion features. In *VISIGRAPP (1: GRAPP)*, pp. 100–109, 2019.

Senecal, S., Nijdam, N. A., Aristidou, A., and Magnenat-Thalmann, N. Salsa dance learning evaluation and motion analysis in gamified virtual reality environment. *Multimedia Tools and Applications*, 79(33):24621–24643, 2020.

Simpson-Litke, R. and Stover, C. Theorizing fundamental music/dance interactions in salsa. *Music Theory Spectrum*, 41(1):74–103, 2019.

Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C. C., and Liu, Z. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11050–11059, 2022.

Siyao, L., Gu, T., Yang, Z., Lin, Z., Liu, Z., Ding, H., Yang, L., and Loy, C. C. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. *arXiv preprint arXiv:2403.18811*, 2024.

Taylor, J. and Richmond, K. Confidence intervals for asr-based tts evaluation. In *Interspeech 2021*, pp. 2791–2795, 2021. doi: 10.21437/Interspeech.2021-2203.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Thomson, R. I. Fluency. *The handbook of English pronunciation*, pp. 209–226, 2015.

Van Gemeren, C., Poppe, R., and Veltkamp, R. C. 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*, pp. 116–133. Springer, 2016.

White, S. Backchannels across cultures: A study of americans and japanese1. *Language in society*, 18(1):59–76, 1989.

Wu, B., Xie, J., Shen, K., Kong, Z., Ren, J., Bai, R., Qu, R., and Shen, L. Mg-motionllm: A unified framework for motion comprehension and generation across multiple granularities. *arXiv preprint arXiv:2504.02478*, 2025.

Wu, Q., Zhao, Y., Wang, Y., Liu, X., Tai, Y.-W., and Tang, C.-K. Motion-agent: A conversational framework for human motion generation with llms. *arXiv preprint arXiv:2405.17013*, 2024.

Xu, L., Lv, X., Yan, Y., Jin, X., Wu, S., Xu, C., Liu, Y., Zhou, Y., Rao, F., Sheng, X., et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22260–22271, 2024.

Yazdian, P. J., Liu, E., Lagasse, R., Mohammadi, H., Cheng, L., and Lim, A. Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv preprint arXiv:2312.12634*, 2023.

Yin, Y., Guo, C., Kaufmann, M., Zarate, J. J., Song, J., and Hilliges, O. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17016–17027, 2023.

Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., and Shan, Y. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14730–14740, 2023.

Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., and Ouyang, W. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7368–7376, 2024.

A. Appendix

A.1. CoMPAS3D: Additional Dataset Details

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 6.

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 6. Pair proficiency levels, annotations and corresponding sequences held out for testing.

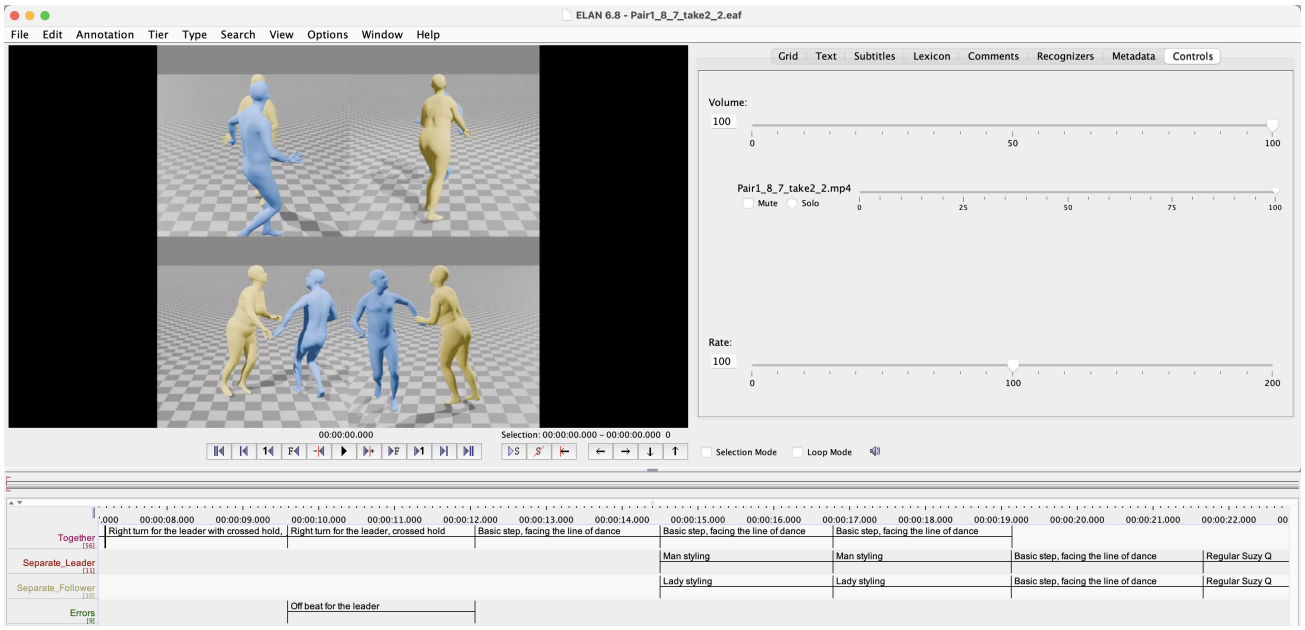

Figure 6. ELAN annotation tool used for segmenting and labeling dance moves in animated SMPL-X representation files. The annotation includes four tracks: Together – when dancers execute the move as a pair; Separate_Leader – when the leader dances solo or adds "Man Styling" to the base move; Separate_Follower – when the follower dances solo or incorporates "Lady Styling"; and Errors – for marking mistakes.

A.2. Annotation Tool

We utilized the ELAN annotation tool (Figure 6) 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 (Simpson-Litke & Stover, 2019). 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.

A.4. Annotation

Moves. Dance move annotations were derived from expert knowledge and standardized salsa terminology (Salsa is Good, n.d.). Each segmented sequence was labeled with base moves and their variations, compiled from a 20-entry dictionary (Table 7). 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 7) 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 8). 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 university copyright office and deemed fair use.

width=

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.
Enchufla	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 7. 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.

Table 8. Songs used in the CoMPAS3D dataset with artist names and tempos.

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

B. Salsa Agent Model Details

B.1. Example Prompts

In this section, we provide further 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:** `<LeaderMotion>` Component: Motion Tokens. **Example Output:** `<Motion_1><Motion_5><Motion_10> ... </Motion_10></Motion_5> </Motion_1></LeaderMotion>`
`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:** `<LeaderScript>` Component: Motion Script. **Example Output:** `0.0s-0.5s: Move your right leg forward. <SEP> 0.5s-1.0s: Left knee bends <SEP> ... </LeaderScript>`

Table 9. Tasks utilizing text (caption) as the primary input modality.

B.2. Training, Hyperparameters and Hardware Details

We list our training hyperparameters in Table 13. 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	Components: Coarse Caption, Motion Script, Role. **Example Prompt:** ### Instruction: Generate a motion sequence based on the description and motion script. ### Input: An intermediate salsa dancer combines footwork and turns with growing confidence. ### Script: <LeaderScript> 0.0s-0.5s: Step forward. <SEP> 0.5s-1.0s: Turn left. <SEP> ... </LeaderScript>	**Response:** <LeaderMotion> Component: Motion Tokens. **Example Output:** <Motion_3><Motion_8><Motion_12> ... </Motion_12></Motion_8> </Motion_3></LeaderMotion>
caption script audio to motion	Components: Coarse Caption, Motion Script, Audio Tokens, Role. **Example Prompt:** ### Instruction: Generate a motion sequence based on description, motion script, and music. ### Input: A professional salsa dancer dazzles with sharp, synchronized, and rhythmic movements. ### Script: <FollowerScript> 0.0s-0.5s: Move right. <SEP> 0.5s-1.0s: Turn. <SEP> ... </FollowerScript> ### Audio: <Audio_120><Audio_121><Audio_122> ...	**Response:** <FollowerMotion> Component: Motion Tokens. **Example Output:** <Motion_15><Motion_20><Motion_25> ... </Motion_25></Motion_20> </Motion_15></FollowerMotion>
caption audio to motionscript	Components: Coarse Caption, Audio Tokens, Role. **Example Prompt:** ### Instruction: Generate a detailed motion script based on the description and music. ### Input: A beginner salsa dancer moves cautiously to the rhythm. ### Audio: <Audio_5><Audio_6><Audio_7> ...	**Response:** <LeaderScript> Component: Motion Script. **Example Output:** 0.0s-0.5s: Step left. <SEP> 0.5s-1.0s: Shift weight. <SEP> 1.0s-1.5s: Turn clockwise. <SEP> ... </LeaderScript>

Table 10. Tasks utilizing text (caption) plus script and/or audio as input.

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

Table 11. Tasks converting between motion scripts and motion tokens.

Task Name	Input	Output
leader to follower	Components: Coarse Caption, Leader Motion Tokens, Audio Tokens. **Example Prompt:** ### 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> ...	**Response:** <FollowerMotion> Component: Motion Tokens. **Example Output:** <Motion_11><Motion_16><Motion_21> ... </Motion_21></Motion_16> </Motion_11></FollowerMotion>
follower to leader	Components: Coarse Caption, Follower Motion Tokens, Audio Tokens. **Example Prompt:** ### 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> ...	**Response:** <LeaderMotion> Component: Motion Tokens. **Example Output:** <Motion_6><Motion_9><Motion_13> ... </Motion_13></Motion_9> </Motion_6></LeaderMotion>
motion completion	Components: Coarse Caption, Partial Motion Tokens (Leader or Follower), Audio Tokens. **Example Prompt:** ### 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> ...	**Response:** <LeaderMotion> Component: Remaining Motion Tokens. **Example Output:** <Motion_8><Motion_12><Motion_16> ... </Motion_16></Motion_12> </Motion_8></LeaderMotion>

Table 12. Tasks converting or predicting between leader and follower motions, including motion completion.

Table 13. 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

C. Classification

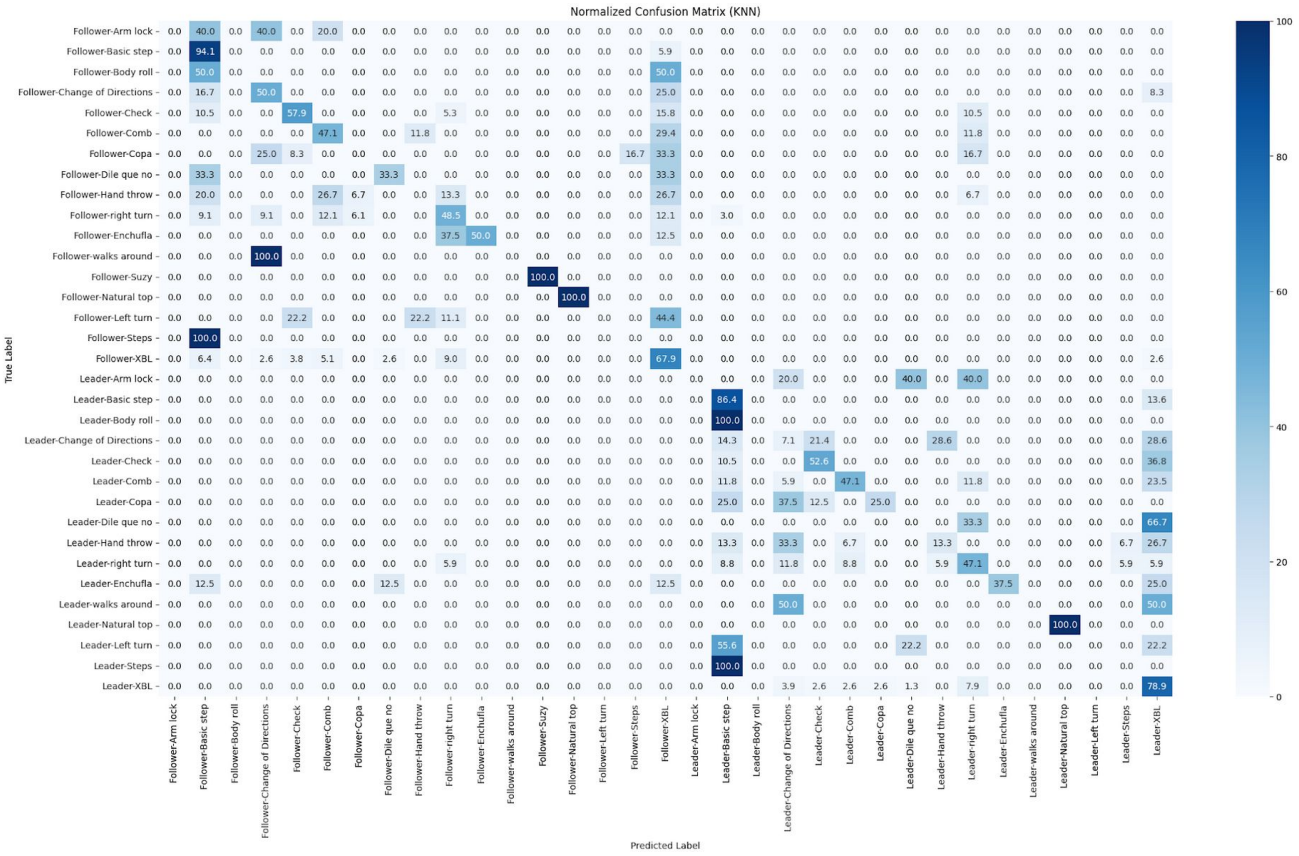

Figure 7. **KNN** confusion matrix results on the test set for k = 5 KNN on the VQ-VAE embedding space. The model over-classifies for the over-represented classes of XBL and basic step.

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. 7. 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. 8. Training time was approximately 7 minutes and ran on a CPU.

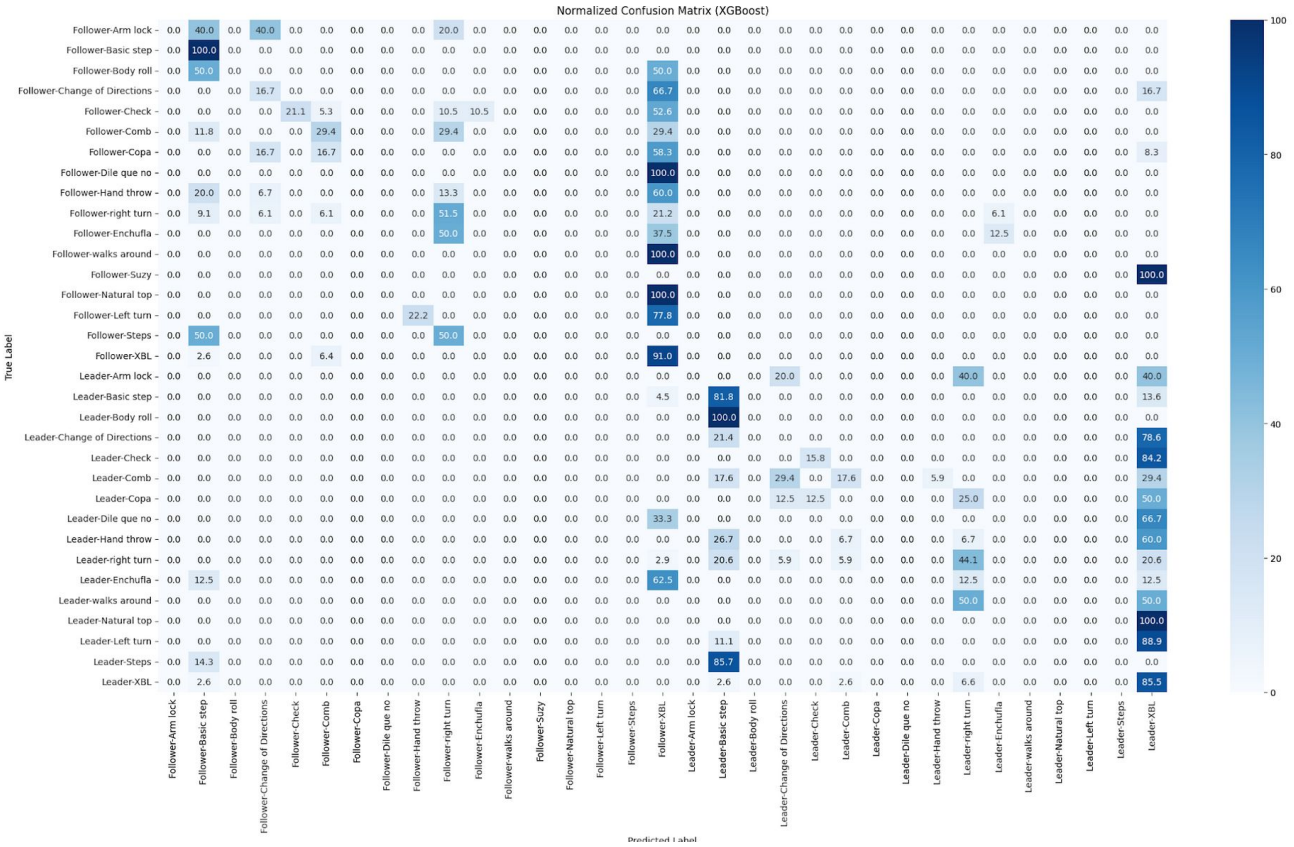

Figure 8. **XG-Boost** confusion matrix results on the test set, conducted on the VQ-VAE embedding space. The model strongly over-classifies for the over-represented classes of XBL and basic step.

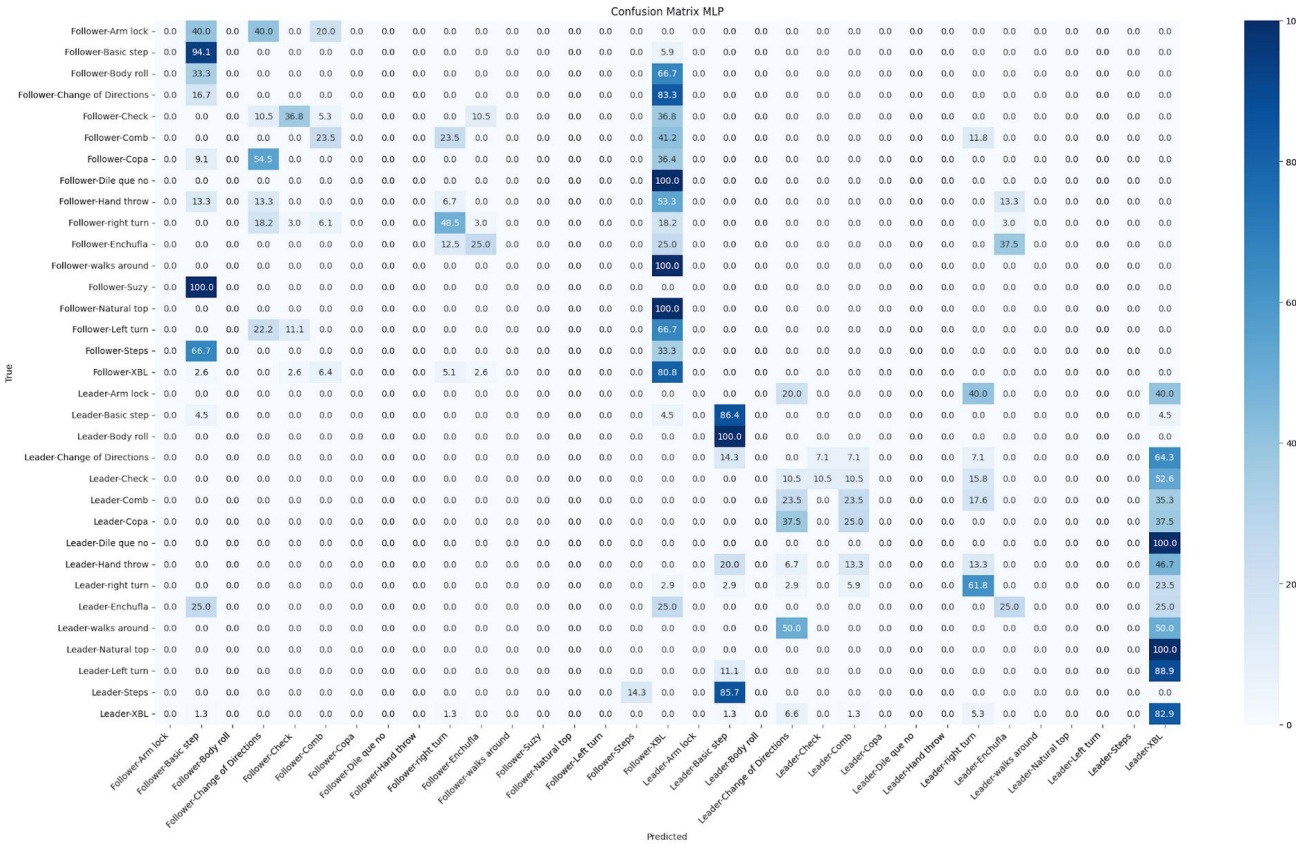

Figure 9. **Multi-layer Perceptron** confusion matrix results on the test set conducted on the VQ-VAE embedding space. The model strongly over-classifies for the over-represented classes of XBL and basic step despite weight balancing in the loss function.

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. 9. Training time was approximately 6 minutes and ran on a single NVIDIA GeForce RTX 4070.