Analyzing Creative Collaboration Discourse in VR Music-Making through Computational NLP

Michael Oehler
Osnabrück University
michael.oehler@uos.de

Leonard Bruns
Osnabrück University
lebruns@uos.de

Benedict Saurbier
Osnabrück University
benedict.saurbier@uos.de

Tray Minh Voong
Osnabrück University
travoong@uos.de
*

Abstract

Collaborative music making in virtual reality depends not only on shared interaction with musical material, but also on how participants verbally negotiate ideas, timing, and repair. This paper presents a transcript-based computational approach for analyzing collaborative discourse in VR music-making sessions in PatchWorld, comparing distributed headset-mediated and co-located direct communication settings under otherwise comparable tasks. Rather than treating communication only as a subjective presence outcome, we model it as an observable process through three families of transcript-derived indicators: interaction intensity, affective-friction dynamics, and creativity-relevant discourse proxies.

Methodologically, the paper combines robust transcript preprocessing for noisy ASR data with chunk-based transformer models for emotion-to-valence estimation and multilingual zero-shot classification of communication modes. Using paired session-level comparisons, we characterize how mediation setting is associated with differences in utterance packaging, local semantic movement, and affective stability. Across five groups, the clearest trend is more compact utterance structure in the co-located direct setting, while semantic novelty and valence-based friction measures show weaker but interpretable directional shifts.

The contribution is primarily methodological: an interpretable NLP-based approach for studying collaborative music-creation processes in VR under realistic data constraints. The findings further suggest that discourse organization can serve as a useful analytical layer for AI-supported research on musical creativity and coordination in immersive environments.

1 INTRODUCTION

Collaborative music-making is inherently interactional: participants continuously negotiate timing, roles, ideas, and repair strategies through both musical and verbal actions. In small ensembles, talk is not merely an auxiliary activity, but a functional resource for coordinating musical structure, resolving uncertainty, and shaping creative decisions (Seddon, 2005; Seddon and Biasutti, 2009). Verbal interaction can therefore be understood as an integral component of collaborative musical practice rather than a by-product of it. Against this background, we ask how two communication settings—distributed headset-mediated and co-located direct communication—shape the verbal coordination of musical ideas and actions during collaborative music-making in PatchWorld.

A growing body of work has examined how musicians use language during collaboration. Studies of jazz improvisation and ensemble rehearsal show that verbal exchanges support real-time negotiation of musical form, role distribution, and evaluation (Seddon, 2005). Communication patterns further depend on ensemble context, reflecting different balances between directive, reflective, and socially regulating utterances (Seddon and Biasutti, 2009). Longitudinal work suggests that such discourse evolves with familiarity, shifting from explicit planning toward increasingly compact and routinized interaction (Ginsborg and Bennett, 2021; Pennill and Timmers, 2022). These findings indicate that the microstructure of discourse is closely tied to how musical coordination and creativity unfold in practice.

Related work on digital musical interaction has shown that interface design, visibility, and social configuration can strongly shape collaborative awareness and interaction. This includes studies of social interaction around tangible music interfaces (Xambó et al., 2017) as well as work on visualisation level and situational visibility in co-located digital musical ensembles (Berthaut and Dahl, 2022).

In multi-user virtual reality (VR), these interaction processes are shaped by additional mediation constraints, including audio transmission, latency, and spatial separation. Research on immersive and extended reality environments highlights the importance of social presence, communication channels, and perceived co-presence for interaction quality (Short et al., 1976; Slater and Wilbur, 1997; Biocca et al., 2003; Makransky et al., 2017). Recent work further suggests that immersive social VR can support rich forms of communication, but that outcomes remain sensitive to technical and contextual factors (Turchet et al., 2021). In collaborative music-making, these constraints are particularly critical, as timing precision and rapid coordination are essential for coherent musical outcomes (Turchet and Casari, 2023).

Prior work on collaborative VR music-making has suggested that communication and spatial configurations can affect perceived communication quality, presence, and engagement (Oehler et al., 2025). In particular, direct co-located communication has been associated with higher perceived communication quality, whereas headset-mediated communication can introduce challenges likely related to latency and audio artifacts. However, such work has primarily focused on subjective measures and has not

*

examined how these conditions shape the observable structure of interaction itself.

This leaves an important gap at the intersection of music, interaction, and computational analysis. While existing work shows that communication conditions influence perceived interaction quality, it remains unclear *how* these conditions affect the discourse processes through which participants negotiate musical ideas, coordinate actions, and resolve problems. In particular, there is limited work connecting (i) linguistic interaction structure, (ii) collaborative musical creativity, and (iii) immersive VR environments. Understanding this relationship may also inform the design of interactive and AI-supported systems for musical collaboration.

To address this gap, we model verbal communication as an observable and temporally structured process using transcript-derived features. We focus on three complementary dimensions: interaction intensity, affective-friction dynamics, and creativity-related discourse proxies capturing how participants propose, develop, and coordinate ideas. The approach combines interpretable session-level indicators with transformer-based emotion classification and zero-shot discourse labeling.

The analysis is designed for realistic constraints, including noisy automatic speech recognition (ASR), small sample sizes, and the absence of full speaker diarization. Using paired within-group comparisons across communication settings, we examine directional patterns in discourse organization. The primary contribution is methodological: a reproducible workflow for extracting interpretable interactional signals from collaborative VR music-making data, accompanied by initial empirical observations concerning coordination and idea development.

1.1 Research question and objectives

This paper asks: *How does communication setting (distributed mediated vs. co-located direct) shape verbal interaction during collaborative music production in VR?*

We operationalize this question through three objectives:

1. quantify differences in communication intensity and utterance packaging,
2. compare creativity-related discourse proxies (e.g., semantic novelty and communication modes),
3. characterize affective and interactional dynamics through valence and friction indicators.

1.2 Working hypotheses

Based on prior work and the communication-setting contrast, we formulate three directional hypotheses:

- **H1 (interaction efficiency):** co-located direct communication is associated with more compact verbal packaging, notably fewer words per utterance.
- **H2 (creative discourse):** co-located direct communication is associated with stronger creativity-oriented language signals, such as higher semantic novelty and a greater balance toward idea-development discourse.
- **H3 (interactional stability):** co-located direct communication is associated with lower global interactional friction (e.g., lower valence variance), while local polarity shifts may remain.

Given the small number of independent groups, these hypotheses are evaluated in an estimation-focused, exploratory framework.

2 METHODS

2.1 Study context and design

This paper analyzes verbal interaction data collected in a controlled multi-user VR music-making study comprising two sessions. The first session focused on onboarding and XR familiarization, while the second session was designed as a collaborative multi-user setting to elicit communication during joint music creation in *PatchWorld*. The present analysis focuses exclusively on the second session and compares two communication settings within groups.

2.2 Participants

The full dataset included 27 participants (14 female, 13 male) forming 9 groups of 3. For the present analysis, a subset of 15 participants in 5 groups was used. Participants' age was $M = 32.4$, $SD = 9.9$ (8 female, 7 male). Prior musical experience and pre-existing relationships among group members were not systematically recorded. All participants received financial compensation for completing the study.

2.3 Inclusion criteria for the present corpus

For the present language-focused analysis, we retained only groups with complete data in two communication settings: (1) a distributed setting, in which participants were located in separate physical rooms and communicated via headset microphone and in-ear headphones, and (2) a co-located setting, in which participants were in the same physical room and communicated directly without microphone or loudspeaker mediation.

To ensure comparability, only groups with complete data in both settings were included. The co-located headset-mediated variant was excluded to reduce condition heterogeneity. Accordingly, the analyzed corpus comprises 10 group-level transcripts (5 groups $\times$ 2 settings; 15 participants).

2.4 Apparatus and collaborative task

All participants used Meta Quest 3 head-mounted displays with controllers. Participants first configured avatars (Ready Player Me) and then joined a shared *PatchWorld* environment (PatchXR, 2025). Groups of three collaboratively created short musical sequences (drum, bass, melody) with a step-sequencer workflow and sample manipulation. The collaborative task comprised two scenarios. In Scenario A, participants generated additional sounds by recording vocal input via a virtual recorder and combining these recordings with available samples. In Scenario B, participants worked with a predefined set of sounds without self-recording. Across both scenarios, the shared task was to create a short musical sequence comprising drum, bass, and melody parts.

In both scenarios, participants used core *PatchWorld* operations, including triggering samples via sequencer connections, editing sample start and end points, and arranging pattern variations. Avatars, sample spheres, and the sequencer are illustrated in Figure 1.

Communication setting was manipulated by condition: (i) distributed mediated communication (different physical

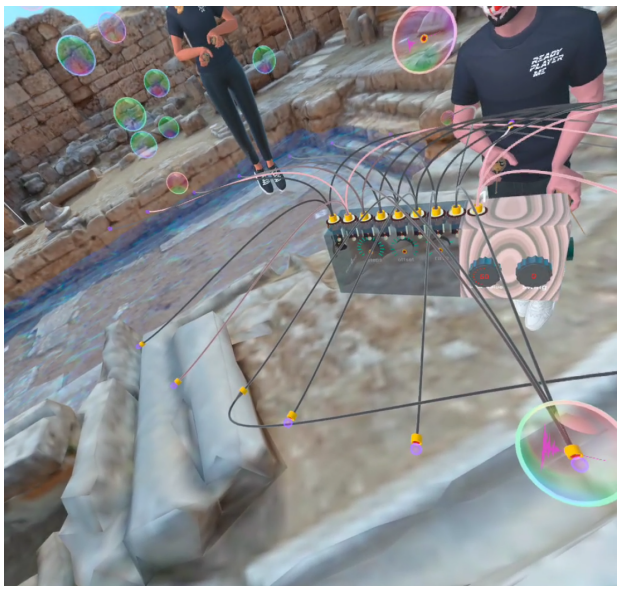

Figure 1: Participants simultaneously working with sample spheres and the step-sequencer in PatchWorld, illustrating the shared musical workspace that required continuous verbal coordination.

rooms; headset microphone and in-ear headphones) and (ii) co-located direct communication (same room; no microphone or loudspeaker mediation). Participants were instructed to communicate as actively as possible in all conditions to reduce communication effort as a confound. The study conductor provided brief task instructions for each scenario and remained available for troubleshooting while minimizing interference with group interaction. Each group completed both task scenarios within a session of approximately 150 minutes. Each scenario lasted approximately 25–30 minutes.

2.5 Condition structure

The study design included multiple communication and spatial configurations. For the present analysis, we retained two target conditions per group:

- **Distributed mediated communication (R1)**: participants in different rooms, verbal interaction via headset microphone and in-ear headphones.
- **Co-located direct communication (R2)**: participants in the same room, verbal interaction without microphone or loudspeaker mediation.

These two conditions form a repeated within-group comparison (R1, R2) for the five retained groups. Order was counterbalanced across groups.

For comparability, analyses were computed on fixed 15-minute transcript windows. These windows were extracted from the active collaborative part of each condition after initial settling-in, so that all sessions contributed equally long segments of task-focused interaction.

2.6 Transcription and text preparation

Audio from the collaborative session was first transcribed automatically using a local installation of OpenAI Whisper (`large-v3`, language set to German) and then conservatively post-corrected. Cleaning targeted obvious ASR artifacts (e.g., repeated filler strings, non-linguistic Unicode fragments, and bracketed meta-notes), while preserving semantically interpretable utterances. As a final quality-control step, all session audio tracks were replayed in full and transcription errors were manually corrected against the original recordings. After this verification pass, transcripts were processed as plain text. For emotion analysis, each session was divided into consecutive, non-overlapping chunks of at most 350 word tokens defined by a regular-expression tokenizer. For zero-shot discourse classification and semantic-novelty analysis, consecutive, non-overlapping chunks of at most 320 whitespace-delimited tokens were used. The final chunk of each session was retained at its remaining length.

2.7 Computational measures

2.7.1 Communication intensity

We computed session-level intensity proxies:

- total word count (n_{words}),
- words per minute,
- estimated syllables per minute (vowel-group heuristic),
- estimated utterances per minute (punctuation and line-break proxy),
- words per utterance.

Words were defined as alphabetic token sequences including German umlauts and ß, with internal apostrophes permitted. Estimated syllables were calculated as the number of contiguous vowel groups per word, with a minimum of one syllable assigned to each non-empty word. Utterances were approximated by splitting transcripts at one or more sentence-final punctuation marks (`.`, `?`, or `!`) or at line breaks and retaining non-empty segments containing at least one letter. Words per utterance was calculated as the total word count divided by the resulting utterance count. These measures quantify how much participants verbalized, how quickly they spoke, and how densely information was packed into utterances.

2.7.2 Emotion-to-valence dynamics

For affective dynamics, transcripts were analyzed in consecutive, non-overlapping chunks of at most 350 word tokens using the GoEmotions-based checkpoint `SamLowe/roberta-base-go_emotions` (Demszky et al., 2020). The model returned a score $s_{t,e}$ for each emotion label e in chunk t. These scores were projected onto a signed valence dimension:

$$v_t = \sum_e w_e s_{t,e},$$

where $w_e = +1$ for joy, amusement, approval, gratitude, love, optimism, pride, relief, and excitement; $w_e = -1$ for anger, annoyance, disapproval, disgust, fear, sadness, grief, remorse, embarrassment, nervousness, confusion, and frustration; and $w_e = 0$ for all remaining labels.

From the resulting chunk-level valence sequence, we computed:

- mean valence, defined as the arithmetic mean of v_t,
- valence variance, calculated across chunks using the population variance,
- negative peaks, defined as chunks with $v_t < -0.05$,

- sign switches, defined as changes in the sign of v_t between adjacent chunks,
- switch rate, defined as the number of sign switches divided by the number of adjacent chunk transitions.

This signed projection intentionally reduces the fine-grained emotion outputs to a simple comparative session-level signal. It should not be interpreted as a comprehensive model of participants' emotional states.

2.7.3 Communication mode and creativity proxies

To characterize *how* groups talked, not only how much, we applied multilingual zero-shot classification using an XLM-RoBERTa model fine-tuned on XNLI (Conneau et al., 2020), implemented via the Hugging Face checkpoint `joeddav/xlm-roberta-large-xnli`. Each transcript chunk was evaluated using German candidate-label formulations and the hypothesis template `Dieser Text ist geprägt durch {}.` Classification was performed in multi-label mode, so that the score for each communication mode was estimated independently and scores were not constrained to sum to one. For readability, the corresponding English label descriptions are listed below:

1. proposing new musical ideas,
2. elaborating or transforming ideas,
3. creative or metaphorical language,
4. timing and coordination talk,
5. technical problem-solving,
6. social filler or small talk.

For each chunk t, a creativity-related discourse proxy was defined as

$$CI_t = \frac{1}{|L^+|} \sum_{\ell \in L^+} s_{t,\ell} - \frac{1}{|L^-|} \sum_{\ell \in L^-} s_{t,\ell},$$

where $s_{t,\ell}$ denotes the classifier score for label ℓ in chunk t. The set L^+ contains idea-generation, elaboration, and creative-language labels, whereas L^- contains coordination, technical, and filler labels. The index captures the relative balance between creative-generative and coordination-oriented discourse modes and should not be interpreted as a direct measure of creative quality or originality. A value of zero indicates equal mean classifier scores for the two label groups. The session-level creativity-related discourse proxy was calculated as the arithmetic mean of CI_t across all chunks in the session.

At session level, we further computed:

- corrected type-token ratio, defined as

$$\mathrm{CTTR} = \frac{V}{\sqrt{2N}},$$

  where V is the number of unique word types and N the total number of word tokens;
- semantic novelty, defined as one minus the mean cosine similarity between TF–IDF vectors of adjacent chunks. TF–IDF representations were fitted separately for each session using unigrams and bigrams with a minimum document frequency of one.

2.7.4 Rationale for metric selection

We selected a compact set of transcript-derived indicators that are (a) interpretable at group level, (b) robust to moderate ASR noise, and (c) comparable across fixed-duration sessions. Intensity features capture interaction load and utterance packaging; valence and friction features capture affective stability versus fluctuation over time; and creativity-oriented language features capture idea-development tendencies beyond raw verbosity. Given the current data constraints (small number of independent groups, no full speaker diarization), we prioritized transparent, reproducible proxies over high-complexity latent models. Transformer components were used where they add clear construct coverage: GoEmotions-based affect labeling for chunk-level emotion profiles (Demszky et al., 2020), and NLI-based zero-shot topic and function labeling with multilingual transfer via XLM-R (Yin et al., 2019; Conneau et al., 2020). Because these models were not specifically fine-tuned for German collaborative music discourse, we interpret their outputs as comparative proxy signals at session level rather than as utterance-level ground truth.

2.8 Statistical strategy

We adopted an estimation-focused analysis strategy at group level. For each metric, condition effects were represented as paired within-group differences:

$$\Delta_g = X_{g,R2} - X_{g,R1}.$$

Primary reporting therefore emphasizes (i) paired trajectories (R1–R2), (ii) the distribution of Δ_g, (iii) median and mean paired deltas, and (iv) direction consistency across groups (number of positive versus negative deltas).

Effect sizes were reported as matched-pairs rank-biserial correlation (RBC), interpreted as a directional magnitude indicator under small-n uncertainty. Hypothesis tests were treated as exploratory. Where reported, exact two-sided Wilcoxon signed-rank tests were used and interpreted cautiously because of coarse p-value resolution in the present design.

To limit multiplicity, primary interpretive emphasis was restricted to a small set of predefined focal outcomes (creativity proxy mean, semantic novelty, and words per utterance), while all additional metrics were treated as descriptive or exploratory. Benjamini–Hochberg correction was applied separately within the intensity, creativity, and friction metric families.

2.9 Reproducibility

All preprocessing and analysis steps were implemented in Python (`pandas`, `numpy`, `scipy`, `transformers`, `scikit-learn`) and exported as CSV artifacts at chunk and session levels. The pipeline is modular, enabling direct replacement of label sets, chunk sizes, and threshold definitions for sensitivity analyses.

3 RESULTS

3.1 Analysis scope and reporting strategy

The present corpus comprises 10 transcripts from 5 groups (paired design: two settings per group). We report paired descriptive statistics as the primary evidence and treat inferential tests as exploratory. Throughout this section, we focus on within-group changes ($\Delta = R2 - R1$), where

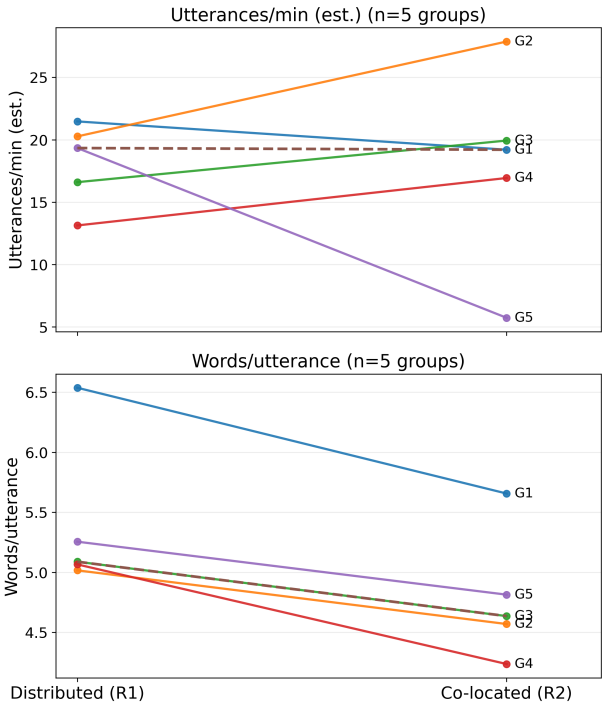

Figure 2: Paired intensity trajectories across settings ($R1 \rightarrow R2$) for each group (G1–G5). Top: estimated utterances per minute. Bottom: words per utterance. Each line represents one group-level paired observation.

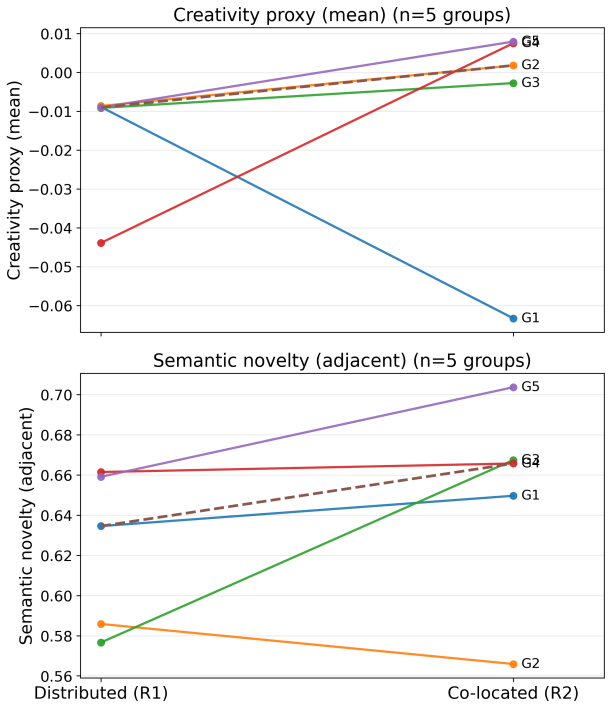

Figure 3: Paired creativity trajectories across settings ($R1 \rightarrow R2$) for each group (G1–G5). Top: creativity proxy (mean; balance between creative-generative and coordination-oriented discourse modes). Bottom: semantic novelty (adjacent chunks).

R1 denotes distributed mediated communication and R2 co-located direct communication. Session-level descriptives are summarized in Table 1; exploratory Wilcoxon tests and effect sizes are reported in Table 2. Paired line plots (spaghetti plots) are shown in Figures 2 and 3, and temporal valence dynamics are shown in Figure 4. For figure readability, we visualize two representative metrics per family that capture complementary aspects of interaction dynamics. Paired results for the session-level outcomes reported in this paper are summarized in the tables.

3.2 Communication intensity

Communication intensity showed mixed directional patterns across groups (Figure 2). At the paired level, median deltas suggested slight increases in words per minute (+5.20), syllables per minute (+9.07), and utterances per minute (+3.33); however, mean deltas were negative for these metrics due to strong between-group variability and at least one low-output session (Table 1). A more consistent directional pattern was found for words per utterance, with a negative paired median delta (-0.45), indicating shorter utterances in $R2$.

Exploratory Wilcoxon tests did not yield statistically robust differences after false-discovery-rate correction (Table 2). The largest directional effect in this family was observed for words per utterance ($p = .0625$, $q = .25$, $RBC = -1.00$), suggesting a potentially meaningful shortening of utterances in the co-located direct condition ($R2$), but not at confirmatory significance levels. These patterns are partially consistent with H1, primarily reflecting shorter utterances under co-located direct communication, suggesting more compact packaging, although effects remain at trend level.

3.3 Creativity proxies and communication mode

For the creativity proxy family (Figure 3), the creativity proxy remained near zero in both settings and showed a small positive shift at paired level (median $\Delta = +0.010$). CTTR showed a small negative shift (median $\Delta = -0.598$), while semantic novelty (adjacent-chunk distance) showed a positive shift (median $\Delta = +0.015$); see Table 1. This combination is compatible with a pattern of somewhat less lexically diverse but more locally shifting discourse content in $R2$, though variability across groups was substantial.

None of the creativity-family tests were significant after false-discovery-rate correction (all $q \geq .4688$; Table 2). Still, effect-size estimates were non-trivial for CTTR ($RBC = -0.60$) and semantic novelty ($RBC = +0.60$), which is useful for planning follow-up studies with higher power. Overall, these results provide limited support for H2, with only partial and inconsistent changes in creativity-related discourse proxies.

3.4 Valence trajectory and friction trends

To provide a descriptive view of ordered valence dynamics, chunk-level valence scores were mapped to normalized session position based on chunk order and grouped into five bins (Figure 4). This representation reflects relative pseudo-time rather than timestamp-aligned audio time. Friction indicators were summarized separately at session level (Table 1). Across normalized position bins, median valence remained close to neutral in both settings, with localized fluctuations and partially overlapping interquartile bands.

At the session-derived friction level, paired descriptives indicated a small positive shift in mean valence (median

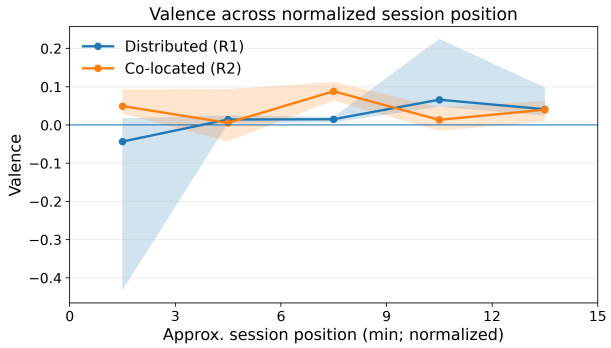

Figure 4: Descriptive valence trajectory across normalized session position for $R1$ and $R2$ (median $\pm$ IQR). Ordered chunks were evenly rescaled to the 15-minute analysis window and assigned to five position bins. The horizontal axis therefore represents approximate pseudo-time based on chunk order, not timestamp-derived three-minute intervals.

$\Delta = +0.002$) and a decrease in valence variance (median $\Delta = -0.038$), while sign-switch rate increased slightly (median $\Delta = +0.333$); see Table 1. Taken together, this pattern suggests marginally more positive average affect in $R2$, accompanied by lower global instability but still present local polarity transitions over time.

Exploratory Wilcoxon tests showed no false-discovery-rate-corrected significance (Table 2). Within this family, the strongest directional signal was observed for lower valence variance ($p = .125$, $RBC = -0.87$), which should be interpreted as trend-level evidence. This pattern is broadly consistent with H3, indicating reduced global fluctuation with continued local variation, although effects remain small and variable across groups.

4 DISCUSSION

This study examined how communication setting shapes verbal interaction during collaborative music production in social VR. Across the three metric families, the results suggest weak directional tendencies rather than confirmatory effects, reflecting the constraints of the paired design. Some measures showed recurrent directional patterns across groups, particularly words per utterance, but the present data do not establish reliable condition effects beyond this sample. The findings are therefore best understood as hypothesis-generating indications of how communication setting may relate to the organization of coordination and idea negotiation during collaborative music-making in VR.

Relative to the working hypotheses, the results are best interpreted as follows. **H1 (interaction efficiency)** received the clearest support at trend level: words per utterance decreased consistently across groups, suggesting more concise and potentially more efficient coordination language under co-located direct communication. **H2 (creative discourse)** was only partially supported: semantic novelty increased, but the aggregate creativity proxy remained close to zero and CTTR decreased slightly, indicating that lexical diversity and local semantic movement capture different aspects of creative interaction and do not necessarily covary in small collaborative sessions. **H3 (interactional stability)** was also partially supported: valence variance decreased (trend toward lower global instability), while sign-switch rate did not decrease, indicating continued local polarity transitions despite reduced overall fluctuation, which may reflect the ongoing need for fine-grained coordination even under more stable communication conditions.

These discourse patterns are relevant to music co-creation because participants were not completing a generic collaborative task: they negotiated the allocation and arrangement of drum, bass, and melody parts, triggered and edited samples, and coordinated pattern variations in a shared sequencer. Within this context, shorter utterances may reflect more immediate coordination around shared musical objects, while changes in adjacent-chunk semantic novelty may reflect shifts between musical ideas, parts, or production parameters. Because utterances were not time-aligned with sequencer actions, these interpretations remain provisional rather than direct evidence about musical outcomes.

4.1 Limitations

Several limitations define the current scope:

- **Text-centered analysis**: the present analysis is primarily text-based. We did not align transcript features with time-synchronous VR interaction logs or musical actions (e.g., sequencer edits, sample manipulations, or controller events). Consequently, conclusions about language–action coupling in collaborative music-making remain preliminary. Moreover, the linguistic measures are not uniquely musical; their music-specific interpretation derives from the collaborative task context and requires validation through alignment with musical actions.
- **Number of independent units**: five paired groups, which limits inferential resolution.
- **Task specificity**: interaction was observed in one concrete musical workflow (step-sequencer collaboration in *PatchWorld*); transfer to other musical practices (e.g., improvisation, conducting, live coding) is not guaranteed.
- **Creativity operationalization**: creativity was approximated via proxy metrics (zero-shot communication-mode scores, CTTR, semantic novelty). These proxies are informative but do not exhaustively represent creative quality.
- **Restricted analytic lens**: the current model family focuses on three domains (intensity, emotion and friction, creativity). Other linguistically relevant phenomena were not modeled explicitly.
- **Model generalization and validation**: the models were not fine-tuned on German conversational, ASR-derived, or music-related discourse. Their outputs may therefore be affected by language and domain mismatch and should be interpreted as relative session-level proxy signals rather than ground-truth classifications. The signed valence projection additionally reduces fine-grained emotion outputs to a coarse comparative dimension. We did not perform manual annotation or agreement checks in the current study, which limits conclusions about model validity at the utterance or chunk level.

4.2 Extended analysis agenda

Following the broader research roadmap, future iterations should expand beyond the current three-domain focus and include richer interactional modeling:

Table 1: Paired descriptive results at session level (5 groups, 15-minute windows). Distributed (R1) and co-located (R2) are condition-wise medians across groups; $\Delta = R2 - R1$ summarizes paired within-group differences. Because the median Δ is calculated from the five group-wise differences, it need not equal the difference between the two condition medians.

Metric	R1 median	R2 median	Median Δ	Mean Δ
Words/min	101.60	92.40	+5.20	-13.39
Syllables/min (est.)	140.40	127.07	+9.07	-20.43
Utterances/min (est.)	19.33	19.20	+3.33	-0.23
Words/utterance	5.09	4.64	-0.45	-0.61
Creativity proxy	-0.0090	+0.0018	+0.0104	+0.0062
CTTR	7.63	7.42	-0.60	-0.22
Semantic novelty (adjacent)	0.635	0.666	+0.015	+0.027
Mean valence	0.028	0.030	+0.002	+0.016
Valence variance	0.041	0.003	-0.038	-0.047
Sign-switch rate	0.333	0.500	+0.333	+0.150

Table 2: Exploratory paired Wilcoxon tests (two-sided) with within-family Benjamini–Hochberg correction. RBC: rank-biserial correlation (directional effect size for paired samples).

Family	Metric	n	Median Δ	W	p	q_{FDR}	RBC
Intensity	Words/min	5	+5.20	6.0	.8125	.8125	-0.20
Intensity	Syllables/min (est.)	5	+9.07	6.0	.8125	.8125	-0.20
Intensity	Utterances/min (est.)	5	+3.33	6.0	.8125	.8125	+0.20
Intensity	Words/utterance	5	-0.45	0.0	.0625	.2500	-1.00
Creativity	Creativity proxy	5	+0.0104	5.0	.6250	.6250	+0.33
Creativity	CTTR	5	-0.5977	3.0	.3125	.4688	-0.60
Creativity	Semantic novelty (adjacent)	5	+0.0151	3.0	.3125	.4688	+0.60
Friction	Mean valence	5	+0.0016	4.0	.4375	.6250	+0.47
Friction	Valence variance	5	-0.0378	1.0	.1250	.3750	-0.87
Friction	Sign-switch rate	5	+0.3333	5.0	.6250	.6250	+0.33

1. **Dialogue-act structure**: add utterance-level dialogue-act annotation or classification using meeting-tested frameworks (MRDA and AMI) and standardized schemes (ISO 24617-2), enabling sequential interaction modeling rather than aggregate counts (Stolcke et al., 2000; Shriberg et al., 2004; Carletta et al., 2005).
2. **Speaker-conditioned conversation modeling**: integrate speaker diarization to distinguish participants and model sequential dependencies such as who responds to whom, how ideas are taken up across turns, and how repair is distributed among group members (Majumder et al., 2019).
3. **Turn-taking and repair dynamics**: combine transcript features with timestamp-accurate audio measures (gaps, overlaps, self-repair and other-repair), grounded in conversation-analytic turn-taking theory (Sacks et al., 1974).
4. **Alignment and coordination metrics**: complement LLM outputs with interpretable measures such as lexical entrainment, linguistic style matching, and interactional alignment (Pickering and Garrod, 2004; Ireland et al., 2011).
5. **Domain adaptation**: adapt language models to PatchWorld discourse using limited-label strategies such as DAPT, TAPT, or parameter-efficient fine-tuning (Gururangan et al., 2020; Hu et al., 2022).

Taken together, these extensions would support a stronger multimodal account of collaboration by distinguishing individual speakers and conversational sequences and by time-aligning utterances with sequencer edits, sample triggers, controller events, communication-quality ratings, and presence-related outcomes.

5 CONCLUSIONS

This paper presented an estimation-focused, transcript-based analysis of verbal interaction in triadic VR music collaboration across two communication settings. While the observed effects are trend-level rather than confirmatory, they point to systematic differences in how discourse is organized under mediated versus co-located communication conditions. In particular, the results suggest more compact utterance structures, modest increases in local semantic variation, and reduced global valence fluctuation in the co-located direct setting.

These findings should be interpreted as hypothesis-generating rather than conclusive. The primary contribution of this work is methodological: it demonstrates a transferable approach for extracting interpretable communication signals from collaborative music-making data under realistic constraints, including noisy ASR and limited speaker information.

More broadly, the results highlight the potential of discourse-based analysis as a complementary perspective on collaborative music-making in immersive environments. Future work should distinguish individual speakers and conversational sequences and time-align utterances with sequencer edits, sample triggers, and controller events. Such integration would allow more direct investigation of

how verbal communication relates to musical coordination, decision-making, and creative development in VR.

AUTHOR DECLARATIONS

The authors declare that they have no conflict of interest (financial or non-financial). The study was positively reviewed and approved by the relevant institutional ethics committee. Informed consent was obtained from all individual participants included in the study. Participants received EUR 150 for participation in all parts of the study. No animals were involved in this research.

REFERENCES

Berthaut, F. and Dahl, S. (2022). The effect of visualisation level and situational visibility in co-located digital musical ensembles. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*.

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators and Virtual Environments*, 12(5):456–480.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2005). The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction (MLMI 2005)*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ginsborg, J. and Bennett, D. (2021). Developing familiarity in a new duo: Rehearsal talk and performance cues. *Frontiers in Psychology*, 12:590987.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL 2020*, pages 8342–8360.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., and Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Makransky, G., Lilleholt, L., and Aaby, A. (2017). Development and validation of the multimodal presence scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, 72:276–285.

Oehler, M., Bruns, L., Saurbier, B., and Voong, T. M. (2025). Impact of spatial and communication environments on collaborative music-making in virtual reality: Insights for music education. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 619–625. IEEE.

PatchXR (2025). Patchworld: Multi-user vr environment for creative music making. https://www.patchxr.com/. Software/platform reference.

Pennill, N. and Timmers, R. (2022). Patterns of verbal interaction in newly formed music ensembles. *Frontiers in Psychology*, 13:987775.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Seddon, F. (2005). Modes of communication during jazz improvisation. *Psychology of Music*, 33(1):47–55.

Seddon, F. and Biasutti, M. (2009). Modes of communication between members of a string quartet. *Small Group Research*, 40(2):115–137.

Short, J., Williams, E., and Christie, B. (1976). *The Social Psychology of Telecommunications*. Wiley, London.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100.

Slater, M. and Wilbur, S. (1997). A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Turchet, L. and Casari, P. (2023). Latency and reliability analysis of a 5g-enabled internet of musical things system. *IEEE Internet of Things Journal*, 11(1):1228–1240.

Turchet, L., Hamilton, R., and Camci, A. (2021). Music in extended realities. *IEEE Access*, 9:15810–15832.

Xambó, A., Hornecker, E., Marshall, P., Jordà, S., Dobbyn, C., and Laney, R. (2017). Exploring social interaction with a tangible music interface. *Interacting with Computers*, 29(2):248–270.

Yin, W., Hay, J., Roth, D., et al. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*.