DIAGNOSING INTERACTION EFFECTS IN POSE PREDICTION: A CROSS-SCENARIO FREQUENCY-DOMAIN METHODOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent pose prediction, the task of forecasting future body poses for multiple people interacting in a shared scene, has become a foundational capability for autonomous systems that share space with people. The standard evaluation methods report a single aggregated error number per model, which cannot separate individual-motion difficulty from the error associated with inter-agent interaction. We address this gap with a three-prong frequency-domain methodology that diagnoses prediction error through complementary spectral decompositions: Fast Fourier Transform, bandpass filtering, and discrete wavelet transform. Applied to both interaction-naive and interaction-aware predictors across different testing scenarios, the three prongs converge on a same pattern: the interaction-associated penalty concentrates below 5 Hz.The results show where interaction stresses pose prediction and suggest multi-agent pose prediction to allocate its interaction-specific modeling capacity in low-to-mid frequencies.

1 INTROUCTION

Human pose prediction aims to forecast future human motion from observed motion history, with applications across a wide range of computer vision domains, including human-robot teaming (HRT). In HRT settings, robots must anticipate their human partners' motions over short future horizons, typically one or two seconds, in order to collaborate efficiently, reduce human cognitive load, and maintain safety. The fundamental task underlying this capability is *multi-agent pose prediction*. While early work often modeled each agent as an isolated subject (Fragkiadaki et al., 2015; Alahi et al., 2016; Martinez et al., 2017; Mao et al., 2019; Aksan et al., 2021), multi-agent pose prediction is not simply the aggregation of multiple independent single-agent forecasts. A central defining factor is the **interaction** among agents. In collaborative environments, agents (e.g., robots or their human partners) do not move independently; instead, they continuously influence and adapt to one another through their relative positions, motion dynamics, and coordination patterns (Figure 1(a)). In other words, each agent's future motion is shaped not only by its own motion history, but also by its interactions with others. Therefore, computationally modeling and evaluating such interactions is imperative for advancing multi-agent pose prediction, which distinguishes this task from conventional single-agent pose prediction.

Recent work has explored model architectural designs that explicitly model interactions among agents (Zheng et al., 2025; Jeong et al., 2024; Peng et al., 2023; Xu et al., 2023; Wang et al., 2021). These studies typically follow a conventional evaluation protocol: benchmarking a proposed architecture against baselines on shared multi-person datasets and reporting a single aggregated error, predominantly using scalar L2-based error metrics such as MPJPE, APE, FDE, and JAE. A lower aggregated error is then often assumed to indicate better interaction modeling. However, this assumption is problematic. In multi-agent settings, a single aggregated error conflates two fundamentally distinct sources of prediction difficulty: (i) individual motion modeling and (ii) inter-agent interaction modeling. A lower error confirms only that the model predicts overall motion more accurately, but reveals nothing about whether the gain comes specifically from improved interaction modeling or from other factors, such as stronger single-agent motion priors. Consequently, the very capability the field claims to improve, namely interaction modeling, is rarely directly measured.

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

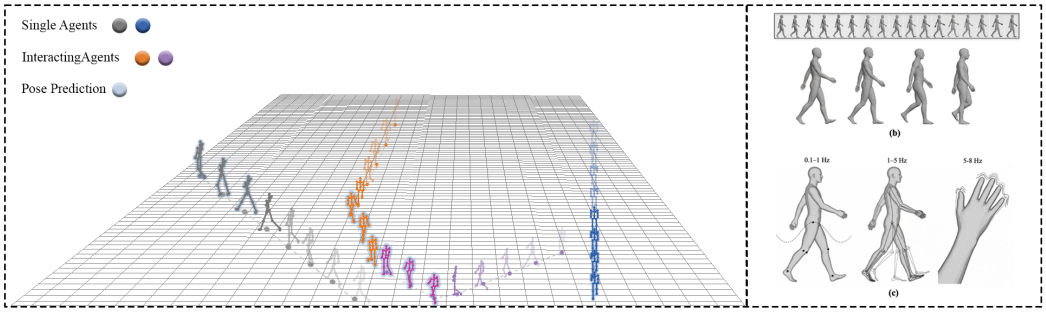

Figure 1: **(a)** Interacting agents change their motion trajectory to influence and adapt one another, making their motion prediction more challenging. Single agents' motion prediction is less complex. **(b)** Human motion in the time domain, viewed by window or by frame. **(c)** Human motion in the frequency domain, frequency bands reveal different motion patterns: The low band (0-1 Hz) captures gait cadence, mid-band (1-5 Hz) covers the bulk of voluntary motion, such as walking rhythm, and high-band (5-10 Hz) covers rapid voluntary motion, such as rapid hand movement.

To isolate and directly measure the effect of interaction, its contribution must be separated from the difficulty of modeling individual motion. This separation requires a controlled reference. A natural way to obtain this reference is to evaluate the same architecture in both single-agent and multi-agent scenarios, which we term *cross-scenario* comparison. Since the single-agent scenario contains no inter-agent interaction, its prediction error primarily reflects the model's capacity for individual motion modeling. Comparing this reference against the multi-agent scenario then allows us to quantify the additional prediction difficulty introduced by interaction. However, designing a valid and interpretable cross-scenario comparison is non-trivial. First, *the comparison must control for motion scale.* Single- and multi-person datasets can differ substantially in motion amplitude. For instance, a "sitting" scenario, where people barely move, is not directly comparable to a "fighting" scenario, where motion is extreme. Without such control, a larger error in the multi-agent scenario may simply incurred by larger motion magnitude rather than a genuine failure to model interaction. Second, *the evaluation metric must go beyond a single scalar summary.* Even after normalizing for motion scale, an aggregated error can only quantify how much interaction adds to overall prediction difficulty; it cannot reveal where that difficulty is concentrated or what types of motion are most disrupted. These two challenges expose **a basic evaluative gap: the field currently lacks an evaluation methodology that can measure how interaction shapes pose-prediction error in a scale-controlled and diagnostically informative way.**

To bridge this gap, we propose a new evaluation methodology that performs cross-scenario comparison in the *frequency domain*, rather than relying only on the conventional *time-domain* view. Our design choice is motivated by a key limitation of time-domain evaluation: it can identify when prediction errors occur by examining errors frame by frame or over temporal windows, but it cannot reveal what types of motion patterns are disrupted by interaction (Figure 1(b)). This limitation is critical because interaction does not affect all motion uniformly. In collaborative settings, agents may simultaneously exhibit slow, sustained coordination and fast, transient reactions. Time-domain analysis collapses these motion scales onto a single temporal axis, making it difficult to determine whether interaction mainly disrupts long-term coordination, rapid reactions, or both. In contrast, frequency-domain analysis decomposes human motion along a scale axis. As shown in Figure 1(c), low-frequency components capture slow and sustained motion patterns, while high-frequency components capture faster and more transient movements. Importantly, the frequency domain itself is not foreign to human pose prediction. A common technique, the Discrete Cosine Transform (DCT), has been widely adopted to compactly represent motion sequences (Jeong et al., 2024; Peng et al., 2023; Wang et al., 2021; Mao et al., 2019; 2020; Dang et al., 2021). However, existing work has primarily used DCT as a basis transformation for data preprocessing. The frequency domain is merely used as a representational space, not as an analytical tool for understanding prediction error. Its potential for revealing interaction-induced error patterns has therefore remained largely unexplored. Therefore, performing cross-scenario comparison in this domain allows us to move beyond asking

when interaction increases prediction error, and instead ask "How does inter-agent interaction fundamentally reshape prediction error across different motion scales?"

To the best of our knowledge, this is the first evaluation methodology that comprehensively diagnoses pose-prediction error in the frequency domain. In summary, our contributions are threefold:

• **We design a three-prong cross-scenario evaluation methodology for pose prediction.** Our methodology consists of three complementary frequency-domain decompositions, each examining prediction error from a different methodological angle. Together, they enable scale-controlled and diagnostically informative comparison of prediction error across paired scenarios, including single-agent versus multi-agent settings.

• **We perform a cross-architecture, cross-dataset study revealing how interaction stresses pose prediction.** We design an interaction-naive probe to expose interaction's effect: evaluating three representative single-agent predictors across multiple datasets and scenarios. We then validate the findings using interaction-aware predictors under a controlled same-backbone comparison. Across these settings, our three-prong analyses reveal a consistent frequency-localized signature of interaction-associated error.

• **We discuss architectural implications for future multi-agent pose prediction.** Our three-prong analyses show that interaction-related errors concentrate primarily in the low-to-mid frequency range with distinct timing and amplitude characteristics. We discuss how these findings motivate model architectures to allocate interaction-modeling capacity selectively across frequency bands, and suggest modeling in frequency-domain as a promising direction for future multi-agent predictors.

2 Related Work

Human Pose Prediction. Human pose prediction involves predicting future human motion from observed motion history. In the early stage, methods were developed for single-agent prediction: from establishing the autoregressive paradigm and the zero-velocity baseline (Martinez et al., 2017; Fragkiadaki et al., 2015), to graph convolutional networks (Mao et al., 2019), transformer-based models (Aksan et al., 2021; Mao et al., 2020), and lightweight MLP designs (Guo et al., 2023). However, for various computer vision downstream tasks, pose predictions need to be multi-agent. The defining challenge, as stated in the introduction, is interaction modeling. Most recent architectural effort targets this challenge directly, through cross-agent attention modules (Wang et al., 2021), body-part-level interaction self-attention (Peng et al., 2023), and dedicated interaction perception modules (Zheng et al., 2025; Jeong et al., 2024). Despite the novelty of these architectural designs, the standard evaluation protocol remains aggregated scalar error metrics (MPJPE, APE, FDE, JAE). As discussed in Section 1, this protocol cannot isolate interaction modeling from individual motion modeling, and therefore cannot directly measure the effect of interaction modeling. Under this background, we construct our three-prong evaluation methodology for understanding the interaction-induced error and revealing where in motion structure those errors accumulate.

Understanding Human Motion in Frequency Domain. Transforming human motion data from time-series into frequency-domain representations has a long history in human motion analysis with various techniques. Discrete Cosine Transform (DCT) is by far the most common frequency-domain technique in pose prediction, as it converts complex human motion data to a compact representation. DCT-as-preprocessing convention became a standard since its introduction (Mao et al., 2019) and subsequent validations (Zheng et al., 2025; Jeong et al., 2024; Peng et al., 2023; Wang et al., 2021; Dang et al., 2021), but its use remains purely representational. It is used as a basis transformation for data preprocessing, not as a tool for diagnosing prediction error. Discrete Wavelet Transform (DWT) has been used in human pose estimation and prediction, where its time-localized, multi-resolution structure is well suited to aperiodic human motion (Gao et al., 2025; Ren et al., 2025; Feng et al., 2024). Fast Fourier Transform (FFT)-based spectral analysis is a standard tool in biomechanics for characterizing the frequency content of human gait, balance, and tremor signals (Winter, 2009). In machine learning, FFT-based decomposition has been applied to trajectory prediction in autonomous driving and pedestrian forecasting (Xia et al., 2025; Yao et al., 2024; Wong et al., 2022), generally as a feature representation rather than as an error diagnostic. While frequency-domain methods effectively characterize raw motion content, they have not been used to evaluate prediction error. In

this work, we construct our evaluation methodology via frequency-domain analysis to investigate what kinds of motion patterns inter-agent interaction disrupts.

3 THE 3-PRONG METHODOLOGY

The three prongs analyze the predictions from three complementary views: frequency (Fourier), time-domain bands (filtering), and time-frequency (wavelets). Each prong addresses a specific limitation of the previous one, providing a complete picture of where and how the model fails.

3.1 PRONG 1: FAST FOURIER TRANSFORM ANALYSIS

Prong 1 measures error directly in the frequency domain. For each sequence, we compute the discrete Fourier transform (DFT) of the prediction and ground truth across the prediction window, report three spectral errors, and aggregate across sequences. The procedure provides the canonical spectral reference: it makes no assumptions about band structure and exposes per-frequency-bin error directly.

Let $G_s(f)$ and $P_s(f)$ denote the per-bin DFT coefficients of ground truth and prediction for sequence s (per joint and axis), and let $E_s = P_s - G_s$. We report the Error Vector Magnitude (EVM), which normalizes the error spectrum by the ground-truth spectrum:

$$\mathrm{EVM}(f) = \sqrt{\frac{\langle |E_s|^2 \rangle}{\langle |G_s|^2 \rangle}}. \tag{1}$$

Here $\langle \cdot \rangle_s$ denotes the ensemble mean over sequences. Zero means a perfect match. 1 is not an upper bound; a value > 1 means the prediction is worse than predicting zero in that bin.

While EVM captures total error, it conflates amplitude and timing mismatches. To disentangle these, we report two complementary diagnostics:

- **Magnitude EVM**: Measures amplitude mismatch, isolating energy undershoot or overshoot. Letting $r(f) = \langle |P_s|^2 \rangle_s / \langle |G_s|^2 \rangle_s$, we compute $\mathrm{EVM}_{\mathrm{mag}}(f) = 1 - \sqrt{r(f)}$.
- **Phase EVM**: Measures timing error, reflecting the error if the prediction had the correct amplitude. Letting $A(f) = \langle \cos(\arg P_s - \arg G_s) \rangle_\phi$ denote phase alignment (averaged only over defined phases), we compute $\mathrm{EVM}_{\mathrm{phase}}(f) = \sqrt{2(1 - A(f))}$.

These diagnostics provide interpretable attributions of the total EVM. However, they serve as complementary metrics rather than a strict additive decomposition of EVM^2, as phase and amplitude errors interact. Appendix A explains the distinction and the averaging requirements.

3.2 PRONG 2: BANDPASS FILTER ANALYSIS

Prong 2 measures error in physical time-domain units (mm of MPJPE) per band. We apply a zero-phase bandpass filter to both the prediction and the ground truth and compute standard MPJPE on the filtered signals. This procedure addresses two limitations of Prong 1: (i) spectral coefficients are difficult to interpret as physical prediction error, and (ii) for shorter sequences, the FFT windowing assumption of periodicity is poorly justified. For a given band b, joint j, and sequence s, we compute the per-band error and matching ground-truth magnitude:

$$e^{(2)}_{bjs} = \frac{1}{h} \sum_{t=1}^{h} \|\hat{\mathbf{x}}^{(b)}_{jts} - \mathbf{x}^{(b)}_{jts}\|_2, \quad g^{(2)}_{bjs} = \frac{1}{h} \sum_{t=1}^{h} \|\mathbf{x}^{(b)}_{jts}\|_2, \tag{2}$$

where h is prediction horizon. We report the **Per-band Relative Error**, $R^{(2)}_b = \sum_{s,j} e^{(2)}_{bjs} / \sum_{s,j} g^{(2)}_{bjs}$. This dimensionless ratio directly answers "how many millimeters of error live in each band?" relative to the ground-truth motion. Because it is amplitude-invariant, it allows fair comparison across scenarios with different motion energy profiles.

3.3 PRONG 3: DISCRETE WAVELET TRANSFORM ANALYSIS

Prong 3 decomposes the prediction window using multi-level DWT and reports a per-coefficient analog of MPJPE in each sub-band. Wavelets are a natural complement to bandpass filtering at short, non-stationary horizons: each DWT coefficient is localized in both scale (frequency band) and time (within the window), addressing a limitation of Prong 1 — the periodicity assumption of the FFT and a limitation of Prong 2 — finite-window filter transients near band edges.

For each sub-band b, joint j, and sequence s, we stack the per-axis wavelet coefficients into a tensor of shape $(N_b, 3)$, where N_b is the number of coefficients. We compute the per-coefficient Euclidean error and ground-truth magnitude:

$$e^{(3)}_{bjs} = \frac{1}{N_b}\sum_{k=1}^{N_b} \|\hat{\mathbf{c}}_{bjsk} - \mathbf{c}_{bjsk}\|_2, \quad g^{(3)}_{bjs} = \frac{1}{N_b}\sum_{k=1}^{N_b} \|\mathbf{c}_{bjsk}\|_2, \tag{3}$$

where $\mathbf{c}_{bjsk} \in \mathbb{R}^3$ is the wavelet coefficient at time index k. This is the wavelet-coefficient analog of MPJPE. We then compute the **Per-coefficient Relative Error**, $R^{(3)}_b = \sum_{s,j} e^{(3)}_{bjs} / \sum_{s,j} g^{(3)}_{bjs}$, following the same dimensionless ratio logic as Prong 2, but localized in time-frequency space.

4 EXPERIMENTS

We design two sets of experiments to serve different diagnostic purposes: *Discovery* and *Validation*. In Discovery, we use an **interaction-naive probe** to make the interaction-specific effect more visible. In Validation, we use **interaction-aware predictors** to test whether the observed error signatures persist when predictors are explicitly designed to learn interactions.

4.1 DISCOVERY: INTERACTION-NAIVE PROBE

To isolate the influence of interaction-specific effect, we need a baseline that knows absolutely nothing about it. We term this the interaction-naive probe. It is a *single-agent* pose predictor trained exclusively on *single-agent* motion data.

The rationale is straightforward: if a model has never been trained on interactions, nor has the capacity to learn interaction, the model knows nothing about interaction. Therefore, the gap between its single-agent and multi-agent performance maximally reflects the error signature associated with interaction, rather than potentially diluted by interaction-specific model design or prior exposure to it.

To fulfill this strict requirement, the predictor must be a single-agent predictor that has no modules designed to learn interaction, and the training data must be completely free of even implicit interaction. We construct this training set by filtering out sequences where a single actor simulates an absent partner (e.g., fighting, being pushed, dodging) using BABEL (Punnakkal et al., 2021) annotations. We use AMASS-derived data (Mahmood et al., 2019) preprocessed via the DIP pipeline (Huang et al., 2018) with updated MoSh++ fitting[1].

We selected three architecturally distinct single-agent predictors to ensure the findings are not model-specific: *Seq2seq* (Martinez et al., 2017), a recurrent encoder-decoder (RNN family); *ST-Transformer* (Aksan et al., 2021), a spatio-temporal transformer (Attention family); and *HisRepItself (HRI)* (Mao et al., 2020), a graph convolutional network (GCN) operating in DCT space.

To thoroughly test the interaction effect, we construct three test scenarios from KIT dataset (Mandery et al., 2015): (i) *single-agent without interaction* (in-distribution baseline), (ii) *multi-agent interaction* (the primary measurement), and (iii) *single-agent novel motion* (a distribution control testing whether novel dynamic single-agent motion produces the same error signature). We further validate our findings across datasets (two testing scenarios from CMU (Carnegie Mellon University, 2003)) and analyze graded interaction levels (three testing scenarios with graded interaction intensity performed by same 59 subjects from BioMotion (Troje, 2002)) in Appendix I.

[1]We re-processed data with MoSh++ rather than the legacy MoSh fits used by DIP. The legacy Mosh fits introduced high-frequency artifacts in multi-person sequences that disappear under MoSh++, ensuring our spectral comparison is not confounded by fitting noise.

4.2 VALIDATION: INTERACTION-AWARE PREDICTOR EVALUATION

If interaction associates specific, localized prediction errors (discovered by the probe), then explicitly modeling interaction should reduce these errors. We use interaction-aware predictors to test this.

We select ST-Transformer as the backbone and add cross-agent attention blocks, which allow the target person's joint features to attend to the partner's joint features at the same frame. We choose ST-Transformer because its per-frame joint embeddings and separate spatial/temporal attention provide a direct interface to add cross-agent attention. We deliberately implement a standard cross-agent attention mechanism. This simple design is sufficient to test our hypothesis and ensures we do not accidentally introduce unintended variables that complicate the interpretation of the results.

If compare the expanded model directly with the original backbone, two effects will be conflated: extra model capacity and interaction modeling. Therefore, we implement a controlled ablation, and compare two models sharing the same expanded architecture and parameter count. The *full predictor* receives real partner motion information, whereas the *ablated predictor* receives learned null tokens (no partner information). Both are trained and evaluated on the InterHuman dataset (Liang et al., 2024) using same target sequences. This same-backbone comparison directly tests whether interaction modeling reduces the frequency-specific errors predicted by the interaction-naive probe. Full results are in

Scope of inference. The interaction-naive probe identifies where error increases when interaction is introduced to interaction-naive predictors; the interaction-aware comparison tests whether partner information reduces error in the implicated bands. This work does not attempt to establish an interaction-only causal effect—rather is to introduce a unique evaluation methodology and provide informative diagnostic insights as a starting point for future works to discover more toward interaction and frequency domain. Although multiple design components (the single-agent novel motion control(parkour), cross-dataset results, and graded interaction comparisons) support an interaction-associated error signature, the probe cannot assign the entire penalty solely to interaction, since the solo and interaction test sequences differ in motion content, so we interpret the measured penalty as an interaction-associated error signature.

4.3 IMPLEMENTATION DETAILS

All experiments are conducted on Nvidia Ada6000 GPUs with 120 observed frames at $f_s = 60$ Hz. The three single-agent predictors generate 60-frame forecasts and are evaluated zero-shot on multi-agent scenarios. We apply all three prongs at $h \in \{24, 60\}$: the $h = 24$ evaluation uses the first 24 frames of the same 60-frame forecast, while $h = 60$ uses the full forecast.

The interaction-aware full and ablated predictors share an expanded architecture with 92,359 parameters, compared with 79,479 in the original solo backbone. Both are trained with seeds $\{42, 1337, 2024\}$. At $h = 60$, we evaluate all three prongs on the full 60-frame forecast. We additionally applied Prong 3 at $h = 120$ with a level 4 sym4 decomposition to obtain finer coarse-scale resolution: A_4 spanning 0–1.875,Hz and D_4 spanning 1.875–3.75,Hz, whereas the three-level decomposition at $h = 60$ combines 0–3.75,Hz in A_3.

All predictions are root-centered at the pelvis, yaw-normalized, and detrended. We apply our three-prong evaluation methodology to each (scenario, horizon, architecture) combination: Prong 1 (FFT): EVM, including magnitude EVM and phase EVM evaluated on 60-frame horizons, visualized at 0–12 Hz; Prong 2 (Band Pass): Butterworth bandpass MPJPE with bands at low [0,1) Hz, mid [1,5) Hz, high [5,10) Hz, and noise [10,30] Hz on 60-frame horizons; The band pass filters the observed seed concatenated with the ground-truth target or prediction, then score only the extracted forecast frames; Appendix E details the filter and edge handling. Prong 3 (DWT): Wavelet coefficient error with $L = 2$ at $h = 24$, $L = 3$ at $h = 60$, and $L = 4$ at $h = 120$. Relative-error ratios are aggregated as a ratio-of-means across joints and windows. Confidence intervals are 95% percentile bootstrap intervals from 2000 recording-level resamples.

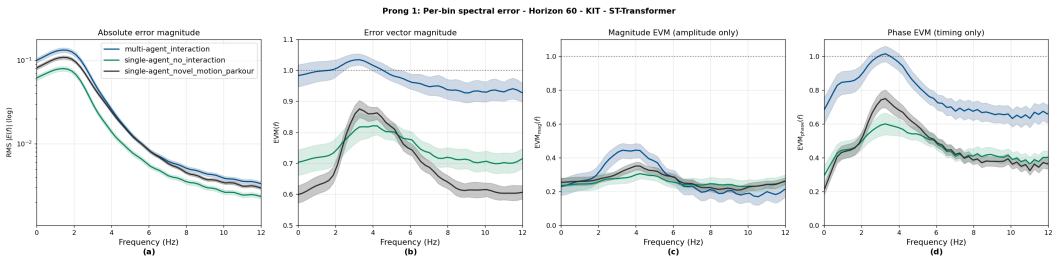

Figure 2: Prong 1 for the KIT ST-Transformer probe at $h = 60$. Panels show **(a)** absolute error magnitude $|E(f)|$, log scale. **(b)**: $\mathrm{EVM}(f)$, **(c)**: Magnitude-only EVM (scaling error). **(d)**: Phase-only EVM (timing error). Other architectures, CMU, and BIOMO are shown in Appendix J.

Table 1: Interaction-naive probe, Prong 2, KIT, $h = 60$. Numbers are the penalty Δ over the in-distribution non-interaction baseline. LOW, MID, and HIGH denote 0–1, 1–5, and 5–10 Hz. A star denotes a pointwise 95% bootstrap interval excluding zero. $h = 24$ results are in Appendix I.

Model	Test scenario	LOW	MID	HIGH
STT	KIT interaction	$+0.233^*$	$+0.173^*$	$+0.057^*$
	CMU interaction	$+0.098^*$	$+0.079^*$	-0.015
	KIT parkour	$+0.016$	$+0.047^*$	$+0.046^*$
HRI	KIT interaction	$+0.189^*$	$+0.106^*$	-0.000
	CMU interaction	$+0.095^*$	$+0.090^*$	$+0.008^*$
	KIT parkour	$+0.024$	-0.030	-0.015^*
Seq2seq	KIT interaction	$+0.321^*$	$+0.207^*$	$+0.025^*$
	CMU interaction	$+0.125^*$	$+0.087^*$	-0.014^*
	KIT parkour	$+0.094^*$	$+0.056^*$	$+0.015$

5 Results and Analysis

5.1 Interaction-naive Probe

Interaction sequences saturate normalized error; Timing error is the interaction-associated signature. EVM approaches saturation on interaction sequences in low-to-mid frequency band ($\mathrm{EVM} \approx 0.95$–$1.0$ over 0–5 Hz; Fig.2(b)), whereas the non-interaction baseline and the dynamic single-agent control (parkour) peak near 3–4 Hz and recover. The two diagnostics (Fig. 2(c)(d)) attribute this penalty differently. Phase EVM—the timing error—roughly doubles under interaction at 0–2 Hz (≈0.6–0.85 vs. 0.25–0.45 for the baseline) and remains elevated above 5 Hz. This timing error varies from sequence to sequence and is not a systematic delay—further phase analysis shows interaction has reduced phase locking, while the mean phase-offset cosine remains close to one (App.J). A dynamic single-agent control (parkour) does not reproduce it, its phase error generally remains closer to the non-interaction. Magnitude EVM shows that all predictions undershoot the ground-truth amplitude, but interaction specifically deepens this undershoot at 2.5–4.5 Hz (+0.14 to +0.24 over baseline).

> **Finding 1.** Multi-agent interaction saturates the normalized spectral error of single-agent-trained predictors in low-to-mid frequency bands across 0–5 Hz. Among the tested scenarios, the dominate interaction-associated failure is timing that is inconsistent from sequence to sequence, rather than through a systematic delay. Interaction also deepens amplitude undershoot most strongly at 2.5–4.5 Hz.

The penalty concentrates in the low-to-mid band, decays monotonically with frequency. At horizon 60, the interaction penalty (Δ Error/GT relative to the per-scene single-agent baseline) is largest in the low and mid bands. Table 1 shows the interaction penalty – the per-band relative error between an interaction scenario and the non-interaction baseline – positive values indicate

Table 2: Interaction-naive probe, Prong 3, $h = 60$, sym4, $L = 3$. Numbers are the penalty Δ over the in-distribution non-interaction baseline. Sub-bands are A_3: 0–3.75, D_3: 3.75–7.5, D_2: 7.5–15, and D_1: 15–30 Hz. A star denotes bootstrap CI excluding zero. $h = 24$ results are in Appendix I.

Model	Test set	A_3	D_3	D_2	D_1
STT	KIT interaction	$+0.262^*$	$+0.184^*$	$+0.130^*$	$+0.024^*$
	CMU interaction	$+0.136^*$	$+0.057^*$	$+0.052^*$	$+0.026^*$
	KIT parkour	-0.080^*	$+0.003$	-0.008	$+0.001$
HRI	KIT interaction	$+0.193^*$	$+0.106^*$	$+0.076^*$	$+0.032^*$
	CMU interaction	$+0.154^*$	$+0.090^*$	$+0.081^*$	$+0.044^*$
	KIT parkour	-0.119^*	-0.056^*	-0.053^*	-0.090^*
Seq2seq	KIT interaction	$+0.320^*$	$+0.227^*$	$+0.174^*$	$+0.031^*$
	CMU interaction	$+0.188^*$	$+0.077^*$	$+0.062^*$	$+0.025^*$
	KIT parkour	-0.030	$+0.040^*$	$+0.021$	$+0.002$

that interaction makes prediction harder in that band. For ST-Transformer, the interaction penalty is $+0.233$ (low-band) and $+0.173$ (mid-band) for KIT, and $+0.098$ and $+0.079$ for CMU. In contrast, the high-band penalties are much smaller. The same pattern holds across all 12 (architecture × dataset × horizon) combinations: the interaction penalty is largest in low-band and decays monotonically through mid to high bands. Prong 3 (Tab. 2) shows the same ordering across wavelet components. The dynamic single-agent control (parkour) does not reproduce the same pattern.

> **Finding 2.** Multi-agent scenarios exhibit the largest interaction penalty at coarse scales, moderate at mid frequency. The high-band (5-10 Hz) penalty is either small or absent.

Cross-Prong Synthesis The three prongs were designed to be sensitive to different aspects of the spectral error structure: Prong 1 reports per-bin spectral error and its decomposition into timing and amplitude components; Prong 2 reports per-band relative error in the original physical units (mm); and Prong 3 reports per-sub-band time-localized wavelet error. These three views provide independent evidence for the same finding: the multi-agent prediction penalty concentrates in the low-to-mid frequency band (0–5 Hz). Prong 1 places the largest timing gap at 0–2 Hz, and elevated amplitude gap at 2.5–4.5 Hz; Prong 2 places the largest band penalty at 0–5 Hz with low band slightly larger than mid band in all 12 (architecture × dataset × horizon) combinations; Prong 3 places the largest excess in the lowest sub-band (0–3.75 Hz at $h = 60$).

5.2 Interaction-Aware Predictor Evaluation

The full and ablated Prong 1 largely overlap; small differences in phase-only and magnitude diagnostics do not show a consistent improvement across metrics. The auxiliary phase and amplitude-consistency plots also remain close. Prong 1 therefore provides no clear separation in this comparison. It neither establishes equality nor supports a strong claim that phase alignment has improved. Results are in Appendix J.

Low-frequency improvements recur across training seeds. Table 3 reports the full-minus-ablated differences averaged equally across three seeds at $h = 60$. The full predictor's Prong 2 low-band error decreases for every seed, with a mean difference of -0.0163 and an across-seed 95% interval of $[-0.0310, -0.0017]$ that excludes zero. Prong 3 gives the same directional result: the lowest sub-band has a mean difference of -0.0139, with interval $[-0.0255, -0.0022]$. The Prong 2 mid-band mean is -0.0061, and the Prong 3 D_3 mean is -0.0037, both bootstrap CI intervals exclude zero, but their across-seed intervals include zero. We therefore regard the middle-band evidence as less stable across training runs but still remains informative.

Table 3: Interaction-aware comparison, Δ Error/GT (full $-$ ablated) at $h = 60$, mean over three seeds. Bootstrap CI: 2000 paired resamples over test recordings, seeds held fixed. Across-seed CI: $t_{0.975,\,2}$ interval over the three seed deltas. * interval excludes 0.

Prong	Band / component (Hz)	Mean Δ	Bootstrap 95% CI	Across-seed 95% CI
P2	0–1	-0.0163	$[-0.0207, -0.0124]^*$	$[-0.0310, -0.0017]^*$
	1–5	-0.0061	$[-0.0092, -0.0028]^*$	$[-0.0129, +0.0007]$
	5–10	-0.0003	$[-0.0019, +0.0011]$	$[-0.0075, +0.0069]$
	10–30	$+0.0004$	$[-0.0002, +0.0011]$	$[-0.0032, +0.0041]$
P3	A3 (0–3.75)	-0.0139	$[-0.0173, -0.0102]^*$	$[-0.0255, -0.0022]^*$
	D3 (3.75–7.5)	-0.0037	$[-0.0059, -0.0012]^*$	$[-0.0134, +0.0060]$
	D2 (7.5–15)	-0.0010	$[-0.0026, +0.0006]$	$[-0.0065, +0.0045]$
	D1 (15–30)	$+0.0001$	$[-0.0009, +0.0011]$	$[-0.0019, +0.0022]$

Finding 3. The interaction-aware model lowers low-band Prong 2 error and coarse-scale Prong 3 error compared to an ablated interaction-naive model with the same architectural design and capacity.

6 Discussion

Architectural insights. From the three-prong evaluation on two sets of experiments, we draw several architectural implications for multi-agent pose prediction: (1) Our results indicate that the multi-agent prediction penalty primarily concentrates below 5 Hz and is small or absent toward higher frequency bands. These results motivate future work to test whether multi-agent architectures benefit from allocating more interaction-modeling capacity to the low-to-mid band.(2) Within the low-to-mid band, timing error and amplitude error sit in different sub-ranges (Prong 1): timing error dominates at 0–2 Hz, consistent with slow inter-agent synchronization and coordination, whereas amplitude error peaks at 2.5–4.5 Hz, a range associated with coordinated voluntary motion; a multi-agent architecture should target both. (3) The frequency domain remains under-utilized as a space for explicit interaction modeling. Existing works already use DCT representations, but interaction modeling still takes place in the spatial or temporal domain. If interaction is band-localized, a natural architectural direction for future work is for frequency-domain interaction modules to operate selectively on low-to-mid-frequency coefficients, and potentially use distinct mechanisms for the timing-dominated and amplitude-dominated sub-bands.

Limitations. Several limitations live in our current experiment design: (1) The interaction-aware comparison is evaluated with one backbone so its extension to other architecture families and interaction modules remain to be tested. (2) Root-centering, yaw normalization, and detrending make the reported evaluation to local pose, not global trajectories. (3) Finally, our finding is descriptive rather than causal: the spectral analyses locate the error but do not identify its underlying mechanism, the frequency-selective architectural implications therefore remain hypotheses for future testing.

7 Conclusion

In this paper we presented a three-prong frequency-domain evaluation methodology for diagnosing where interaction-associated error concentrates in pose prediction. Applied across single-agent and multi-agent predictors and datasets, the methodology reveals a consistent pattern: the multi-agent prediction penalty concentrates below 5 Hz. We hope this methodology serves as a foundational tool for evaluating interaction modeling pose prediction, and the empirical pattern we surface motivates new architectures to allocate interaction-specific modeling capacity in low-to-mid frequency band.

Reproducibility statement Section 3 and Appendix A describe the proposed three-prong methodology. Section 4 describe the experimental design with implementation details. Section 4.1 and Appendix D F G documents the datasets, splits, and preprocessing steps. The supplementary material contains anonymized code and scripts for reproducing the reported results.

AI USE STATEMENT

In this work, we used "Claude Opus 5.5" and "ChatGPT GPT-5.6 Sol" to suggest grammar and readability edits to the manuscript. The authors reviewed and revised all suggested text. We did not use generative AI to develop the research questions or methods, implement the methods, process or analyze data, interpret results, or formulate claims or proofs. The authors take full responsibility for the paper's content.

REFERENCES

Advanced Computing Center for the Arts and Design. ACCAD MoCap Dataset. URL https://accad.osu.edu/research/motion-lab/mocap-system-and-data.

Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, June 2015.

Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International conference on 3D vision (3DV)*, pp. 565–574. IEEE, 2021.

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Carnegie Mellon University. CMU MoCap Dataset, 2003. URL http://mocap.cs.cmu.edu.

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11467–11476, October 2021.

Yuming Feng, Zhiyang Dou, Ling-Hao Chen, Yuan Liu, Tianyu Li, Jingbo Wang, Zeyu Cao, Wenping Wang, Taku Komura, and Lingjie Liu. Motionwavelet: Human motion prediction via wavelet manifold learning. *arXiv preprint arXiv:2411.16964*, 2024.

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pp. 4346–4354, 2015.

Shengchuan Gao, Shuo Wang, Yabiao Wang, and Ran Yi. WaveAR: Wavelet-aware continuous autoregressive diffusion for accurate human motion prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kW2u5szHb6.

Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. MoVi: A large multipurpose motion and video dataset. *arXiv preprint arXiv: 2003.01888*, 2020.

Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 4809–4819, 2023.

Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of hand: Perception of finger motion from reduced marker sets. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pp. 79–86, 2012. doi: 10.1145/2159616.2159629.

Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser learning to reconstruct human pose from sparseinertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37(6):185:1–185:15, November 2018.

Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1617–1628, 2024.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pp. 1–21, 2024.

Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, November 2014. doi: 10.1145/2661229.2661273. URL http://doi.acm.org/10.1145/2661229.2661273.

Eyes JAPAN Co. Ltd. Eyes Japan MoCap Dataset. URL http://mocapdata.com.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pp. 5442–5451, October 2019.

Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The KIT whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pp. 329–336, 2015.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9489–9497, 2019.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pp. 474–489. Springer, 2020.

Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2891–2900, 2017.

M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007.

Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17121–17130, 2023.

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 722–731, 2021.

Junlong Ren, Gangjian Zhang, Honghao Fu, Pengcheng Wu, and Hao Wang. Wamo: Wavelet-enhanced multi-frequency trajectory analysis for fine-grained text-motion retrieval. *arXiv preprint arXiv:2508.03343*, 2025.

L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, March 2010.

Nikolaus F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):2–2, September 2002. doi: 10.1167/2.5.2.

Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total Capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017.

Simon Fraser University and National University of Singapore. SFU Motion Capture Database. URL http://mocap.cs.sfu.ca/.

Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34: 6036–6049, 2021.

David A Winter. *Biomechanics and motor control of human movement*. John wiley & sons, 2009.

Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, Yibo Yang, and Xinge You. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *European Conference on Computer Vision*, pp. 682–700. Springer, 2022.

Beihao Xia, Conghao Wong, Duanquan Xu, Qinmu Peng, and Xinge You. Another vertical view: A hierarchical network for heterogeneous trajectory prediction via spectrums. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *The Eleventh International Conference on Learning Representations*, 2023.

Pengfei Yao, Yinglong Zhu, Huikun Bi, Tianlu Mao, and Zhaoqi Wang. Trajclip: Pedestrian trajectory prediction method using contrastive learning and idempotent networks. *Advances in Neural Information Processing Systems*, 37:77023–77037, 2024.

Yuanhong Zheng, Ruixuan Yu, and Jian Sun. Efficient multi-person motion prediction by lightweight spatial and temporal interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10844–10853, 2025.

A Metric Details and Scope

A.1 Fourier identities and timing–amplitude diagnostics

Prong 1: per-axis DFT metrics. Fix a frequency bin, joint, and coordinate axis, and let s index sequences. The DFT metrics in Section 3.1 average over sequences. For positive ground-truth ensemble power, define

$$r = \frac{\langle |P_s|^2 \rangle_s}{\langle |G_s|^2 \rangle_s}. \tag{4}$$

When prediction power is also positive, let

$$\gamma_{\mathrm{DFT}} = \frac{\langle G_s^* P_s \rangle_s}{\sqrt{\langle |G_s|^2 \rangle_s \langle |P_s|^2 \rangle_s}}, \qquad C_{\mathrm{DFT}} = |\gamma_{\mathrm{DFT}}|^2, \qquad \theta_{\mathrm{DFT}} = \arg \gamma_{\mathrm{DFT}}. \tag{5}$$

Expanding the squared error gives

$$\langle |P_s - G_s|^2 \rangle_s = \langle |P_s|^2 \rangle_s + \langle |G_s|^2 \rangle_s - 2 \operatorname{Re} \langle G_s^* P_s \rangle_s, \tag{6}$$

$$\mathrm{EVM}_{\mathrm{DFT}}^2 = 1 + r - 2\sqrt{r}\, \operatorname{Re} \gamma_{\mathrm{DFT}} = 1 + r - 2\sqrt{r C_{\mathrm{DFT}}} \cos \theta_{\mathrm{DFT}}. \tag{7}$$

The final expression applies when the angle is defined. Replacing $\operatorname{Re} \gamma_{\mathrm{DFT}}$ by $|\gamma_{\mathrm{DFT}}|$ gives a lower bound at fixed r and C_{DFT}, with equality when γ_{DFT} is real and nonnegative. A zero prediction has $\mathrm{EVM}_{\mathrm{DFT}} = 1$ and no defined cross-spectrum phase.

The reported magnitude diagnostic, $\mathrm{EVM}_{\mathrm{mag}} = 1 - \sqrt{r}$, is signed: positive values indicate ensemble power undershoot and negative values indicate overshoot. Equal ensemble powers do not imply matching coefficient magnitudes in every sequence.

The reported phase diagnostic instead averages valid coefficient phases without amplitude weighting. Let $\mathcal{V}$ contain the sequences for which both coefficient phases are defined, and write $\Delta\phi_s = \arg P_s - \arg G_s$. When $\mathcal{V}$ is nonempty,

$$A_{\mathrm{DFT}} = \frac{1}{|\mathcal{V}|} \sum_{s \in \mathcal{V}} \cos \Delta\phi_s, \qquad \mathrm{EVM}_{\mathrm{phase}}^2 = 2(1 - A_{\mathrm{DFT}}) = \frac{1}{|\mathcal{V}|} \sum_{s \in \mathcal{V}} \left| \frac{P_s}{|P_s|} - \frac{G_s}{|G_s|} \right|^2. \tag{8}$$

This phase diagnostic is undefined if $\mathcal{V}$ is empty. It differs from θ_{DFT}, which comes from an amplitude-weighted cross-spectrum. The signed magnitude diagnostic and phase EVM therefore interpret aspects of total EVM but do not form an additive decomposition of it.

Multitaper record-level decomposition. To examine timing and amplitude consistency across joints and sequences, let i index sequence–joint records, excluding joints rigidly attached to the root. At each frequency, form $S_{gg}^{(i)}$, $S_{pp}^{(i)}$, and $S_{gp}^{(i)}$ by averaging the corresponding auto- and cross-products over the five tapers and summing over the three spatial axes. Use the convention $S_{gp} = G^* P$, and let $\langle \cdot \rangle_i$ denote the mean over records. Define

$$r_{\mathrm{MT}} = \frac{\langle S_{pp} \rangle_i}{\langle S_{gg} \rangle_i}, \qquad \Gamma = \frac{\langle |S_{gp}| \rangle_i}{\sqrt{\langle S_{gg} \rangle_i \langle S_{pp} \rangle_i}}, \qquad A_{\mathrm{MT}} = \frac{\mathrm{Re}\langle S_{gp} \rangle_i}{\langle |S_{gp}| \rangle_i}. \tag{9}$$

These ratios are defined when their denominators are positive. Unlike A_{DFT}, A_{MT} weights each record's cross-spectrum phase by $|S_{gp}^{(i)}|$; it also uses axes and tapers combined within each record. Thus the two phase averages should not be assigned the same numerical value.

The record-level alignment separates into phase locking and a mean phase offset:

$$A_{\mathrm{MT}} = L \cos \varphi, \qquad L = \frac{|\langle S_{gp} \rangle_i|}{\langle |S_{gp}| \rangle_i}, \qquad \varphi = \arg \langle S_{gp} \rangle_i. \tag{10}$$

Here $L \in [0, 1]$ measures the consistency of cross-spectrum phase *across records*; it can equal one even when every record has the same nonzero phase offset. The angle φ describes that shared offset. A consistent time lag would produce an approximately linear phase shift with frequency, rather than being established by $\cos \varphi$ at one frequency. If $\langle S_{gp} \rangle_i = 0$, the angle is undefined but $A_{\mathrm{MT}} = 0$ whenever its denominator is positive.

Using the same multitaper estimates in the error and normalization gives the exact identity

$$\mathrm{EVM}_{\mathrm{MT}}^2 = \frac{\langle S_{pp} + S_{gg} - 2\,\mathrm{Re}\, S_{gp} \rangle_i}{\langle S_{gg} \rangle_i} = 1 + r_{\mathrm{MT}} - 2\sqrt{r_{\mathrm{MT}}}\, A_{\mathrm{MT}} \Gamma. \tag{11}$$

This is the multitaper, record-pooled counterpart of the DFT identity above; it does not equate $\mathrm{EVM}_{\mathrm{MT}}$ with the per-axis $\mathrm{EVM}_{\mathrm{DFT}}$ reported in Prong 1. Algebraically, setting $A_{\mathrm{MT}} = \Gamma = 1$ leaves $(1 - \sqrt{r_{\mathrm{MT}}})^2$, while setting $r_{\mathrm{MT}} = \Gamma = 1$ leaves $2(1 - A_{\mathrm{MT}})$. These limits explain the forms of the main-text diagnostics, although their averaging rules differ.

The remaining factor separates exactly as

$$\Gamma^2 = \Phi_{\mathrm{gain}} \Phi_{\mathrm{within}}, \qquad \Phi_{\mathrm{gain}} = \frac{\langle \sqrt{S_{gg} S_{pp}} \rangle_i^2}{\langle S_{gg} \rangle_i \langle S_{pp} \rangle_i}, \qquad \Phi_{\mathrm{within}} = \left(\frac{\langle |S_{gp}| \rangle_i}{\langle \sqrt{S_{gg} S_{pp}} \rangle_i} \right)^2. \tag{12}$$

Both factors lie in $[0, 1]$. The first measures how consistently ground-truth and prediction powers scale *across records*. It equals one when their amplitude ratio is common across records with positive power, regardless of whether that common gain is correct. The second measures cross-spectrum consistency *within records*, across axes and taper estimates. It can decrease because of phase variation, gain variation, or other mismatch within a record; it is not solely an amplitude measure. Finite taper counts also affect its estimate.

For these multitaper quantities, magnitude-squared coherence is

$$C_{\mathrm{MT}} = \frac{|\langle S_{gp} \rangle_i|^2}{\langle S_{gg} \rangle_i \langle S_{pp} \rangle_i} = L^2 \Gamma^2 = L^2 \Phi_{\mathrm{gain}} \Phi_{\mathrm{within}}. \tag{13}$$

This factorization separates between-record phase variation from two sources of cross-spectrum magnitude loss. In particular, coherence is not a phase-only statistic: if $P_s = a_s G_s$ with positive but varying a_s, the defined coefficient phase differences vanish, yet coherence can be below one.

Interpretation and uncertainty. The multitaper ratios are invariant to common rescaling of prediction and ground truth. If record phase offsets follow a Gaussian jitter model and their magnitudes are independent of those offsets, $L \approx \exp(-\sigma_\theta^2/2)$; the corresponding single-frequency timing scatter is approximately $\sigma_t \approx \sqrt{-2 \ln L}/(2\pi f)$, subject to phase wrapping. Likewise, under a lognormal model for amplitude ratios with ground-truth-power-weighted log standard deviation s, $\Phi_{\mathrm{gain}} \approx \exp(-s^2)$. Thus $\Phi_{\mathrm{gain}} = 0.5$ corresponds to a one-standard-deviation spread of about $e^{\sqrt{\ln 2}} \approx 2.3$ around the *central gain*, which need not equal one.

Confidence bands are 95% sequence-level bootstrap intervals from 1000 resamples, keeping records from each sampled sequence together. Finite samples can bias estimated phase locking upward, especially when true locking is weak.

A.2 Pooling and normalization

For either Prong 2 or Prong 3, write e_{bjs} and g_{bjs} for the corresponding error and ground-truth magnitudes defined in Section 3.2 or Section 3.3. When $D_b = \sum_{s,j} g_{bjs} > 0$, the reported ratio is

$$R_b = \frac{\sum_{s,j} e_{bjs}}{D_b}. \tag{14}$$

For pairs with $g_{bjs} > 0$, let $r_{bjs} = e_{bjs}/g_{bjs}$. Then the exact pooling identity is

$$R_b = \sum_{\{s,j:\, g_{bjs}>0\}} \frac{g_{bjs}}{D_b}\, r_{bjs} + \frac{\sum_{\{s,j:\, g_{bjs}=0\}} e_{bjs}}{D_b}. \tag{15}$$

Thus pooling weights defined pairwise ratios by ground-truth mean norm, not squared energy. Errors where ground-truth magnitude is zero still contribute to the pooled numerator. If $D_b = 0$, the relative error is undefined.

Because both prongs use matching mean-of-Euclidean-norm definitions for e_{bjs} and g_{bjs}, a zero prediction gives $R_b = 1$ when $D_b > 0$, and a perfect prediction gives $R_b = 0$. Scaling prediction and ground truth together by the same factor leaves R_b unchanged. Scaling only selected sequences changes their contributions to the pooled ratio. Consequently, the ratio is dimensionless but still depends on the dataset's distribution of motion magnitudes. In Prong 2, $e^{(2)}_{bjs}$ is a time-domain error in millimeters; $R^{(2)}_b$ expresses that error relative to ground-truth band magnitude. Prong 3 averages errors in wavelet-coefficient space, so $R^{(3)}_b$ is not a time-domain MPJPE or a directly interchangeable bandpass result.

A.3 Temporal support and wavelet boundaries

A scoring horizon alone does not fully specify either transform. Detrending support, filter input length and padding, wavelet boundary extension, and retained coefficients can affect the result. Filtering a longer trajectory and scoring its first 60 frames, for example, need not equal filtering only those 60 frames. Horizon comparisons should state both the transform's input support and the scoring interval.

The reported sym4 decomposition uses three levels for $h = 60$. The $h = 24$ results use two levels and serve as a shorter-horizon sensitivity analysis. For an input of N samples, the usual maximum useful DWT level is

$$L_{\max} = \left\lfloor \log_2 \frac{N}{\ell - 1} \right\rfloor, \tag{16}$$

where sym4 has filter length $\ell = 8$. If the transform is applied only to the prediction window, $L_{\max} = 3$ for $N = 60$ but $L_{\max} = 1$ for $N = 24$. A two-level transform of only 24 samples can therefore be strongly affected by boundary extension. This limit must instead be evaluated using the longer input length if coefficients were computed from a longer trajectory.

At a 60 Hz sampling rate, the nominal $h = 24$, two-level approximation band is 0–7.5 Hz, whereas the $h = 60$, three-level approximation band is 0–3.75 Hz. They should not be treated as the same frequency band. The main conclusions use $h = 60$. Wavelet localization changes how finite-window effects appear; it does not eliminate boundary effects.

The 10–30 Hz range is reported where supplied, without assuming that all its content is measurement noise. Prong 1 plots stop at 12 Hz for display. That plotting limit does not impose a 12 Hz cutoff on DWT components: at a 60 Hz sampling rate, D_2 has a nominal 7.5–15 Hz band.

B Aggregation: Ratio-of-Means with Bootstrap CI

Both Prongs 2 and 3 produce per-(band, joint, sequence) error and magnitude quantities. We aggregate these into a single dimensionless headline number per (scenario, band, horizon) using the *ratio of means*:

$$\text{Relative Error Ratio}_{\text{scen},b,h} = \frac{\sum_{s,j} \text{err}_{\text{scen},b,j,s,h}}{\sum_{s,j} \text{mag}_{\text{scen},b,j,s,h}},$$

where err is either the band-MPJPE (Prong 2) or the coefficient error magnitude (Prong 3), and mag is the matching ground-truth magnitude. Confidence intervals are computed by bootstrap resampling sequences with 1000 resamples, treating sequences (not joints, not frames) as the independent unit of observation.

Why ratio-of-means rather than mean-of-ratios. Mean-of-ratios assigns equal weight to every sequence regardless of its motion magnitude, which up-weights low-magnitude sequences whose ratios are noisier (small denominators). Ratio-of-means weights each sequence by its motion magnitude, which is the same weighting implicit in standard MPJPE reporting on the unfiltered signal. It also collapses cleanly to the test-set-wide error/signal ratio, which is the quantity practitioners care about.

Why mean-of-L2-norms rather than RMS. A natural alternative denominator is the per-band root-mean-square (RMS) of the ground-truth signal, which has the appealing Parseval property that band-wise squared energies add up to the total squared energy. We use mean-of-L2-norms instead for one reason: our error numerator follows the MPJPE convention of mean-of-Euclidean-norms over time (not RMS over time). Mixing a mean-of-norms numerator with an RMS denominator would, by Jensen's inequality, give a ratio strictly less than 1 even when the prediction exactly matches the ground truth, breaking the clean interpretation of the ratio as "fraction of GT signal that is error." We retain RMS-based per-band energy as a reference statistic in all prong 1's result figures for readers who want a Parseval-consistent description of how motion distributes across bands; the headline ratio uses mean-of-norms throughout.

Position vs. velocity magnitude. Our denominator uses position magnitude (per-joint distance from the centered, yaw-normalized pelvis frame) rather than velocity magnitude. After preprocessing, this measures per-joint excursion rather than per-joint motion: a limb held at a fixed offset has a large position magnitude but no motion to predict, which would inflate the low-band ratio. We verify that the low-band result is not driven by static-pose offset by repeating the analysis with velocity-based magnitudes (∇_t applied before the L2 norm); the position and velocity ratios agree to within bootstrap error in the low band, confirming that genuine motion content (rather than offset) drives the low-band penalty.

C A Methodological Finding: Amplitude-Confounded Cross-Scenario MPJPE

A separate methodological observation falls out of the Prong 2 analysis. Naive band-wise MPJPE comparisons across scenarios are confounded by ground-truth motion amplitude: a band with less motion in one scenario will produce a smaller absolute MPJPE in that band, not because the model is doing better but because there is less to get wrong. Two of the most striking effects in raw MPJPE on our data turn out to be amplitude artifacts.

The KIT high-band illusion. In raw MPJPE at $h = 60$, the high-band error for `2a_kit` is 1.06 mm, against 1.92 mm for `non_interaction`. Naively, this would say the model predicts KIT high-frequency motion better than baseline. But `2a_kit` has approximately half the high-band ground-truth magnitude (0.98 mm vs. 1.88 mm). Once normalized, the conclusion inverts: KIT's amplitude-normalized high-band error is 1.083, larger than baseline's 1.021

The CMU "catastrophic everywhere" illusion. In raw MPJPE, `2a_cmu` appears dramatically worse than baseline in every band (26.4 mm vs. 16.0 mm in mid, 3.0 mm vs. 1.9 mm in high). But CMU also carries 1.5–2$\times$ the ground-truth motion magnitude of baseline. Under amplitude normalization, the gap shrinks substantially: the mid-band penalty is $+0.18$ rather than the $+10$ mm absolute gap suggested.

Implication for cross-scenario evaluation. Cross-scenario per-band MPJPE comparisons that do not control for ground-truth motion amplitude can produce spurious effects in either direction. We argue, on the basis of these two examples, that any future work comparing per-band errors across scenarios with different motion content (which is the typical case for any single-agent vs. multi-agent

comparison) should use amplitude-normalized ratio-of-means with explicit confidence intervals. We return to this prescription in Section 6.

D AMASS-BABEL FILTERING PROCEDURE

The interaction-naive probe design (discussed in Section4) requires the training data to be free of any form of interaction, even of the implicit kind. AMASS, as a large-scale single-agent motion dataset, routinely contains sequences of a single subject performing interaction-like actions (e.g., fighting, being pushed, dodging). Such motions are shaped by partners – even imaginary ones. Training on such motions would contaminate the probe by interaction priors and defeat the purpose of the design.

We construct the interaction-free single-agent training corpus from AMASS in two stages: subset selection and interaction filtering.

Subset Selection. AMASS aggregates a wide range of motion capture datasets, not all of which are appropriate for full-body pose prediction. We exclude subsets focused on hand-only or finger-only motion, as these capture motion at a granularity unrelated to our research question on inter-agent interaction. The remaining 17 subsets are: [ACCAD(Advanced Computing Center for the Arts and Design), BioMotion(Troje, 2002), BMLmovi(Ghorbani et al., 2020), CMU(Carnegie Mellon University, 2003), DFaust67(Bogo et al., 2017), EKUT(Mandery et al., 2015), EyesJapan(Ltd.), HDM05(Müller et al., 2007), HEva(Sigal et al., 2010), JointLimit(Akhter & Black, 2015), KIT(Mandery et al., 2015), MPI_mosh(Loper et al., 2014), SFU(University & of Singapore), SSM(Loper et al., 2014), TCD_handMocap(Hoyet et al., 2012), TotalCapture(Trumble et al., 2017), Transition(Loper et al., 2014)]. This selection is broader than the union of training corpora used by the three single-agent predictors in our study [ref each predictor and their original training sets], ensuring that our probe is exposed to a strictly larger and more diverse pool of single-agent motion than any predictor encountered in its original publication.

Interaction Filtering. A non-trivial portion of AMASS sequences depict a single subject performing actions whose motion is implicitly shaped by an (absent) partner, for example, fighting, being pushed, or dodging. Although only one body is recorded, the motion contains interaction priors that would contaminate an interaction-naive probe. We therefore filter out such sequences using the following procedure:

1. Keyword set. We define a set of seed keywords describing interaction-driven actions (e.g., fight, push, dodge, block, throw, etc.) and expand each seed with semantically similar terms to improve recall.
2. Annotation matching. We match each AMASS sequence against its BABEL(Punnakkal et al., 2021) annotation labels. Sequences whose labels match any keyword in the expanded set are flagged as simulated-interaction.
3. Exclusion of ambiguous sequences. Sequences for which BABEL provides no label, a `null` label, or a label that is too vague to classify (i.e., could plausibly describe either interaction or non-interaction motion) are conservatively excluded from the non-interaction corpus to prevent contamination.

The Simulated-Interaction Subset. The interaction filtering procedure above isolates 381 sequences in which a single subject performs strongly interaction-driven actions (fighting, dodging, etc.). We treat this as a separate evaluation set rather than discard it, since it offers a unique intermediate condition: the motion is shaped by interaction, but only one body is present. Evaluating the probe on this subset allows us to test whether interaction's frequency-domain signature persists in the absence of a second agent.

E BAND DEFINITION

We adopt a four-band split motivated by the physiological structure of human motion rather than by signal-vs-noise considerations. The low band (0-1 Hz) captures gait cadence and slow postural changes. The mid band (1-5 Hz) covers the bulk of voluntary motion, including the inter-agent timing

and coordination band. The high band (5-10 Hz) covers rapid voluntary motion such as fast end-effector trajectories, action transitions. The noise band (10-30 Hz) is dominated by measurement noise and filter artifacts are reported for completeness. We verify that the headline result is not sensitive to small shifts of the mid-band edges (alternate splits 1-5, 1.5-6 all yield the same qualitative pattern). Another split commonly used in signal-processing contexts (e.g., 0-2/ 2-8/ 8-30 Hz, motivated by separating articulated motion from sensor noise at $\sim$8 Hz) would lump the coordination band (1-5 Hz) together with the fast-motion band (5-8 Hz) under a single mid-band label.

F Filter design, edge effects and the seed runway

We use a 4th-order Butterworth bandpass filter applied with zero-phase `filtfilt`, yielding an effective 8th-order roll-off of approximately 48 dB/octave. Edge bands degenerate appropriately: when the lower cutoff is zero we use a lowpass filter, and when the upper cutoff exceeds Nyquist we use a highpass filter. The 4th-order choice is conservative; higher orders would sharpen band edges further but introduce longer transients into 60-frame windows. We verify that the choice does not affect the headline result. A bandpass filter introduces transient distortions at the start of any finite signal. To prevent these transients from corrupting the prediction window, we filter on the long sequence

$$\mathbf{x}_{\text{long}} = [\mathbf{x}_{\text{seed}};\ \mathbf{x}_{\text{target}}]$$

constructed from the conditioning seed (the input to the predictor) concatenated with the target window, and we extract the prediction window only after filtering. The same construction is applied to the prediction:

$$\hat{\mathbf{x}}_{\text{long}} = [\mathbf{x}_{\text{seed}};\ \hat{\mathbf{x}}_{\text{pred}}],$$

using the ground-truth seed for both, so that filter transients are identical in the ground-truth and prediction signals and exclude per-sequence linear drift when we compute the error. After filtering, we extract the first h frames of the prediction window from each filtered signal and compute MPJPE between them. This cancellation property is the reason for using the ground-truth seed in $\hat{\mathbf{x}}_{\text{long}}$ rather than the model's own past.

We report results at horizons $h \in \{24, 60\}$. The same filter is used for all horizons; only the extracted window length differs.

G DWT Decomposition level and wavelet family

For horizon $h = 60$ we use a 3-level decomposition, yielding four sub-bands at $f_s = 60$ Hz (`approx_3`: 0-3.75 Hz, `detail_3`: 3.75-7.5 Hz, `detail_2`: 7.5-15 Hz, `detail_1`: 15-30 Hz). For horizon $h = 24$ we use a 2-level decomposition (`approx_2`: 0-7.5 Hz, `detail_2`: 7.5-15 Hz, `detail_1`: 15-30 Hz). For horizon $h = 120$ we use a 3-level decomposition (`approx_4`: 0-1.875 Hz, `detail_4`: 1.875-3.75 Hz, `detail_3`: 3.75-7.5 Hz, `detail_2`: 7.5-15 Hz, `detail_1`: 15-30 Hz). The level count is chosen so that each sub-band retains enough coefficients for the per-coefficient mean to be statistically meaningful; deeper decompositions at $h = 24$ leave the coarsest sub-band with too few coefficients to estimate reliably.

We use `sym4` (a near-symmetric Daubechies variant, 8-tap) as the primary choice. Symlets are well suited to piecewise-smooth signals such as pose trajectories: they preserve more of the temporal structure than asymmetric Daubechies wavelets while remaining short enough to support multi-level decomposition at our horizons. We verify robustness against `db4` and `coif3`; `coif3` (18-tap) does not support 3-level decomposition at $h = 60$ and is consequently dropped from that comparison.

H Uncertainty and Reporting

The point estimate pools window-level errors and ground-truth magnitudes. Resampling must preserve all dependent observations within an independent unit, including any overlapping windows from the same recording. Pooling and the bootstrap unit are separate choices: resampling recordings need not change the window-weighted point estimate.

For scenario probes with unmatched examples, each dataset is resampled independently and the difference between recomputed pooled ratios is retained. Different sample counts alone do not decide

whether pairing is possible; pairing depends on matched observations. Probe intervals describe variation in the evaluation data for fixed checkpoints, not variation across independent training runs.

For interaction-aware comparisons, paired draws retain the full and ablated predictions together. With a shared test set across seeds, each replicate uses the same sampled units across all seeds. It computes a pooled ratio for each model and seed, subtracts full minus ablated, and averages the three differences. The seed-averaged analysis uses 2,000 replicates and 212 resampling units.

Across-seed intervals are calculated separately. For the three contrasts Δ_k, with sample standard deviation s_Δ, the interval is

$$\bar{\Delta} \pm t_{0.975,2} \frac{s_\Delta}{\sqrt{3}}, \qquad t_{0.975,2} \simeq 4.303. \tag{17}$$

This interval describes the mean training-run contrast on the fixed test set under the Student-t assumptions. With only three seeds, those assumptions are difficult to assess. It is not a combined interval over both new training runs and new test recordings. The test-sequence-bootstrap interval instead conditions on the three observed model pairs.

All displayed intervals are pointwise. They do not adjust for the number of bands, horizons, wavelets, or weighting choices examined. Since these choices were explored using the results, interval exclusions are reported descriptively rather than as confirmatory discoveries. A band-by-model contrast is required to test whether an improvement is larger at one scale than another. Similarly, a direct strong-minus-weak contrast is required for the graded BIOMO comparison.

I Additional Numerical Results - Interaction-naive Probe

Tables 4–5 contain the probe estimates and intervals. Positive differences indicate greater error than the corresponding non-interaction baseline. Stars follow unrounded interval decisions; a displayed endpoint rounded to zero can still carry a star. These Prong 2 probe tables cover 0–10 Hz, while the Prong 3 tables include all decomposition levels.

J Additional Fourier Results

The following figures show the architecture, scenario, and seed views. They support qualitative comparisons of the displayed estimates; overlap or separation of marginal ribbons is not itself a test of a paired difference. The phase-locking and amplitude-consistency panels are descriptive diagnostics.

Interaction-aware predictors. The full and ablated Prong 1 curves (Fig. 8) largely overlap for all three seeds. Small differences in phase-only and magnitude diagnostics do not show a consistent improvement across metrics. The auxiliary phase and amplitude-consistency plots also remain close in Fig. 9. Prong 1 therefore provides no clear visual separation in this comparison. It neither establishes equality nor supports a strong claim that phase alignment has improved.

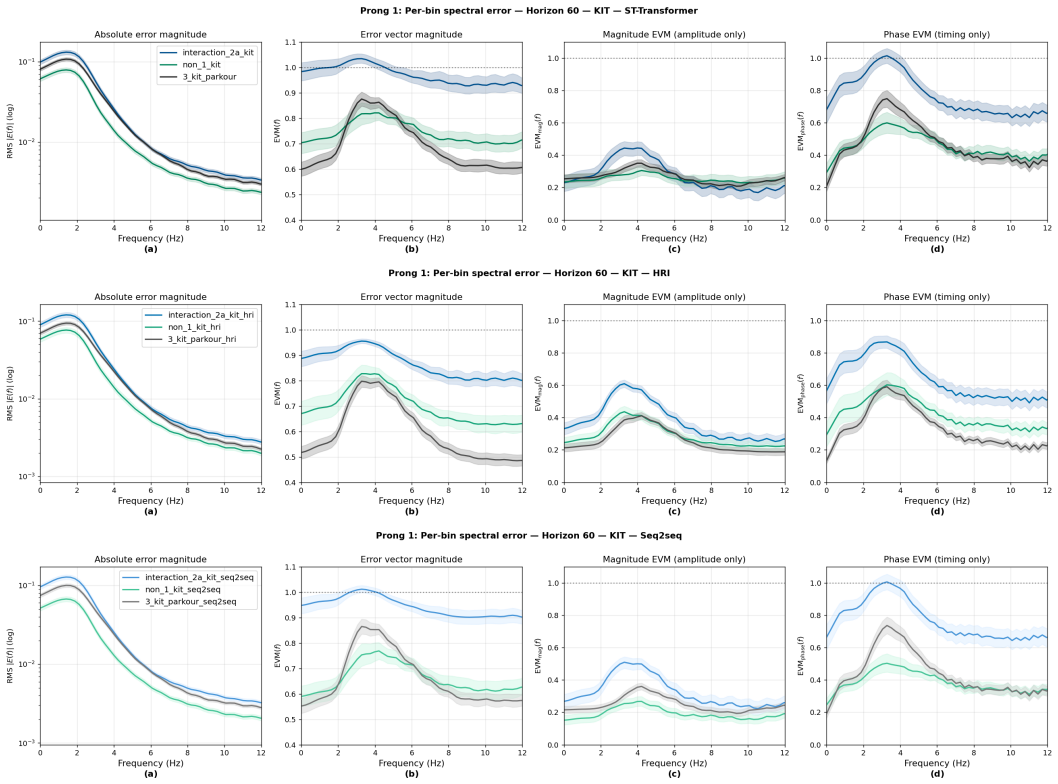

Figure 3: Prong 1 on KIT for ST-Transformer, HRI, and Seq2seq. Interaction increases phase-only error relative to the non-interaction baseline and generally relative to parkour. Magnitude-diagnostic differences vary by architecture.

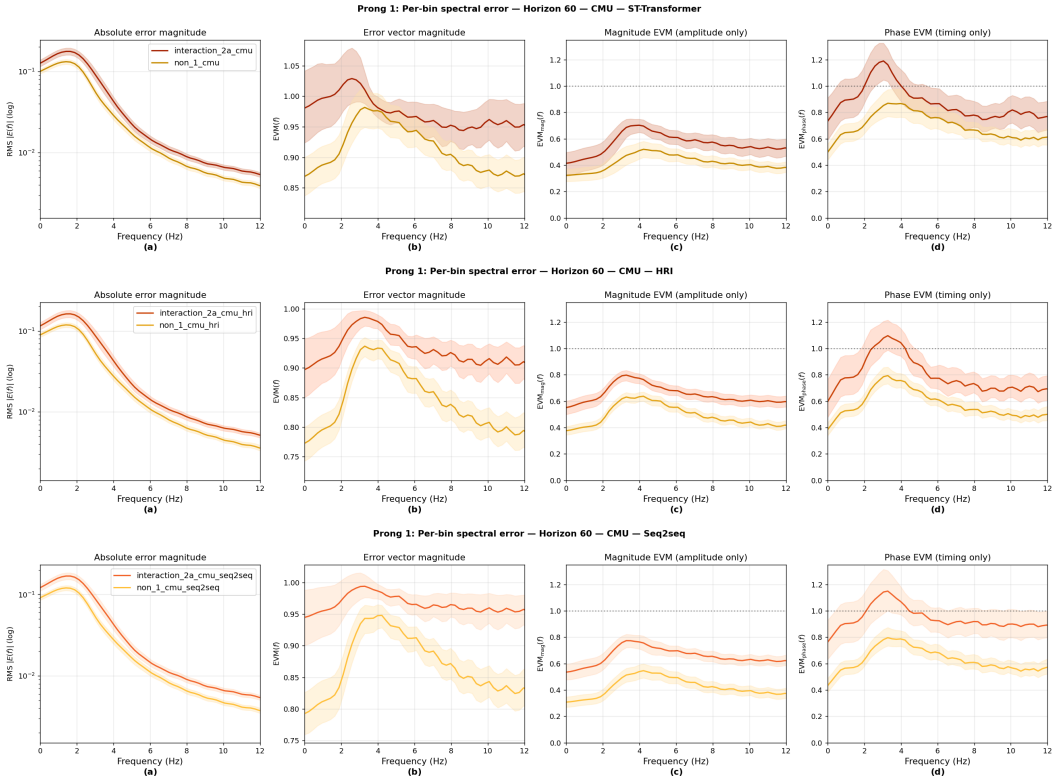

Figure 4: Prong 1 on CMU for the three probe architectures. Both magnitude and phase diagnostics are elevated under interaction. The plots do not by themselves establish significance for per-frequency differences.

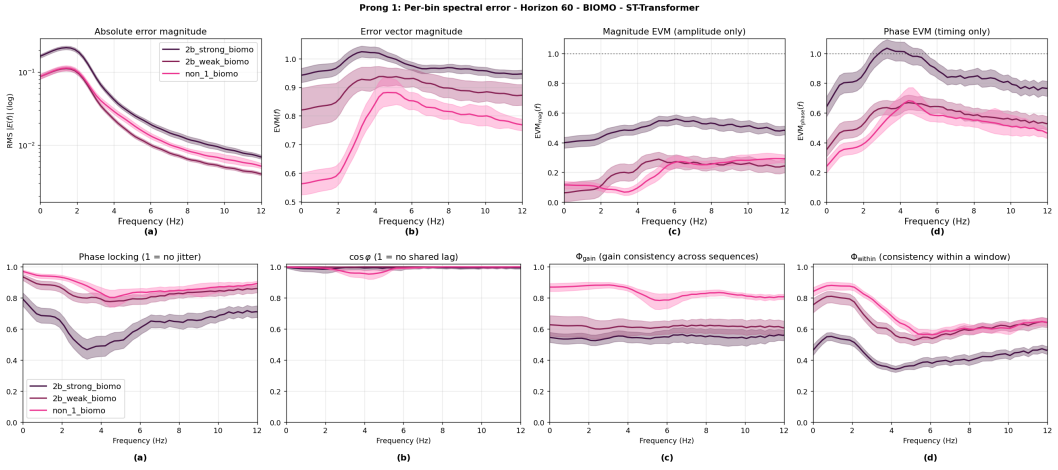

Figure 5: BIOMO Fourier results and auxiliary diagnostics for ST-Transformer. The strong-interaction and weak-interaction condition have a similar shape for all three EVMs compared to their non-interaction baseline. The strong condition has larger phase error and lower phase locking than the weak-interaction and non-interaction conditions. The amplitude undershoot is uniformly deeper for the strong condition, both strong and weak condition deepens the undershoot at 2.5-4.5 Hz.

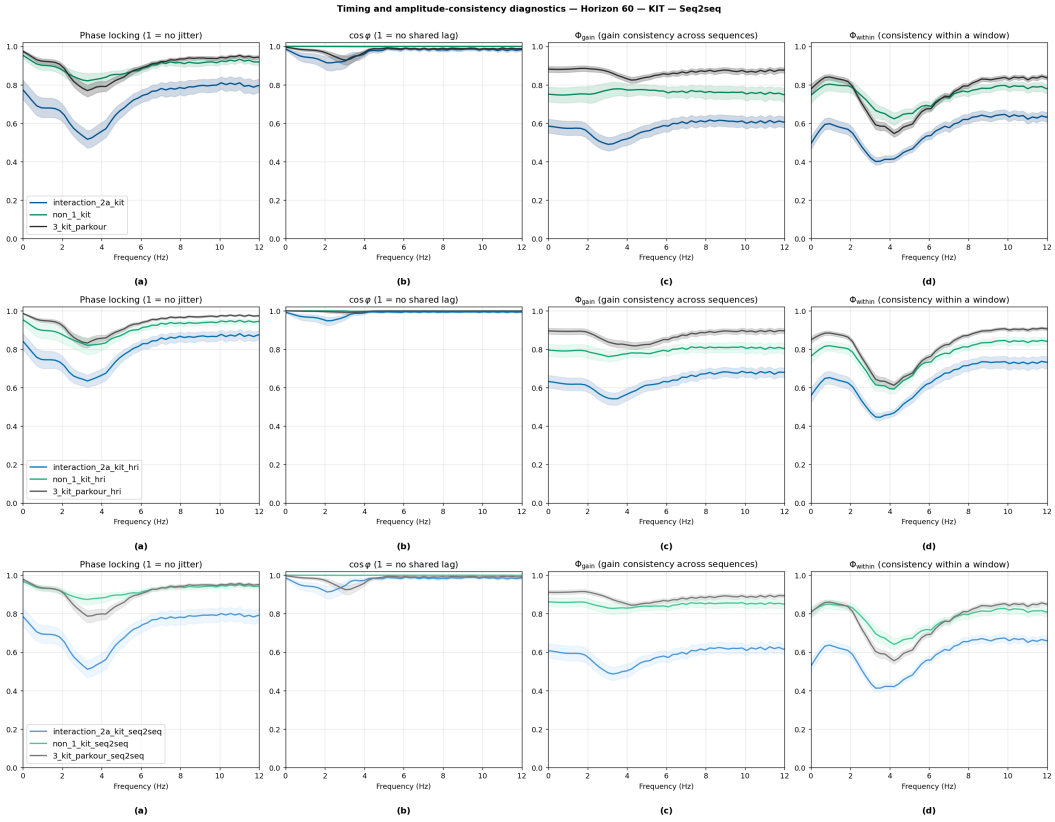

Figure 6: KIT auxiliary diagnostics for the three probe architectures: phase locking, mean phase-offset cosine, and the supplied gain- and within-window consistency measures. Interaction has reduced phase locking, while the mean phase-offset cosine remains close to one.

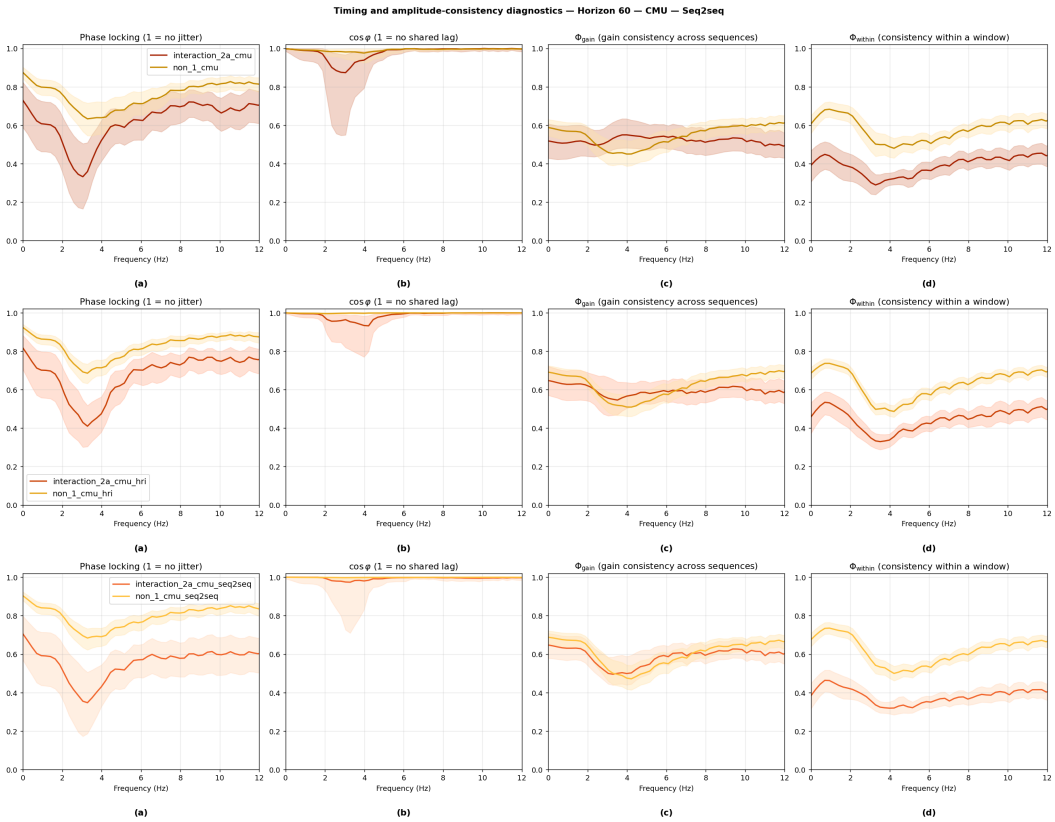

Figure 7: CMU auxiliary diagnostics. Interaction reduces phase locking. Gain-consistency curves cross for some architectures, so no uniform ordering is claimed for that diagnostic.

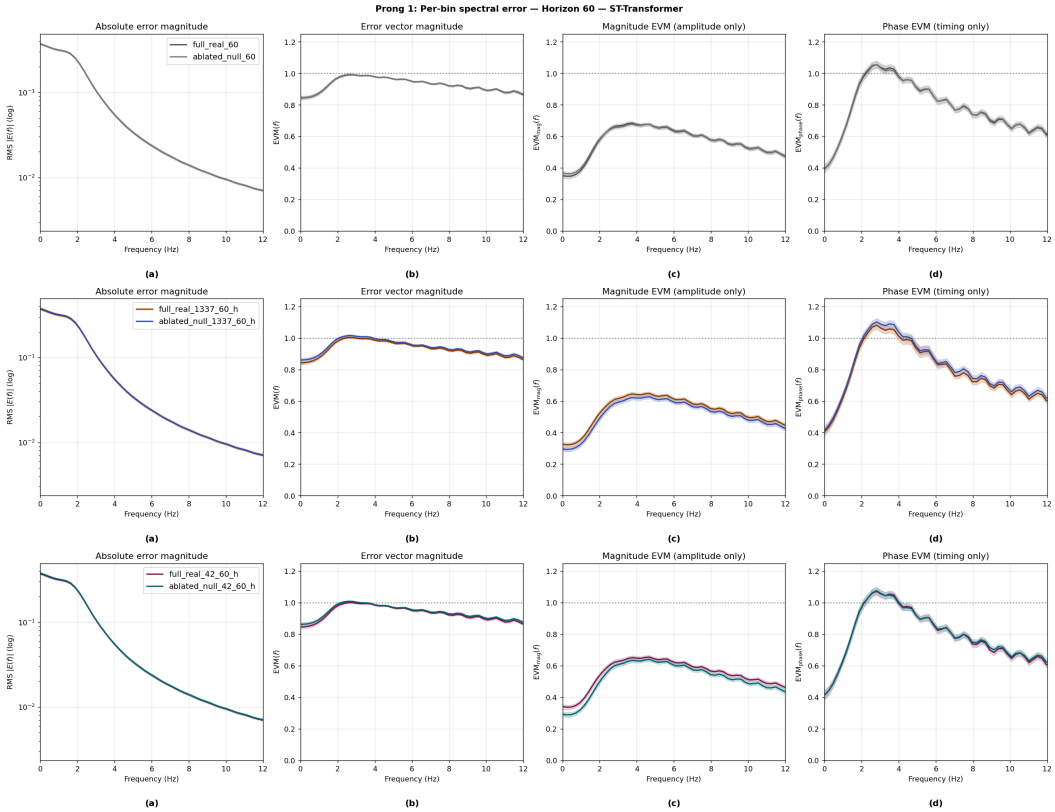

Figure 8: Interaction-aware Prong 1 results for seeds 2024, 1337, and 42, from top to bottom. Full and ablated curves largely overlap. The plots provide no consistent visual improvement across all diagnostics.

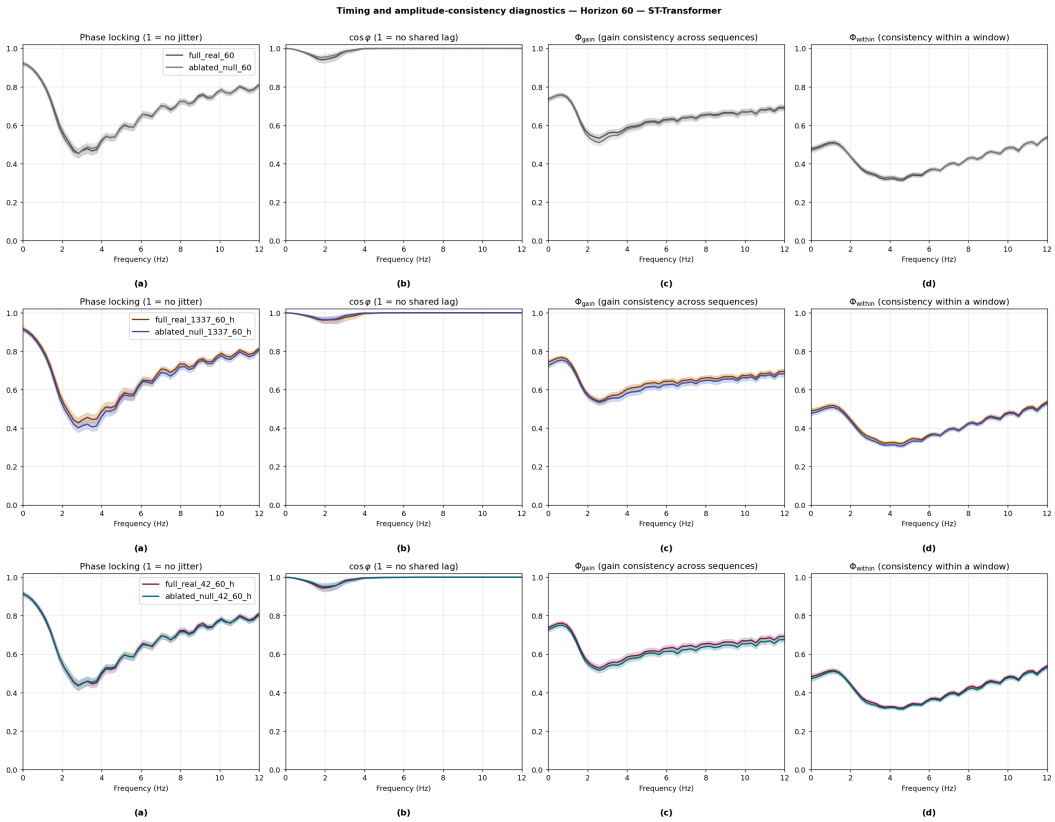

Figure 9: Interaction-aware auxiliary diagnostics for seeds 2024, 1337, and 42. Full and ablated curves remain close. This does not establish equivalence or a specific phase-alignment mechanism.

Table 4: Interaction-naive probe, Prong 2. Numbers are the penalty Δ over the in-distribution non-interaction baseline, with pointwise 95% bootstrap CIs. * CI excludes zero.

			LOW (0–1 Hz)		MID (1–5 Hz)		HIGH (5–10 Hz)	
Model	Test Scenario	h	Δ	95% CI	Δ	95% CI	Δ	95% CI
STT	KIT interaction	24	$+0.186^*$	$[+0.148, +0.225]$	$+0.161^*$	$[+0.125, +0.198]$	$+0.051^*$	$[+0.020, +0.082]$
		60	$+0.233^*$	$[+0.193, +0.276]$	$+0.173^*$	$[+0.138, +0.209]$	$+0.057^*$	$[+0.034, +0.080]$
	KIT parkour	24	-0.008	$[-0.049, +0.033]$	$+0.004$	$[-0.031, +0.040]$	$+0.017$	$[-0.017, +0.052]$
		60	$+0.016$	$[-0.022, +0.055]$	$+0.047^*$	$[+0.008, +0.085]$	$+0.046^*$	$[+0.023, +0.069]$
	CMU interaction	24	$+0.152^*$	$[+0.102, +0.210]$	$+0.102^*$	$[+0.057, +0.151]$	$+0.004$	$[-0.018, +0.027]$
		60	$+0.098^*$	$[+0.039, +0.153]$	$+0.079^*$	$[+0.045, +0.117]$	-0.015	$[-0.030, +0.000]$
	BIOMO weak	24	$+0.141^*$	$[+0.093, +0.194]$	$+0.171^*$	$[+0.131, +0.211]$	$+0.006$	$[-0.019, +0.031]$
		60	$+0.176^*$	$[+0.124, +0.227]$	$+0.206^*$	$[+0.165, +0.247]$	-0.012	$[-0.031, +0.008]$
	BIOMO strong	24	$+0.216^*$	$[+0.176, +0.252]$	$+0.283^*$	$[+0.247, +0.317]$	$+0.021$	$[-0.001, +0.044]$
		60	$+0.229^*$	$[+0.193, +0.265]$	$+0.284^*$	$[+0.247, +0.316]$	-0.023^*	$[-0.039, -0.005]$
HRI	KIT interaction	24	$+0.158^*$	$[+0.126, +0.192]$	$+0.095^*$	$[+0.065, +0.124]$	$+0.005$	$[-0.011, +0.022]$
		60	$+0.189^*$	$[+0.154, +0.222]$	$+0.106^*$	$[+0.078, +0.134]$	0.000	$[-0.008, +0.007]$
	KIT parkour	24	-0.002	$[-0.033, +0.029]$	-0.036^*	$[-0.066, -0.004]$	-0.021^*	$[-0.041, -0.003]$
		60	$+0.024$	$[-0.007, +0.055]$	-0.030	$[-0.062, +0.001]$	-0.015^*	$[-0.024, -0.006]$
	CMU interaction	24	$+0.151^*$	$[+0.097, +0.208]$	$+0.096^*$	$[+0.060, +0.133]$	$+0.017^*$	$[+0.001, +0.033]$
		60	$+0.095^*$	$[+0.043, +0.152]$	$+0.090^*$	$[+0.063, +0.118]$	$+0.008^*$	$[+0.001, +0.015]$
Seq2seq	KIT interaction	24	$+0.287^*$	$[+0.254, +0.319]$	$+0.167^*$	$[+0.131, +0.201]$	$+0.017$	$[-0.014, +0.045]$
		60	$+0.321^*$	$[+0.286, +0.356]$	$+0.207^*$	$[+0.172, +0.240]$	$+0.025^*$	$[+0.008, +0.042]$
	KIT parkour	24	$+0.069^*$	$[+0.036, +0.103]$	-0.006	$[-0.042, +0.030]$	-0.029	$[-0.063, +0.003]$
		60	$+0.094^*$	$[+0.062, +0.127]$	$+0.056^*$	$[+0.014, +0.096]$	$+0.015$	$[-0.003, +0.032]$
	CMU interaction	24	$+0.172^*$	$[+0.109, +0.239]$	$+0.087^*$	$[+0.046, +0.130]$	-0.009	$[-0.034, +0.014]$
		60	$+0.125^*$	$[+0.059, +0.199]$	$+0.087^*$	$[+0.055, +0.119]$	-0.014^*	$[-0.026, -0.003]$

Table 5: Interaction-naive probe, Prong 3, sym4, $h = 24$ ($L = 2$) and $h = 60$ ($L = 3$). Numbers are the penalty Δ over the in-distribution non-interaction baseline, with pointwise 95% bootstrap CIs. At $h = 24$ the decomposition has depth $L = 2$, so its approximation band A_2 (0–7.5 Hz) spans both A_3 and D_3. * CI excludes zero.

			A_3 (0–3.75 Hz)		D_3 (3.75–7.5 Hz)		D_2 (7.5–15 Hz)		D_1 (15–30 Hz)	
Model	Test scenario	h	Δ	95% CI	Δ	95% CI	Δ	95% CI	Δ	95% CI
STT	KIT interaction	24	A_2: $+0.238^*$	$[+0.189, +0.284]$			$+0.120^*$	$[+0.089, +0.148]$	$+0.019^*$	$[+0.002, +0.037]$
		60	$+0.262^*$	$[+0.217, +0.306]$	$+0.184^*$	$[+0.154, +0.215]$	$+0.130^*$	$[+0.104, +0.155]$	$+0.024^*$	$[+0.014, +0.033]$
	KIT parkour	24	A_2: -0.081^*	$[-0.126, -0.039]$			-0.046^*	$[-0.075, -0.018]$	-0.021^*	$[-0.042, -0.001]$
		60	-0.080^*	$[-0.122, -0.039]$	$+0.003$	$[-0.028, +0.033]$	-0.008	$[-0.033, +0.017]$	$+0.001$	$[-0.011, +0.013]$
	CMU interaction	24	A_2: $+0.181^*$	$[+0.135, +0.225]$			$+0.095^*$	$[+0.067, +0.124]$	$+0.058^*$	$[+0.042, +0.074]$
		60	$+0.136^*$	$[+0.088, +0.189]$	$+0.057^*$	$[+0.033, +0.081]$	$+0.052^*$	$[+0.032, +0.071]$	$+0.026^*$	$[+0.017, +0.034]$
	BIOMO weak	24	A_2: $+0.265^*$	$[+0.207, +0.323]$			$+0.123^*$	$[+0.096, +0.152]$	$+0.044^*$	$[+0.031, +0.056]$
		60	$+0.297^*$	$[+0.234, +0.367]$	$+0.156^*$	$[+0.112, +0.199]$	$+0.082^*$	$[+0.059, +0.107]$	$+0.015^*$	$[+0.006, +0.023]$
	BIOMO strong	24	A_2: $+0.387^*$	$[+0.348, +0.424]$			$+0.175^*$	$[+0.153, +0.196]$	$+0.050^*$	$[+0.040, +0.061]$
		60	$+0.387^*$	$[+0.351, +0.423]$	$+0.200^*$	$[+0.169, +0.229]$	$+0.097^*$	$[+0.080, +0.115]$	$+0.017^*$	$[+0.009, +0.024]$
HRI	KIT interaction	24	A_2: $+0.186^*$	$[+0.148, +0.223]$			$+0.089^*$	$[+0.062, +0.114]$	$+0.046^*$	$[+0.016, +0.079]$
		60	$+0.193^*$	$[+0.157, +0.231]$	$+0.106^*$	$[+0.081, +0.131]$	$+0.076^*$	$[+0.055, +0.101]$	$+0.032^*$	$[+0.004, +0.059]$
	KIT parkour	24	A_2: -0.092^*	$[-0.130, -0.056]$			-0.060^*	$[-0.087, -0.033]$	-0.084^*	$[-0.118, -0.050]$
		60	-0.119^*	$[-0.155, -0.082]$	-0.056^*	$[-0.082, -0.027]$	-0.053^*	$[-0.079, -0.025]$	-0.090^*	$[-0.120, -0.059]$
	CMU interaction	24	A_2: $+0.188^*$	$[+0.141, +0.236]$			$+0.105^*$	$[+0.077, +0.132]$	$+0.071^*$	$[+0.054, +0.087]$
		60	$+0.154^*$	$[+0.114, +0.199]$	$+0.090^*$	$[+0.068, +0.112]$	$+0.081^*$	$[+0.061, +0.101]$	$+0.044^*$	$[+0.035, +0.053]$
Seq2seq	KIT interaction	24	A_2: $+0.292^*$	$[+0.249, +0.334]$			$+0.157^*$	$[+0.128, +0.186]$	$+0.034^*$	$[+0.022, +0.047]$
		60	$+0.320^*$	$[+0.278, +0.361]$	$+0.227^*$	$[+0.196, +0.257]$	$+0.174^*$	$[+0.149, +0.197]$	$+0.031^*$	$[+0.024, +0.038]$
	KIT parkour	24	A_2: -0.046^*	$[-0.095, -0.004]$			-0.031	$[-0.061, +0.001]$	-0.012	$[-0.027, +0.001]$
		60	-0.030	$[-0.074, +0.012]$	$+0.040^*$	$[+0.008, +0.071]$	$+0.021$	$[-0.005, +0.046]$	$+0.002$	$[-0.005, +0.009]$
	CMU interaction	24	A_2: $+0.192^*$	$[+0.143, +0.241]$			$+0.093^*$	$[+0.068, +0.119]$	$+0.043^*$	$[+0.030, +0.056]$
		60	$+0.188^*$	$[+0.146, +0.230]$	$+0.077^*$	$[+0.055, +0.101]$	$+0.062^*$	$[+0.044, +0.080]$	$+0.025^*$	$[+0.018, +0.033]$