

A Design Space for Cross-Modal Packet Loss Concealment in the Musical Metaverse

Cristina Rottondi

*Dept. of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
cristina.rottondi@polito.it*

Filippo Daniotti

*Dept. of Information Eng. and Computer Science
University of Trento
Trento, Italy
filippo.daniotti@unitn.it*

Omran Ayoub

*University of Applied Sciences and Arts of Southern Switzerland
Lugano, Switzerland
omran.ayoub@supsi.ch*

Luca Turchet

*Dept. of Information Eng. and Computer Science
University of Trento
Trento, Italy
luca.turchet@unitn.it*

Abstract—The Musical Metaverse (MM) enables real-time, immersive interaction among geographically distributed musicians and audiences through the exchange of heterogeneous multimodal data streams, including audio, visual, gestural, haptic, and neurophysiological signals. Ensuring robustness against packet loss under the stringent latency constraints of musical performance remains a major challenge. Existing Packet Loss Concealment (PLC) techniques are unimodal, thus do not exploit the correlations naturally arising among multiple modalities. This paper introduces a multimodal perspective on PLC for the MM. We propose a scenario-driven design space spanning six representative interaction scenarios, relating participant interactions, modality composition, temporal constraints, and reconstruction strategies. We complement the proposed design space with an analytical latency and computational model derived from representative PLC implementations and MM timing requirements. The illustrative analysis shows that latency-critical performer interactions require lightweight PLC techniques, whereas audience-oriented and delay-tolerant interactions can accommodate more sophisticated neural and cross-modal reconstruction strategies under the adopted assumptions. These quantitative results define plausible operating regions rather than predict implementation performance. More broadly, the results suggest that PLC in the MM should be approached as a scenario-dependent systems problem, where the choice of reconstruction strategy is jointly determined by interaction patterns, modality composition, and latency constraints.

Index Terms—Cross-modal Packet Loss Concealment; Musical Metaverse; Low-latency Streaming;

I. INTRODUCTION

The emergence of the Musical Metaverse (MM) is redefining the way music is created, performed, and experienced, enabling geographically distributed musicians and audiences

This work was supported by the MUSMET project funded by the EIC Pathfinder Open scheme of the European Union (grant agreement n. 101184379) and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the European Innovation Council can be held responsible for them.

to interact in real time within immersive, multisensory environments [1], [2]. These systems rely on the continuous exchange of heterogeneous data streams, including audio, gestures, visual information, haptic feedback, and neurophysiological signals. Ensuring the reliability and temporal consistency of such multimodal communication is critical, particularly in music performance contexts where even small impairments can disrupt synchronization, expressivity, and user experience [3].

A key challenge in MM systems is the occurrence of packet losses, which are unavoidable in wireless and large-scale distributed environments. To address this issue, several methods have been investigated, typically classified as either sender-based or receiver-based [4]. While the former introduce redundancy to the transmitted stream, the latter aim to produce a perceptually similar replacement for the missing segment, and are referred to as Packet Loss Concealment (PLC) algorithms. In traditional Networked Music Performance (NMP) systems, PLC has been addressed within a *unimodal* framework, focusing exclusively on audio signal reconstruction [5]. However, this approach may be insufficient in the MM context, where multiple tightly coupled modalities contribute jointly to the perception of musical interaction. In particular, current PLC techniques do not exploit the inherent redundancy and correlation across modalities, nor do they account for the heterogeneous latency constraints and perceptual priorities that characterize multimodal immersive environments.

This gap highlights the need for a new generation of PLC methods that are:

- *cross-modal*, leveraging correlations between audio, gesture, haptic, and visual data and exploiting such relationships for cross-inference;
- *context-aware*, adapting to different interaction scenarios (e.g., musician–musician vs musician–audience);
- *latency-aware*, operating under strict real-time constraints for musical synchronization;

- *perception-driven*, prioritizing perceptual coherence over exact signal reconstruction.

A fundamental characteristic of the MM is that, despite its multimodal nature, musical interaction remains primarily anchored in audio [6], [7]. While visual, gestural, haptic, and physiological modalities enrich communication and contribute to immersion, they complement rather than replace the central role of sound in musical performance. Consequently, PLC strategies should not treat all modalities uniformly: preserving audio continuity and synchronization remains the primary objective, while the remaining modalities can support reconstruction, compensate for missing information, or enhance perceptual coherence according to their respective temporal and perceptual roles.

More fundamentally, the diversity of interaction patterns envisioned for the MM suggests that no single PLC strategy is likely to be universally appropriate. Instead, the choice of reconstruction technique should depend on the interaction scenario, the participating modalities, and the associated latency and perceptual constraints. This observation motivates the scenario-driven perspective adopted throughout this paper.

Despite the growing interest in multimodal streaming and immersive applications [8], a systematic framework for understanding and designing cross-modal PLC techniques for the MM is currently missing. In particular, there is no unified design space that relates interaction scenarios, modality composition, latency requirements, and reconstruction strategies, nor a quantitative analysis of the computational and temporal trade-offs associated with different PLC approaches. Consequently, there is currently little guidance on which PLC strategies are suitable for different MM scenarios and how latency constraints influence the feasibility of increasingly sophisticated cross-modal reconstruction techniques.

This paper addresses these challenges by providing both conceptual and quantitative contributions to the design of cross-modal PLC systems for the MM:

- *Scenario-driven design space for cross-modal PLC*: We propose a classification framework for cross-modal PLC strategies across six representative MM scenarios, identifying the relationships between participant interactions, modality composition, temporal constraints, and reconstruction methods. The resulting design space highlights the evolution from traditional unimodal PLC towards cross-modal and hierarchical reconstruction approaches.
- *Analytical characterization of PLC functions*: We derive representative computational, memory, and latency requirements for different classes of PLC algorithms by combining resource estimates from the state-of-the-art PLC literature with representative MM processing pipelines. This characterization provides a common reference for comparing classical signal-processing methods, neural networks (NN)-based PLC techniques, and cross-modal inference approaches.
- *Analytical design-space exploration*: Using the proposed latency and computational models, we analyze the operating regions in which classical, neural, and cross-modal

PLC strategies remain feasible under representative MM scenarios.

The remainder of the paper is organized as follows. Section II reviews the literature on audio PLC, motion reconstruction, and cross-modal inference. Section III introduces the proposed scenario-driven design space for cross-modal PLC strategies across representative MM applications. Section IV provides a quantitative characterization of PLC functions together with a design-space analysis of the trade-offs between latency, computational complexity, and reconstruction strategies. Finally, Section V concludes the paper and outlines future research directions.

II. RELATED WORK

Existing research on PLC can be broadly organized into three streams relevant to MM applications: audio PLC, motion and gesture reconstruction in networked environments, and multimodal reconstruction techniques exploiting correlations across modalities.

A. Audio Packet Loss Concealment

Audio PLC is the most mature research area and has been extensively investigated in both speech and music communication systems. Classical approaches can be broadly categorized into interpolation and extrapolation methods. Interpolation techniques estimate missing samples using both past and future observations surrounding a loss interval and are therefore suitable when additional delay can be tolerated. Extrapolation techniques instead predict missing samples using only previously received information, making them particularly suitable for real-time applications with stringent latency constraints [4].

Traditional PLC methods include waveform repetition, linear prediction, waveform substitution, and frequency-domain reconstruction techniques [4]. More recently, deep learning approaches have significantly advanced the state of the art. Examples related to speech include PLCNet [9], tPLCNet [10], and BS-PLCNet [11], as well as systems developed for the ICASSP Audio PLC Challenge [12]. These approaches achieve improved perceptual quality by learning temporal patterns directly from audio signals. Despite their effectiveness, however, they remain fundamentally unimodal, focusing exclusively on reconstructing missing audio packets without exploiting information available from other modalities.

Within musical applications, PLC research remains comparatively limited. Existing work has primarily focused on audio signal recovery for networked music performance systems, where maintaining synchronization and low latency is critical [13]. Examples include waveform substitution approaches [14] and, more recently, hybrid approaches such as PARCnet [15], as well as subsequent work exploring alternative loss functions for training PARCnet, notably Tilt Loss [16]. Recent initiatives such as the IEEE Internet of Sounds Music PLC Challenge [5] have highlighted the need for PLC techniques specifically tailored to musical content, which exhibits different temporal and spectral characteristics compared to speech. At the same time, the lack of reliable objective metrics for assessing

the quality of music PLC algorithms remains a significant challenge [17], [18].

B. Motion and Gesture Reconstruction in Networked Environments

A second body of literature addresses the reconstruction of motion and gesture information in distributed interactive systems and networked virtual environments. Rather than being explicitly framed as PLC, this research focuses on compensating for delayed or missing updates of avatars, tracked body parts, or interaction events [19].

Classical solutions include dead reckoning and predictive state estimation techniques [20]. These methods estimate missing motion states from previously observed trajectories using interpolation, extrapolation, Kalman filtering, or biomechanical constraints. More recently, Machine Learning (ML)-based approaches have been proposed for reconstructing human motion from sparse observations and incomplete tracking data [21]–[24]. These methods exploit spatial and temporal correlations to infer plausible movement trajectories even when sensor measurements are partially unavailable. Although developed independently from the PLC literature, they address a conceptually similar problem: estimating missing information under timing constraints. Consequently, they provide valuable insights for reconstructing gesture and motion streams in MM environments, particularly in scenarios involving distributed musical performance and embodied interaction.

C. Cross-Modal Reconstruction and Inference

A growing body of research has explored the use of one modality to compensate for missing or degraded information in another modality [25]. In audiovisual communications, visual information has been used to improve speech enhancement and speech reconstruction by exploiting correlations between lip movements and acoustic signals [26]. Similarly, audio-driven gesture synthesis techniques generate body movements from speech signals, while gesture-based approaches infer timing, articulation, and expressive characteristics of sound-producing actions [27], [28].

Recent advances in multimodal ML have further enabled cross-modal inference and missing-modality reconstruction through self-supervised learning, multimodal transformers, and generative models [29]. These approaches leverage correlations across modalities to reconstruct unavailable information from observed signals. Examples include audiovisual representation learning, multimodal data imputation, and cross-modal generation systems capable of predicting visual, auditory, or physiological information from complementary sensory inputs.

While these methods demonstrate the feasibility of cross-modal reconstruction, they are rarely studied in the context of PLC and are typically evaluated in offline or non-real-time settings. Furthermore, most existing approaches focus on pairs of modalities rather than the heterogeneous multimodal ecosystems characteristic of MM applications.

D. Research Gap

The literature reviewed above reveals a fragmented landscape. Audio PLC research has produced highly effective unimodal reconstruction techniques, motion reconstruction research has developed predictive methods for recovering missing avatar and gesture information, and multimodal ML has demonstrated the feasibility of cross-modal inference. However, these research streams have largely evolved independently.

To the best of our knowledge, no prior work has proposed a unified classification scheme and design space for PLC strategies that jointly considers audio, gesture, visual, haptic, and neurophysiological modalities under the stringent latency constraints of MM environments. Furthermore, existing studies do not address how multimodal PLC functions can be orchestrated and deployed across distributed cloud–edge infrastructures while balancing reconstruction quality, latency, and computational cost. This gap motivates the scenario-driven design space presented in this paper.

Notably, unlike general multimodal reconstruction frameworks that primarily ask whether a missing modality can be inferred from the available ones, our proposed design space treats packet loss as an online system problem. It jointly indexes the direction of interaction, modality-specific timing budget, causal or look-ahead requirements, computational placement, and the perceptual priority of the protected stream. Its main contribution is therefore a decision framework for determining when unimodal or cross-modal reconstruction is deployable in an MM communication path.

III. DESIGN SPACE OF CROSS-MODAL PLC TECHNIQUES

A. Requirements of MM Scenarios

This section describes the design space of cross-modal PLC techniques grounded in the characteristics identified across six MM scenarios, where heterogeneous data streams (audio, gestures/visuals, EEG, and haptics) are tightly coupled under strict temporal constraints, making PLC design highly dependent on interaction patterns and perceptual requirements. The scenarios are defined as follows:

- **Scenario 1 – Audience-only interaction:** Audience members participate in a virtual concert based on pre-recorded music. Interactions occur among audience members and with the system through voice, gestures, and neurophysiological signals, which are used to adapt the immersive experience.
- **Scenario 2 – Musicians with virtual instruments:** Distributed musicians perform together using purely virtual instruments in a shared virtual environment. The interaction is limited to performer-performer communication, with real-time exchange of audio, control, and gesture data.
- **Scenario 3 – Musicians with real instruments (mixed reality):** Musicians perform using physical instruments while interacting through a mixed reality environment.

TABLE I
CROSS-MODAL PACKET LOSS CONCEALMENT (PLC) ACROSS MUSICAL METAVERSE SCENARIOS

Scenario	Interactions	Data Streams	Timing Sensitivity	PLC Targets	PLC Strategies
1	Audience–Audience; Audience–System	Audio playback, voice, gestures, EEG, haptics, avatars	Medium	audience-generated voice/gestures/EEG signals (uplink/downlink); audio, avatar and haptics (downlink only)	Audio interpolation; gesture smoothing; EEG temporal filtering; cross-modal inference (audio/gestures)
2	Musician–Musician	Instrument audio, control signals, gestures, avatars	Very High	Audio, control signals, gestures (uplink/downlink)	Ultra-low latency audio extrapolation; predictive control reconstruction; kinematic gesture prediction; audio-driven gesture recovery
3	Musician–Musician	High-quality (spatialized) audio, gestures, visuals	Very High	Audio, gestures (uplink/downlink)	Ultra-low latency audio extrapolation; kinematic gesture prediction; perceptual masking
4	Musicians–Musicians; Musicians–Audience; Audience–Audience	Combined streams from Scenarios 1 & 2	Very High (musicians), Medium (audience)	Combined streams from Scenarios 1 & 2	Techniques from Scenarios 1 & 2, cross-modal emotion inference, cross-modal PLC (audio ↔ gesture/haptics)
5	Musicians–Musicians; Musicians–Audience; Audience–Audience	Combined streams from Scenarios 1 & 3	Very High (musicians), Medium (audience)	Combined streams from Scenarios 1 & 3	Techniques from Scenarios 1 & 3, cross-modal emotion inference, cross-modal PLC (audio ↔ gesture/haptics)
6	Musicians–Musicians; Musicians–Audience; Audience–Audience	Combined streams from Scenarios 1, 2 & 3	Very High (musicians), Medium (audience)	Combined streams from Scenarios 1, 2 & 3	Techniques from Scenarios 1, 2 & 3, cross-modal emotion inference, cross-modal PLC (audio ↔ gesture/haptics)

Real-world audio signals are combined with virtual representations, requiring tight synchronization between physical and virtual domains.

- **Scenario 4 – Musicians and audience (virtual instruments):** This scenario extends Scenario 2 by including audience members, who can interact with each other and with the musicians. It introduces bidirectional performer-audience communication in a fully virtual setting.
- **Scenario 5 – Musicians (real instruments) and audience (virtual reality):** Musicians perform using real instruments in mixed reality, while audiences attend the concert in virtual reality. This scenario combines real-world performance with virtual audience interaction.
- **Scenario 6 – Fully integrated multimodal interaction:** The most complex scenario, combining real and virtual instruments, musicians, and audiences. All participants interact through multiple modalities (audio, gestures, haptics, and physiological signals), requiring scalable, synchronized, and immersive communication.

Table I analyzes each of the six scenarios under four key dimensions that determine PLC strategies:

- **Interaction matrix:** identifies whether communication occurs among musicians, audiences, or both.
- **Modality composition:** describes the types of data streams involved (audio, gesture, EEG, haptics, visuals).
- **Temporal criticality:** captures latency and synchronization requirements.
- **Reconstruction strategy:** distinguishes between uni-modal and cross-modal approaches.

1) *Audience-Centric PLC (Scenario 1):* In Scenario 1, interactions are limited to audience–audience and audience–system exchanges, with no live musical performance. Data streams include voice, gestures, EEG signals, and haptic feedback, along with downlink audio playback. Due to the absence of tight performer synchronization loops, temporal constraints are moderate. PLC is mainly applied to: *i)* uplink audience signals (e.g., missing gesture or voice packets), and *ii)* downlink immersive rendering streams.

This scenario supports *moderate-complexity PLC strategies*, such as:

- interpolation and smoothing for gesture and EEG signals;
- audio concealment via inter or extrapolation;
- simple cross-modal inference (e.g., reconstructing gestures from audio events, such as hand clapping, or vice-versa).

2) *Performer-Centric PLC (Scenarios 2 and 3):* Scenarios 2 and 3 involve musician–musician interactions, forming a tightly coupled real-time loop. Scenario 2 considers virtual instruments, while Scenario 3 involves real instruments in mixed reality. These scenarios impose *strict latency constraints* (on the order of tens of milliseconds), making PLC particularly challenging. The most critical streams are audio signals, control data (e.g., MIDI-like events), and gesture streams. PLC must therefore rely on *ultra-low-latency techniques*, including: *i)* time-domain audio extrapolation; *ii)* predictive reconstruction of control signals, *iii)* kinematic prediction of gestures.

Cross-modal PLC is feasible but limited to lightweight mappings (e.g., audio-driven gesture prediction), as complex models would violate latency constraints. In Scenario 3, ad-

ditional challenges arise from spatial audio and physical realism, which favor perceptually motivated PLC techniques over computationally intensive signal reconstruction. In particular, auditory masking could be exploited to conceal packet-loss artifacts when they occur simultaneously with louder sounds or salient musical events [30], [31]: rather than reconstructing the exact missing waveform, the system would ensure that any resulting distortion remains below the listener’s perceptual threshold.

3) *Hybrid Performer–Audience PLC (Scenarios 4 and 5)*: Scenarios 4 and 5 introduce interactions between musicians and audiences, combining the requirements of Scenarios 1, 2, and 3. These scenarios exhibit heterogeneous temporal constraints: musician streams require ultra-low latency, whereas audience streams tolerate higher delays. This leads to a hierarchical PLC strategy, where low-latency, lightweight PLC is applied to musician streams, whereas more complex, possibly ML-based PLC is applied to audience-related streams. Additionally, edge-based processing can be used to reconstruct aggregated audience signals (e.g., crowd submixes), reducing bandwidth requirements and enabling scalable PLC.

4) *Fully Cross-modal PLC (Scenario 6)*: Scenario 6 represents the most complex case, integrating all modalities and interaction types. The system must handle audio, gesture, EEG, haptic, and visual streams, taking into account multi-directional interactions among all participants and strict synchronization across modalities. Temporal requirements are heterogeneous: audio and haptics are highly time-critical, visual and gesture data have moderate constraints, while EEG signals are relatively tolerant to delay.

In this scenario, cross-modal PLC becomes essential. Reconstruction strategies include:

- audio-to-gesture and gesture-to-audio inference;
- audio-driven haptic synthesis;
- perceptual masking guided by visual continuity;
- adaptive prioritization based on user state (e.g., EEG-derived attention).

Moreover, PLC is envisioned to be implemented as a *multi-layer system* spanning device, edge, and cloud levels, balancing latency and computational complexity.

B. PLC Strategy Design Space

Based on the scenarios discussed above, we derive a PLC strategy design space characterized by a trade-off between latency constraints and reconstruction complexity. Scenarios with stringent real-time requirements (e.g., musician–musician interaction) tend to favor lightweight PLC methods, whereas scenarios with greater latency tolerance (e.g., audience-facing interactions) can accommodate more computationally intensive approaches, including ML-based and cross-modal reconstruction techniques. Furthermore, PLC in MM systems must prioritize *perceptual coherence* over exact signal reconstruction. In many cases, maintaining consistency across modalities yields a better user experience than perfectly reconstructing a single stream. Nevertheless, perceptual coherence should not

be interpreted as implying equal importance across modalities. Because musical interaction is fundamentally conveyed through sound, audio reconstruction generally remains the highest priority, whereas gesture, visual, haptic, and physiological modalities primarily contribute to preserving the continuity, intelligibility, and immersion of the overall musical experience.

Across all scenarios, PLC techniques can be classified into three main categories:

- **Unimodal PLC**: Traditional PLC techniques operate independently on each modality. Examples include audio (waveform interpolation, spectral reconstruction), gesture/visual (trajectory interpolation, filtering), haptics (signal interpolation or event substitution), and EEG (temporal smoothing). These methods are essential for low-latency scenarios but limited in their ability to exploit multimodal redundancy.
- **Cross-Modal PLC**: Cross-modal PLC leverages correlations between modalities to reconstruct missing data. Examples include audio to gesture (such as reconstructing motion from rhythmic patterns), gesture to audio (such as predicting timing and dynamics), audio to haptics (such as mapping signal energy to tactile feedback), and visual to audio (such as masking glitches using visual continuity). This class is particularly relevant in complex MM scenarios, where perceptual coherence across modalities is more important than exact reconstruction.
- **Hierarchical PLC**: Given the distributed architecture of MM systems, PLC can be implemented at multiple levels: *device level* (ultra-low latency, lightweight signal-based PLC), *edge level* (moderate latency, cross-modal and model-based PLC), and *cloud level* (high-complexity reconstruction for non-critical streams). This hierarchical approach enables adaptive trade-offs between latency and reconstruction quality.

Taken together, the proposed design space suggests that PLC should no longer be viewed as a single reconstruction problem, but rather as a family of scenario-dependent design problems. Different MM applications naturally require different concealment strategies according to their interaction patterns, modality composition, and temporal constraints. Consequently, selecting an appropriate PLC strategy becomes an architectural design decision rather than merely an algorithmic one.

It’s worth noting that the taxonomy and feasibility model are conceptual contributions, whereas the device–edge–cloud mapping is an engineering recommendation derived from them. It should therefore be interpreted as a testable architectural hypothesis whose final placement decisions depend on measured network, hardware, privacy, and workload conditions.

IV. QUANTITATIVE DESIGN-SPACE ANALYSIS

The previous section introduced the qualitative design space of cross-modal PLC in the MM. In this section, we complement it with an analytical model that estimates the feasibility of representative PLC strategies under different latency and

computational constraints. Rather than benchmarking specific implementations, the objective is to identify the operating regions in which different classes of PLC algorithms are expected to remain feasible across the proposed scenarios.

A. Assumptions

The analysis uses representative PLC processing delays derived from the audio PLC literature and from the relative complexity of non-audio reconstruction tasks. Classical PLC methods, such as repetition, interpolation, waveform substitution, and Linear Predictive Coding (LPC)-based extrapolation, operate directly on the received signal and are commonly used in low-latency audio streaming because they require negligible buffering and limited computation [4]. Lightweight NN-based PLC methods, such as PLCNet and tPLCNet, are explicitly designed for real-time operation with low latency and short temporal context [9], [10]. More complex neural architectures, such as BS-PLCNet, achieve stronger reconstruction performance but at substantially higher computational cost; for example, the original BS-PLCNet model reports 8.95 GFLOP per inference [11]. At a 20-ms inference cadence (50 inferences/s), this corresponds to 0.4475 TFLOP/s; Table II rounds this throughput proxy to 0.50 TFLOPS, while a 10-ms cadence would approximately double it. The ICASSP 2024 Audio Deep PLC Challenge further confirms that modern NN-based PLC systems are evaluated under strict real-time constraints [12].

For non-audio modalities, standardized PLC benchmarks are not yet available. We therefore use conservative representative values based on the expected complexity of the corresponding reconstruction tasks: gesture and avatar-state concealment rely on interpolation, dead reckoning, Kalman filtering, or trajectory prediction; haptic concealment relies on low-dimensional interpolation or event prediction; EEG and biometric concealment rely on temporal smoothing or autoregressive imputation. Cross-modal PLC is modeled as an additional fusion step that exploits correlations across modalities and is assigned the same latency class as advanced NN-based PLC.

Table II summarizes the representative characteristics of the PLC methods adopted in the analytical design-space exploration. The reported values are derived from the literature and are intended as design-space parameters rather than implementation-specific benchmarks. Their purpose is to distinguish between lightweight, latency-critical PLC functions and more computationally demanding neural and cross-modal reconstruction approaches.

The frame and look-ahead ranges in Table II describe algorithm classes rather than a single common setting. In the calculations, classical unimodal audio PLC uses a 1.33-ms packet and zero look-ahead (causal extrapolation), lightweight and advanced audio profiles use their stated processing delays with zero look-ahead, and cross-modal fusion uses one future frame, a 10-ms audio frame, 5-ms gesture frame, 1-ms haptic frame, or 100-ms EEG frame as applicable.

B. Latency Budget Computation

A PLC function is useful only if the complete transmission and processing chain remains compatible with the latency budget of the modality it protects. Following the end-to-end latency decomposition commonly adopted for NMP systems [13], we introduce the *baseline end-to-end latency* associated with modality m , denoted by T_m^{base} . This quantity represents the latency experienced by the modality before PLC is applied and includes all modality-dependent acquisition, processing, communication, and rendering stages.

For audio streams, T_m^{base} corresponds to the classical NMP latency decomposition, including acquisition, encoding, packetization, network transmission, de-jitter buffering, decoding, and playback delays. For the remaining modalities (gesture/avatar, haptic, and physiological streams), T_m^{base} encompasses the corresponding sensing, serialization, communication, buffering, deserialization, and rendering stages.

The end-to-end latency experienced by modality m when protected by PLC strategy p is therefore modeled as

$$T_{p,m}^{\text{E2E}} = T_m^{\text{base}} + D_p + T_p^{\text{look}}, \quad (1)$$

where D_p denotes the processing delay introduced by PLC strategy p , while T_p^{look} denotes the additional *look-ahead delay* required by the concealment algorithm. The latter represents the time that the receiver must wait before reconstruction can start and depends exclusively on the selected PLC strategy.

Extrapolation-based PLC methods estimate missing information exclusively from previously received packets and can therefore reconstruct a lost packet immediately after its detection. Hence, $T_p^{\text{look}} = 0$. Conversely, interpolation-based methods exploit both past and future packets to reconstruct the missing information. They must therefore wait until one or more subsequent packets have been received before reconstruction can begin. The corresponding look-ahead delay is modeled as $T_p^{\text{look}} = q_p T_m^{\text{frame}}$, where T_m^{frame} denotes the duration of one frame for modality m , and $q_p \in \mathbb{N}$ is the number of future frames required by PLC strategy p . For example, a linear interpolation algorithm requiring the next correctly received frame has $q_p = 1$, whereas more sophisticated bidirectional neural architectures may require $q_p > 1$. Since packet delay variation is absorbed by the de-jitter buffer already included in T_m^{base} , no additional jitter term is introduced.

The residual latency available for PLC depends on both the interaction type and the protected modality and is computed as

$$R_{i,m} = L_{i,m} - T_m^{\text{base}}, \quad (2)$$

where $L_{i,m}$ denotes the end-to-end latency budget associated with modality m during interaction type i (musician–musician, musician–audience, or audience–audience). A unimodal PLC strategy p is considered feasible only if

$$D_p + T_p^{\text{look}} \leq R_{i,m}. \quad (3)$$

TABLE II
CHARACTERISTICS OF THE PLC METHODS ADOPTED IN THE QUANTITATIVE DESIGN-SPACE ANALYSIS.

Category	Representative methods	CPU (vCPU)	RAM (GB)	GPU (TFLOPS)	Delay (ms)	Look-ahead (frames)	Frame (ms)	Placement	References
Audio	Waveform repetition, interpolation, LPC extrapolation	0.05	0.10	0	0.5	0–1	1.33	Device/Edge	[4], [13]
Audio	PLCNet, tPLCNet	0.25	0.50	0	3	0–1	10–20	Edge	[9], [10]
Audio	BS-PLCNet, ICASSP PLC Challenge	1.00	1.00	0.50	10	0	10–20	Edge/Cloud	[11], [12]
Gesture / Avatar	Dead reckoning, Kalman filtering, sparse pose reconstruction	0.10	0.25	0	1	0	5	Device/Edge	[20]–[22], [24]
Haptics	Event prediction, interpolation	0.05	0.10	0	1	0–1	1	Device/Edge	[32]
EEG / Biometrics	Temporal smoothing, autoregressive imputation	0.05	0.10	0	20	0	100	Edge	[33]
Cross-modal	Fusion and inference	1.00	1.00	0.50	10	1	$\max(T^{\text{frame}})$	Edge/Cloud	[25], [29]

Table III reports the representative component-level latency assumptions adopted to derive the baseline end-to-end latency of each modality. Following [34], a representative network transport delay of 15 ms is assumed throughout the analysis. This value is consistent with metropolitan/inter-city deployments spanning approximately 500 km and remains compatible with the one-way latency budgets commonly adopted for real-time NMP applications. The 30-ms musician–musician budget reflects the stringent interactive-performance regime discussed in NMP studies [3], [13]. The larger audience-facing budgets are explicit design assumptions reflecting lower synchronization criticality. Non-audio component values are likewise conservative engineering assumptions motivated by the processing chains described above. Table IV reports the interaction-dependent latency budgets adopted throughout the numerical analysis.

Based on the residual latency budgets derived above, we evaluate the feasibility of each PLC strategy for every interaction type and modality participating in the six MM scenarios. For a unimodal PLC strategy p , we define the feasibility ratio as

$$\gamma_{p,i,m} = \frac{D_p + q_p T_m^{\text{frame}}}{R_{i,m}}. \quad (4)$$

Values $\gamma_{p,i,m} < 1$ indicate that the PLC strategy fits within the residual latency budget, whereas $\gamma_{p,i,m} \geq 1$ indicates that the concealment delay exceeds the available budget.

For cross-modal PLC, concealment is performed by jointly exploiting a set of correlated modalities $\mathcal{F} = \{m_1, m_2, \dots\}$. Since reconstruction must preserve the temporal consistency of all participating streams, feasibility is evaluated conservatively with respect to the most stringent residual latency budget among the modalities involved:

$$\gamma_{p,i}^{\text{CM}}(\mathcal{F}) = \frac{D_p + q_p \max_{m \in \mathcal{F}} T_m^{\text{frame}}}{\min_{m \in \mathcal{F}} R_{i,m}}, \quad (5)$$

where the numerator accounts for both the PLC processing delay and the largest look-ahead delay required within the set

of involved modalities¹, while the denominator corresponds to the smallest residual latency budget among the participating modalities. Consequently, a cross-modal PLC strategy is feasible whenever $\gamma_{p,i}^{\text{CM}}(\mathcal{F}) < 1$.

C. Latency Feasibility Analysis

Table V reports the feasibility ratio defined in (4) for the representative PLC techniques introduced in Table II. The table is organized according to the six MM scenarios described in Section III. For each scenario, separate columns are provided for the interaction types it involves (musician–musician, musician–audience, and audience–audience).

The resulting feasibility matrix reveals that the applicability of PLC strategies depends jointly on the interaction scenario, the interaction type, and the modalities involved in the reconstruction process. Audience–audience interactions, which characterize Scenario 1 and the audience-facing components of Scenarios 4 and 5, retain comparatively large residual latency budgets after accounting for acquisition, processing, communication, and rendering delays. Consequently, these interactions can accommodate advanced neural and cross-modal PLC techniques.

A markedly different behavior is observed for musician–musician interactions in Scenarios 2 and 3. Here, the stringent latency constraints imposed by real-time musical performance leave only a few milliseconds available for PLC, particularly for audio/control and gesture streams. As a result, classical extrapolation techniques and lightweight neural PLC remain feasible, whereas more computationally demanding approaches rapidly exceed the available latency budget. Multimodal PLC is therefore feasible only when based on lightweight fusion mechanisms requiring little or no look-ahead.

The hybrid Scenarios 4 and 5 illustrate the benefits of distinguishing between interaction types within the same application. Performer-facing streams remain subject to strict timing constraints and therefore require lightweight PLC, whereas

¹For simplicity, we conservatively assume that the look-ahead delay is determined by the slowest modality participating in the fusion set.

TABLE III
REPRESENTATIVE COMPONENT-LEVEL LATENCY MODEL USED TO DERIVE THE BASELINE END-TO-END DELAY (IN MS) OF EACH MODALITY.

Latency contribution	Audio/control	Gesture/avatar state	Haptic	EEG/biometrics
Acquisition / sensing	0.5	2.0	1.0	10.0
Encoding / local processing	1.0	1.0	0.5	20.0
Packetization / serialization	2.5	1.0	0.5	1.0
Network transit	15.0	15.0	15.0	15.0
De-jitter / buffering	5.0	2.0	1.0	5.0
Decoding / deserialization	1.0	1.0	0.5	1.0
Rendering / playback / actuation	0.5	4.0	2.0	5.0
Baseline latency T_m^{base}	25.5	26.0	20.5	57.0

TABLE IV
REPRESENTATIVE END-TO-END LATENCY BUDGETS (MS) ADOPTED FOR EACH MODALITY UNDER THE THREE INTERACTION TYPES CONSIDERED IN THE QUANTITATIVE ANALYSIS.

Interaction type	Audio/control	Gesture/avatar	Haptic	EEG/biometrics
Musician–musician	30	30	30	–
Musician–audience	80	80	100	500
Audience–audience	100	100	100	500

musician–audience and audience–audience interactions can exploit more sophisticated cross-modal reconstruction strategies without violating the corresponding latency budgets.

Finally, Scenario 6 represents the most demanding operating condition, combining all interaction types and modalities within the same application. While unimodal PLC remains feasible for latency-critical streams through lightweight concealment strategies, the feasibility of multimodal PLC strongly depends on the selected fusion set. In particular, pairwise fusion involving latency-compatible modalities remains feasible in several cases, whereas larger multimodal fusion sets are increasingly constrained by the most stringent latency budget among the participating streams.

These observations quantitatively support the hierarchical PLC design principles introduced in Section III, showing that the choice of reconstruction strategy should jointly consider the interaction type, the modalities being fused, and the residual latency available for concealment.

D. Scenario-Level Computational Requirements

The latency feasibility analysis presented in the previous subsection identifies the set of PLC strategies that satisfy the timing constraints associated with each modality, interaction type, and MM scenario. Among all feasible strategies, a system designer would typically select the most sophisticated reconstruction method that remains compatible with the available residual latency budget in order to maximize the expected reconstruction quality. Since standardized perceptual quality metrics for cross-modal PLC are not yet available, we assume that PLC strategies can be ordered according to their reconstruction quality, following the progression discussed in Section II, namely classical PLC, lightweight NN-based PLC, advanced NN-based PLC, and cross-modal PLC, where increasingly sophisticated strategies exploit progressively more complex prediction and inference mechanisms.

For each interaction type i in scenario s , the selected unimodal PLC strategy for each modality m corresponds to the most advanced profile satisfying the latency feasibility condition in (4). Similarly, among all feasible cross-modal fusion sets, the one involving the largest number of modalities is selected, thereby maximizing the amount of information that can be exploited for PLC.

Using the representative models introduced above, we analytically estimate the computational resources required by each scenario s by selecting the most sophisticated PLC strategy that satisfies the corresponding latency constraints:

$$C_s = \sum_{i \in \mathcal{I}_s} \left(\sum_{m \in \mathcal{M}_{s,i}} C_{p_{s,i,m}^*} + \sum_{f \in \mathcal{F}_{s,i}^*} C_f \right), \quad (6)$$

where $\mathcal{I}_s$ denotes the set of interaction types present in scenario s , $\mathcal{M}_{s,i}$ is the set of modalities participating in interaction type i , $p_{s,i,m}^*$ is the selected unimodal PLC strategy protecting modality m , $\mathcal{F}_{s,i}^*$ is the set containing the selected cross-modal fusion strategy for interaction i , and C_f denotes its computational requirement. Equivalent expressions are adopted to derive the aggregate RAM and GPU requirements.

Equation (6) reports one PLC instance per interaction type and should not be read as a deployment-wide capacity estimate. If $N_{s,i}$ independent participant streams of type i are processed concurrently, the corresponding terms are multiplied by $N_{s,i}$ in the absence of batching or shared inference. Practical scaling may be sublinear when models and accelerators are shared, or superlinear once scheduling and memory contention appear; concurrency-aware dimensioning is therefore left to implementation-level evaluation.

Figure 1 summarizes the resulting resource requirements. The three stacked bars associated with each scenario respectively report the CPU, RAM, and GPU resources required by the selected PLC strategies, while the colors identify the contribution of the different interaction types. Hatched regions correspond to cross-modal PLC functions, whereas solid regions denote unimodal concealment techniques.

The figure confirms the trends already observed in the feasibility analysis. Scenario 1 is dominated by audience–audience interactions, whose relaxed latency constraints allow the deployment of advanced neural PLC together with multimodal fusion, resulting in comparatively high resource requirements. Scenarios 2 and 3 exhibit the lowest computational demand

TABLE V
FEASIBILITY RATIOS OF REPRESENTATIVE UNIMODAL AND CROSS-MODAL PLC STRATEGIES ACROSS MM SCENARIOS AND INTERACTION TYPES.

Modality / fusion set	PLC strategy	S1 A-A	S2 M-M	S3 M-M	S4 M-M	S4 M-A	S4 A-A	S5 M-M	S5 M-A	S5 A-A	S6 M-M	S6 M-A	S6 A-A
Audio/control	Classical audio PLC	0.01	0.11	0.11	0.11	0.01	0.01	0.11	0.01	0.01	0.11	0.01	0.01
Audio/control	Lightweight neural PLC	0.04	0.67	0.67	0.67	0.06	0.04	0.67	0.06	0.04	0.67	0.06	0.04
Audio/control	Advanced neural PLC	0.13	2.22	2.22	2.22	0.18	0.13	2.22	0.18	0.13	2.22	0.18	0.13
Gesture/avatar	Dead reckoning / trajectory PLC	0.01	0.25	0.25	0.25	0.02	0.01	0.25	0.02	0.01	0.25	0.02	0.01
Haptics	Haptic event PLC	0.01	—	—	—	0.01	0.01	—	0.01	0.01	0.11	0.01	0.01
EEG/biometrics	Physiological signal PLC	0.05	—	—	—	0.05	0.05	—	0.05	0.05	—	0.05	0.05
Cross-modal	Fusion A + G	0.27	5.00	5.00	5.00	0.37	0.27	5.00	0.37	0.27	5.00	0.37	0.27
Cross-modal	Fusion A + H	0.27	—	—	—	0.37	0.27	—	0.37	0.27	4.44	0.37	0.27
Cross-modal	Fusion A + E	1.48	—	—	—	2.02	1.48	—	2.02	1.48	—	2.02	1.48
Cross-modal	Fusion G + H	0.20	—	—	—	0.28	0.20	—	0.28	0.20	3.75	0.28	0.20
Cross-modal	Fusion G + E	1.49	—	—	—	2.04	1.49	—	2.04	1.49	—	2.04	1.49
Cross-modal	Fusion H + E	1.38	—	—	—	1.38	1.38	—	1.38	1.38	—	1.38	1.38
Cross-modal	Fusion A + G + H	0.27	—	—	—	0.37	0.27	—	0.37	0.27	5.00	0.37	0.27
Cross-modal	Fusion A + G + E	1.49	—	—	—	2.04	1.49	—	2.04	1.49	—	2.04	1.49
Cross-modal	Fusion A + H + E	1.48	—	—	—	2.02	1.48	—	2.02	1.48	—	2.02	1.48
Cross-modal	Fusion G + H + E	1.49	—	—	—	2.04	1.49	—	2.04	1.49	—	2.04	1.49
Cross-modal	Fusion A + G + H + E	1.49	—	—	—	2.04	1.49	—	2.04	1.49	—	2.04	1.49

Note: M-M = musician-musician; M-A = musician-audience; A-A = audience-audience. Fusion sets use A = audio/control, G = gesture/avatar, H = haptic, and E = EEG/biometrics. Scenario 6 uniquely assumes a bidirectional performer haptic channel as part of its fully integrated configuration; Scenarios 2–5 do not include haptics in their musician-musician paths.

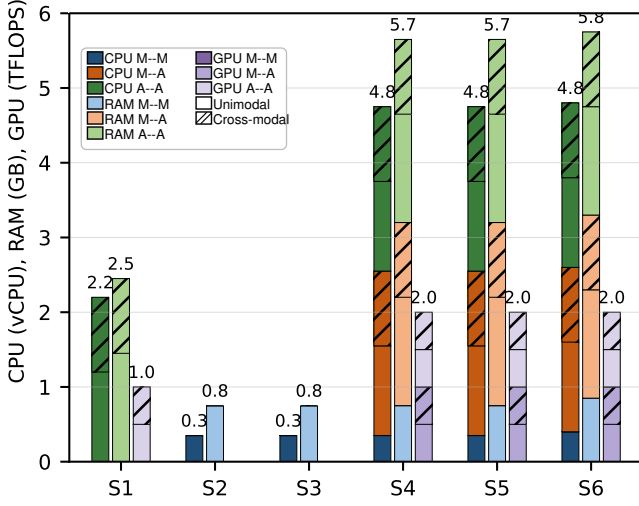


Fig. 1. Scenario-level resource requirements obtained by selecting, for each interaction type, the most advanced feasible PLC strategy satisfying the latency constraints. For each scenario, the three stacked bars report the aggregate CPU (left), RAM (center), and GPU (right) requirements. Colors identify the interaction type, while darker, medium, and lighter shades respectively denote CPU, GPU, and RAM resources. Hatched segments represent the additional computational overhead introduced by cross-modal PLC.

because their stringent musician-musician latency constraints restrict the selection to lightweight concealment techniques, preventing the deployment of cross-modal fusion. Scenarios 4 and 5 combine performer- and audience-oriented interactions, producing high resource requirements due to the coexistence of lightweight unimodal PLC for latency-critical streams and more sophisticated cross-modal processing for delay-tolerant interactions. Finally, Scenario 6 exhibits the largest overall resource demand because all interaction types coexist and multiple cross-modal fusion functions become feasible for musician-audience and audience-audience communications, despite the strict latency constraints that still limit musician-musician interactions.

Notably, the analytical design-space exploration presented here should be interpreted as a conceptual evaluation framework rather than an experimental benchmark. Although the representative latency and computational models are grounded in the PLC literature, actual implementations will inevitably exhibit hardware-, software-, and deployment-specific variations. Consequently, the results are intended to guide architectural decisions and identify feasible operating regions rather than predict the exact performance of individual PLC systems.

E. Future Research Directions

The proposed design space provides a conceptual and analytical foundation for cross-modal PLC in the MM, but several research challenges remain open. First, concrete cross-modal PLC algorithms should be developed and experimentally evaluated. While the present work identifies representative fusion strategies, practical implementations exploiting audio, gesture, haptic, and physiological signals for real-time reconstruction are still largely unexplored.

A second challenge concerns the evaluation methodology. Existing PLC benchmarks and objective quality metrics have been developed almost exclusively for unimodal audio applications and therefore fail to capture the perceptual consistency of multimodal experiences. New evaluation frameworks should jointly consider reconstruction quality, cross-modal synchronization, user immersion, and task performance, complementing objective metrics with subjective listening and user studies.

Another promising research direction is the development of adaptive PLC policies capable of dynamically selecting reconstruction strategies according to network conditions, interaction type, and user context. Such mechanisms could exploit real-time measurements of packet loss, latency, and available computational resources to balance reconstruction quality against latency and resource consumption.

Finally, the analytical models presented in this paper should be validated through a trace-driven simulation and a proof-of-

concept deployment on a representative MM platform, followed by controlled user evaluation. Integrating cross-modal PLC into distributed cloud-edge architectures and testing it under realistic, bursty network impairments will provide evidence about the practical trade-offs among latency, computational cost, scalability, privacy, and perceptual quality.

V. CONCLUSION

This paper investigated PLC for the MM and proposed a scenario-driven design space for cross-modal PLC strategies spanning six representative MM scenarios. The analysis shows that PLC in the MM should be regarded as a scenario-dependent systems problem rather than a purely signal reconstruction task, with different interaction contexts requiring different reconstruction strategies according to their modality composition, latency constraints, and perceptual priorities. The proposed design space is complemented by an analytical model relating latency constraints, computational requirements, and PLC strategies.

Overall, the proposed design space and quantitative analysis provide practical guidelines for selecting PLC strategies according to the interaction type, protected modalities, and available latency budget, supporting the development of adaptive, scenario-aware PLC architectures for future MM systems.

REFERENCES

- [1] L. Turchet, "Musical Metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, pp. 1–17, 2023.
- [2] C. Rinaldi and C. Centofanti, "The musical metaverse: Advancements and applications in networked immersive audio," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–7.
- [3] R. Hupke, D. Jan, N. Werner, and J. Peissig, "Latency and quality-of-experience analysis of a networked music performance framework for realistic interaction," in *Audio Engineering Society Convention 152*. Audio Engineering Society, 2022, pp. 1–10.
- [4] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE network*, vol. 12, no. 5, pp. 40–48, 1998.
- [5] A. I. Mezza, A. Bernardini, and C. Rinaldi, "The IEEE-IS 2 2025 Music Packet Loss Concealment Challenge," in *2025 IEEE 6th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2025, pp. 1–10.
- [6] A. Çamcı and R. Hamilton, "Audio-first VR: New perspectives on musical experiences in virtual environments," *Journal of New Music Research*, vol. 49, no. 1, pp. 1–7, 2020.
- [7] A. Boem, V. Casado, L. Bucciarelli, and L. Turchet, "Plug the guitar in before you put the headset on: Co-designing a mixed reality collaborative application with professional musicians," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026, pp. 1–20.
- [8] H. Wang, H. Dong, and A. El Saddik, "Immersive multimedia communication: state-of-the-art on extended reality streaming," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 7, pp. 1–33, 2025.
- [9] Baiyun Liu and Qi Song and Mingxue Yang and Wuwen Yuan and Tianbao Wang, "PLCNet: Real-time Packet Loss Concealment with Semi-supervised Generative Adversarial Network," in *Interspeech 2022*, 2022, pp. 575–579.
- [10] N. L. Westhausen and B. T. Meyer, "tplcnnet: Real-time deep packet loss concealment in the time domain using a short temporal context," in *Proc. Interspeech*, 2022, pp. 2903–2907.
- [11] Z. Zhang, X. Xia, C. Huang, Y. Xiao, and L. Xie, "BS-PLCNet 2: Two-stage Band-split Packet Loss Concealment Network with Intra-model Knowledge Distillation," in *Interspeech 2024*, 2024, pp. 1750–1754.
- [12] L. Diener, S. Branets, A. Saabas, and R. Cutler, "The icassp 2024 audio deep packet loss concealment grand challenge," *IEEE Open Journal of Signal Processing*, vol. 6, pp. 231–237, 2025.
- [13] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.
- [14] M. Fink and U. Zölzer, "Low-delay error concealment with low computational overhead for audio over ip applications," in *Proceedings of the International Conference on Digital Audio Effects*, 2014, pp. 309–316.
- [15] A. I. Mezza, M. Amerena, A. Bernardini, and A. Sarti, "Hybrid packet loss concealment for real-time networked music applications," *IEEE Open Journal of Signal Processing*, 2023.
- [16] F. Daniotti, L. Vignati, and L. Turchet, "Tilt loss: a perceptual loss function to improve music packet loss concealment," *Journal on Audio, Speech, and Music Processing*, 2026.
- [17] A. F. Khalifeh, A.-K. Al-Tamimi, and K. A. Darabkh, "Perceptual evaluation of audio quality under lossy networks," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2017, pp. 939–943.
- [18] L. Vignati and L. Turchet, "On the lack of a perceptually-motivated evaluation metric for packet loss concealment in networked music performances," *Journal of the Audio Engineering Society*, 2025.
- [19] X. Gu, Y. Yuan, J. Yang, and L. Li, "AI-empowered pose reconstruction for real-time synthesis of remote metaverse avatars," in *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 2024, pp. 86–93.
- [20] N. La Porta, L. Minardi, and M. Papandrea, "Fine-grained human activity recognition through dead-reckoning and temporal," in *Sensor-Based Activity Recognition and Artificial Intelligence: 9th International Workshop, iWOAR 2024, Potsdam, Germany, September 26–27, 2024, Proceedings*. Springer Nature, 2025, p. 3.
- [21] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 4, pp. 1–11, 2016.
- [22] G. Averta, M. Iuculano, P. Salaris, and M. Bianchi, "Optimal reconstruction of human motion from scarce multimodal data," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 5, pp. 833–842, 2022.
- [23] X. Chen, X. Jiang, L. Zhan, S. Guo, Q. Ruan, G. Luo, M. Liao, and Y. Qin, "Full-body human motion reconstruction with sparse joint tracking using flexible sensors," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 2, pp. 1–19, 2023.
- [24] J. L. Ponton, H. Yun, A. Aristidou, C. Andujar, and N. Pelechano, "Sparseposer: Real-time full-body motion reconstruction from sparse data," *ACM Transactions on Graphics*, vol. 43, no. 1, pp. 1–14, 2023.
- [25] X. Wei, Y. Yao, H. Wang, and L. Zhou, "Perception-aware cross-modal signal reconstruction: From audio-haptic to visual," *IEEE Transactions on Multimedia*, vol. 25, pp. 5527–5538, 2022.
- [26] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.
- [27] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow, "Style-controllable speech-driven gesture synthesis using normalising flows," in *Computer graphics forum*, vol. 39, no. 2. Wiley Online Library, 2020, pp. 487–496.
- [28] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman, "Audio to body dynamics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7574–7583.
- [29] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [30] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.
- [31] K. C. Pohlmann, *Principles of digital audio*. McGraw-Hill Professional, 2000.
- [32] E. Steinbach, S. Hirche, M. Ernst, F. Brandi, R. Chaudhari, J. Kammerl, and I. Vitorias, "Haptic communications," *Proceedings of the IEEE*, vol. 100, no. 4, pp. 937–956, 2012.
- [33] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.
- [34] L. Severi, M. Sacchetto, A. Bianco, C. Rottondi, A. Knapinska, and P. Lechowicz, "Demonstration of a networked music performance experience with mevo," *arXiv preprint arXiv:2404.09665*, 2024.