(Networked) Islands in the Stream: A Real-time Immersive Streamed Experience

Jacob Cooper*†, Jack Hardwick†, Anastasiia Voronina‡, Andrew Chadwick*, Katarzyna Sochaczewska*,
Laney Haywood*, Gavin Kearney*† and Helena Daffern*†
*AudioLab, University of York, UK
†CoSTAR Live Lab, University of York, UK
‡XR Stories, University of York, UK
Contact: jacob.cooper@york.ac.uk [0000-0001-6166-2479], jack.hardwick@york.ac.uk [0009-0009-0781-4406],
ana.voronina@york.ac.uk, andrew.chadwick@york.ac.uk, k.sochaczewska@york.ac.uk [0000-0001-5080-3360],
laney.haywood@york.ac.uk, gavin.kearney@york.ac.uk [0000-0002-0692-236X],
helena.daffern@york.ac.uk [0000-0001-5838-0120]

***Abstract*—This paper presents an applied framework for distributed Networked Music Performance (NMP) that is optimised for musicians and utilises commercially available software. We discuss a novel deployment to a live, remote audience in which a geographically distributed four-piece band performed in a shared virtual environment using audio streaming, motion-tracking and game engine technologies. This relies upon bespoke motion-tracking hardware to capture the movements of each performer and their instruments, rendering them as avatars and props. The framework utilised a high-quality digital twin of BBC's Maida Vale Studio 4, developed by the AudioLab, University of York, to serve as the virtual environment. Performers were physically distributed across approximately 3 miles in York, with the audience 30 miles away at CoSTAR Live Lab's performance venue in West Yorkshire. Audio was streamed to and from performer endpoints via the Dante protocol, with one of the York locations serving as the 'master' site, streaming audio objects and cinematic game engine visuals to the audience site. This 'master' site collated, synchronised and distributed motion and face-tracking data from all four sites, as well as synchronising the audio and visuals for the outbound stream. A public audience of 45 viewed the performance on an LED screen and experienced a spatial audio mix within a 28 channel surround sound L-Acoustics L-ISA system. Performers were also able to observe realtime game engine visuals from a fixed perspective. This paper provides a comprehensive description of the audiovisual workflows used and how they can be scaled for future immersive performances.**

***Index Terms*—immersive audio, streaming media, audio-visual systems, spatial audio, digital representation, metaverse, avatars, virtual environments, immersive experience, motion capture**

I. Introduction

Following the COVID-19 pandemic, there has been increased interest in, and development of, Networked Music Performance (NMP) systems. There is demand from high profile performers to find solutions to high cost, logistically challenging and climate damaging rehearsal processes from geographically separated performers. Challenges around the streaming of high definition video at latencies as low as the audio remain significant, with eXtended Reality (XR) solutions being explored that utilise avatar representations to reduce bandwidth requirements. Alongside a developing culture for NMP rehearsal, there is increasing appetite to explore NMP workflows for performances that are distributed to remote audiences. These experiences maintain the shared experience of the live concert, even with the performers physically at a different location, with Eurovision finals watch parties being a popular example. Online platforms such as Fortnite [1] are regularly used to deliver pre-recorded 'concerts' from artists such as Ariana Grande [2] and Metallica [3] through motion capture and animation workflows. Meanwhile, ABBA Voyage [4] is a current state-of-the-art example utilising motion capture workflows to render avatars to a live audience with spatial audio and a live band on stage performing to pre-recorded and rendered audiovisual content.

There is an opportunity to exploit the technological developments and theoretical understanding of NMP to provide new innovative experiences for performance and audiences, testing and pushing the boundaries of the musical metaverse. An XR NMP distributed live performance is presented in this paper, highlighting the importance of artists using this technology to engage with live audiences.

This work was supported by CoSTAR (Convergent Screen Technologies and Performance in Realtime) Live Lab, funded by the Arts and Humanities Research Council, grant reference AH/Y001079/1.

II. Background

NMP has developed with different solutions for different contexts, often depending on the needs of the user (e.g. numbers of musicians, locations, performance goals) and their networking solutions. The one-way latency at which performance becomes impaired is estimated to be 20 - 30 ms [5]. LOLA was one of the first software solutions for NMP, designed to run on a specialised network to provide high quality low-latency audio and video streaming, prioritising audio to support musicians [6]. Other solutions have focused on providing musicians a platform from which to rehearse from distributed positions at home. This application brings additional challenges of limitations of hardware and software and terrestrial internet. There is ongoing work looking to democratise NMP through improved network protocols [7], especially adopting solutions such as Starlink [8] and 5G [9]

to address these issues, although they have not yet provided a robust low-latency solution. The MUSMET project gives an overview of networking architectures tailored for a Musical Metaverse [10].

A. NMP with spatial audio

Since robust, low-latency protocols have been established, there has been a shift to consider the quality of audio being transmitted. Spatial audio is now being implemented as an important tool to contribute to musicians' experiences when playing together. This has been utilised within Jacktrip with gesturally controlled spatial location being explored [11], while Bonza Music [12] has placement of musicians in virtual spaces integrated into its GUI. When assessing different tools that integrate immersive binaural audio for 'online jamming', Müller et al. found that latencies of spatial audio plug-ins are not prohibitive and could be further optimised [13].

Whilst spatial audio rendering is possible for performers, the audio format to the audience must also be considered. The method of audio rendering is dependent on delivery format for the audience (stereo loudspeakers, headphones, loudspeaker array) and the creative aims. Channel-based, object-based and scene-based audio rendering all offer their own advantages depending on the context.

B. NMP in Virtual Reality (VR)

The visual element of NMP systems is a significant challenge, with video streaming increasing bandwidth demands and increasing latency. Reduced video quality may be a practical solution for 'at home' applications, however for performance NMP systems, avatar representation of musicians can provide a solution as well as creative opportunities. Avatar representations of performers and instruments can be achieved using game engine workflows and realtime tracking data from Head Mounted Displays (HMDs), optical motion tracking, markerless motion tracking and IMUs. In 2020, Loveridge highlighted the potential of VR as an alternative to teleconferencing in NMP workflows [14]. A VR NMP experience has since been realised in a digital model of BBC Maida Vale, allowing a four-piece band to record a track physically separated from each other, seeing each other's avatars and hearing themselves positioned in the digital twin [15]. Further work looked at the viability of extending this model to a server-based spatial audio rendering system to deliver to audiences at wireless VR endpoints, however, motion-sound latency needed optimisation [16]. A systematic review of VR music concerts identified a key feature of live music attendance being the interaction between the performer and audience, with this aspect currently under-explored in VR research [17].

C. NMP and hybrid experience in XR

It was reported in 2021 that most XR music experiences had focused on standalone rather than networked, and individual rather than social experiences [18]. Examples of hybrid musical performances utilizing XR that bring people together (either performers, audience or both) are increasing, including one notable implementation that maintained a shared audience experience that is remote from the live performers [19]. This utilised low-latency streaming protocols, multi-track audio streams and avatar representations of musicians, placing the audience in an XR environment and allowing them to interact with performers whilst hearing spatialised audio. Another study presented a distributed live a capella vocal consort to a live audience with four of the singers in person and the other four of the octet performing from a studio 50 miles away [20].

Researchers in New York have developed a framework for realtime streaming and synchronisation of different data-types through the Holodeck project, a bespoke infrastructure for networked performance [21] that provides an exhibition node whereby audiences can experience the performance through XR devices [22]. This brings together the different elements of NMP, maximising data efficiency through the use of motion capture and embedding immersive audio. This work notes the need to democratise these technologies and workflows to lower the barrier of entry to artists.

D. Shared experience in a musical metaverse

The importance of shared experience and live performance is becoming increasingly apparent, evidenced by increased ticket sales for live events and research identifying the benefits and effects of shared experiences [23], [24]. The sonic implications of a shared live sound experience have also been explored [25]. There is a growing focus on combining distributed performances with live audiences, seeking to maximise the potential and flexibility of the musical metaverse [26] that has been accelerated by rapid advancements in NMP frameworks.

III. System Specification & Design Overview

Whilst there are many examples of 'immersive concerts', very few utilise NMP workflows and are truly live, negating the need for accurate and robust visual workflows. Here a specification is defined to deliver a novel experience, with a system design that can achieve this for a real-world implementation.

A. Specification

The *Real-time Immersive Streamed Experiences for Motion Capture Performance* (RISE MoCaP) project set out to deliver an XR framework that satisfies three key criteria. Firstly, it should be optimised for live musicians; this means that motion tracking hardware must capture performers' movements with as much accuracy as possible, without affecting their ability to perform. Any performer-to-performer latency must also be such that performers can play together unaffected. Secondly, this system must geographically separate performers from each other and the audience, with both the audience and performers observing the performance from within a virtual environment. Finally, this should facilitate streaming multi-channel audio and visuals to an audience at a typical broadcast latency, with a simple return telecommunication allowing the performers to hear applause from the audience between songs.

This framework can be used to deliver a novel experience for both performers and audience members. Through testing

in a live scenario, areas for improvement can be identified. Further research can also be conducted to understand how audiences engage with this delivery format.

B. Implementation

This framework was implemented and tested in front of a live audience in a real-world scenario, with Yorkshire-based artist Marnie Glum [27] performing. This utilised four sites in York that were physically separated (4 rooms across 3 separate buildings, spanning a geographical distance of approximately 3 miles) and the performance venue at CoSTAR Live Lab approximately 30 miles away at Production Park, South Kirkby, West Yorkshire. Each of the York sites was used for a single performer in the band, as shown in Table I.

TABLE I
LOCATION AND INSTRUMENTS FOR EACH PERFORMER.

Site	Location	Instrument
1	The Guildhall	Lead Vocals, Acoustic Guitar
2	AudioLab Listening Room	Backing Vocals, Electric Guitar
3	AudioLab Studio	Drums
4	Ron Cooke Hub	Bass, Synthesiser

C. Design Overview

Fig. 1 shows a top level overview of the framework that was designed to facilitate this immersive performance. Low-latency audio between performers was delivered using the Dante protocol over a Campus Area Network (CAN) spanning the City of York. This is a layered network with external connectivity provided by its 10,000 Mb/s Janet connection [28]. Whilst low-latency means of connecting performers over a Wide Area Network (WAN) exist, a CAN was used to control latency. Dante was used for local monitor mixes and the stream to the audience. One of the York sites doubled as the 'master' site where the audio for the stream was ingested. Fig. 2 shows the signal flow for the 'master' system.

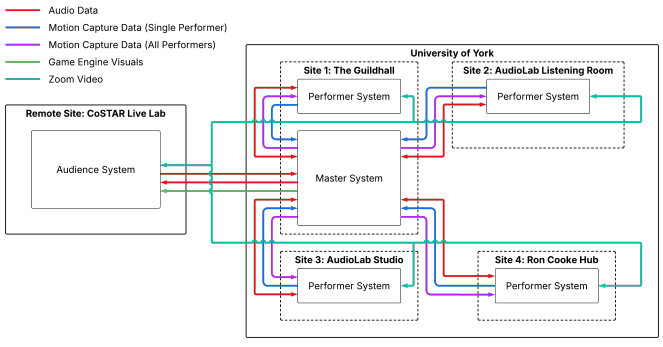


Fig. 1. Top level signal flow diagram for the framework showing how data passes from each performer system into the master system and to the audience.

A hub and spoke approach was used to send and receive motion and face-tracking data. Four motion capture PCs and face-tracking devices sites sent their data to the 'master' PC running Unreal Live Link Hub [29], where the data streams were collated, synchronised and streamed back to the

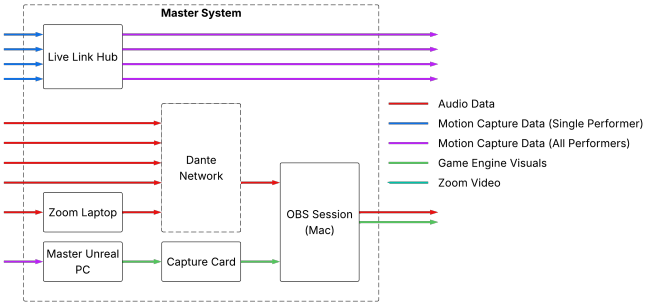


Fig. 2. Signal flow for the master system showing the combination of audio, motion tracking and video data for transmission.

PC rendering the game engine scene for each performer, as shown in Fig. 3. This data was also streamed to a machine rendering a game engine scene with cinematic visuals for the audience. This workflow is discussed in Section VI-F; whilst this implementation is for four performers, this could be scaled for more performers.

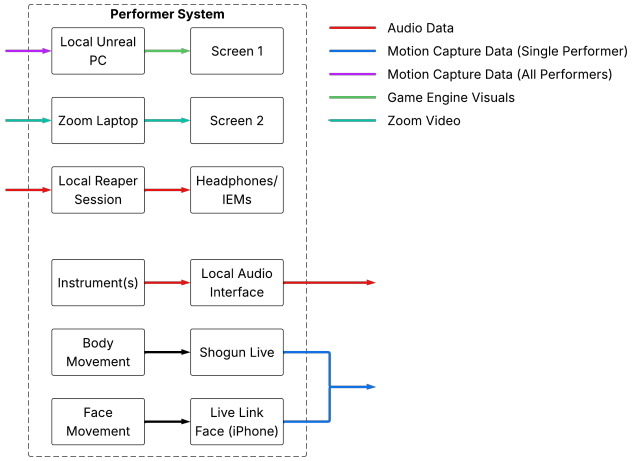


Fig. 3. Signal flow for each of the performer systems showing how audio, motion tracking and video data was used in their performance experience.

Audio sources and a rendered video feed were ingested into a single machine where they were combined and synchronised, before being sent via a single stream to the performance venue. The outbound stream was endpoint agnostic, streaming only audio sources and video data. This would allow the deployment to be scaled to multiple venues, with the only technical requirement being a means of audiovisual playback. Within the performance venue, the rendered video feed was displayed on screens, and audio sources mixed using a 28-channel L-Acoustics L-ISA spatial audio system to create an immersive spatial mix. The connection to the internet at the performance venue was provided through a 1Gb/s link. Video conferencing (Zoom) facilitated a live feed from the audience to each band member, as well as providing a 'behind the scenes' video feed from the band members that could be shown to the audience. It also allowed communication between

the technical team at the performance venue and the technical teams and performers in York.

IV. Streaming System Design

Audio and visual data were combined and synchronised within a single instance of Open Broadcaster Software (OBS) [30] running on a Mac and streamed via the Secure Reliable Transport (SRT) protocol to the performance venue.

A. Sources & synchronisation

Audio data was ingested into OBS via Dante Virtual Soundcard [31] at the master site alongside three visual streams: a cinematic game engine feed (feed one), the game engine perspective of the singer (feed two) and a static real-world camera (feed three). The first was to serve as the main feed, with the others used as a backup. Feed two meant that there was a backup game engine scene even if the PC running the main feed crashed, and feed three ensured that the audience could still observe visuals if the PC facilitating distribution of the motion-tracking data crashed. Video input was possible via the Network Device Interface (NDI) protocol and hardware capture cards. Audio and visual sources were synchronised using offsets within OBS.

OBS contains multichannel audio support up to 7.1; this allows recording and streaming of 8 discrete channels of audio. However, when streaming, channel 4 (traditionally the sub channel) is band-limited due to the compression format. This setup therefore provided only 7 full-range audio channels.

B. Streaming via SRT

The SRT protocol was used within OBS to facilitate the stream; this is a packaged form of the User Datagram Protocol (UDP). The end-to-end latency from the master site to the performance venue was less than 2.5 seconds, in line with that of typical broadcast; tested stability was favoured over optimising performer to audience latency. The stream ran at 30 fps with a bitrate of 7.36 Mbps. OBS does not report jitter, but during a trial performance, OBS reported 'frames missed due to rendering lag', 'skipped frames due to encoding lag' and 'dropped frames (network)' all as 0; these measures were not recorded during the actual performance.

V. Audio System Design

A robust audio framework was designed to facilitate the setup shown in Figs. 1-3. For the purposes of illustration, the Dante network has been shown as part of the 'master system', however in practice this sits outside of both systems, facilitating the audio routing from each site to the other.

A. Local and streaming systems

Each performer system consisted of a single audio interface and PC running a Reaper session, connected to the Dante network. A common channel routing, shown in Table II, was used across all four machines routing the audio from each site to the local PC. This Reaper session was used to give performers a stereo monitor mix.

TABLE II
Common channel list for audio system.

Dante Ch.	Site	Instrument	Capture	OBS Ch.
1	1	Lead Vox	Head-mounted mic	1
2	1	Ac. Gtr	DI	2
3	1	Zoom L	Line In	NONE
4	1	Zoom R	Line In	NONE
5	2	B Vox	Head-mounted mic	3 (Shared)
6	2	Elec. Gtr	DI	5
7	3	Drum L	Stereo mixdown	7
8	3	Drum R	Stereo mixdown	8
9	3	Kick	Desk send	NONE
10	3	Drum Talkback	Mic on stand	NONE
11	4	Synth	DI	3 (Shared)
12	4	Bass	DI	6
13	4	Bass Talkback	Mic on stand	NONE

This allowed for local control of each monitor mix whilst minimising latency across the network by using as few Dante devices in the signal chain as possible. Where possible, latencies of Dante devices were set to their lowest possible setting; latencies were unique to the source and receiver combination used. Whilst there was a limitation on the number of channels streamed to the performance venue, there was no meaningful limitation on number channels streamed over the CAN.

The 'master' site also facilitated the ingestion of audio from a laptop running Zoom; this allowed the musicians to hear an audio feed from the desk at the performance venue. The drums were pre-mixed locally and sent over the Dante network, with the Kick being sent separately as an auxiliary send to allow for improved performer monitor mixing. The setup also included talkback microphones between the non-singing performers.

The PC responsible for streaming the data used the same shared Dante routing; OBS was used to select the channels required for the stream, as shown in Table II. The restriction of 7 channels meant that channel 3 was shared between backing vocals and synth; during the musical performance, there was only one brief moment where both elements were used at once.

B. Performance venue

The received tracks were mixed on an Allen & Heath SQ6 console using its built-in processing and effects. Within L-ISA Controller [32], sound sources were distributed with varying spatial separation and width, corresponding to the conventional positions of the musicians on stage. The direct audio from each musician was positionally static during the performance, keeping the mix appropriate to a live concert setting. Acoustics were approximated during the breaks between songs, while during the performance itself, the aim was to create a creative and immersive mix. This was achieved through the placement of stereo audio effects including a doubler, reverberation and delay. These were distributed around and above the audience using automated movement such as 'entropy', generating a sense of envelopment, immersion and gentle movement of textures, while preserving the centre-focused character of the performance. Depth in the audio scene was established through parameters including simulated

distance (rolling off high frequencies as perceived distance increases), and early/cluster/late components of the reverberation processor in L-ISA Controller. Presets for audio effects and reverberation were changed between songs according to their respective tempo and densities.

Automations for opening and closing the sound scene were activated twice during the performance. These employed the common technique of widening the sound scene to create contrast between stasis and movement, and between narrow and wide, which aided the perception of change relative to a wholly stationary sound stage. Similarly, the creative processing was disabled between songs to establish contrast, and to match the acoustics of the virtual room which was a relatively acoustically dead studio space.

Microphones were used by the event hosts to introduce the performance; a downmix of these microphones and a stereo pair of room mics was used to provide the band with auditory feedback from the audience.

VI. Visual System Design

A visual system was designed to place the four performers in a digital twin of BBC Maida Vale Studio 4 within Unreal Engine [33]. This modelled performers as MetaHumans and instruments as props that were controlled in real time with motion-tracking and face-tracking data. Motion-tracking data was collated, synchronised and distributed from the master site; here it was used to render the cinematic game engine scene that was streamed to the audience.

A. Environment rendering

An existing model of Maida Vale Studio 4 was optimised to allow it to render more efficiently whilst preserving realistic visuals; this involved cleaning up the topology and reducing the number of materials used. The number of materials was reduced from 180 to 11 using trim sheets (grouping materials into categories) in Blender [34]. The texture sheet for the plastic material, for instance, had stripes of different colours that had varying roughness and metalness on the associated maps. Topology clean-up was focused on making the model watertight, reducing unnecessary vertices and removing non-planar faces. Thus, the polygon count was reduced from 32.2k to 1.7k. The final model is shown in Fig. 4.

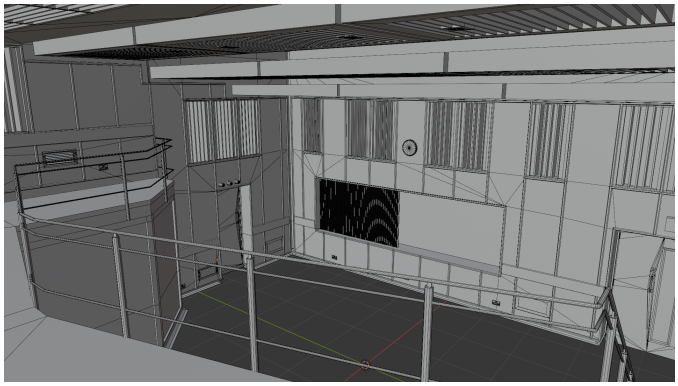


Fig. 4. The 3D model of Maida Vale Studio 4 was created using SketchUp and optimised using Blender.

B. MetaHumans

Lifelike avatars were created to resemble the performers' appearances.

1) Face Mesh Creation: Photogrammetric face scans of each band member were performed by recording a short video of each performer's face from all angles. Videos were recorded using the rear cameras of mobile phones due to their availability and minimal lens distortion. The videos were recorded at either 1080p or 4k resolutions and at 24 or 30fps.

Each video was imported into RealityScan 2.1 photogrammetry software [35] and processed following Epic Games' recommended workflow [36]. 4-5 equally-spaced frames of video were extracted per second of video, yielding between 297 and 398 images of each performer. Images with significant motion blur were manually removed, and the remaining images were masked using the automatic AI masking feature in RealityScan.

From each set of images, a point cloud and subsequent detailed face mesh was generated, manually trimmed, simplified, and smoothed. Finally, a texture was generated using the texturing feature of RealityScan. Each mesh was exported and imported into Unreal Engine as a Static Mesh Asset. Fig. 5 shows the process of creating the face mesh.

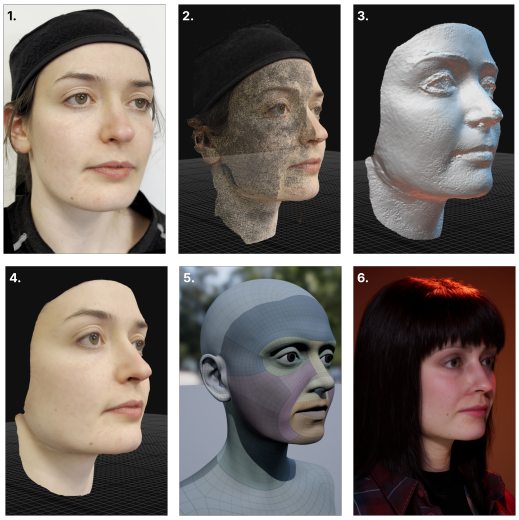


Fig. 5. The creation of the lead singer's MetaHuman. 1: Video, 2: Point cloud generation, 3: Mesh creation, 4: Texturing, 5: MetaHuman Identity creation, and 6: Final avatar.

2) MetaHuman Creation: A MetaHuman Identity asset was created for each performer using their face mesh, in line with the recommended workflow [37]. Only a single mesh per performer was possible using this approach, in a neutral facial pose. As a result, an additional Teeth Pose was not used when creating the MetaHuman Identity. Each avatar was then created in the MetaHuman Creator utilising free MetaHuman hairstyles, matching style and colouring as closely as possible.

Two MetaHuman asset packs [38], [39] containing various top, bottom, and shoe assets facilitated a mix-and-match approach for clothing. The avatars' skin meshes were masked to prevent them clipping through clothing and subtly rescaled to minimise visible clipping around cuffs and collars.

3) MetaHuman Retargeting: A full Range Of Motion (ROM) capture was recorded for each musician in order to build a bespoke retarget mapping each performer's Vicon skeleton to their MetaHuman's skeleton, which would be applied to their live motion capture data during the performance.

Vicon Shogun Post software was used to create the retargets [40]. An FBX file of each MetaHuman's Skeletal Mesh asset was imported; the Vicon and MetaHuman skeletons were then aligned using the built-in automatic MetaHuman retargeting script and recommended workflow [41]. This initial retarget was further refined to fix five recurring postural issues in the MetaHuman skeleton:

1) **Forward pelvis translation:** the MetaHuman skeleton's pelvis was translated forward, resulting in the legs leaning forward and the torso leaning back. The pelvis was translated rearward and the leg and foot bones were translated forward to mitigate this.
2) **Downward-sloping clavicles:** while the Vicon skeleton clavicles are generally sloped downwards from the neck to the shoulders, the MetaHuman skeleton expects more parallel shoulders. The MetaHuman clavicles were rotated upwards to raise the shoulders. The upper arm bones were then rotated to closely match hand position between the two skeletons, given the need for convincing hand-prop interaction.
3) **Forward head posture:** the neck bones of the MetaHuman skeleton were forward-tilting, resulting in MetaHumans having 'tech neck'. Neck bones were rotated and translated rearwards to ensure realistic head posture.
4) **Toe deformation:** the performers' shoe covers regularly unhooked causing the toes in the Vicon and MetaHuman skeletons to point upwards unnaturally. The toe bones in the MetaHuman skeleton were therefore offset downwards to maintain realistic foot positions.
5) **Forearm rotation:** all three degrees of freedom in the forearm bone in the MetaHuman skeleton were unlocked, in order to more closely align the MetaHuman and Vicon skeleton in the lower arms and wrists.

The retargets were then recalculated and exported for real time use within Shogun Live [42].

C. Instrument modelling

While pre-recorded human animations can be adjusted to account for arm space and bending of joints, live recordings do not have post-production flexibility. To match the musician's movements as precisely as possible, instruments were made adjustable and fit the shapes of those used by the performers.

Instrument models were sourced from online 3D asset sources [43]–[46] and customised to fit. In the case of the drum kit, this involved splitting the cymbal stands into separate parts using Maya [47]. This allowed independent control over position, height and cymbal rotation, accounting for the rotation axis by manually adjusting in Unreal. This process was repeated for each element of the drum kit.

D. Motion tracking

Each performer was located in a different optical motion capture volume, as shown in Fig. 6. While all volumes used Vicon Vero 2.2 cameras [48], the trackable area and camera count for each volume varied and is shown in Table III.

Fig. 6. Marnie Glum performing live in motion capture from 4 sites.

TABLE III
MOTION CAPTURE VOLUME SPECIFICATIONS.

Location	Dimensions	# of Cameras
The Guildhall	7m x 4m	18
AudioLab Listening Room	3m x 3m	10
AudioLab Studio	2.5m x 2m	18
Ron Cooke Hub	6m x 6m	16

1) Performer capture & calibration: Vicon's Production 3 Finger markerset [49], consisting of 63 markers per subject, was used for all musicians. Preliminary testing revealed that high numbers of finger markers make the performance of dextrously demanding instruments difficult; the 3 finger markerset provided the best compromise between capture fidelity and performability. Custom gloves were fabricated utilising off-the-shelf medical gloves to hold the finger markers in position. Fingertips were removed and incisions made to allow flexible markers to protrude through the glove.

2) Prop capture & calibration: The performers' instruments were tracked as props in Shogun. As these had no intrinsic connection to the performers' avatars in the virtual environment, accurate synchronisation between the tracked physical object and rendered model was essential to maintain believable interactions between avatars and their instruments.

Guitars and bass were tracked by rigidly mounting markers to the instruments using 3D printed G clamps. These clamps included an articulated felt pad on one side and a fixed rubber pad on the other to ensure they remained stable while not damaging the instrument. The clamps had a clip mount for a set of up to three markers to be attached, and three clamps were mounted to each instrument at points chosen so as not to interfere with performance. Figure 7 shows the G clamp assembly, and the mounting positions on a guitar.

At the prototyping stage, drumsticks and brushes were tracked by taping markers to them, with one marker attached to the bottom of the stick and several taped around the shoulder at various angles of rotation. The drummer didn't find this disruptive, hence it was also used for the performance.

For tracking cymbals, a "crown" assembly was mounted to the top of the bell, which remained aligned to the plane of the cymbal as it deflected without detrimentally affecting its sound.



Fig. 7. An exploded view of the G clamp design, and the placement of three clamps on an electric guitar.

A calibration process was developed to ensure alignment of the tracked guitars and cymbals with their virtual representations. The guitars were laid flat with the neck aligned with the Z (forward) axis, and the Unreal prop rotated to match this orientation. Some small post-hoc adjustments were made to ensure alignment with the avatars' fretting hands to account for position offsets imparted by the retargeting process.

The drums were not tracked as they move very little when played; the benefit of capturing this movement was outweighed by the added complexity of adding markers to the drums. The position of each virtual drum was calibrated by placing the calibrated sticks perpendicular to the centre of each drum head in turn, allowing their positions to be replicated in Unreal.

The quantity of reflective surfaces on the drum kit proved to be a complicating factor in tracking; reflections on chrome hardware and bronze cymbals can result in large numbers of "ghost" markers appearing and disappearing as the system attempts to resolve their positions.

This problem was addressed in two ways. Firstly, the amount of visible chrome in the area around the performer was minimised. The musical style called for covering the drums with cloths, and where chrome hoops needed to be visible these were covered with matt black tape. Secondly, during the masking phase of calibrating the volume, the cymbals were manually moved through their entire range of motion, and the resulting areas masked, excepting regions where the cymbal tracker, drum stick or performer might be. This significantly reduced the quantity of ghost markers without unnecessarily restricting the tracking volume around the kit.

E. Facial capture

When determining a suitable face tracking solution for capturing the musicians, a trade-off was identified between the fidelity of the capture and invasiveness for performers. Given the need to minimally obstruct the musicians' ability to perform, a stand-mounted rather than head-mounted face tracking solution was used. The Live Link Face app [50] for iOS devices captured the faces of three of the performers, while the drummer's face was not tracked. The fixed position of each device was chosen to maximise the extent to which the performer's face filled the frame, while also allowing some natural range of movement without losing tracking. Head-mounted vocal microphones were used for the performance as initial testing using bulky stand-mounted microphones revealed that microphone occlusion negatively affected tracking performance around the mouth. While using stand-mounted capture devices was non-invasive, tracking quality suffered when they were off-axis from the camera, which occasionally resulted in visibly poor tracking during the final performance.

F. Data collation & distribution

Preliminary testing in connecting multiple motion capture volumes highlighted the logistical challenge of managing data flow when numerous sites and systems are involved. In this case, each of the five Unreal Engine render nodes connected to seven Live Link sources; four Vicon optical systems using the Vicon Live Link plugin for Unreal Engine [51], and three Live Link Face app instances.

To mitigate this challenge, the Live Link Hub application [29] was used as the central node of a hub-and-spoke model for streaming motion tracking and face tracking data to the five Unreal render nodes. Fig. 8 shows the flow of tracking data from Live Link sources to Unreal render machines.

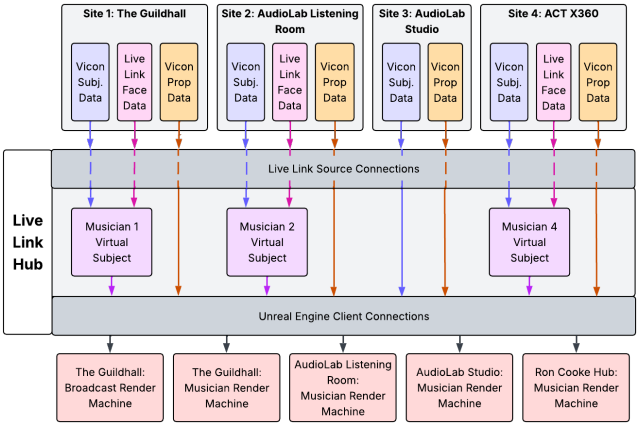


Fig. 8. Data flow from Vicon & Live Link Face sources to the Live Link Hub application, and onwards to Unreal render machine clients.

Virtual Subjects within Live Link Hub were used to combine face and body tracking data for each performer. As face-tracking data from the Live Link Face app incurred higher latency than the Vicon motion tracking systems, additional delay was added to the Vicon data in Live Link Hub in order to ensure that face and body were aligned for each musician. Each musician was asked to blink and clap simultaneously to provide a synchronisation cue in both data streams to use for alignment, which was then used to visually set the time offset of the Vicon tracking data to synchronise with the face tracking data. Instrument tracking data was delayed in line with subject tracking data. Vicon data from all sources was offset by 50 ms.

G. Visual feedback for performers

Each performer observed a locally rendered game engine scene from a fixed perspective on a screen in front of them. No synchronisation was done between audio and video for the performers to ensure the audio latency was as low as possible. Performers could also see a live video feed from the audience at the performance venue via the Zoom call.

H. Deployment to audience

Seven Cinematic Camera actors, shown in Fig. 9, were placed in the broadcast Unreal Engine scene, consisting of one view of each performer and three views showing all the performers; two low-angled views from the corners of the studio, and one elevated view from the virtual mezzanine. The frame rate of the Unreal project fluctuated due to GPU load.



Fig. 9. The seven camera angles shown during the performance. 1-4: Performer views, 5: Low view 1, 6: Low view 2, and 7: Mezzanine view.

Visuals were observed on a 4m x 2.5m LED wall and two 85-inch TV screens (both 30 fps) in the performance venue, shown in Fig. 10. Audio and visuals across all screens were synchronised. The Zoom call facilitated a split-screen view which allowed all four performers to be displayed on the TV screens only. This was used before and after the performance to allow the audience to see the real-life performers; it was not used at any point during the performance.

VII. Discussion

A five-song set was performed to a live audience of 45; performers were instructed to treat this like a normal gig where possible. The audience took part in an audience experience study which is the subject of a separate publication, however some provisional observations are noted where relevant to the system's performance. This study used a mixed-methods design to investigate audience experience; it combined subjective questionnaires, behavioural video data, and continuous physiological measurement to examine emotional, social, and overall audience experience within a live venue environment. In particular, it aimed to examine core themes surrounding audience emotion, absorption, social connection, presence/immersion, physiological arousal, and interpersonal synchrony. All audience members took part in the behavioural and subjective components of the study. A subset of 15 participants wore Polar H10 heart-rate sensors for the duration of the performance, enabling continuous physiological measurement.



Fig. 10. Performance shown on an LED wall and two TV screens within the performance venue.

The performance served as a solid proof-of-concept for this framework and technology; no major technical problems were observed. Tracking was generally stable throughout the performance, with the motion tracking hardware developed allowing for much more accurate instrument tracking than in previous iterations of this work. However, four recurring issues with subject tracking were identified. Whilst tracking accuracy numbers are not reported in Shogun, the software shows which markers were occluded at any given time.

Firstly, the MetaHumans' hands sometimes appeared warped. Review of the captured footage indicates this was due to the Vicon solver incorrectly interpreting the knuckle markers in the 3-Finger markerset as the index and pinky finger markers instead. In turn, the solver understood the finger markers as closer to the wrist than expected, therefore the hand bones were incorrectly solved to align the finger markers with the real world position of the knuckles.

Secondly, two of the performers' knees occasionally briefly hyperextended backwards during the performance. Locations of the knee markers were correct after the performance, therefore this issue was likely caused by poor camera coverage in those regions of the volumes.

Thirdly, tracking of one musician's leg was lost during and after kneeling down. When she returned to a standing position, her right leg in the Vicon skeleton was raised and pointed forwards. This was likely due to her guitar occluding her waist markers resulting in the leg markers being incorrectly labelled once standing again. She had to leave and re-enter the volume without her instrument to correct this. The performer was asked not to kneel during the performance.

Finally, tracking of the drum sticks was frequently lost for intervals of up to 1 second during the performance. This was likely due to the marker configuration and occlusion of the drum sticks by the drummer's hands.

Audience feedback highlighted hand and drum stick tracking as particular areas for improvement. Some audience members also noted that the realism of the avatars could be improved. This would have been aided by a more robust, head-mounted face-tracking solution, as well as tracking the face of the drummer.

Virtual camera angles were set during the performance soundcheck. However, once the performance started, the musicians moved significantly more, drifting close to the camera or out of frame altogether. This was a significant indicator that the musicians had a sense that they were performing to an audience, but did affect the visual output.

Considering the performer experience: some prop calibration had to be done within each instance of Unreal, meaning that the visual calibration in the streamed scene did not necessarily match that of the local scenes. The streamed scene was treated as the priority here.

VIII. Further Work

In this instance robustness and audio quality were prioritised; future work would look at alternative methods for streaming that would lower performer to audience latency. Synchronous streaming of audio and motion capture data (rather than video data) could also lower bandwidth requirements. However, in doing this the technical requirements of the endpoint would be increased; the current implementation requires no specialist hardware or expertise (beyond that of a standard music venue) at the endpoint, and as such allows this to be deployed to a range of spaces, including grassroots venues. Low-latency audio streaming solutions to stream from performer-to-performer over a WAN, would allow implementation across a truly distributed network.

The framework could also expand on the OBS streaming workflow to stream a greater number of audio objects to the venue. Whilst the audience were still able to enjoy a high-quality mix in this instance, other band line-ups, particularly with the growing trend to incorporate backing tracks in live performances, would require higher channel counts. There could also be a greater focus on performer audio. The current setup facilitated stereo monitor mixes, however the ideal implementation would use binaural, head-tracked monitor mixes with real-world recorded Impulse Responses (IRs).

Visual quality could have benefitted from improved motion tracking, especially more robust methods for tracking drum sticks and more optical cameras in the tracking volumes. By observing the occluded markers in the performance capture, optimal camera locations could be selected. More time in the calibration stage would also improve tracking results. The issue of accurate finger tracking for musicians still remains; there is yet to be a high fidelity mechanism of tracking musicians' fingers that works live and can be used with guitars. Additional face-tracking solutions should also be considered, including using low-profile, head-mounted cameras. It is important with all tracking hardware used that it does not impact the performers' abilities to perform.

Whilst each performer had their own visual feedback via an Unreal session, this could be improved through either deployment to a larger screen, or a head-mounted display (HMD); this would allow a greater sense of immersion in the performance and improved interaction between the band. Use of an HMD will, however, have an impact on the performers' ability to perform. An improved workflow for streaming visual data to performers' local Unreal sessions would involve sending calibrated data from a centralised machine, so any positional offsets would be applied across all Unreal sessions.

IX. Conclusion

This paper presented a framework that successfully delivered a distributed NMP between York and a remote venue. Streaming of audio objects allowed for an immersive mix to be delivered to the audience. Meanwhile, bespoke motion-tracking hardware allowed an unaffected performance whilst producing significant steps forward in the quality of visuals rendered for live musical performers. A test-bed has been presented for further development with applications in making live performance more accessible for audiences and allowing artists to reach audiences across the globe.

Acknowledgment

The authors would like to thank Joe Rees-Jones his hard work and support on the technical aspects of the project. The event itself was a huge undertaking which would not have been possible without the wider project team, comprising of 17 individuals across the University of York and the facilities of the School of Arts & Creative Technologies and XR Stories. We would also like to express our gratitude to Vicon Motion Systems Limited and L-Acoustics for their support in providing hardware to make this project possible.

The project was brought to life by a performance from Marnie Glum; sincere appreciation goes to Charlotte Lynch, Sol Edwards, Lola Simms and Emma Liu for their patience throughout and for a truly accomplished musical performance.

References

[1] Epic Games, "Fortnite," 2017, accessed: Jun. 29, 2026. [Online]. Available: https://www.epicgames.com/fortnite

[2] Epic Games, "Fortnite presents the rift tour featuring Ariana Grande," Virtual Live Event Performance, 2021, in-game interactive concert. Executed: Aug. 6–8, 2021. [Online]. Available: https://www.epicgames.com/fortnite

[3] Epic Games, "Metallica: Fuel. fire. fury." Virtual Live Event Performance, 2024, in-game interactive concert within Fortnite Festival framework. Executed: Jun. 22–23, 2024. [Online]. Available: https://www.epicgames.com/fortnite

[4] ABBA Voyage. (2026) Abba voyage official website. Accessed: Jun. 29, 2026. [Online]. Available: https://www.abbavoyage.com

[5] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

[6] B. Redman, "The potential of videoconferencing and low-latency (lola) technology for instrumental music teaching the potential of video-conferencing and low-latency (lola) technology for instrumental music teaching," *"Music & Practice"*, 07 2020.

[7] L. Borgianni, S. Giordano, and C. Chafe, "Democratizing networked music performance with a rl-based sd-wan," *IEEE Communications Magazine*, vol. 62, no. 12, pp. 26–32, 2024.

[8] L. Turchet and P. Casari, "The internet of musical things meets satellites: Evaluating starlink support for networked music performances in rural areas," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–8.

[9] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5g-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 4691–4709, 2024.

[10] L. Turchet, A. Boem, O. Ayoub, F. Malandrino, J. Llorca, C. Fischione, and C. Rottondi, "Towards a networking architecture for the musical metaverse: The musmet vision," 10 2025.

[11] R. Hoy and D. Van Nort, "A technological and methodological ecosystem for dynamic virtual acoustics in telematic performance contexts," in *Proceedings of the 16th International Audio Mostly Conference*, ser. AM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 169–174. [Online]. Available: https://doi.org/10.1145/3478384.3478425

[12] Bonza Music. (2026) Bonza music official website. Accessed: Jun. 29, 2026. [Online]. Available: https://bonzamusic.com

[13] P. Müller, R. Haefeli, J. Schütt, and M. Ziegler, "Telemersive audio systems for online jamming," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–15.

[14] B. Loveridge, "Networked music performance in virtual reality: current perspectives," *Journal of Network Music and Arts*, vol. 2, no. 1, p. 2, 2020.

[15] P. Cairns, A. Hunt, D. Johnston, J. Cooper, B. Lee, H. Daffern, and G. Kearney, "Evaluation of metaverse music performance with bbc maida vale recording studios," *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 313–325, Jun. 2023, this is an author-produced version of the published paper. Uploaded in accordance with the University's Research Publications and Open Access policy.

[16] P. Cairns, H. Daffern, and G. Kearney, "Investigation of server–based spatial audio for metaverse concert distribution," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–8.

[17] J. Park, Y. Choi, and K. M. Lee, "Research trends in virtual reality music concert technology: A systematic literature review," *IEEE transactions on visualization and computer graphics*, vol. PP, 03 2024.

[18] L. Turchet, R. Hamilton, and A. Çamci, "Music in extended realities," *IEEE Access*, vol. 9, pp. 15 810–15 832, 2021.

[19] N. Amarie, G. Fadda, M. Murroni, M. Alexandru, V. Popescu, D. Giusto, and A. L. Dumbravă, "Immersive application for real-time interactive music performances using spatial audio," in *2025 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2025, pp. 1–4.

[20] P. Cairns, T. Rudzki, J. Cooper, A. Hunt, K. Steele, G. Acosta Martínez, A. Chadwick, H. Daffern, and G. Kearney, "Singer and audience evaluations of a networked immersive audio concert," *Journal of the Audio Engineering Society*, vol. 72, pp. 467–478, 07 2024.

[21] A. F. Genovese, Z. Nguyen, M. Gospodarek, R. Pahle, C. Brenner, and A. Roginska, "Holodeck: A research framework for distributed multimedia concert performances," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–10.

[22] A. F. Genovese, M. Gospodarek, Z. Nguyen, R. Pahle, and A. Roginska, "Locally adapted immersive environments for distributed music performances in mixed reality," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–10.

[23] W. Tschacher, S. Greenwood, S. Ramakrishnan, T. Martin, M. Wald-Fuhrmann, C. Seibert, C. Weining, and D. Meier, "Audience synchronies in live concerts illustrate the embodiment of music experience," *Scientific Reports*, vol. 13, 10 2023.

[24] S. Karlsson-Brown and D. Knox, "Why go to pop concerts? the motivations behind live music attendance," *Musicae Scientiae*, vol. 21, 05 2016.

[25] C. J. Dahlie, J. Mulder, S. Pisfil, and N. Reeder, "Live sound matters," in *Researching Live Music: Methods and Practices*. London: Focal Press, 2021.

[26] L. Turchet, "Musical metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, vol. 27, 01 2023.

[27] Marnie Glum. (2026) Marnie glum official website. Accessed: Jun. 29, 2026. [Online]. Available: https://www.marnieglum.com/

[28] University of York. The campus network. IT Services, University of York. Accessed: 2026-09-14. [Online]. Available: https://www.york.ac.uk/it-services/contact/dco/its/network/

[29] Epic Games. Live link hub in unreal engine. Epic Games Developer. [Online]. Available: https://dev.epicgames.com/documentation/unreal-engine/live-link-hub-in-unreal-engine

[30] The OBS Project, "Obs studio," 2026, accessed: Jun. 29, 2026. [Online]. Available: https://obsproject.com

[31] Audinate, "Dante virtual soundcard," 2026, accessed: Jun. 29, 2026. [Online]. Available: https://www.getdante.com/products/software-essentials/dante-virtual-soundcard/

[32] L-Acoustics, "L-isa controller," 2026, accessed: Jun. 29, 2026. [Online]. Available: https://www.l-acoustics.com/software/

[33] Epic Games, "Unreal engine," 2024, accessed: Jun. 29, 2026. [Online]. Available: https://www.unrealengine.com

[34] The Blender Foundation, "Blender," 2024, accessed: Jun. 29, 2026. [Online]. Available: https://www.blender.org

[35] Epic Games. Realityscan — 3d models from images and laser scans. RealityScan. [Online]. Available: https://www.realityscan.com/

[36] Epic Games. Scan yourself for mesh to metahuman — tutorial. Epic Games Developer. [Online]. Available: https://dev.epicgames.com/community/learning/tutorials/Ovd2/realityscan-scan-yourself-for-mesh-to-metahuman

[37] Epic Games, "From mesh | metahuman documentation," https://dev.epicgames.com/documentation/metahuman/from-mesh, Jun. 2026.

[38] CGTailor, "Parametric casual clothing for metahuman," https://www.fab.com/listings/879807fb-632a-42ca-9cb3-811d2e62896b, Oct. 2024.

[39] CGTailor, "Streetwear outfit pack for metahuman," https://www.fab.com/listings/9d0807e7-83aa-4fb5-8035-7cf3d2038579, Jul. 2025.

[40] Vicon Motion Systems, "Shogun post," 2026, accessed: Jun. 30, 2026. [Online]. Available: https://www.vicon.com/software/shogun/

[41] "MetaHuman and Vicon real-time motion capture tutorial," accessed: Sep. 14, 2026. [Online]. Available: https://www.youtube.com/watch?v=7LplGd6nKG8

[42] Vicon. Vfx motion capture software — shogun by vicon. Vicon. [Online]. Available: https://www.vicon.com/software/shogun/

[43] Fab Repository, "Acoustic guitar 3d model," 2026, accessed: Jun. 29, 2026. [Online]. Available: https://www.fab.com/s/e45f7f5740cc

[44] Free3D Repository, "Gibson es-335 3d model," 2026, accessed: Jun. 29, 2026. [Online]. Available: https://www.free3d.com/3d-model/gibson-es-335-816888.html

[45] Sketchfab Repository, "Drum kit 3d model," 2026, accessed: Jun. 29, 2026. [Online]. Available: https://www.sketchfab.com/3d-models/drum-kit-99218427e7a2424ea8afc9b8c668a5c7

[46] Sketchfab Repository, "Bass 3d model," 2026, originally accessed via direct link on Sketchfab; URL no longer available as of Jun. 2026.

[47] Autodesk, "Maya," 2024, accessed: Jun. 29, 2026. [Online]. Available: https://www.autodesk.com/products/maya/

[48] Vicon. Vicon Vero — Advanced Super Wide Motion Capture Camera. Vicon. [Online]. Available: https://www.vicon.com/hardware/cameras/vero/

[49] Vicon Motion Systems, *Place Markers on a Performer*, shogun live 1.19 documentation ed., Vicon Help Center, 2026, accessed: Jun. 30, 2026. [Online]. Available: https://help.vicon.com/space/ShogunLive119/850134562/Place+markers+on+a+performer

[50] Unreal Engine. Live link face app - app store. App Store. [Online]. Available: https://apps.apple.com/gb/app/live-link-face/id1495370836

[51] Vicon, "Unreal Engine | Third Party Plugins | Vicon Motion Capture," accessed: Aug. 26, 2026. [Online]. Available: https://www.vicon.com/software/plugins-integrations/unreal-engine/