TRAIL'26 Workshop - TReC (Lausanne)

- **Full Name of the team leader (Researcher)**

Thierry Ravet

- **Contact Email**

thierry.ravet@umons.ac.be

- **Project Title**

LAIGO: Learning Assembly Instructions from Gestures and Objects

Human-Centered Spatial AI for Mixed Reality Assembly Guidance

- **Profile of the team leader(s) & Expected Team Composition**

The project is proposed by researchers involved in XR and Spatial Computing research activities related to the Wal4XR ecosystem, a collaborative initiative supporting research and innovation in Extended Reality technologies across Walloon academic and research centers. The team leader is a researcher at ISIA Lab and Le CLICK (UMONS), with experience in computer vision, motion analysis, XR systems, spatial computing, and Human-Computer Interaction.

The project combines expertise in Mixed Reality development, spatial computing, multimodal perception, motion analysis, and procedural XR interaction. Additional contributors from Human-Computer Interaction, ergonomics, pedagogy, cognitive sciences, or learning sciences would further strengthen the project by supporting the design and evaluation of procedural guidance strategies, usability, cognitive load management, and learning-oriented XR interaction.

The objective is to assemble a multidisciplinary team capable of addressing both the technical and human-centered dimensions of XR procedural assistance systems while maintaining a realistic prototype-oriented scope.

- **Have you already identified potential team members for your project?**

Yes

- **List the team members you have identified and briefly describe their profiles/roles (e.g., expertise, affiliation, expected contribution).**

The current core team already includes six contributors: one project coordinator, three AI researchers specializing respectively in motion analysis and AI systems, and two 3D/XR developers. The project still seeks an additional contributor with expertise in computer vision and object detection / 6DOF tracking.

Evelyne Meurisse, Thomas Baldassarre, and Alexandre Philippon are PhD candidates at ISIA Lab of the University of Mons (UMONS), whose research combines AI-based modelling, Human-Computer Interaction, and human-state estimation. Evelyne focuses on quantifying physical effort for agentic AI coaching in resistance training; Thomas works on adaptive virtual coaching agents in VR through real-time state estimation; and Alexandre works on deep learning for acoustic and psychoacoustic modelling, especially HRTF estimation and evaluation, and brings experience in applying scientific results in XR contexts.

Dorian Van Nieuwenhove and Bryan Bartoloni, both working at the MiiL (UCLouvain), have extensive experience in Unity development for XR applications and in building AR and MR proof-of-concept demonstrators. Within the TReC project, they would contribute to the development of the Mixed Reality tutorial prototype, including XR integration, 3D asset management, interaction design, AI component

integration, and the delivery of a stable and functional demonstrator. While Dorian would primarily focus on the implementation of XR features, interaction mechanisms, and Unity-based AI integration, Bryan would bring additional senior expertise in full-stack development, software architecture, and system integration to ensure the overall coherence and sustainability of the solution.

- **Domain of Application**

Industry 4.0

- **Scientific Theme**

Human-AI Interaction

- **Abstract**

This project explores how expert demonstrations of industrial assembly tasks can be transformed into interactive Mixed Reality (MR) guidance. Instead of manually writing step-by-step instructions, the system aims to record an expert performing an assembly procedure, analyse what objects are manipulated and how the gestures are performed, and convert this information into an Extended Reality (XR) tutorial that can later guide a learner.

The workshop will evaluate which perception and AI components are useful for this workflow. This includes 6DOF tracking of assembly parts, analysis of object-state transitions, markerless motion capture, hand and finger tracking, and temporal segmentation of procedural actions. These signals will be combined to identify assembly steps, detect key interaction events, and structure the expert demonstration into reusable procedural guidance.

The resulting prototype will be implemented as an MR assembly assistant. During the learning phase, the system will reuse the same perception pipeline to monitor the learner, validate completed steps, detect missing or incorrect actions, and display spatial guidance such as object highlights, directional cues, and ghosted assembly states. The project will therefore assess both the technical feasibility of real-time XR procedural assistance and the usability of this guidance for procedural learning.

- **Background information & Problem Statement**

Figure 1 — Existing modules available for the project. Left: ISIA Lab calibration and synchronized multi-video acquisition system. Right: MiiL AR/MR assembly guidance prototype using a physical scale model.

Industrial assembly knowledge is often transmitted through demonstrations, observation, and tacit spatial reasoning that are difficult to formalize in conventional documentation. Much of this expertise depends on gestures, sequencing, object positioning, and contextual adjustments that are only partially captured by written instructions or video recordings.[11]

XR is a natural medium for this problem because assembly instructions are spatial by nature: they must be understood in relation to real parts, object positions, user actions, and the physical workspace [12]. In industrial contexts, this relevance is reinforced by the frequent availability of 3D asset files, which can be reused to represent parts, assembly states, target positions, and spatial guidance directly in the user's

environment. The MiiL's experience with experimental software setups for manually encoding tutorials and presenting them in Mixed Reality provides an existing basis for exploring this approach. (Fig. 1, right)

The more this tutorial-creation process can be automated, and the more the underlying technology can remain transparent to the user, the more useful it becomes for industrial training and assistance. AI is therefore relevant because recent advances in Spatial AI, 6DOF object tracking, computer vision, motion analysis, and procedural action understanding can help identify manipulated objects, gestures, spatial relations, and procedural events from demonstrations. At ISIA Lab, the available calibration, synchronization, and multimodal capture infrastructure already makes it possible to test how these heterogeneous signals can be combined in practice for the implementation described below (Fig. 1, left).

- **Project Objectives & Concrete Implementation**

1. Project Objectives

The project aims to evaluate the transferability of recent Spatial AI, tracking, and motion analysis approaches to XR assembly scenarios while identifying the most relevant approaches for multimodal perception and real-time procedural guidance. This evaluation is intended to validate, through a proof-of-concept MR pipeline, the relevance of these AI algorithms for developing reusable XR pipelines for tutorial generation and procedural learning validation based on multimodal perception signals and procedural reconstruction logic. The project also investigates how XR procedural guidance can support task execution, spatial comprehension, and procedural learning in assembly contexts.

2. Concrete Implementation

The concrete implementation follows a two-phase MR workflow. In a recording phase, an expert performs an assembly task while the system captures object poses, hand and body motion, and relevant interaction events. These multimodal signals are then used to structure the demonstration into a procedural tutorial that can be represented spatially in Mixed Reality.

In a learning phase, the tutorial is reused to guide a learner through the same assembly task. The system compares the learner's actions and assembly states with the recorded procedure, enabling step validation and contextual MR feedback. The implementation is therefore divided into three complementary work packages: perception and benchmarking, tutorial generation during recording, and learner guidance during execution, with one optional exploratory work package for human-centered evaluation.

All work packages will rely on a common experimental setup combining a Meta Quest 3 headset for egocentric stereoscopic video capture and hand tracking, a ZED 2 RGB-D stereo camera for exocentric video capture and body pose detection, and a Unity 3D MR application. Sensor and interaction data will be streamed to a PC acting as a server, where the AI components will run and return perception or procedural information to the MR application.

The project is structured into work packages that correspond to the main stages of the proposed workflow.

WP1 — Multimodal Perception & Benchmarking

This work package first benchmarks AI-based perception, tracking, and motion-analysis algorithms on selected sequences from the retained datasets [1][2][5][6][7], using the evaluation metrics reported in the corresponding papers, while also measuring inference time to assess real-time compatibility. The most relevant approaches will then be tested on data recorded with the WP2 prototype to identify which signals are robust and exploitable for XR tutorial generation and procedural validation under XR constraints.

A first component focuses on object-level perception and 6DOF tracking of manipulated assembly parts. Candidate approaches will include recent object detection, segmentation, pose-estimation, and tracking methods such as FoundationPose [8], GigaPose [9], and SAM-6D [10]. The evaluation will use BOP-related benchmarks [6], especially the recent BOP-H3 datasets: HOT3D [7], HANDAL [5], or HOPEv2 [6]. The objective is to evaluate how robustly these approaches can identify parts, estimate their poses, and reconstruct trajectories, assembly states, and spatial transitions under XR constraints.

A second component focuses on markerless motion capture and gesture analysis. It combines external body pose estimation with hand and finger tracking computed by the HMD to reconstruct assembly gestures and hand-object interactions. For gesture recognition, the project will use Assembly101 [1] or AssemblyHands [2] pose annotations as methodological references and will evaluate representative models such as MS-G3D [3] and HandFormer [4], both already used for action recognition.

WP2 — MR Tutorial Generation Pipeline

WP2 implements the recording-phase pipeline, from data acquisition to the transformation of expert demonstrations into structured MR tutorial content. During recording, the system captures the expert's hand movements, head orientation, object poses, and spatial configuration of the assembly parts using the HMD, RGB-D capture, and the modules selected in WP1.

The outputs of the AI-based action recognition and perception systems are then processed by a rule-based procedure to infer key interaction events, such as grasping, positioning, joining, or fastening parts, and to segment the continuous demonstration into logical assembly steps. The pipeline generates the spatial and temporal metadata required to create the MR tutorial, including step timing, manipulated objects, target poses, validation conditions, and the information required to support later MR guidance.

WP3 — MR Learning Validation Pipeline

WP3 addresses the learning phase by first implementing the MR guidance elements used to support the learner during the assembly task. This includes the presentation of visual cues, object highlights, directional indications, and ghosted assembly states anchored in the user workspace, with particular attention to spatial readability, information density, visibility preservation, and cognitive load.

In a second step, the same perception module is reused to monitor the learner's actions in real time, validate assembly states and procedural transitions, and detect missing or incorrect actions. Attention is given to latency and robustness so that the validation remains responsive during procedural learning tasks.

WP4 — Optional Exploratory Human-Centered Evaluation

Depending on workshop progress, an exploratory human-centered evaluation may compare the XR guidance approach with more conventional supports such as video or paper-based instructions. This evaluation focuses on collecting first observations regarding procedural comprehension, usability, gesture reproduction, cognitive load, information overload, and user perception of XR procedural guidance in procedural learning situations.

- **Do you plan to deliver, as an outcome of your project, a reusable "brick" for the TRAIL Factory (https://factory.trail.ac/en/home_page) that could later be transferred and converted into a company process?**

Not sure

- **Briefly describe what the brick would be and its intended users.**

The reusable technological brick will consist of an AI evaluation and integration methodology for XR assembly workflows: a benchmarked and documented set of perception and action-understanding

components designed to transform multimodal recordings into exploitable procedural signals. It will compare candidate approaches for object detection and segmentation, 6DOF tracking, body motion reconstruction, hand-object interaction analysis, finger tracking, and temporal action segmentation, with the aim of identifying which combinations can robustly support the structuring of assembly procedures.

The proof of concept will demonstrate how these AI outputs can be connected to existing XR authoring and guidance pipelines. The contribution is therefore not the MR interface alone, but the link between perception outputs and tutorial-generation logic: identifying manipulated objects, detecting relevant action events, segmenting demonstrations into steps, generating procedural metadata, and reusing these outputs for later learner guidance or validation.

Its transferability should be understood at this methodological and integration level rather than as a directly deployable industrial product: such deployment would still require embedding or edge-optimizing the AI components and adapting the pipeline from video-passthrough MR headsets to optical see-through MR glasses. The project will therefore document model performance, inference time, and hardware dependencies as part of the transferability analysis.

- **Project Dataset**

The project will use Assembly101 [1] and AssemblyHands [2] as references for procedural activity understanding and hand-pose-based gesture recognition. Assembly101 provides multi-view assembly videos with fine- and coarse-grained action annotations, while AssemblyHands provides 3D hand-pose annotations derived from Assembly101.

For object-level perception and 6DOF tracking, the project will rely on BOP-related benchmarks [6], with particular attention to BOP-H3 datasets: HOT3D [7] for egocentric hand-object interaction, HANDAL [5] for manipulable objects with pose annotations and reconstructions, and HOPEv2 [6] for object pose estimation with onboarding data.

- **Detailed work plan**

Before the workshop, the team will prepare a predefined baseline assembly scenario using physical scale-model kits composed of multiple parts to assemble and disassemble (Fig. 1, right), together with their associated 3D assets. Prerecorded demonstrations and manually encoded procedural baselines will also be provided as fallback resources in case of hardware, calibration, or tracking issues. This preparation will support the benchmarking activities, reduce dependency risks, and allow the different work packages to progress in parallel. This is part of the WAL4XR project initial objectives and will be done regardless of the outcome of our project's submission to the TReC.

Day 1

Finalization of the predefined assembly scenario, including the physical setup, associated 3D assets, baseline procedural description, and fallback recordings. Definition of the learning and recording pipeline structures.

Day 2–Day 5

WP1 — Multimodal Perception & Benchmarking

Initial setup and first benchmark of the perception components, starting from existing software modules:

- Adaptation of existing modules: a calibrated multi-video-stream acquisition module and an analysis server.
- Selection and preparation of benchmark sequences from the retained datasets.
- Implementation of the candidate models retained for object detection, segmentation, 6DOF tracking, gesture recognition, and action segmentation.
- First benchmark on controlled recordings, including:
  - accuracy and robustness to occlusions for 6DOF tracking;
  - confusion matrix analysis for gesture recognition;

    - inference time and compatibility with real-time MR constraints.
- Preliminary fusion of perception outputs and identification of the most exploitable signals for tutorial generation and learner validation.

WP2 — MR Tutorial Generation Pipeline

Initial implementation of the recording and tutorial-generation pipeline:

- Implementation of recording-control XR interfaces, including record, start, stop, and delete functions.
- Implementation of the data-recording module.
- First version of a rule-based segmentation procedure based on detecting when parts are grasped and placed on the assembly frame. This step is considered low-risk, as it relies on explicit object manipulation events in the predefined scenario.

WP3 — MR Learning Validation Pipeline

Initial implementation of the learning and validation pipeline:

- Implementation of a real-time validation procedure for gestures and assembly actions.
- Preparation of visual assistance assets for assembly understanding, including elements for representing action cues, target poses, motion trajectories, and virtual replicas of the parts to be assembled.

Day 6–Day 9

WP1 — Multimodal Perception & Benchmarking

Second benchmark and integration-oriented evaluation:

- Evaluation of the selected algorithms in realistic assembly conditions using the predefined scenario.
- Comparison of the most relevant combinations of perception outputs for tutorial generation and real-time learner validation.
- Selection of the perception outputs to be integrated into the two MR pipelines.

WP2 — MR Tutorial Generation Pipeline

Implementation and evaluation of the tutorial-generation process:

- Implementation of the tutorial-creation procedure with server access.
- Refinement of the rule-based segmentation procedure using the outputs of gesture analysis and object tracking, with support from WP1 contributors after completion of the benchmark.
- Evaluation of the recording process and of the generated tutorial structure.

WP3 — MR Learning Validation Pipeline

Implementation and evaluation of the learner-validation process:

- Implementation of the action-validation procedure with server access.
- Testing and improvement of the real-time validation procedure, with support from WP1 contributors after completion of the benchmark.
- Evaluation of the learning process and of the responsiveness of the MR guidance.

WP4 — Optional Exploratory Human-Centered Evaluation

If time permits, the prototype will be used to collect preliminary observations on user experience:

- First observations regarding procedural comprehension and gesture reproduction during MR-guided tasks.
- Observation of usability, cognitive load, and user perception of MR procedural guidance.
- Preliminary comparison between MR guidance and more conventional instructional supports.

Day 9–Day 10: Finalization

Preparation of the final presentation and demonstration.

- **Bibliographic references**

[1] Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., & Yao, A. "Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[2] Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., & Keskin, C. "AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[3] Liu, Z., Zhang, H., Chen, Z., Wang, Z., & Ouyang, W. "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[4] Shamil, M. S., Chatterjee, D., Sener, F., Ma, S., & Yao, A. "On the Utility of 3D Hand Poses for Action Recognition." European Conference on Computer Vision (ECCV), 2024.

[5] Guo, A., Wen, B., Yuan, J., Tremblay, J., Tyree, S., Smith, J., & Birchfield, S. "HANDAL: A Dataset of Real-World Manipulable Object Categories with Pose Annotations, Affordances, and Reconstructions." IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023.

[6] Nguyen, V. N., Hodan, T., & BOP Challenge organizers. "BOP Challenge 2024 on Model-Based and Model-Free 6D Object Pose Estimation." CVPR Workshops, 2025.

[7] Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., Engel, J. J., & Hodan, T. "HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

[8] Wen, B., Yang, W., Kautz, J., & Birchfield, S. "FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[9] Nguyen, V. N., Hu, Y., Xiao, Y., Salzmann, M., & Lepetit, V. "GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[10] Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., & Li, Y. "SAM-6D: Segment Anything Model Meets Zero-Shot 6D Object Pose Estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[11] Burggraef, P., Adlon, T., Steinberg, F., et al. "Automatic Generation of Assembly Instructions by Analyzing Process Recordings: A Concept Overview." Procedia CIRP, 2024.

[12] Kyaw, A. H., Ma, H., Zivkovic, S., and Sabin, J. 2026. Augmented Assembly: Object Recognition and Hand Tracking for Adaptive Assembly Instructions in Augmented Reality. In Proceedings of the Twentieth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '26). Association for Computing Machinery, New York, NY, USA, Article 94, 1–8.

- **Does the project include multidisciplinarity between STEM & SSH?**

No

- **We confirm that the Team Leader will be present for the full duration of TReC'26 if the project is selected (August 24th - September 4th, 2026, Lausanne, Switzerland)**

I/We agree and confirm