EUROXR 2026

Future Directions:

Language Augmentation *in XR*

Valencia

Arantza del Pozo
adelpozo@vicomtech.org

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE

graphicsvision.ai



vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Current
Technological Context

Exploiting Frontier LLMs for interaction in XR

Core Approach 1

In-Context Learning (ICL)

Tailored conversational assistance with zero parameter updates

Core Approach 2

Grounded RAG

Retrieval over domain knowledge bases for factual guidance

Core Approach 3

Agentic Architectures

Autonomous tool calling linking intent to environment states and data

LUMINOUS CORE APPLICATIONS

Personalized Interaction

Dynamic persona and context-aware dialogue.

Feedback Generation

Live performance feedback synthesized from gameplay telemetry.

Tutoring in VR Training

Mapping runtime state variables directly to pedagogical next steps.

Environment Actuation

Natural commands mapped to 3D representations (BIM / CAD).

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Limitations

Limitations

LIMITATION 1

Cloud Dependency

→ Future Direction 1

Cloud LLM routing compromises latency and privacy, while headsets lack enough compute and local on-premise serving still lags behind cloud speeds.

LIMITATION 2

Reliability Gap

→ Future Direction 2

Performance varies across domain applications; hallucinations and unhandled edge cases threaten simulation safety.

LIMITATION 3

Spatial-Semantic Ambiguity

→ Future Direction 3

There is a lack of standardized spatial-semantic schemas that LLMs can read and write; words often have zero meaning without spatial context, gaze tracking or a clear target resolution system ("this valve", "over there").

LIMITATION 4

Language Bias & Accessibility

→ Future Direction 4

LLM models penalize regional and low-resource European languages; speech and cognitive assumptions exclude impaired users.

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Future Directions

Future Direction 1: Compute Continuum

CURRENT BARRIERS

1. Cloud dependency breaks interaction

Cloud voice streaming excedes conversational latency tresholds, risks offline failure and raises compliance barriers in sensitive environments.

2. Headset hardware not yet ready

Strict thermal budgets and RAM contention make standalone 3B–8B LLM model execution unfeasible

3. On-premise reasoning latency lags behind Cloud Flash

Standard local servers lack hyperscaler silicon clusters. Open models with reasoning loops are beyond Cloud Flash.

Cloud LLMs break real-time dialogue and privacy; while local edge setups must still match Cloud Flash speeds

THE FUTURE VISION

The Hierarchical Compute Continuum

Confining headset compute to real-time gaze raycasts, hand tracking and wake-word detection, safeguarding rendering framerates.

Deploying compact models on local networks for direct function calling, bypassing verbose reasoning.

Reserving larger models and complex reasoning synthesis for non-blocking asynchronous streaming tasks.

Achieving real-time conversational presence while preserving technological sovereignty.

Future Direction 2: Trust & Safety

LUMINOUS PROJECT EVALUATIONS

Performance Gap

Ongoing evaluations across various application tasks demonstrate solid baseline performance, yet remain below 100% dependability.

The Core Question:

In immersive training or critical simulation environments, how can we guarantee that the remaining unhandled failure modes do not cause real-world damage?

1. Constrained action decoding

Enforcing strict grammars and finite state machine guardrails so that assistants cannot trigger unauthorized, out-of-context or physically hazardous virtual actions.

2. Self-confidence scoring

Empowering assistants to recognize uncertainty and explicitly ask the user for clarification instead of hallucinating ambiguous procedural steps.

3. Hallucination auditing

Integrating lightweight on-device verification layers that validate agent output against grounded spatial physics and verified digital twin states before execution.

Future Direction 3: Spatial-Semantic Grounding

CURRENT BARRIERS

Lack of Universal spatial-semantic standards

Current 3D engines and formats model geometry meshes or static architectures (BIM), lacking unified schemas for interactive object affordances and dynamic simulation states.

Unresolved multimodal deictic ambiguity

Users naturally say “Put this valve there”. Open-source models cannot yet bind continuous 3D eye-gaze cones and pointing vectors to speech timestamps reliably.

Core Barrier: Lack of standard scene semantics bridging 3D world physics, deictic gestures and language.

THE UNIFIED VISION

Enabling assistants to perceive, point to and manipulate any virtual entity unambiguously

Perceive state & affordances

Formal ontological schemas encoding dynamic topological hierarchies, state preconditions and functional physical affordances across any XR engine.

Fuse gaze, gesture & language

Fusing high-frequency eye-tracking gaze cones and hand pointing vectors directly to speech acoustic timestamps for sub-millisecond deictic resolution.

Manipulate scene graphs

Mapping resolved 3D entities directly into deterministic engine action APIs (Unity, Unreal, OpenXR), lifting open models to commercial capabilities.

Future Direction 4: Diversity & Inclusion

CURRENT BARRIERS

1. Low-resource language penalties

Multilingual models suffer token inflation and weaker reasoning in low-resource tongues, while translation cascades add multi-second latency and strip cultural nuance.

2. Acoustic fragility

Standard commercial ASR breaks down for speakers with speech impairments (dysarthria, aphasia) or strong non-standard regional dialects.

3. Accessibility exclusions

Overly dense verbal outputs overwhelm users with intellectual or learning disabilities, while Deaf users lack native 3D sign-language interfaces.

Core Barrier: Monolithic language and voice pipelines exclude non-dominant languages and diverse users.

THE UNIFIED VISION

Sovereign, diverse & accessible interfaces

Language to 3D action mapping

Native speech maps directly to virtual engine actions via lightweight language adaptation, avoiding English detour latency and ending token inflation.

Atypical speech & dialect adaptation

Acoustic models adapted for strong regional dialects, accents and pathological speech (dysarthria), supporting clinical XR and regional equity.

Easy-to-read simplification & sign-avatars

Dynamic real-time easy-to-read text simplification and two-way hand-skeleton sign language avatars ensuring zero user exclusion.

Guaranteeing native digital identity, cultural sovereignty and universal accessibility in XR.

vicomtech
MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

Conclusions

Conclusions

Four strategic pillars for Language Augmention in XR

1

Compute Continuum

Meeting conversational turn-taking thresholds shifting from cloud APIs to on-device sensor binding and proximate edge tool-calling SLMs, preserving EU data sovereignty.

2

Trust & Safety

Bridging the reliability gap in mission-critical XR through constrained decoding, uncertainty estimation and state auditing to eliminate unhandled failures.

3

Spatial-Semantic Grounding

Establishing open spatial-semantic standards and multimodal fusion to enable XR assistants to perceive object affordances, resolve deictic gaze and gestures, and manipulate dynamic scene graphs unambiguously.

4

Diversity & Inclusion

Guaranteeing cultural sovereignty and universal inclusion by mapping native speech directly to 3D actions, adapting to atypical speech and regional dialects and delivering real-time cognitive and sign-language interfaces.





vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Thank you!

www.vicomtech.org

Vicomtech is the owner of the information shown in this document. It is classified as restricted. This information shall not be used out of the scope of the document itself.