

Product Requirements Document

HumanityAI'd Museum AR Smart Glass Guide — v1.0

July 2026

CONFIDENTIAL — HUMANITYAI'D

1. Product overview

The Museum AR Smart Glass Guide turns a museum visit into a hands-free, personalized, multilingual tour. Visitors wear lightweight AR glasses (RayNeo X3 Pro); the glasses stream what the visitor is looking at to an **on-premises AI server**, which recognizes the exhibit in front of them and instantly plays a professionally narrated audio guide in the visitor's language, while contextual text and imagery can be shown on the glasses' display.

Everything — visual recognition (CLIP + FAISS), translation content, and neural text-to-speech (Chatterbox Multilingual) — runs **inside the museum's own network**. No visitor images or behavioural data leave the building: a decisive advantage for government and cultural-sector procurement in the region.

2. Problem statement

Today's pain	Consequence
Handheld audio guides require numbered keypads and constant map-checking	Visitors look at devices, not art
Guided group tours run on fixed schedules and 1-2 languages	Lost revenue from non-Arabic/English speakers
Mobile apps demand downloads, drain phones, and depend on visitor hardware	< 15% typical adoption
Analytics on visitor flow and engagement are coarse or absent	Curators cannot measure which exhibits work

3. Target users & personas

- **The Visitor** — international tourist; wants zero-friction, own-language narration that starts automatically when they look at an exhibit.
- **The Curator / Content Manager** — owns exhibit content; needs a web panel to add exhibits, photos, translations, and to review/approve AI-generated narration before it goes live.
- **The Operations Manager** — issues and collects glasses at the front desk; needs device/session management and daily analytics.
- **The Museum Director / Digital Transformation lead** — buys on visitor experience, data sovereignty, and measurable engagement.

4. Core user journeys

1. **Wear & go:** visitor receives glasses, staff starts a session with language preference → visitor walks; when they look at an exhibit for ~2 s, recognition fires and narration plays in their language. No buttons.
2. **Curated tours:** predefined routes (e.g., "Highlights in 60 minutes") with stop-by-stop guidance and tour progress tracking.
3. **Content lifecycle:** curator uploads 1–3 photos per exhibit → system builds visual index automatically → writes (or auto-generates) descriptions → generates TTS narration per language → reviews and approves each clip → live.
4. **Analytics review:** dwell time per exhibit, most/least visited pieces, audio play-through rates, per-session tour summaries.

5. Functional requirements

5.1 Visitor experience (glasses app)

- F-1 Hands-free exhibit recognition from live camera stream (WebRTC/H.264; target < 1.5 s from gaze to result)
- F-2 Automatic audio narration playback in session language; 8 launch languages: **en, ar, fr, es, zh, ru, hi, de** (Arabic incl. RTL text display)
- F-3 On-lens display of exhibit title and key facts; adjustable via web-based glasses settings panel
- F-4 Offline-tolerant behaviour: reconnect logic when Wi-Fi drops; session resumes

5.2 Recognition service

- F-5 CLIP ViT-L/14 image embeddings + FAISS similarity index; multiple photos per exhibit (angles) for robustness
- F-6 Recognition threshold configurable per deployment; returns exhibit + translation + approved audio URL in one response
- F-7 Index updates live on image upload — no restart or re-training required

5.3 Content management (admin panel + API)

- F-8 CRUD for museums → galleries → exhibits (categories, inventory numbers, artist, period, tags)
- F-9 Multi-image upload with primary/angle designation; automatic embedding
- F-10 Translation records per exhibit per language: title, short & full description, fun fact, audio script
- F-11 One-click TTS generation per exhibit across all languages (female/male voices); **approval workflow** — narration is served to visitors only after explicit review approval
- F-12 Tours: ordered stops, multilingual tour metadata

5.4 Sessions & analytics

- F-13 Session lifecycle (start/end, language, device), event logging (view, dwell, audio played)
- F-14 Analytics: overview dashboard, popular exhibits, per-session visit summary with dwell seconds
- F-15 Prometheus metrics + Grafana dashboards for ops (GPU, latency, request rates)

5.5 Platform & deployment

- F-16 Fully on-premises Docker deployment (single GPU server, NVIDIA runtime); all services containerized (API, PostgreSQL, Redis, nginx, Prometheus, Grafana)
- F-17 API-key security for all endpoints; admin login for panel; per-key rate limits
- F-18 Data persistence across restarts for images, FAISS index, audio, and DB (named volumes)

6. Non-functional requirements

Requirement	Target
Recognition latency (frame → result)	< 1.5 s p95
Recognition precision on enrolled exhibits	> 95% with 1–3 photos/exhibit

Requirement	Target
Narration quality	Neural TTS, museum-tuned loudness (−18 LUFS), Opus 80 kbps
Concurrent visitors per GPU server (8 GB)	25–40 active sessions
Availability	99.5% during opening hours; watchdog + auto-restart
Privacy	No cloud egress of visitor imagery; sessions pseudonymous
Languages	8 at launch; TTS engine supports 23 (expansion-ready)

7. Proven reference deployment (July 2026)

An end-to-end pilot runs today at the HumanityAI'd office ("HumanityAI'd Office" museum instance): **10 exhibits, 12 enrolled photos, 8 languages, 80 approved narrations**, generated and served by the exact stack above on a single RTX 4060 Ti (8 GB). Enrolment-to-live time for a new exhibit: **under 5 minutes** including narration generation.

8. Out of scope (v1)

- Visitor-owned-device (BYOD phone) mode — roadmap v1.5
- Indoor positioning / wayfinding overlays — roadmap v2
- Face recognition of visitors (deliberately excluded — privacy posture)

9. Success metrics

- $\geq 60\%$ of equipped visitors complete ≥ 5 exhibit narrations
- Average dwell time on narrated exhibits +30% vs. baseline
- Content team can enrol a new exhibit in < 10 minutes without vendor help
- NPS of glasses experience ≥ 50

10. Dependencies & risks

Risk	Mitigation
GPU memory contention (recognition + TTS on one 8 GB card)	TTS generation is an offline/admin operation; runtime serving is pre-generated audio. 16 GB GPU recommended for large catalogues
Glasses hardware supply (RayNeo)	App isolates vendor SDK; portable to other Android-based AR glasses
Long Chinese scripts increase TTS memory	Script-length guideline (≤ 60 s narration) enforced in authoring UI
Museum Wi-Fi quality	Site survey in POC phase; dedicated SSID recommended