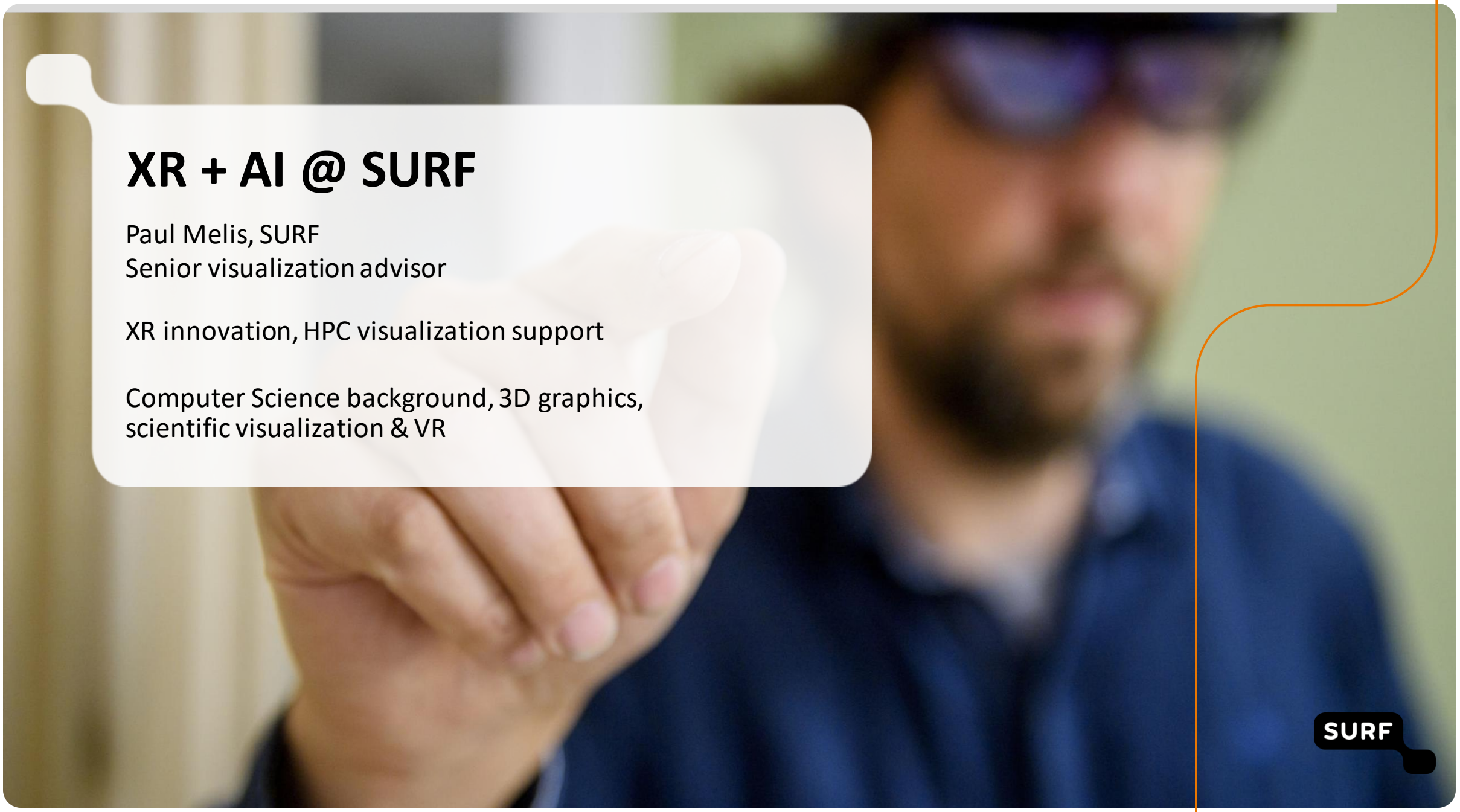


XR + AI @ SURF

Paul Melis, SURF
Senior visualization advisor

XR innovation, HPC visualization support

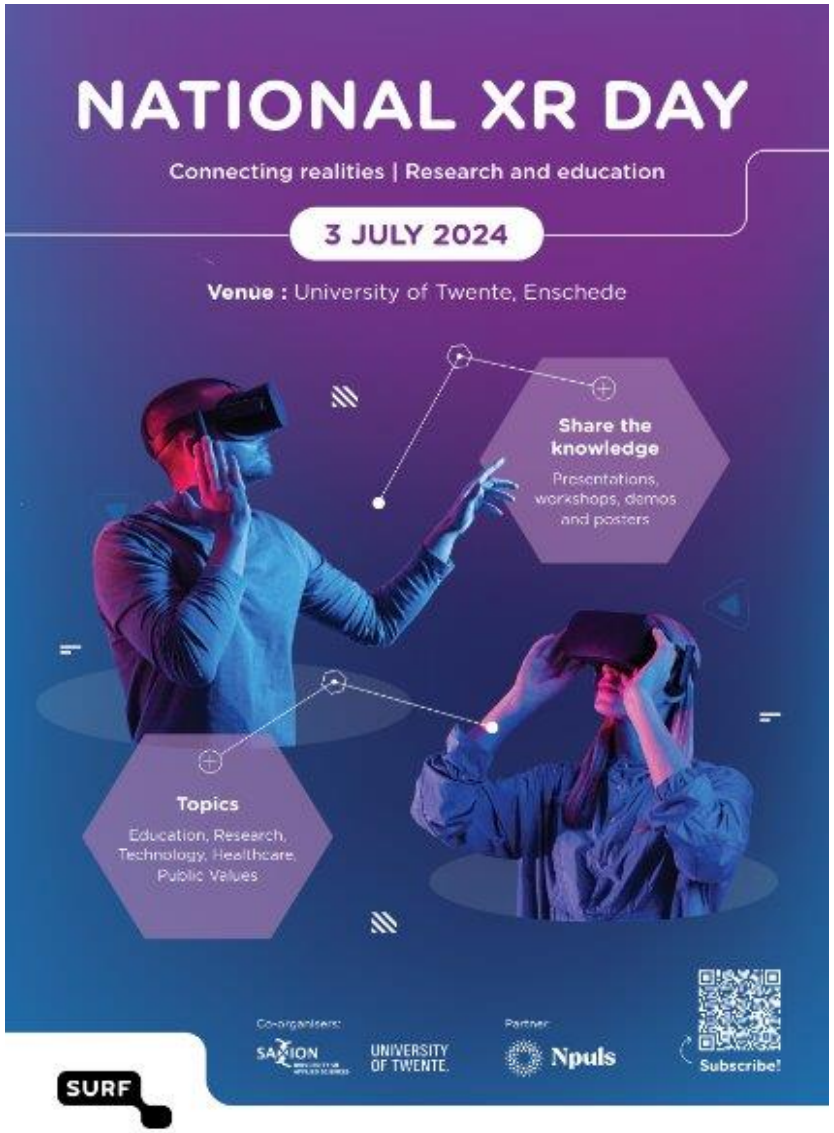
Computer Science background, 3D graphics,
scientific visualization & VR

About SURF

- SURF Cooperative
  - Collaborative organisation for ICT in Dutch education and research
  - 110+ Dutch member institutes
    (Universities, University medical centers, HBOs, MBOs, …, KNMI, RIVM, …)
- SURF B.V.
  - 650+ employees
  - Services, expertise and innovation in support of research and education in NL
  - Offices in Utrecht and Amsterdam (ASP)
- Central role, not educators/researchers ourselves, but facilitate and collaborate with
- Example infrastructure & services:
  - High-speed SURFnet network between institutes
  - High-Performance Computing & Visualization group (HPCV).
    Dutch National supercomputer Snellius + other HPC infrastructure.
    Support for HPC and visualisation
  - Cloud computing & Data storage
- Public values focus

SURF XR activities

- Futuring
  - Trend watching
  - Links to other developments, such as AI
  - Device assessment, market watch
- Infrastructure for XR
  - Experiments
- Procurement of XR hardware, software & services (initial process)
- Community
  - National XR Day (XR for edu & research in NL)
  - XR Developer Network
- Public values
  - Responsible XR



Let's ask Meta's Llama3 8B...

What are uses of AI for extended reality workflows?

AI plays a significant role in enhancing Extended Reality (XR) workflows by automating tasks, improving accuracy, and increasing efficiency. Here are some uses of AI in XR:

1. **Content Creation**: AI algorithms can generate 3D models, textures, and animations for VR and AR experiences, reducing the need for human artists.

2. **Virtual Character Animation**: AI-powered animation systems create realistic character movements, facial expressions, and voice responses, making interactions more natural and engaging.

3. **Object Recognition and Tracking**: AI-driven computer vision enables real-time object recognition, tracking, and interaction in AR and VR environments.

4. **Audio Processing**: AI algorithms enhance audio quality, remove noise, and provide real-time language translation for XR experiences.

5. **Scene Reconstruction**: AI-based reconstruction techniques create 3D models of physical spaces from 2D images or video footage, simplifying XR content creation.

6. **Physics-Based Simulation**: AI-powered physics engines simulate realistic interactions between virtual objects and the environment, enhancing immersion in VR and AR experiences.

7. **Predictive Analytics**: AI-driven analytics predict user behavior, preferences, and emotional responses to XR content, enabling personalized experiences and improving engagement.

8. **Content Optimization**: AI algorithms optimize XR content for various devices, platforms, and network conditions, ensuring seamless playback and reduced lag.

9. **User Interface Design**: AI-powered design tools generate intuitive UI elements, such as menus and buttons, for XR interfaces, streamlining user interaction.

10. **Testing and Quality Assurance**: AI-driven testing frameworks automate XR content testing, detecting issues and improving overall quality before deployment.

11. **Real-time Rendering**: AI-accelerated rendering techniques speed up the processing of complex graphics in XR environments, reducing latency and increasing frame rates.

12. **Audio Mixing and Mastering**: AI-powered audio mixing and mastering tools enhance the sound quality of XR content, ensuring a professional-grade audio experience.

13. **Virtual Try-On**: AI-driven virtual try-on technology allows users to digitally try on clothing, accessories, or makeup in AR environments.

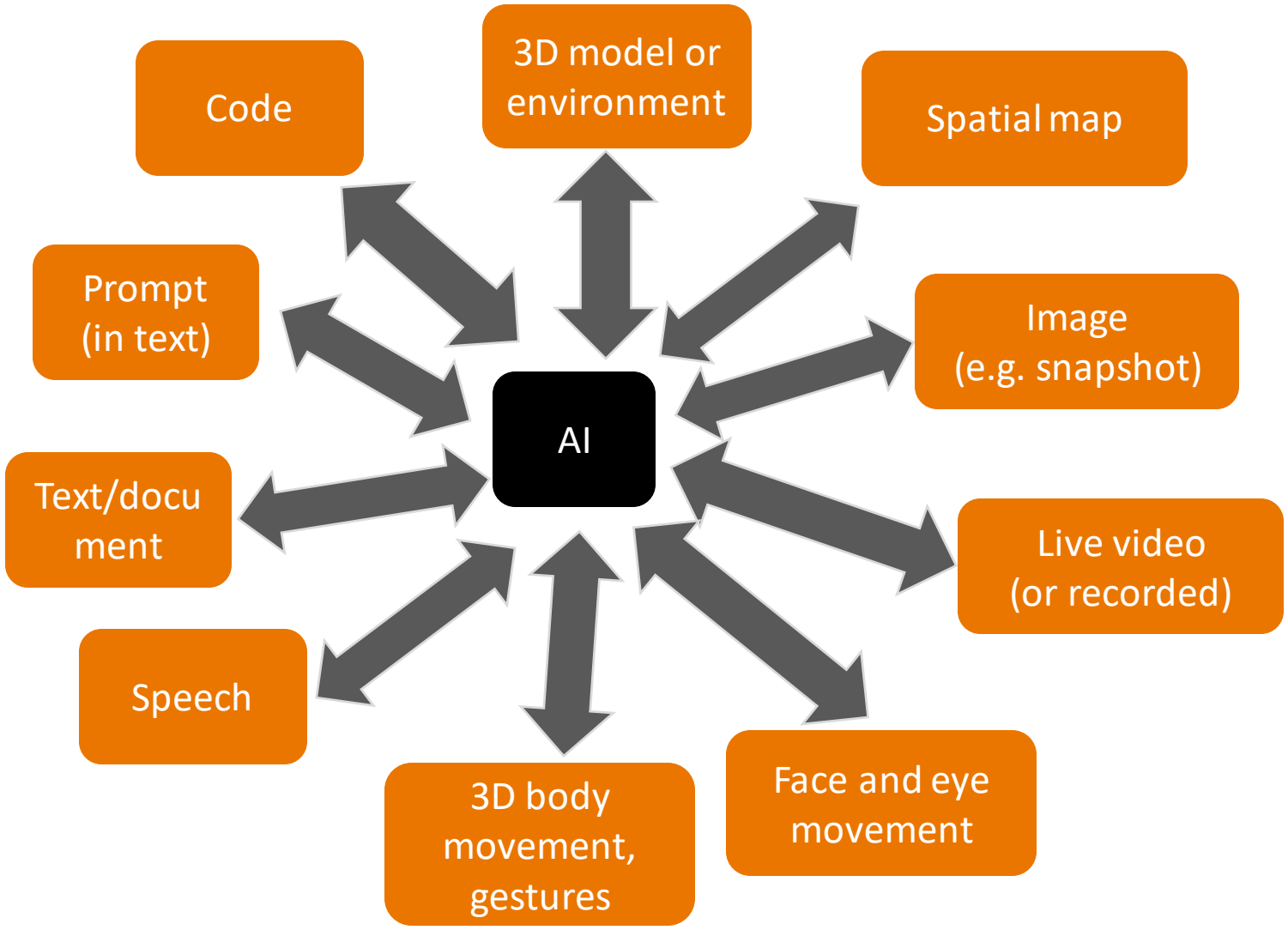
14. **Facial Recognition and Emotion Analysis**: AI-based facial recognition and emotion analysis detect user emotions, recognizing and responding to their emotional state in XR experiences.

15. **Personalized Recommendations**: AI-powered recommendation systems suggest personalized content, products, or services based on users' preferences and behaviors in XR environments.

By integrating AI into XR workflows, developers can create more immersive, interactive, and engaging experiences that better connect with users.

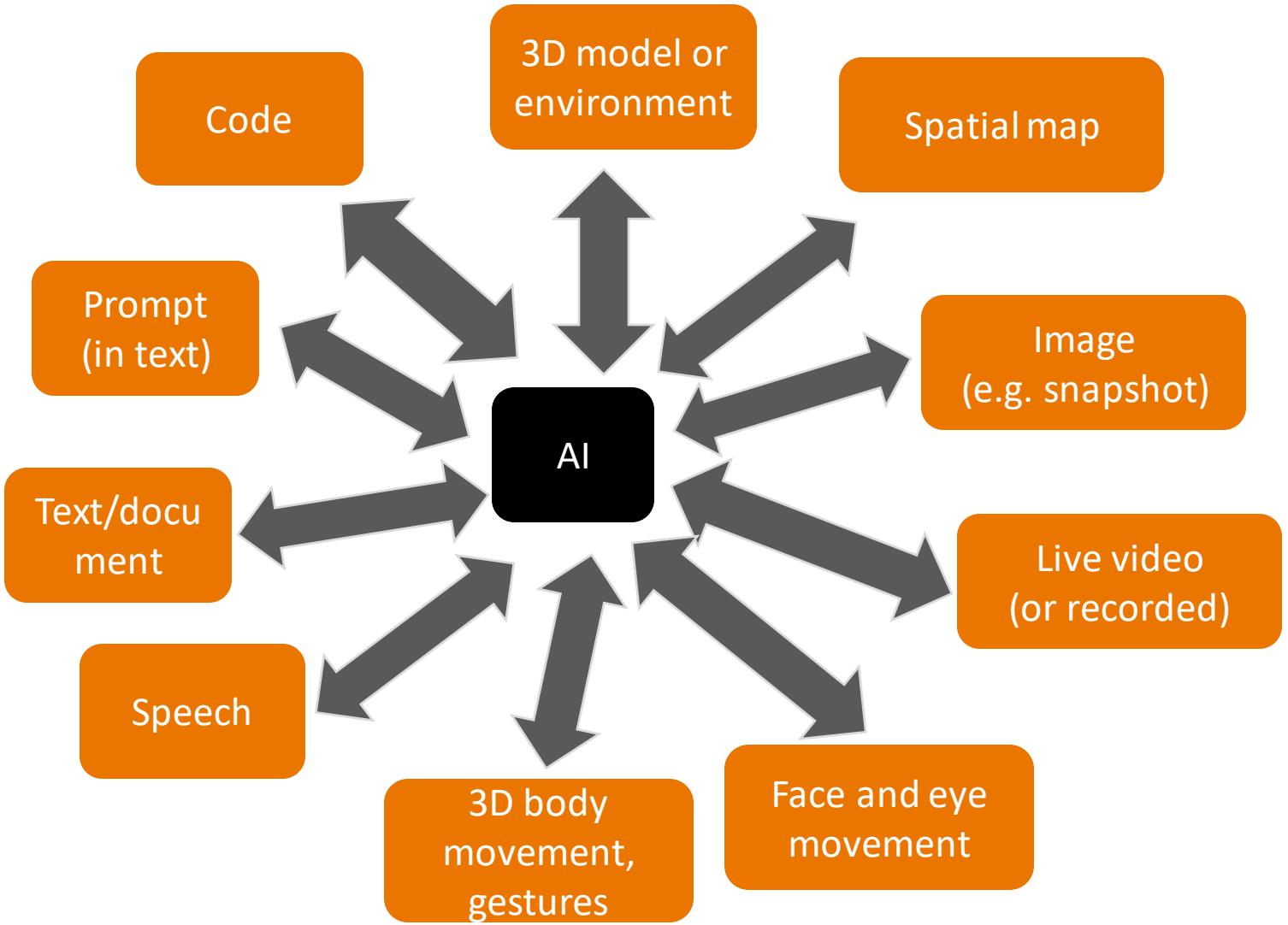
SURF

Interesting types of data with XR use cases



- Currently a lot of AI already used in devices, e.g. hand and controller tracking at the OS level
- Lots of interesting end-user use cases, examples:
  - Description to 3D environment (input: speech)
  - Generate code for XR actions (input: speech)
  - Add creative virtual overlay to physical room (input: 3D scan)
  - Custom gesture recognition (input: hand/body movement)
  - Ambient intelligence (input: preferably all data)
- The data, and ideas (!), are there, but lacks…
  - … access (i.e. protected by vendor)
  - … (real-time) performance
  - … effective algorithms

Local AI processing, i.e. on the XR device itself?



Reasons for local/on-device AI:

- Reduced latency
- Standalone apps
- Customized models
- Lower cost?
- Privacy
- Security

Is on-device AI possible?

```
melis@blackbox 11:20:~$ ./ollama run llama3:latest
>>> What are uses of AI for extended reality workflows?
...
```

Ollama prompt GPU memory usage (desktop)

```
File  Edit  View  Search  Terminal  Help
Every 1.0s: nvidia-smi                                                  blackbox: Wed May 2

Wed May 22 11:31:48 2024
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 550.78                 Driver Version: 550.78         CUDA Version: 12.4     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA TITAN V                 Off |   00000000:01:00.0  On |                  N/A |
| 30%   48C    P2             42W /  250W |    5920MiB /  12288MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|    0   N/A  N/A       584      G   /usr/lib/Xorg                                236MiB |
|    0   N/A  N/A     14025      G   ...erProcess --variations-seed-version        48MiB |
|    0   N/A  N/A     15106      G   /usr/lib/thunderbird/thunderbird              13MiB |
|    0   N/A  N/A     15816      G   /usr/lib/firefox/firefox                     391MiB |
|    0   N/A  N/A     17122      G   ...AAAAAAAACAAAAAAAAAA= --shared-files        48M |
|    0   N/A  N/A     34584      C   ...unners/cuda_v11/ollama_llama_server      5172M |
+-----------------------------------------------------------------------------------------+
```

META QUEST 3:

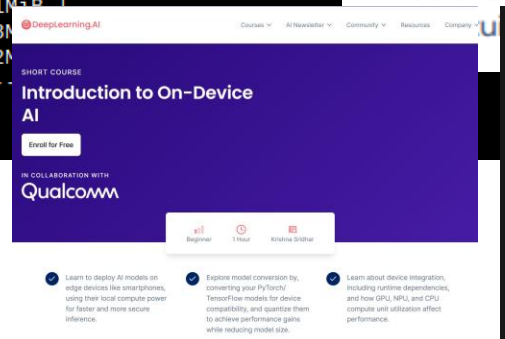
TOTAL 8 GB RAM, SHARED BY CPU AND GPU

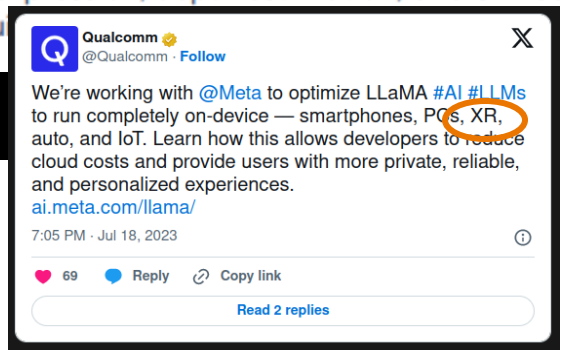
Qualcomm snapdragon XR2+ GEN 2 (platform product brief)

Unleashed on-device AI

Our 8x more performant AI[2], support for 12+ concurrent cameras, and a dedicated XR acceleration block enable unprecedented interaction concurrency essential for MR experiences.

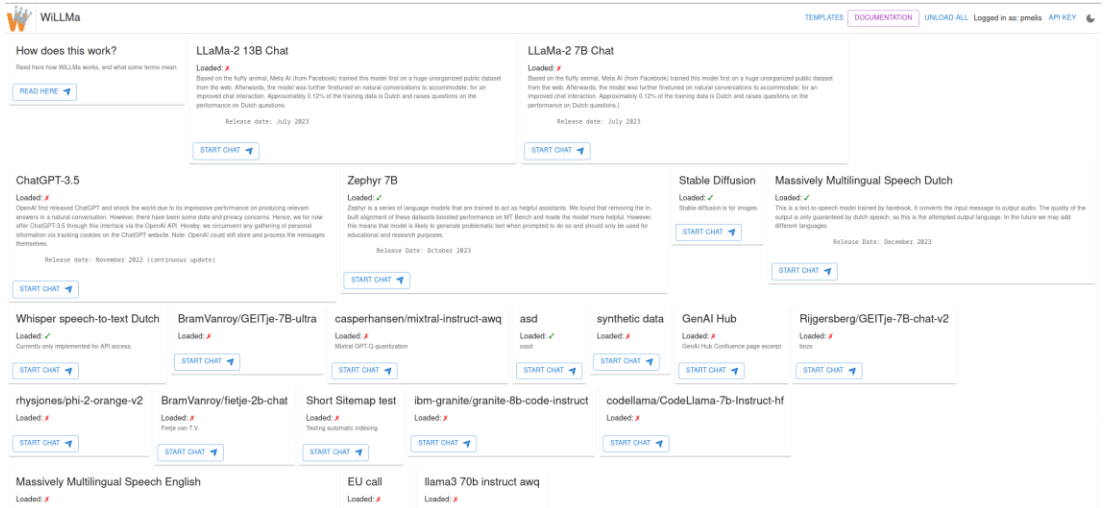
- Support for 12+ concurrent cameras for video see-through, controller, eye, hand, head, and face tracking, and depth sensing
- Dedicated hardware-accelerated computer vision block for enhanced, low-power, and high-accuracy perception algorithms, including head tracking, 3D reconstruction, and more
- Improved INT8 acceleration, thanks to fused microarchitecture and larger on-device memory
- Perception concurrency includes hand, head, controller, facial expression, depth estimation, and 3D reconstruction





WiLLMa – a generative AI platform by SURF

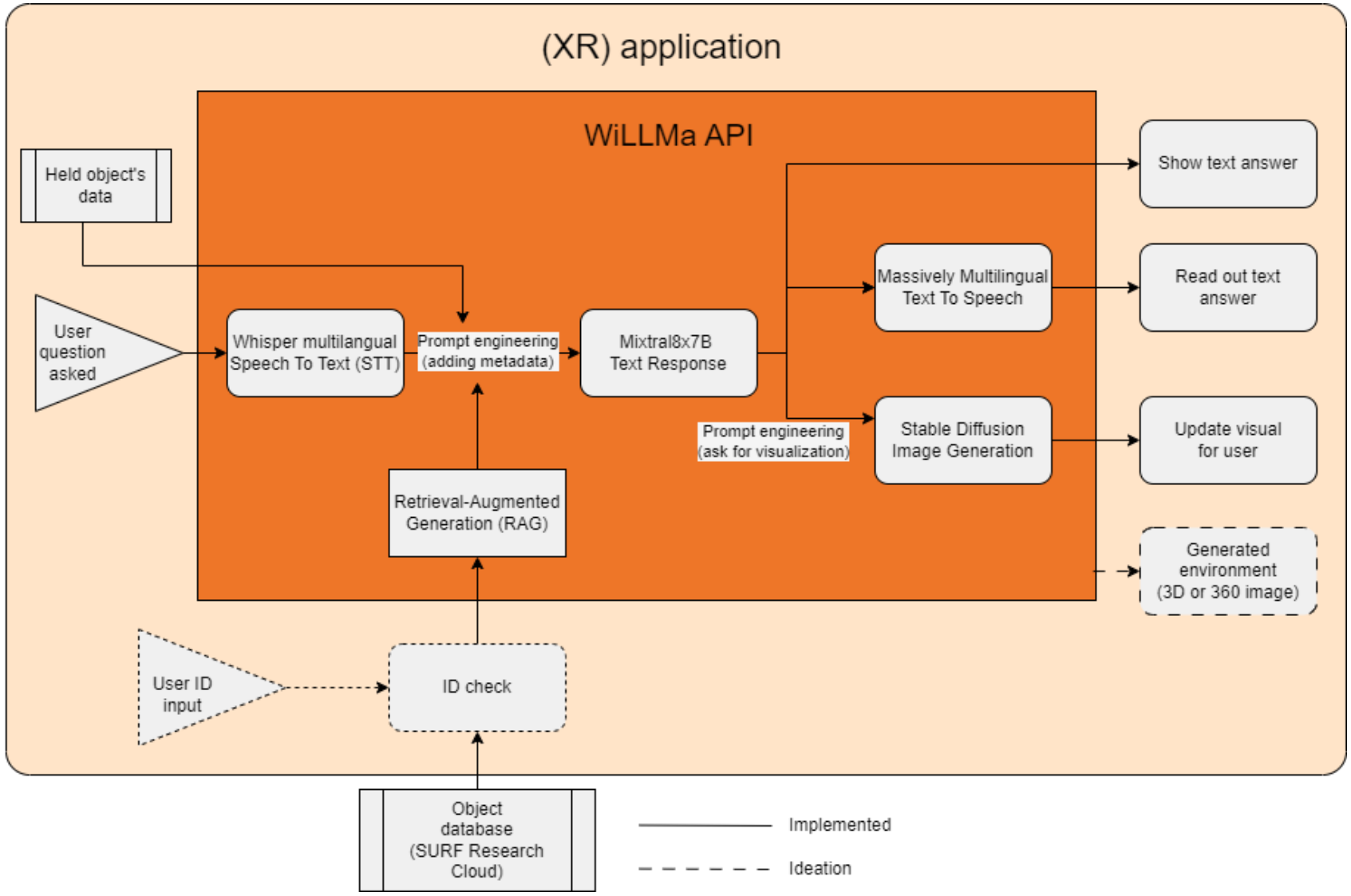
- Developed and maintained by the SURF Machine Learning team
- Provides a selection of (open-source) AI models from Hugging Face
  - Conversational/instructional/coding LLMs
  - Text-to-speech/speech-to-text (Dutch)
  - Image generation (StableDiffusion)
  - Possibly models uploaded by users in the future
- Model use either through API or command-line interface
- Advantages:
  - Easy way to experiment with AI models for end-users
  - Possibilities to integrate with other SURF infrastructure
  - Develop platform along our community's goals
  - Digital sovereignty



Digital heritage in XR (proof-of-concept)

"What is the history of this artifact?"

Under the hood...



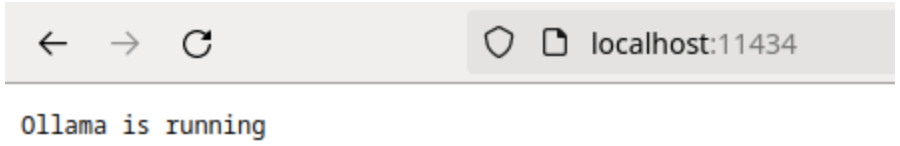
Intermezzo: quick-and-dirty local model + API using ollama
(Serve from machine with decent GPU)

No TTS/STT or image generation in ollama (currently)

```
# Step 0 – Install ollama
See https://ollama.com/
```

```
# Step 1 - Retrieve an LLM (one-time action)
$ ./ollama pull llama3:8b
pulling manifest
pulling 6a0746a1ec1a... 100% ... 4.7 GB
pulling 4fa551d4f938... 100% ... 12 KB
pulling 8ab4849b038c... 100% ... 254 B
pulling 577073ffcc6c... 100% ... 110 B
pulling 3f8eb4da87fa... 100% ... 485 B
verifying sha256 digest
writing manifest
removing any unused layers
success
```

```
# Step 2 - Start server (default TCP port 11434)
$ OLLAMA_HOST=0.0.0.0:11434 ./ollama serve
<keep running>
```




```
# Step 3 – Use API (e.g. from remote system)
$ curl http://localhost:11434/api/generate -d '{"model":"llama3:8b", "prompt": "is vr dead?", "stream": false}'
{
  "model": "llama3:8b",
  "created_at": "2024-05-22T19:02:06.439821826Z",
  "response": "The question that has plagued the VR community for years!\n\nTo answer your question: No,
  VR is not dead. In fact, the virtual reality industry has continued to evolve and grow steadily over the
  past few years. …
```

GPT-NL - a proprietary open language model for the Netherlands

- Open large language model for Dutch
- Plus "virtual facility"
  - Will be trained and hosted on Dutch National Supercomputer Snellius, maintained by SURF
  - Usable by academic institutions, researchers, governments, companies
- 2-year project
  - Developing partners: SURF, TNO & NFI
  - Budget: 13.5M€ from ministry of EZK
  - 1st year: development, 2nd year: exploitation
- Goals
  - Digital sovereignty (experience and development of LLMs, responsible use, …)
  - Transparent, fair and testable use of AI
- https://www.surf.nl/en/qa-gpt-nl-dutch-own-open-ai-language-model

Meta
Used 16K GPUs for training Llama3
Llama3 8B took 1.3M GPU-hours
(https://ai.meta.com/blog/meta-llama-3/)

Snellius @ SURF
288 GPUs (currently) + 352 (Q3)

TNO

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

SURF

THANKS FOR THE ATTENTION!

Paul Melis

Paul.melis@surf.nl

www.surf.nl

@PMel3D (Twitter)

Driving innovation together