Trustworthy AI for Accessibility:

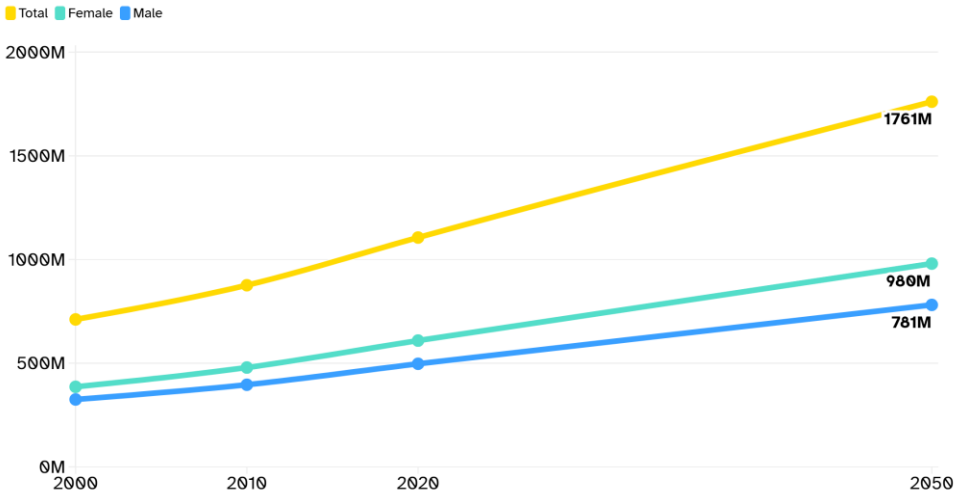
Enhancing Navigation and Scene Understanding Without Sacrificing Privacy

Prof. Hong-Chuan Yang
Professor and Director of Wireless+AI lab (YWAILab)
University of Victoria, Canada
hy@uvic.ca

Acknowledgement: Jinke Jiang, Abida Sultana, NSERC Discovery Grant, Weighton Product Development Fund

Prevalence of Visually Impaired

- Currently, 43 million people (0.5% of global population) are blind.
- Expected to double in 2050 due to population growth and aging.
- Effective accessibility measures are keys to improve their quality of life.



How to cite: Bourne R, et al. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Glob Health. 2020. Accessed via the IAPB Vision Atlas: visionatlas.iapb.org.

University of Victoria

AI Companion for Visually Impaired

- Enhance accessibility with AI technologies
- Voice assistant solutions, using
  - Large language models (LLMs)
  - Vision-language models (VLMs)

  to replace guide dogs/virtual volunteers
- Sample products
  - Ray-Ban Meta glasses[1]
  - Envision glasses [2]
  - AiSee[3] from NUS

University of Victoria

University of Victoria

Limitations of existing solutions

- Existing solutions upload captured images to the cloud for processing
  - Unnatural interaction with insufficient accuracy
  - Poor privacy protection for users
- General-purpose functions only so far
  - Scene description, document reading, question answering, etc.
  - Missing guidance and navigation support

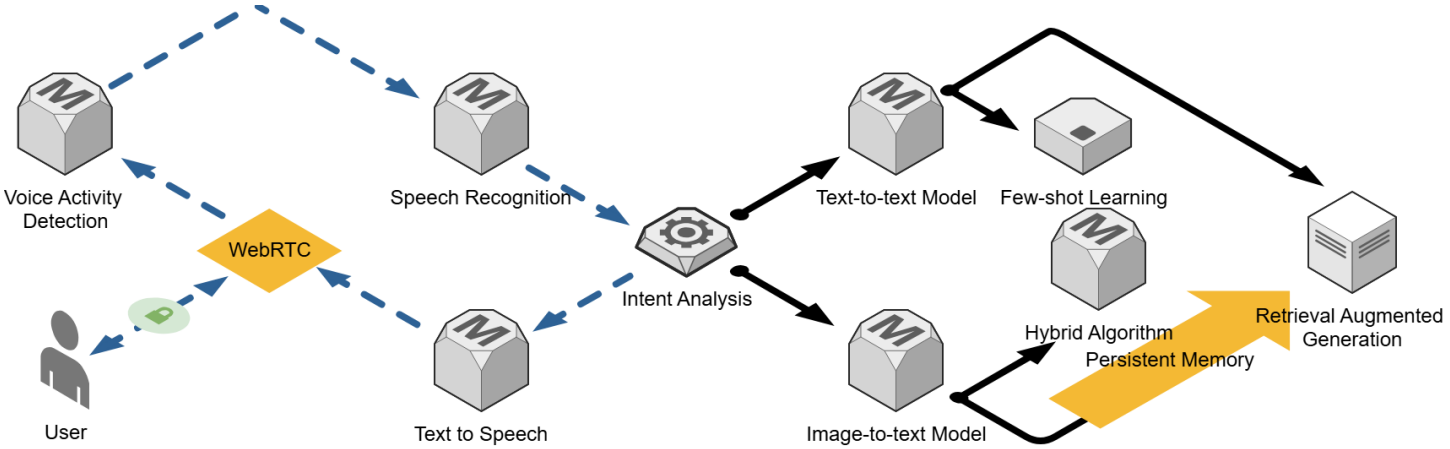
Enhancing accessibility without sacrificing privacy!

University of Victoria

EnSightingAI: Our ongoing effort

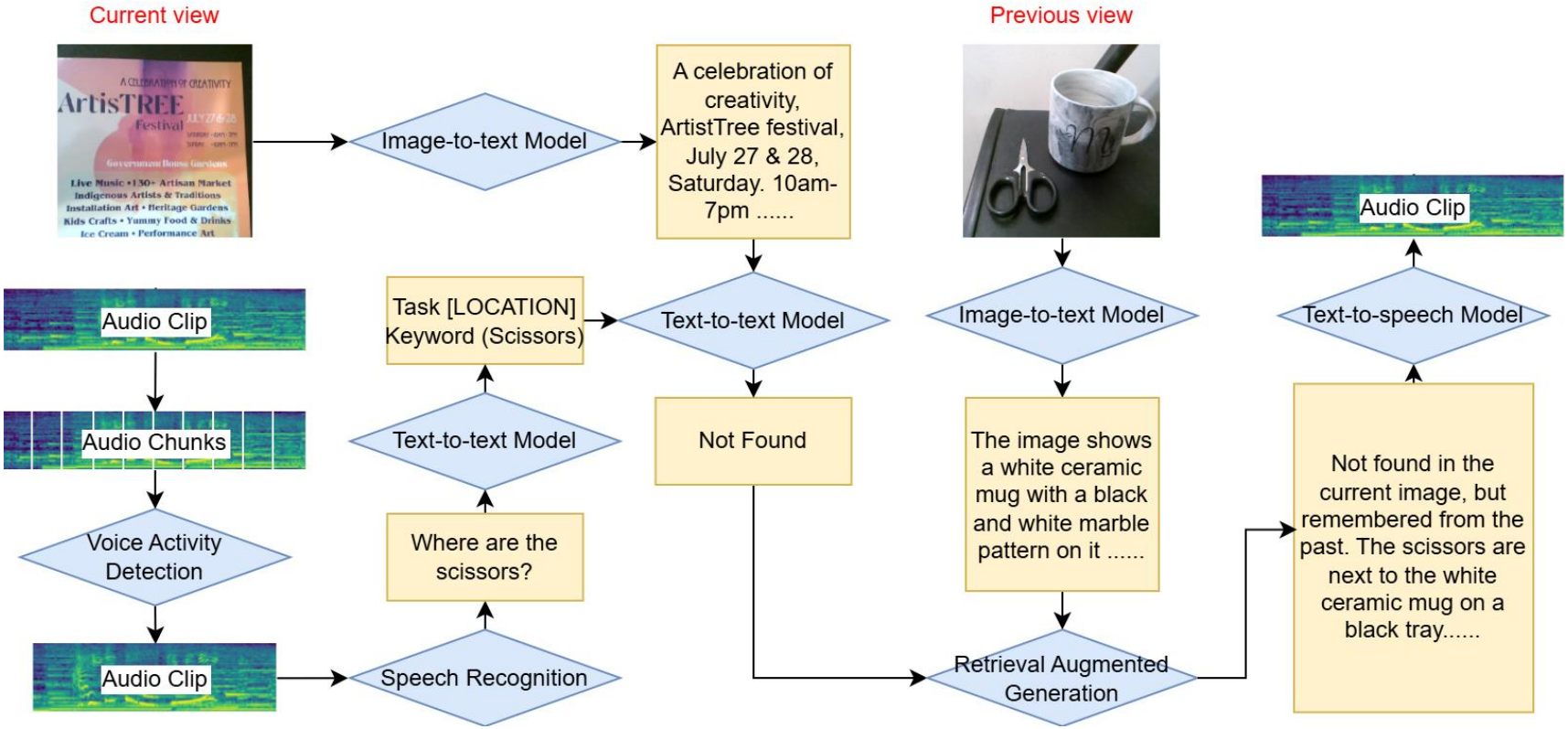
- Goal: stand-alone voice assistant solution with navigation support
- Current implementation:
  - Stand-along deployment based on Nvidia A6000;
  - Extensible architecture using open-source models, e.g. Florence-2[4], Gemma-2[5],...
- Target features:
  - Natural voice interaction with low latency;
  - Locating items from managed memory;
  - Object-reaching guidance and navigation.

Implementation structure



- Real-time communications through WebRTC;
- Interruptible response to reduce latency;
- Retrieval augmented generation (RAG) with memory management.

Object locating with RAG



Challenge: Distance estimation

- Object-reaching guidance requires accurate distance estimation.
- Existing AI models cannot reach human level accuracy in counting and distance calculation [6].

Mean Relative Accuracy (MRA) comparison († = Tiny data set)

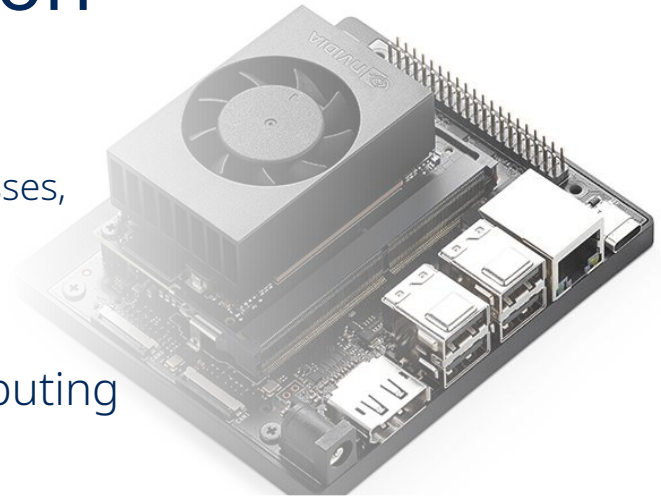
Methods	†Human Level	†Gemini-1.5 Flash	†Gemini-1.5 Pro	†Gemini-2.0 Flash	GPT-4o	Gemini-1.5 Flash	Gemini-1.5 Pro	LongVA-7B	InternVL2-40B	VILA-1.5-40B	LLaVA-NeXT-Video-72B	LLaVA-OneVision-72B	VLM-3R (7B)
Obj. Count	94.3	50.8	49.6	52.4	46.2	49.8	56.2	38	34.9	22.4	48.9	43.5	70.2
Abs. Dist.	47	33.6	28.8	30.6	5.3	30.8	30.9	16.6	26.9	24.8	22.8	23.9	49.4
Obj. Size	60.4	56.5	58.6	66.7	43.8	53.5	64.1	38.9	46.5	48.7	57.4	57.6	69.2
Room Size	45.9	45.2	49.4	31.8	38.2	54.4	43.6	22.2	31.8	22.7	35.3	37.5	67.1

Challenge: Outdoor solution

- Portable platform:

  Nvidia Jetson Orin™ NX Super 16G module with glasses, Bluetooth headphone, and portable camera

- Performance loss due to quantization
- Slower response because of low-power computing



Nvidia A6000	Nvidia Jetson Orin			
PyTorch-fp16	ONNX-int8-CUDA	ONNX-int8-CPU	PyTorch-fp16	TensorRT-fp16
1.004~1.006s	13.6~14.2s	8.3~8.9s	3.58~3.68s	1.46-1.48s

3*562*750 image, 40W power level

University of Victoria

Concluding remarks

- Existing general AI models needs to be customized for specific downstream tasks.
- Modular design allows for extensibility, flexibility, and potentially better performance.
- Downstream metrics key evaluate the readiness of AI application.

References

[1] Ray-Ban | Meta AI glasses. (n.d.). https://www.ray-ban.com/canada/en/rayban-meta-ai-glasses

[2] Envision Glasses. (n.d.). https://www.letsenvision.com/glasses/home

[3] AISEE – Augmented Human Lab. (n.d.). https://ahlab.org/project/aisee/

[4] Xiao B, Wu H, Xu W, et al. Florence-2: Advancing a unified representation for a variety of vision tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 4818-4829.

[5] Team G, Riviere M, Pathak S, et al. Gemma 2: Improving open language models at a practical size[J]. arXiv preprint arXiv:2408.00118, 2024.

[6] VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction. (n.d.). https://vlm-3r.github.io/

University of Victoria

Thank You

Visit us at:

https://oac.ywailab.uvic.ca