



Conversational AI with Gemma on Modalix MLSoC

SiMa.ai Value Proposition

- Turns equipment into intelligent assistants.
- Fully on-device intelligence & built-in privacy.
- Low latency, cloud-free operation.
- Server-class AI within embedded power budgets.
- Sees, listens, and reasons in real time, providing insight to guide users or autonomous systems.

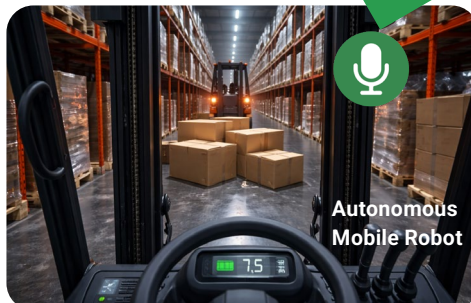
Example Prompts & Responses:

"What appears to have happened in this scene, and what evidence suggests this?"



"The scene depicts a severe storm, likely a hurricane, has damaged a large solar panel array. Evidence: The panels are scattered and broken, indicating high winds cause the damage. The surrounding landscape also shows debris and destruction, further supporting the storm's impact."

"Is there enough room for me to steer around these boxes?"



"It appears very tight! Judging from the image, there isn't much room to maneuver the forklift. The boxes in the foreground and the stacked shelves created a significant obstruction. Steering around them in a forklift would likely be difficult and potentially hazardous."

"What do you see in this image, and what do you think could have happened?"



"The image shows two identical conveyor belts. The left belt is pristine and new, while the right belt is heavily soiled with debris and shows signs of wear and tear. It appears the right belt was likely used to transport a material that caused significant damage or contamination. The contrast highlights the difference in wear and tear after a period of heavy use and potential failure."

Cloud-Free Conversational AI for Embedded Systems

SiMa Modalix brings conversational AI into embedded Physical AI systems by executing language, speech, vision, and retrieval pipelines entirely on-device, for example, the Google Gemma model.

- Multimodal conversational AI that sees, listens, and reasons—running entirely on-device.
- Real-time speech, vision, and language understanding without cloud dependency.
- Context-aware, private, and deterministic—designed for real-world Physical AI systems.
- Lower system cost and complexity compared to GPU-based edge deployments.

Google Gemma Powered by SiMa Modalix

Cloud-Class Multimodal AI with Sub-10W Physical AI Deployment

Technical Details

Runs Gemma 3 or other LLMs fully on-device, compiled with LLiMa for MLA acceleration—no cloud or host required.

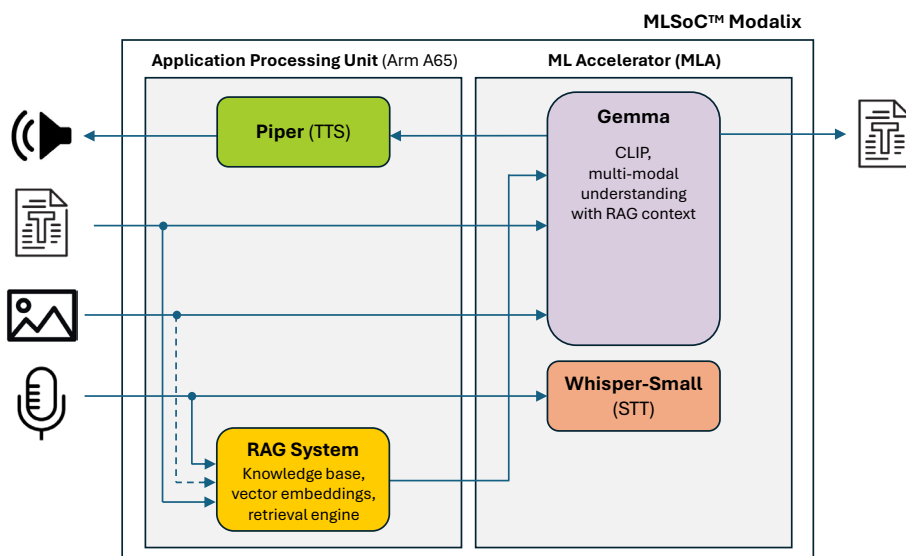
Integrates Whisper (STT) and Piper (TTS) for real-time speech recognition and synthesis.

Enables vision-language reasoning via CLIP-based image embeddings fused into Gemma context.

Executes RAG locally with on-device embeddings and similarity search—no retraining required.

Achieves ~0.5 s TTFT and over 13 tokens/sec under a sub-10W power envelope.

Utilizes x8 Arm Cortex A65 APU + MLA + DSP/CVU for concurrent multimodal inference.



Convert Hugging Face models into optimized, binaries to run on SiMa Modalix SoCs using SiMa's LLiMa framework, integrated into the Palette SDK

Hardware

Our Silicon



Modalix

Our Boards



SoM



SoM Dev



PCIe HHL

Software



Hugging Face huggingface.co/simaaai

Model Zoo sima.ai/model-browser/



About SiMa.ai

SiMa.ai is a leader in Physical AI, delivering a purpose-built, software-centric platform that brings best-in-class performance, power efficiency, and ease of use to Physical AI applications. Focused on scaling Physical AI across robotics, automotive, industrial automation, aerospace & defense, smart vision, and healthcare, SiMa.ai is led by seasoned technologists and backed by top-tier investors. Headquartered in San Jose, California. Learn more at www.sima.ai.

