



Multimodal Enterprise-Scale Applications in the Generative AI Era

Mumtaz Vauhkonen, Ph.D

Senior Director

Skyworks Solutions

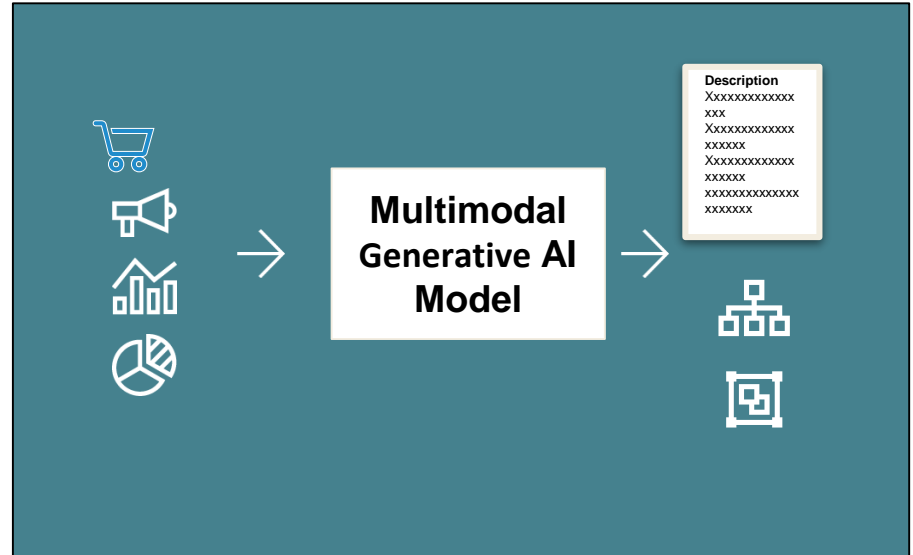
Agenda

- What is multimodal AI?
- Typical pipeline for multimodal AI application
- The data and features: The input and expected outputs
- Applications of multimodal AI
- Challenges and areas of focus
- Q&A

What is multimodal AI?

AI Models that are capable of taking multiple types of data as input and producing output in a single or multiple types of data outputs.

Very similar to humans taking in multiple sensory inputs for processing and making their decisions, AI models also benefit from multimodal inputs in multiple scenarios.



Multimodal AI evolution

Lidar



Cameras

Radar

Ultrasound

Autonomous vehicles have been tapping into multimodal AI for the past decade

Multimodal Transformer Models

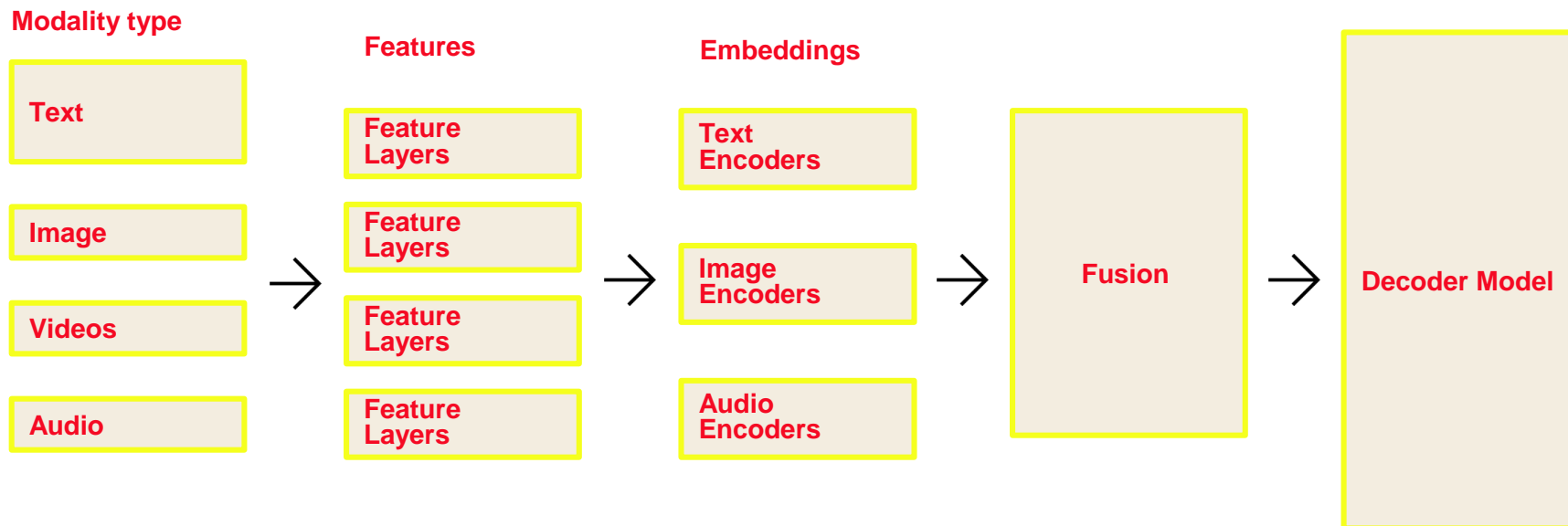
GPT-4o

Gemini 2.5pro

Llama 4 scout and Maverick

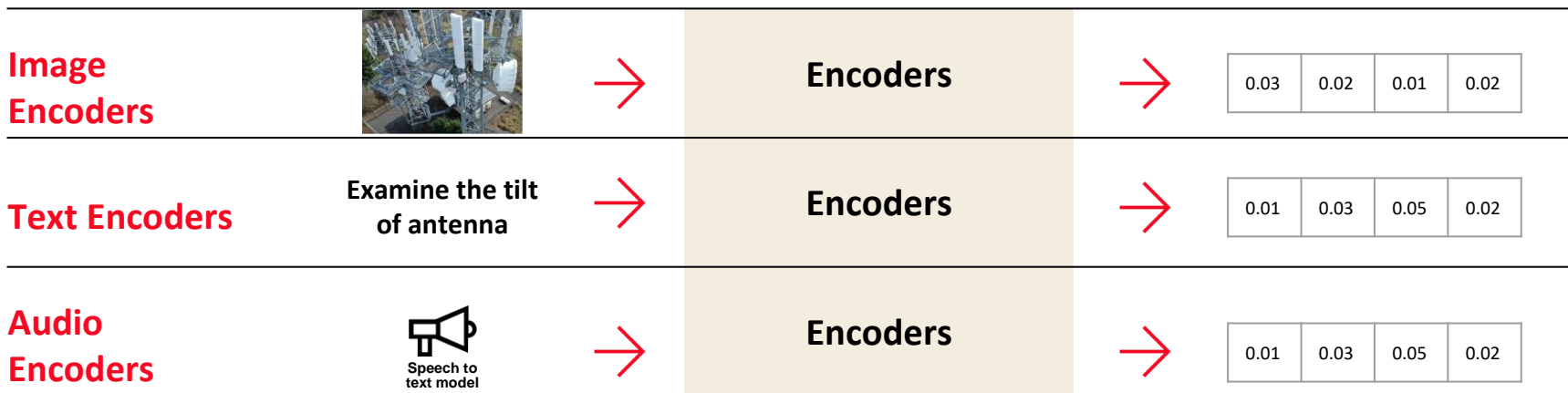


How do current multimodal GenAI models work?



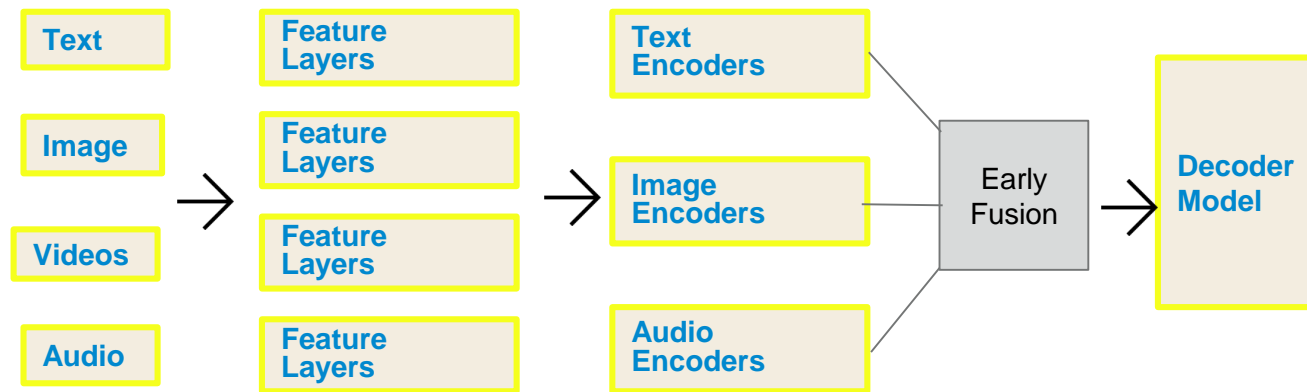
Embeddings for multimodal models

Each modality has its own embeddings that are connected with other modalities



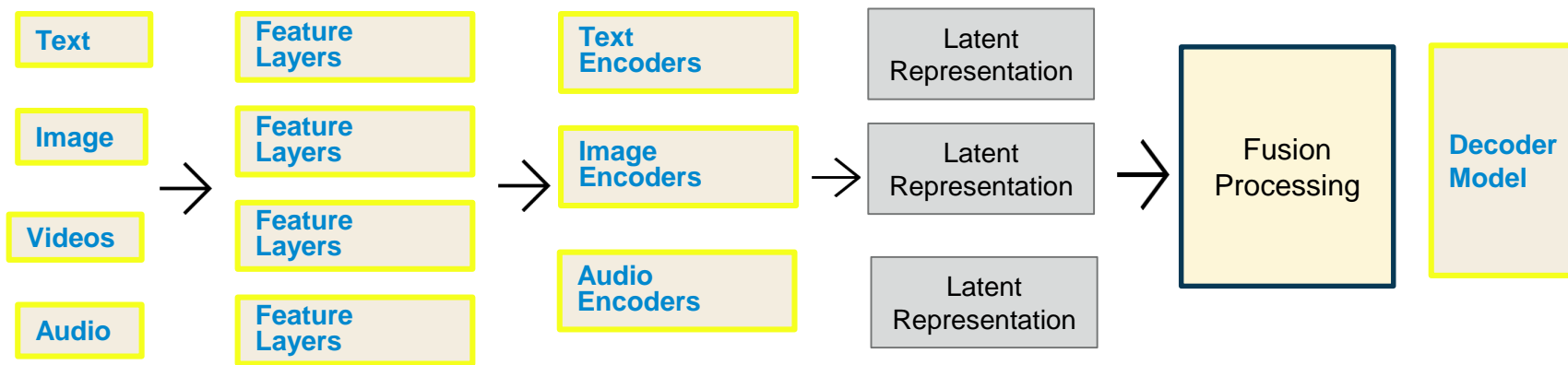
- Multiple strategies to fuse the embeddings from the modalities
 - **Early fusion:** Integrate all the modalities before passing as inputs to the model.
 - **Mid-cycle fusion:** Embeddings are projected into a latent space where the latent representations are fused and passed on to the model.
 - **Late fusion:** Process the raw inputs independently and combine the output.
 - **Hybrid fusion:** Integrating all above fusion approaches in each stage as required for desired outputs.
- Main fusion techniques include: **element wise addition, concatenation or some attention mechanism.**
- Measuring which modalities fuse at what stage for optimal outcomes is a step that can enhance performance⁽¹⁾.

Fusion of multimodalities: Early fusion



Early Fusion is feature vector level fusion before the decision happens in the model.

Fusion of multimodalities: Mid fusion



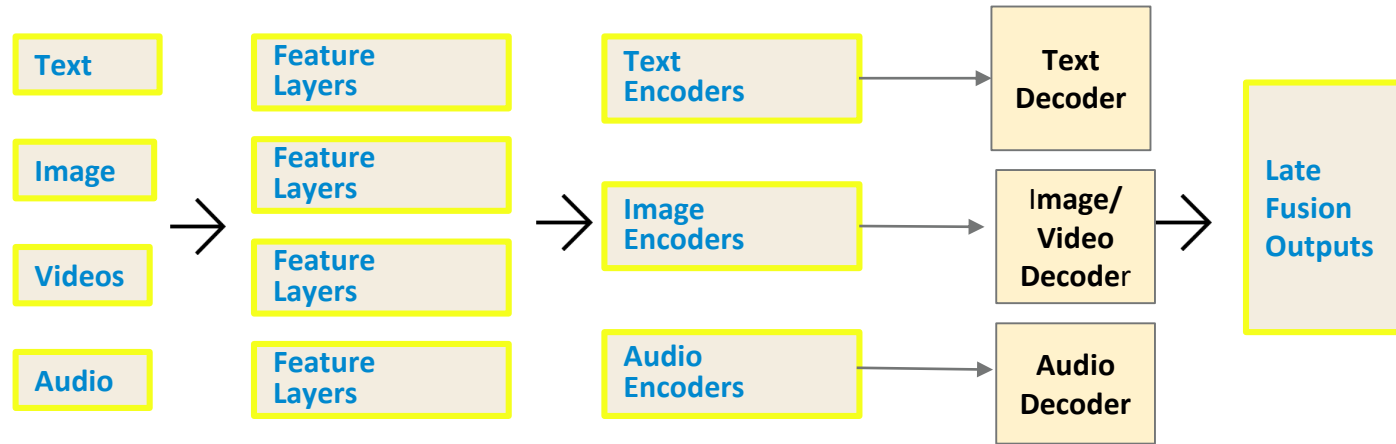
Mid Fusion creates latent representations and fuses those representations using element wise addition, concatenation or some attention mechanisms before sending to decision modules ⁽²⁾.

Latent Representation means reducing high dimensional data into lower dimensions preserving the patterns, features and relationships present in data.

Example: Images -> Instead of every pixel, preserve representations of features.

Text -> Vectorizing a document or sentence and reducing dimensions.

Fusion of multimodalities: Late fusion



In late fusion outputs are combined at the decision level using various techniques like averaging the outputs, concatenation or using another model for final output.

Some sample use cases

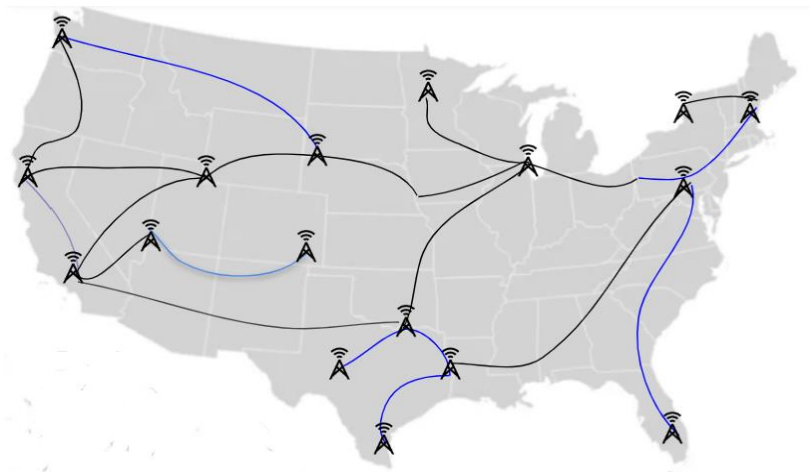
- Technical trouble shooting assistive bots: Assistive bots for trouble shooting equipment, machinery, infrastructure installments, semiconductor circuits etc.
- Health care assistants: Physical therapy AI assistants for patients to guide movements.
- Maps and navigation tools. Recommendations based on image, voice and location data.
- Complex document processing chat bots: Documents with mixed modality content such as embedded videos, presentations and tabular data.
- Manufacturing pipeline for semiconductor products: die imperfections, outlier and contamination detections, tracking material flow etc.

General Requirements

- Parse domain specific technical content
- Identify steps for a given query for resolving an issue
- Bring tickets or documents based on time-based query
- Able to parse and query multimodal content – audio & videos
- Hold previous questions in context when troubleshooting
- Disambiguate based on simple prompts
- Have the ability to connect abbreviations to content

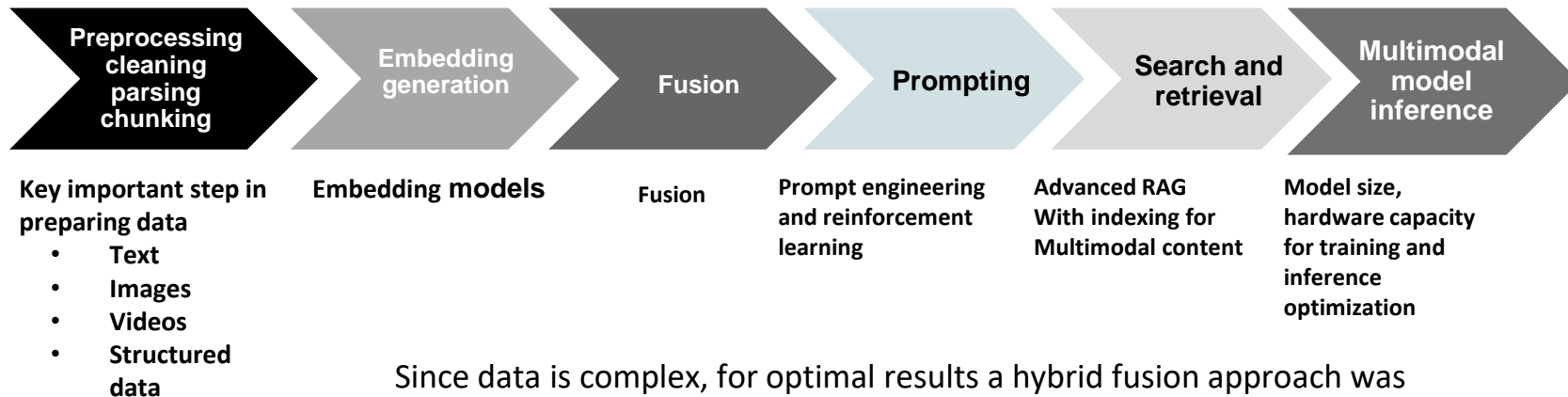
Implementing these requirements for a productive infobot requires a multilayered approach

Telco Equipment troubleshooting - Smart assistant



Multiple types of equipment exist with specification manuals, videos, parts descriptions, etc. for technicians and engineers in assembling and trouble shooting process.

Step-by-step flow of Smart Assist Bot



Since data is complex, for optimal results a hybrid fusion approach was taken by testing which modalities fused best at what stage.

1. Text and work and design flows were fused as early fusion
2. Audio and text as mid fusion
3. Video content as late fusion



- Cleaning: Reducing noise in data is crucial
 - Removing irrelevant content
 - Spelling and grammar correction
 - Adding relevant data: Metadata filtering, dates, key tags
- Image processing with OCR and other models
- Video processing and audio parsing
 - Fusing audio text with image frames
 - Maintaining links

Sample queries:



- How to install QLV?
- Bring me previous tickets with light indicating red in last 6 months for model V722

Methods to optimize

- Optimizing searching and indexing algorithms for multimodal content
- Storing processed outputs
- Creating appropriate internal and external indexing

Site equipment installation for build engineering assist example

Example of late fusion



Multi-modal LLM giving two images and a prompt as input to determine if two buildings are similar.

LLM-generated answer:

After examining both images, it appears that the largest building in each image is indeed quite similar. Both buildings have a rectangular shape with a symmetrical arrangement of windows and balconies on the upper floor. The color scheme seems consistent across both structures, featuring shades of blue or gray for the walls and dark roofing material. There are no visible distinctive features to suggest otherwise.

Based on these observations, my final answer would be: {similar:YES}

Key aspects for scaling to production

- **Data pipelines:** Preparing data, storing each modality in suitable storage systems, and having pipelines that can keep data fresh with high availability.
- **Latency aspects:** Text modalities load faster than image or videos, having suitable pipelines to reduce latency for all modalities is essential. Parallelization proves very beneficial.
- **Deployment of multiple models:** More than one model typically exists in the pipeline. Deploying without generating bottlenecks at each stage makes the experience seamless.
- **Consistency in preprocessing and fusion:** Changing preprocessing method in one modality might impact fusion outcomes. Ensuring compatibility results in desirable outcomes.

Challenges and areas of focus

- Gathering and preparing datasets for fine tuning
- Identifying the right models in the pipeline to integrate multiple types of data
- Aligning domain expertise and AI expertise

Gathering and preparing datasets for fine tuning

For very domain specific tasks, finetuning the embedding models and LLMs becomes necessary along with RAG pipelines.

- Preprocessing multiple modalities poses a big challenge for production pipelines where data refresh rate is high. Having robust preprocessing and data movement pipeline overcomes these challenges.
- Creating a data set for finetuning requires careful balance of multiple modality inputs. Some modalities can be sparse vs others.

Model pipelines: Identifying the right models

- Identifying the right models in the pipeline to integrate multiple types of data
- Experimentation with ever evolving models and staying on top of AI industry releases
- Adequate infrastructure capacity allocation for each model in pipeline to avoid bottle necks
- Adapting right frameworks and libraries to support the deployments

Aligning domain expertise and AI expertise

One of the key things that makes AI applications successful is involving domain or subject matter experts

- Focus on the right types of data: In addition to data creation, creating the relevant data sets require domain expertise
- Right type of questions for fine tuning the outputs and responses

- Multimodal applications for a specific domain need the right fusion strategy depending on the use case for optimal results
- Experimenting and aligning the models in a scalable data pipeline for all modalities is essential for seamless performance
- Engaging domain / subject matter experts all the way from dataset preparation to the right applications is essential for successful development and adaptability of the AI applications

- High Modality Multimodal Transformer: Quantifying Modality & Interaction Heterogeneity for High – Modality representation Learning. Transactions on Machine Learning Research (05/2023). P Liang Et al.
- Multimodal Sensor fusion in latent representation space. Scientific Reports. 13.20.1038/s41598. Piechocki, Et al.
- Information Fusion in Signal and Image Processing, Isabelle Bloch. Wiley.

Thank you