MULTIMODAL FOUNDATION MODELS

ALBERTO ARKADER KOPILER

IMAGE PROCESSING COURSE

IMPA 2023

Prompt: “Generate an image from a beautiful house with sunset background”

Description: “The first generated image depicts a beautiful two-story house with large windows and a wrap-around porch. The house is set against a stunning sunset with a sky painted in hues of orange, pink, and purple. The landscape is serene, featuring a well-manicured lawn and a few scattered trees, which contribute to the overall picturesque quality of the scene. As the evening approaches, the first stars of the night are beginning to twinkle in the sky, adding a touch of magic to the tranquil setting.”

Summary: “The image features a two-story house with expansive windows and a porch, set in a peaceful setting with a manicured lawn and trees, against a backdrop of a vibrant sunset with emerging stars.”

MULTIMODAL FOUNDATION MODELS

What is a Foundation Model?

A **foundation model** is a *pre-trained* deep neural network that forms the backbone for various tasks:

- ✓ Object classification, Object detection, Image segmentation, Image Generation
- ✓ Text summarization, Text generation, Text Translation

The concept of foundation models comes from building upon a base or ‘foundation’ that’s already been built.

MULTIMODAL FOUNDATION MODELS

What is a Foundation Model? Wikipedia

A **foundation model** (also called **base model**)[1] is a large machine learning (ML) model trained on a vast quantity of data at scale (often by self-supervised learning or semi-supervised learning)[2] such that it can be adapted to a wide range of downstream tasks.[3][4]

MULTIMODAL FOUNDATION MODELS

What is a Foundation Model?

These models are trained on massive, diverse datasets to capture textual/audio/visual features that are universal to many different domains. They can then be used to perform specific tasks without the data needed to train a custom model from scratch. This approach leverages neural networks' powerful representational learning capabilities to generalize well across tasks.

MULTIMODAL FOUNDATION MODELS

Why Multimodality?

ML model operated in one data mode:

- Text
  - Translation
  - Language Modeling
- Image
  - Object detection
  - Image classification
- Audio
  - Speech recognition

MULTIMODAL FOUNDATION MODELS

Why Multimodality?

"Natural intelligence is not limited to just a single modality"

- Humans can read and write text
- We can see images and watch videos
- We listen to music to relax and watch out for strange noises to detect danger
- Being able to work with multimodal data is essential for us or any AI to operate in the real world.

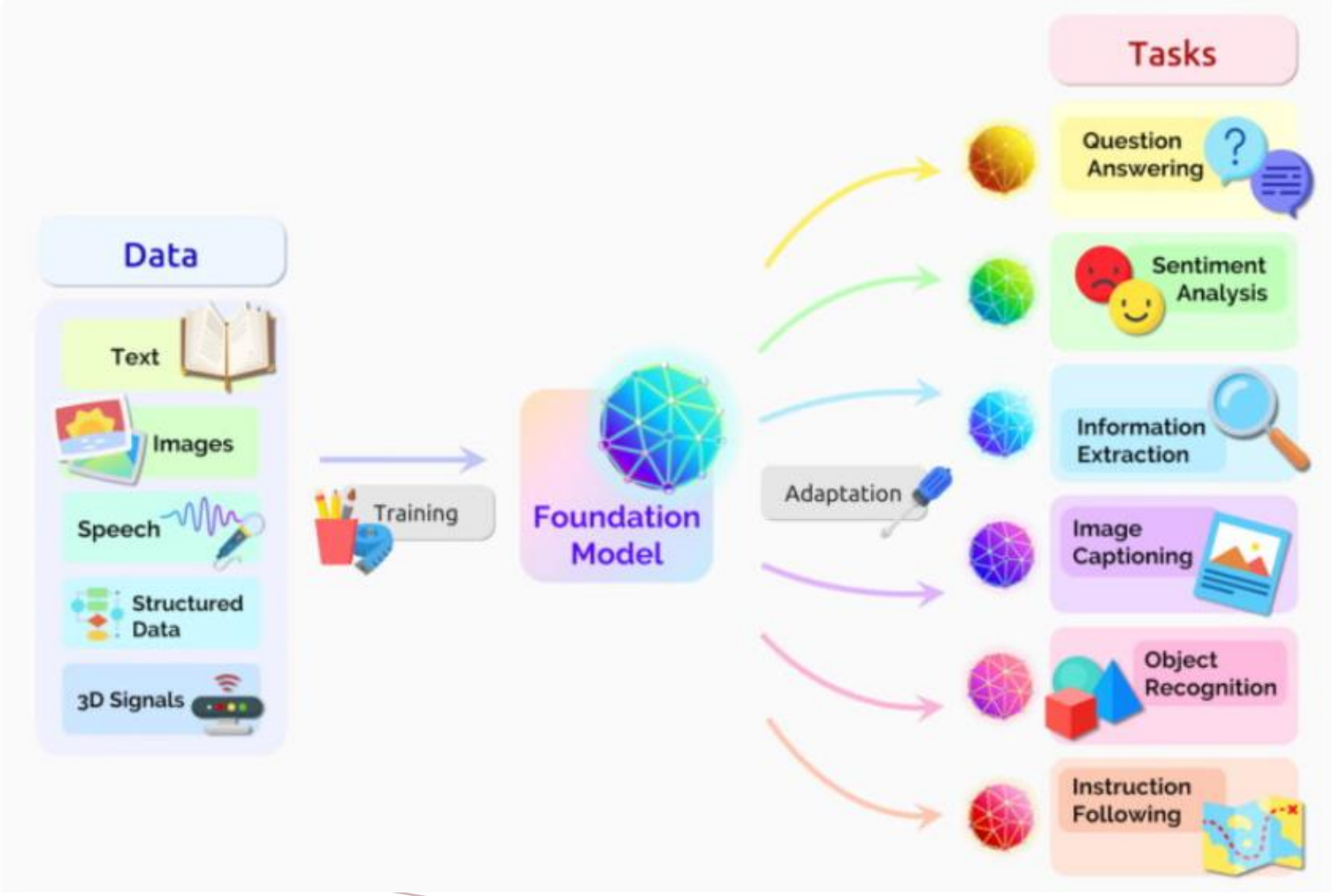
MULTIMODAL FOUNDATION MODELS

Incorporating additional modalities to LLMs (Large Language Models) creates LMMs (Large Multimodal Models)

LLM – LARGE LANGUAGE MODELS →

Multimodality and Large Multimodal Models (**LMM**s)

MULTIMODAL FOUNDATION MODELS



Source: NVIDIA

MULTIMODAL FOUNDATION MODELS

- ✓ MULTIMEDIA
- ✓ TEXT (ANALYSIS e SYNTHESIS (ex. summarization, generation, stylization)
- ✓ AUDIO (ANALYSIS (ex. Whisper) e SYNTHESIS (ex. Generation of Images and Sounds by Voice Command); Human Voice and Music
- ✓ IMAGE (ANALYSIS (ex. segmentation, description, image retrieval, image search) and SYNTHESIS); 2D, 3D, RGB-D
- ✓ VIDEO
- ✓ MODEL'S PIPELINE
- ✓ MODEL DESCRIBES AN IMAGE → MODEL generates an image from a Description

TEXT-TO-VIDEO

- Google: 'Imagen Video' a text-to-video generative AI model capable of producing high-definition videos from a text prompt
- Meta: 'Make-A-Video' an AI system that allows users to turn text prompts into short video clips
- NVIDIA: 'Gen2' – A multimodal Generative AI system that can generate novel videos with text, images or video clips. Realistically and consistently synthesize new videos with nothing but text. It's like filming something new, without filming anything at all

TEXT-TO-AUDIO

- Google: 'AudioLM' an audio generation model that learns to generate realistic speech and piano music by listening to audio-only.
- Meta: 'AudioGen,' an auto-regressive generative model that generates audio samples based on text inputs.

TEXT-TO-3D

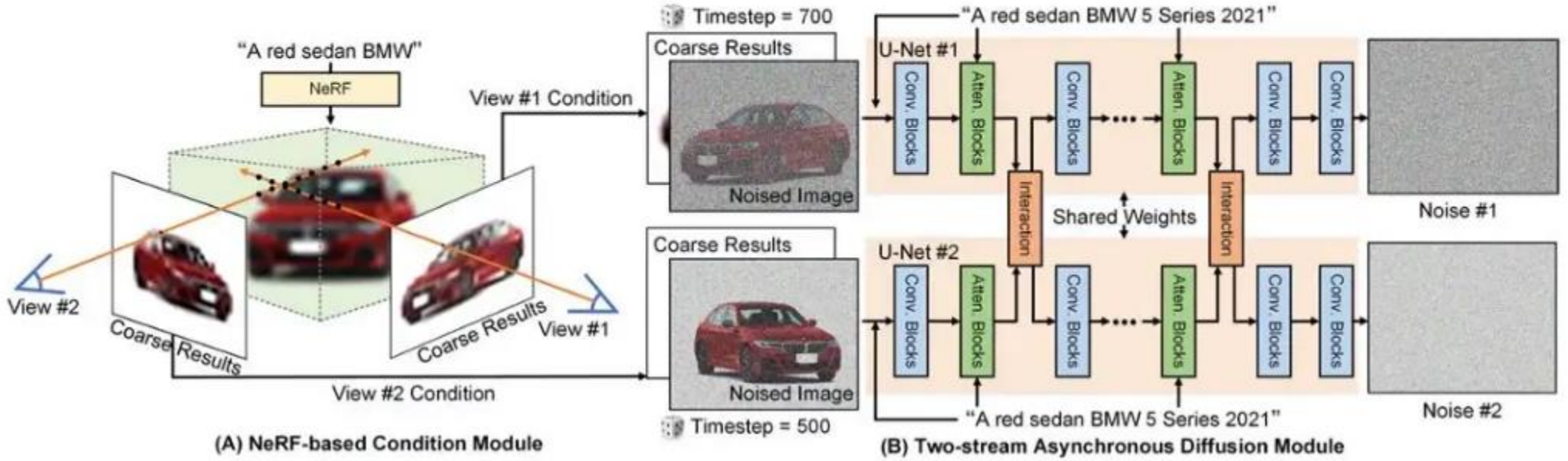
- Google:‘DreamFusion’, discovered a method to produce 3D models based on a user’s word input The new technology, dubbed ‘DreamFusion’, employs 2D Diffusion and is expected to make significant advances in text-to-image generation.

The researchers optimised a randomly-initialised 3D model called ‘NeRF’ (a Neural Radiance Field) via gradient descent so that the renderings in 2D from random angles achieve lesser loss.

TEXT-TO-3D

3DDesigner: Towards Photorealistic 3D Object Generation and Editing with Text-guided Diffusion Models

Text-guided 3D-consistent generation framework (training phase).



MULTIMODAL RACE

- ✓ **Google Deepmind** – Gemini, Med-PaLM Multimodal
- ✓ **OpenAI** – OpenAI noted in their GPT-4V(ision) system card that "incorporating additional modalities (such as image inputs) into LLMs is viewed by some as a key frontier in AI research and development."; GPT-4 Multimodal (DALL-E 3 + GPT-4)
  - Other model: CLIP (Contrastive Language-Image Pretraining) – neural network trained on a variety of (image, text) pairs
- ✓ **stability.ai** – DeepFloyd IF (Stable Diffusion + StableLM + Stable Audio) – No multimodal yet
- ✓ **Meta** – CMEleon, ImageBind, SeamlessM4T
  - Other models: DINOv2, SAM (Segment Anything)
- ✓ **Tecent** - Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration https://arxiv.org/abs/2306.09093

MULTIMODAL RACE

ImageBind (Meta)

first of its kind AI model that is capable of binding data from six modalities (images and video, audio, text, depth, thermal and inertial measurement units (IMUs) at once, without the need for explicit supervision).

https://imagebind.metademolab.com/

https://arxiv.org/pdf/2305.05665.pdf

MULTIMODAL RACE

DINOv2: A Self-supervised Vision Transformer Model (Meta)

A family of foundation models producing universal features suitable for image-level visual tasks (image classification, instance retrieval, video understanding) as well as pixel-level visual tasks (depth estimation, semantic segmentation).

https://dinov2.metademolab.com/

Segment Anything Model (SAM): a new AI model from Meta AI that can "cut out" any object, in any image, with a single click (Meta)

SAM is a promptable segmentation system with zero-shot generalization to unfamiliar objects and images, without the need for additional training.

https://segment-anything.com/

MULTIMODAL RACE

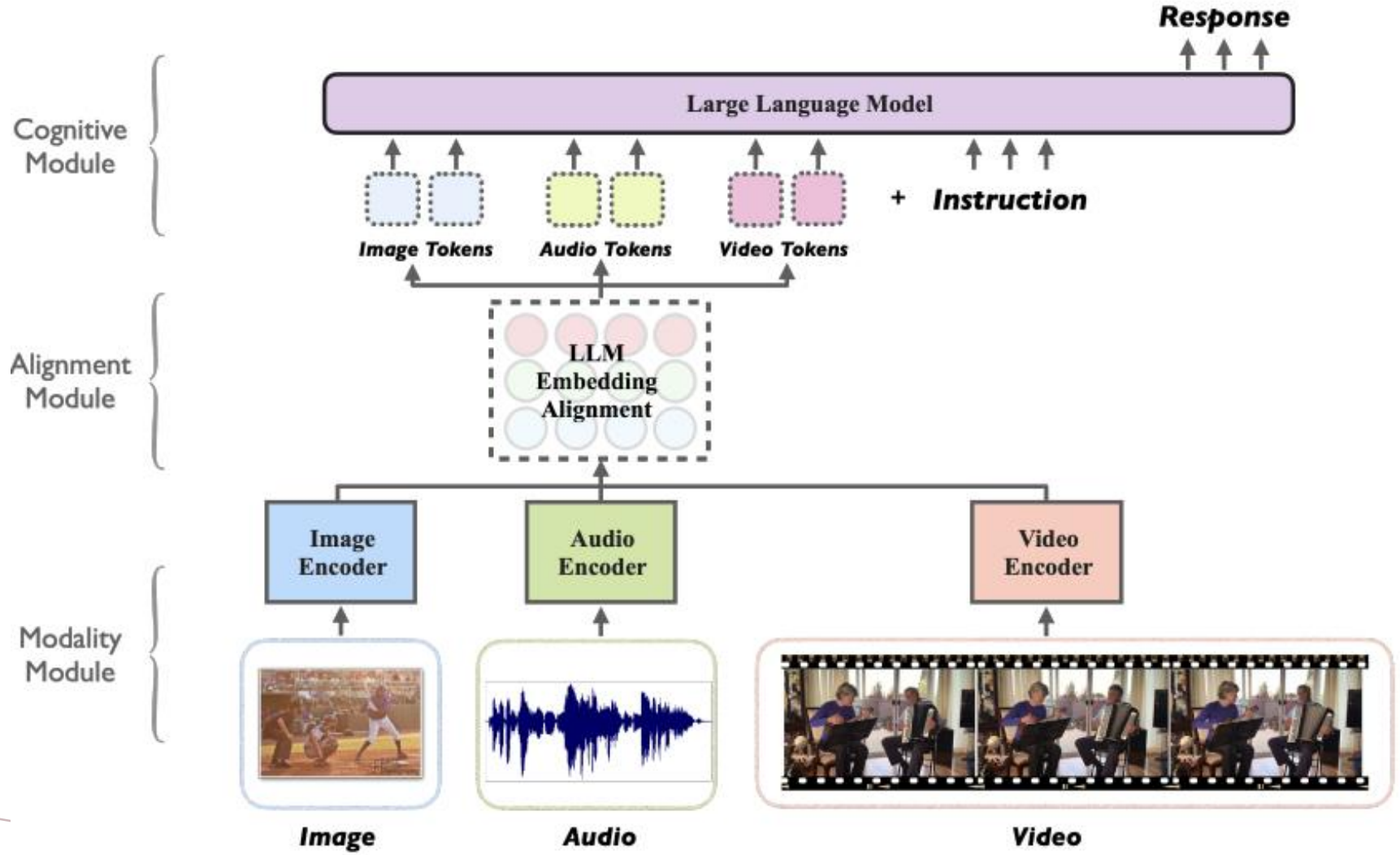


Figure 1: An overview of MACAW-LLM model architecture.

MULTIMODAL RACE

LLaVA: Large Language and Vision Assistant (Microsoft)

https://llava-vl.github.io/

BLIP-2: Scalable Pre-training of Multimodal Foundation Models for the World's First Open-source Multimodal Chatbot (Salesforce)

https://blog.salesforceairesearch.com/blip-2/

https://twitter.com/LiJunnan0409/status/1621649677543440384?ref=blog.salesforceairesearch.com

CONCLUSION

- Many multimodal models have been released simultaneously
- They are called LMMs
- Many of them are multimodal in a pipelined way
- Others are multimodal by design
- Some are Open Source
- Some use LLMs others use Embedding Extractors (Ex.: Clip, Dino)
- Pipeline - You can join different models to achieve your task (Ex.: input image to BLIP2 that outputs a text that is input to Stable Diffusion to generate a new image based on the caption)
- Tasks can be accomplished using Agents

MULTIMODAL FOUNDATION MODELS

THANK YOU!

ALBERTO ARKADER KOPILER
akopiler@gmail.com

Credits: AIM – Analytics India Magazine