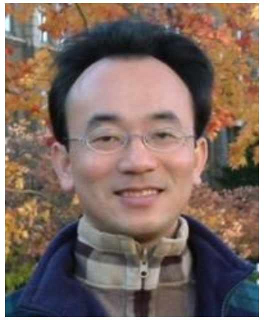
Generative AI for Multimodal Information Generation

1132GAIIA06
MBA, IM, NTPU (M6031) (Spring 2025)
Tue 2, 3, 4 (9:10-12:00) (B3F17)

https://meet.google.com/paj-zhhj-mya



Min-Yuh Day, Ph.D,
Professor
Institute of Information Management, National Taipei University
https://web.ntpu.edu.tw/~myday

2025-04-15

Syllabus

Week	Date	Subject/Topics
1	2025/02/18	Introduction to Generative AI Innovative Applications
2	2025/02/25	Transformers for Natural Language Processing and Computer Vision
3	2025/03/04	Large Language Models (LLMs), NVIDIA Building RAG Agents with LLMs Part I
4	2025/03/11	Case Study on Generative AI Innovative Applications I
5	2025/03/18	NVIDIA Building RAG Agents with LLMs Part II
6	2025/03/25	NVIDIA Building RAG Agents with LLMs Part III

Syllabus

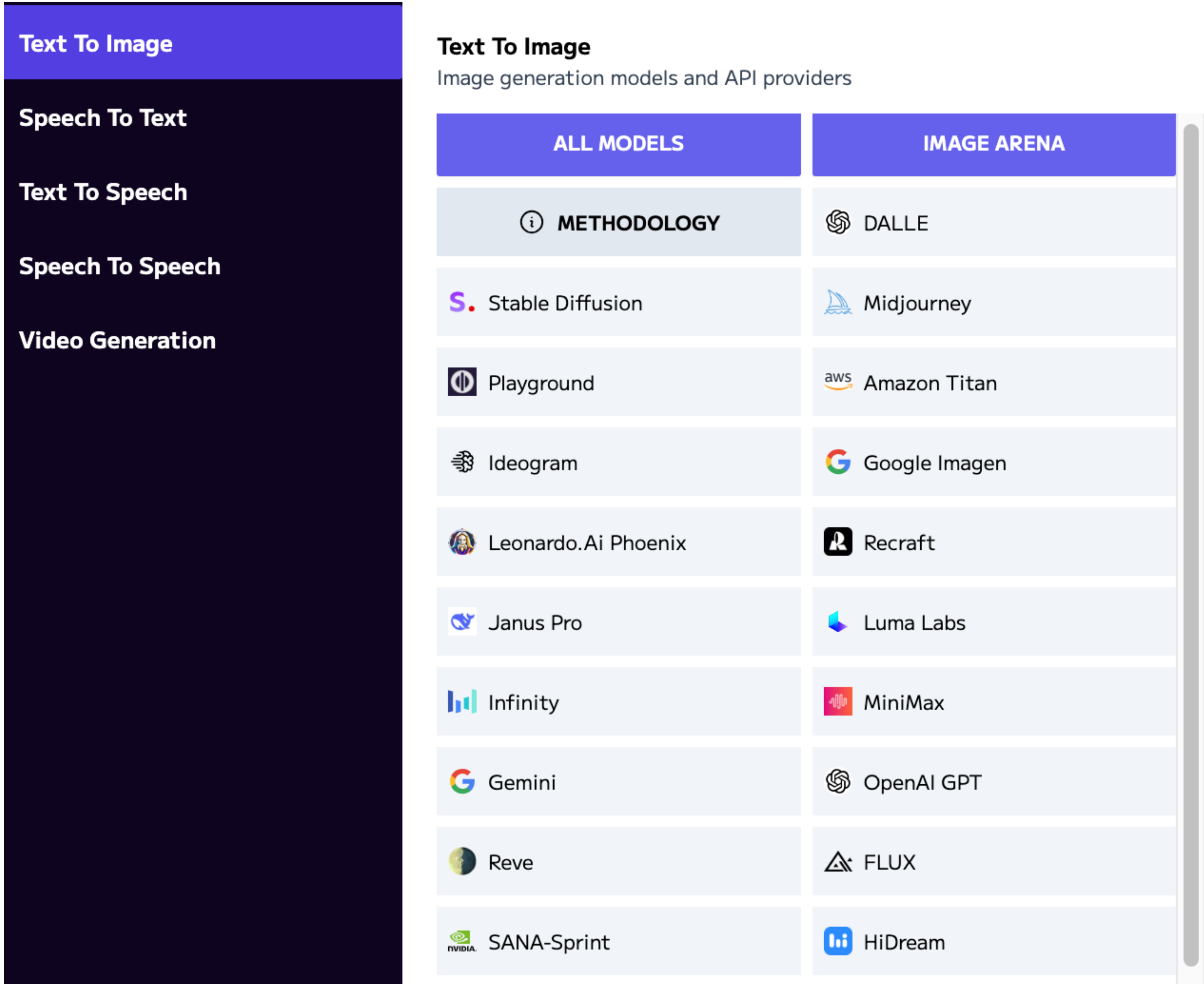
Week	Date	Subject/Topics
7	2025/04/01	Self-Learning
8	2025/04/08	Midterm Project Report
9	2025/04/15	Generative AI for Multimodal Information Generation
10	2025/04/22	NVIDIA Generative AI with Diffusion Models Part I
11	2025/04/29	NVIDIA Generative AI with Diffusion Models Part II
12	2025/05/06	Case Study on Generative AI Innovative Applications II

Syllabus

Week	Date	Subject/Topics
13	2025/05/13	NVIDIA Generative AI with Diffusion Models Part III
14	2025/05/20	AI Agents and Large Multimodal Agents (LMAs)
15	2025/05/27	Final Project Report I
16	2025/06/03	Final Project Report II

Generative AI for Multimodal Information Generation

Language Models
Text Image Speech Video Models



- Text To Image
- Speech To Text
- Text To Speech
- Speech To Speech
- Video Generation

Text To Image

Image generation models and API providers

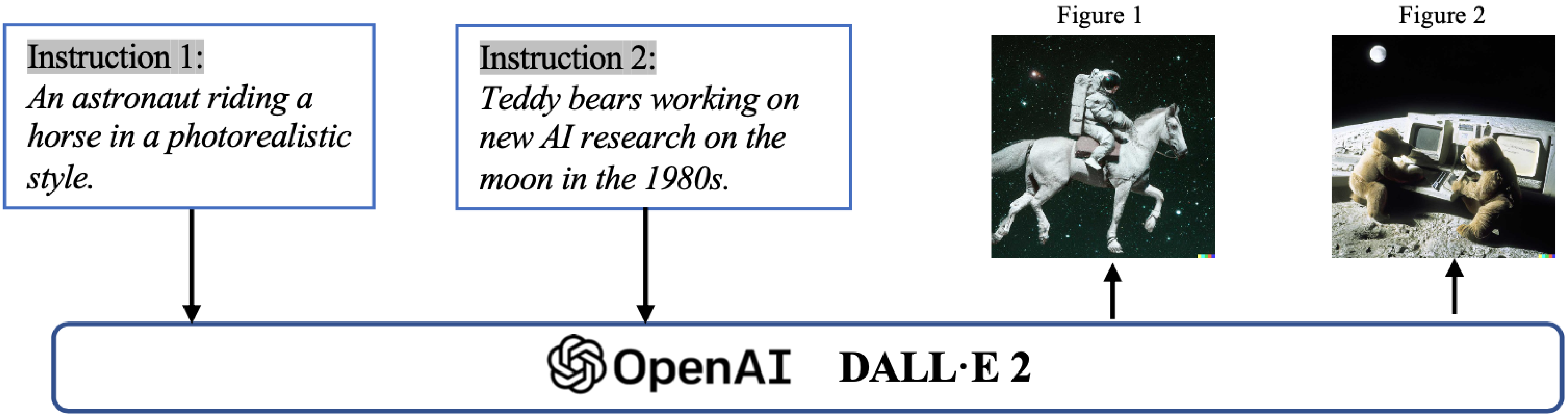
ALL MODELS	IMAGE ARENA
ⓘ METHODOLOGY	DALLE
Stable Diffusion	Midjourney
Playground	Amazon Titan
Ideogram	Google Imagen
Leonardo.Ai Phoenix	Recraft
Janus Pro	Luma Labs
Infinity	MiniMax
Gemini	OpenAI GPT
Reve	FLUX
SANA-Sprint	HiDream

https://artificialanalysis.ai/

Generative AI (Gen AI)

AI Generated Content (AIGC)

Image Generation



Instruction 1:
An astronaut riding a horse in a photorealistic style.

Instruction 2:
Teddy bears working on new AI research on the moon in the 1980s.

Figure 1

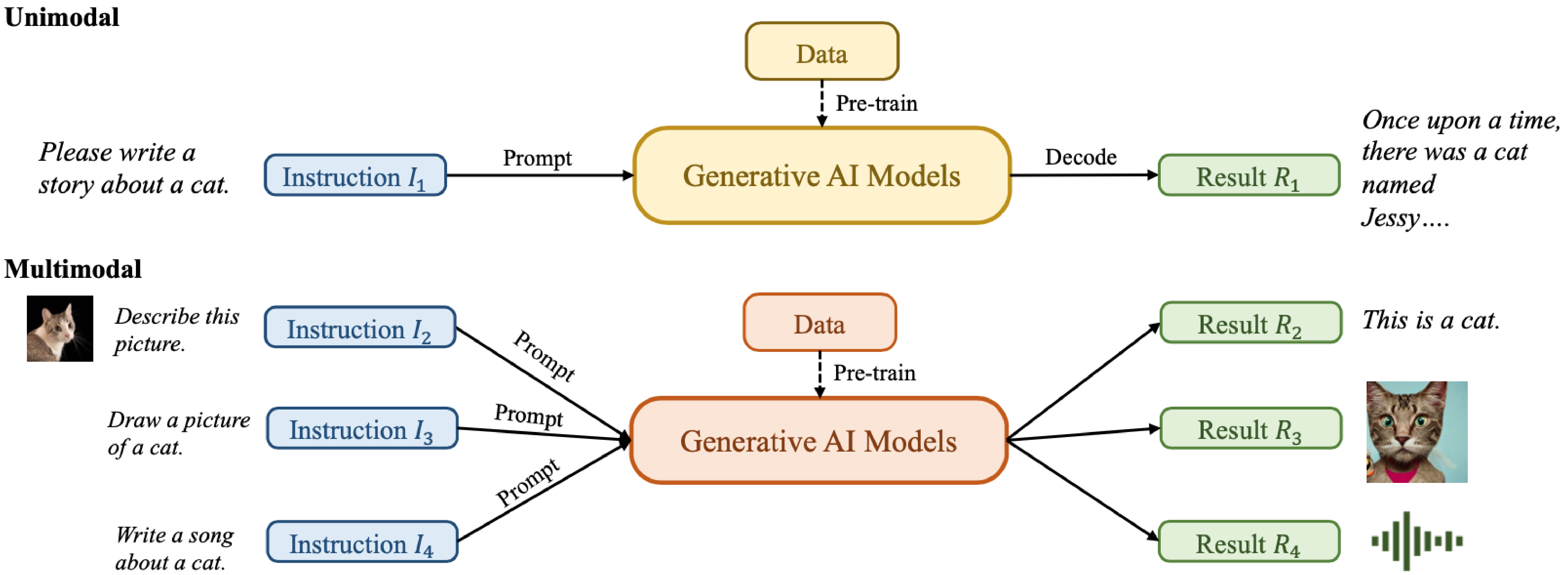
Figure 2

OpenAI DALL·E 2

Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

Generative AI (Gen AI)

AI Generated Content (AIGC)



Unimodal

Please write a story about a cat.

Instruction I_1

Prompt

Data

Pre-train

Generative AI Models

Decode

Result R_1

Once upon a time, there was a cat named Jessy….

Multimodal

Describe this picture.

Instruction I_2

Prompt

Draw a picture of a cat.

Instruction I_3

Prompt

Write a song about a cat.

Instruction I_4

Prompt

Data

Pre-train

Generative AI Models

Result R_2

This is a cat.

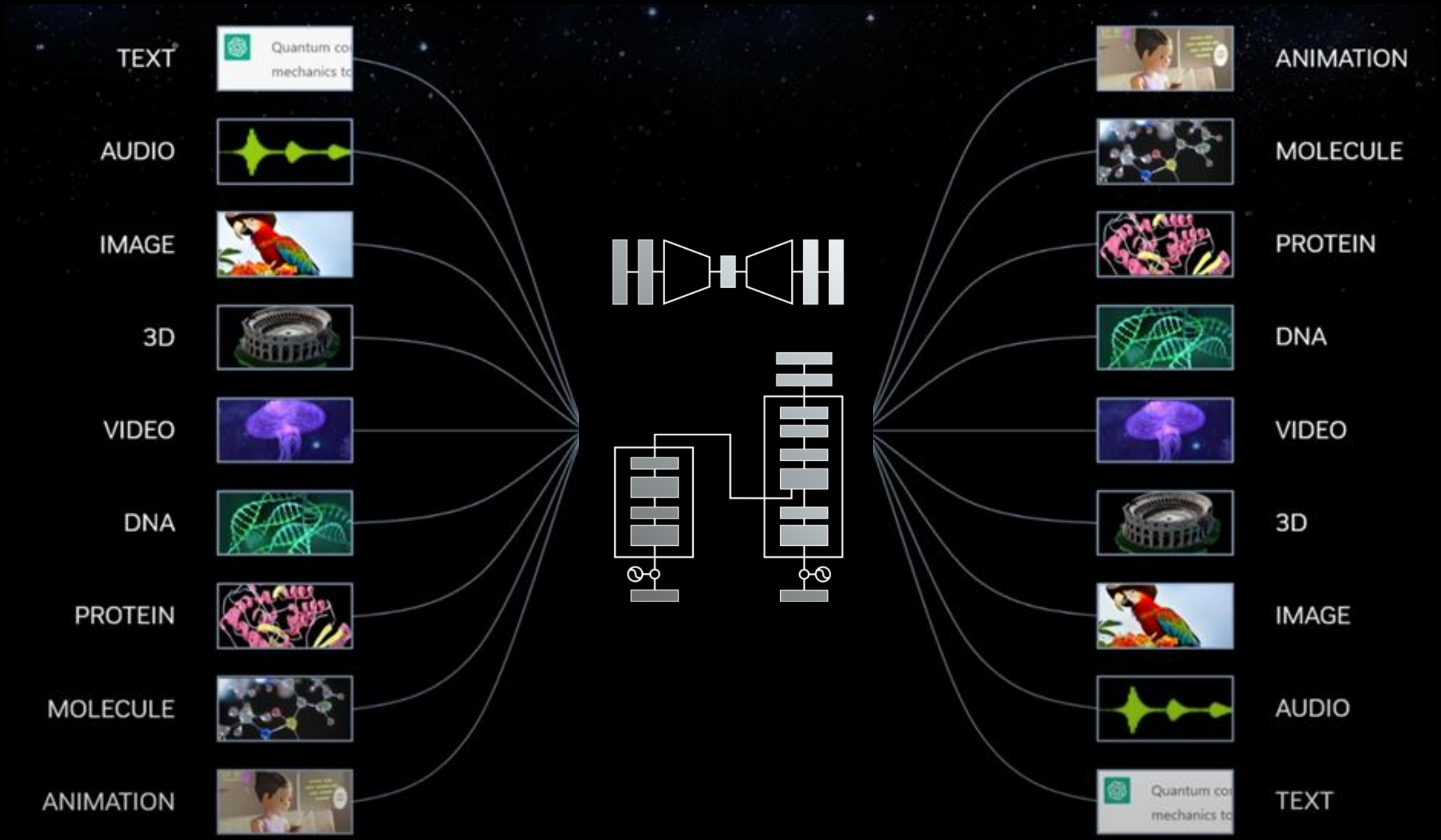
Result R_3

Result R_4

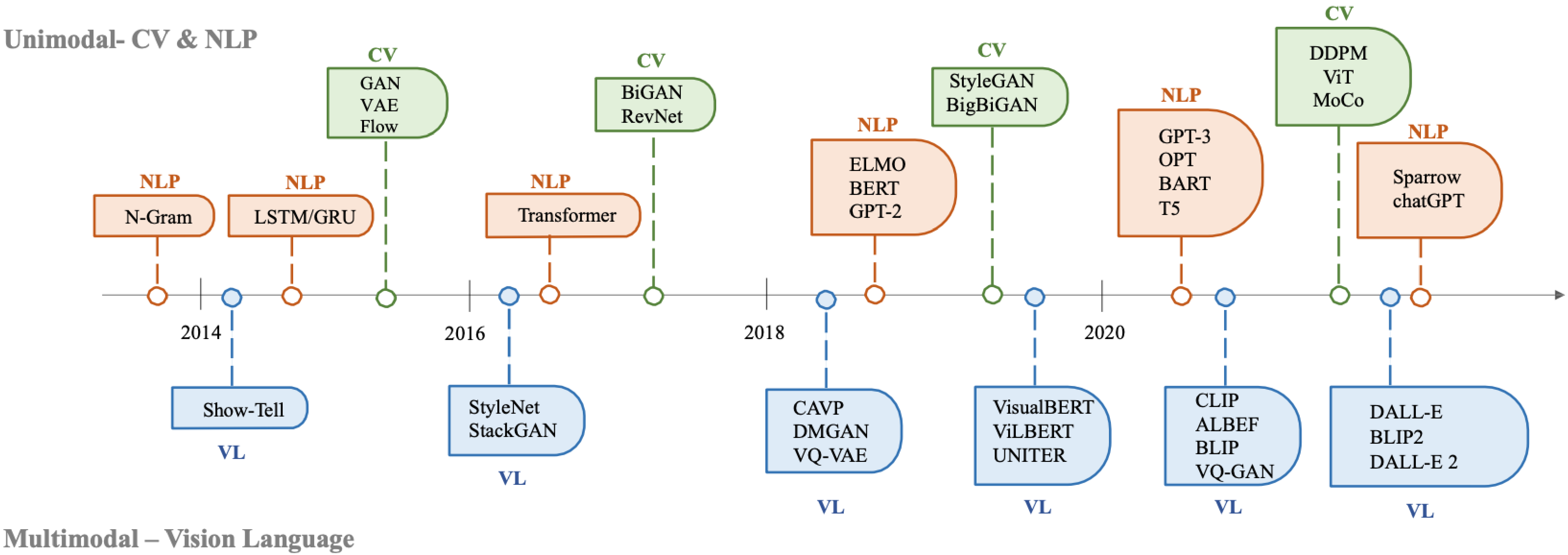
Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

Modular Modalities

Where Can The Transformer Fit?

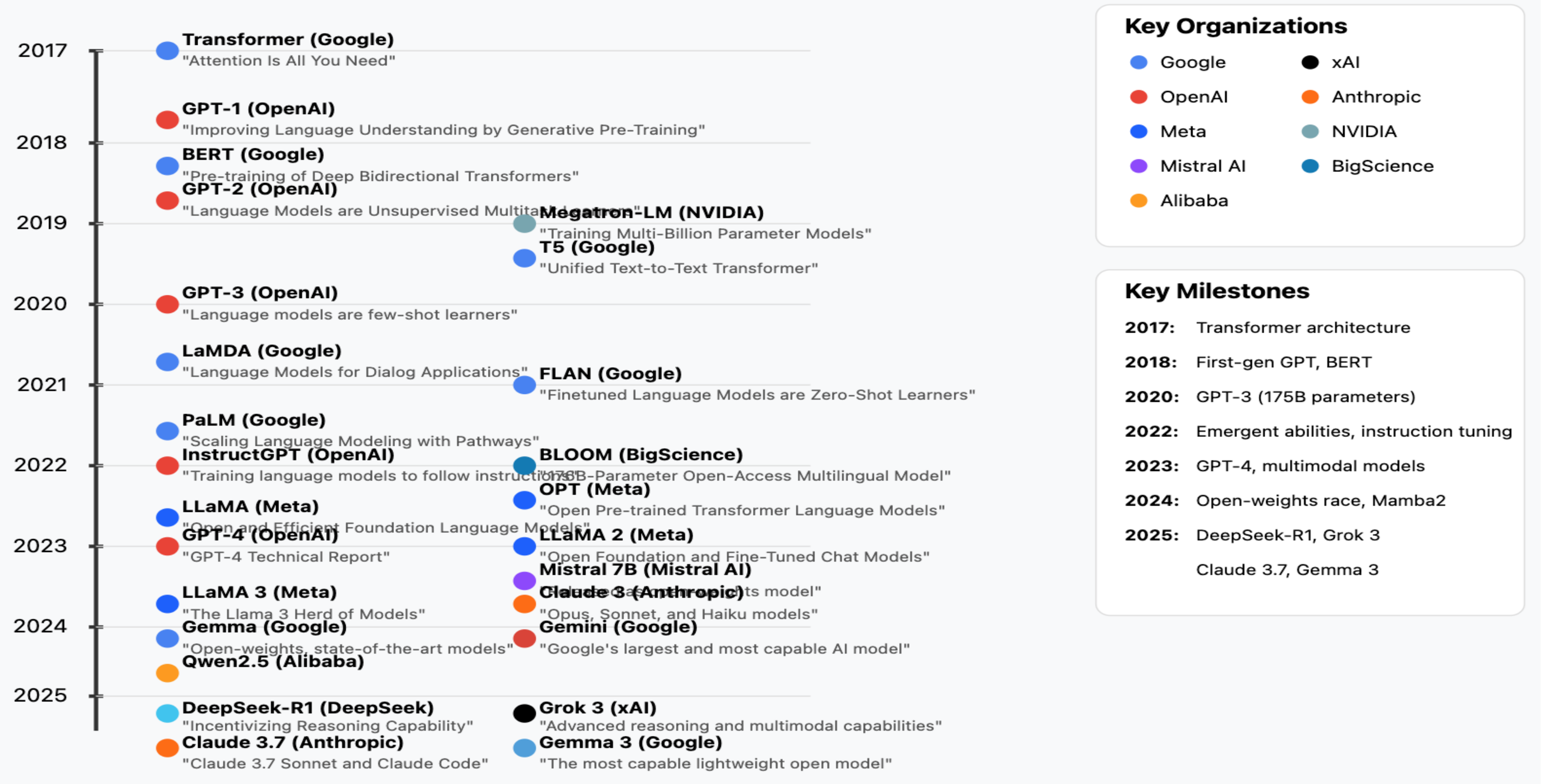


The history of Generative AI in CV, NLP and VL



Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

Generative AI LLMs (2017-2025)



Transformer (Google)
"Attention Is All You Need"

GPT-1 (OpenAI)
"Improving Language Understanding by Generative Pre-Training"

BERT (Google)
"Pre-training of Deep Bidirectional Transformers"

GPT-2 (OpenAI)
"Language Models are Unsupervised Multitask Learners"

Megatron-LM (NVIDIA)
"Training Multi-Billion Parameter Models"

T5 (Google)
"Unified Text-to-Text Transformer"

GPT-3 (OpenAI)
"Language models are few-shot learners"

LaMDA (Google)
"Language Models for Dialog Applications"

FLAN (Google)
"Finetuned Language Models are Zero-Shot Learners"

PaLM (Google)
"Scaling Language Modeling with Pathways"

InstructGPT (OpenAI)
"Training language models to follow instructions"

BLOOM (BigScience)
"176B-Parameter Open-Access Multilingual Model"

OPT (Meta)
"Open Pre-trained Transformer Language Models"

LLaMA (Meta)
"Open and Efficient Foundation Language Models"

GPT-4 (OpenAI)
"GPT-4 Technical Report"

LLaMA 2 (Meta)
"Open Foundation and Fine-Tuned Chat Models"

Mistral 7B (Mistral AI)
"Released as open-weights model"

Claude 3 (Anthropic)
"Opus, Sonnet, and Haiku models"

LLaMA 3 (Meta)
"The Llama 3 Herd of Models"

Gemma (Google)
"Open-weights, state-of-the-art models"

Gemini (Google)
"Google's largest and most capable AI model"

Qwen2.5 (Alibaba)

DeepSeek-R1 (DeepSeek)
"Incentivizing Reasoning Capability"

Grok 3 (xAI)
"Advanced reasoning and multimodal capabilities"

Claude 3.7 (Anthropic)
"Claude 3.7 Sonnet and Claude Code"

Gemma 3 (Google)
"The most capable lightweight open model"

Key Organizations

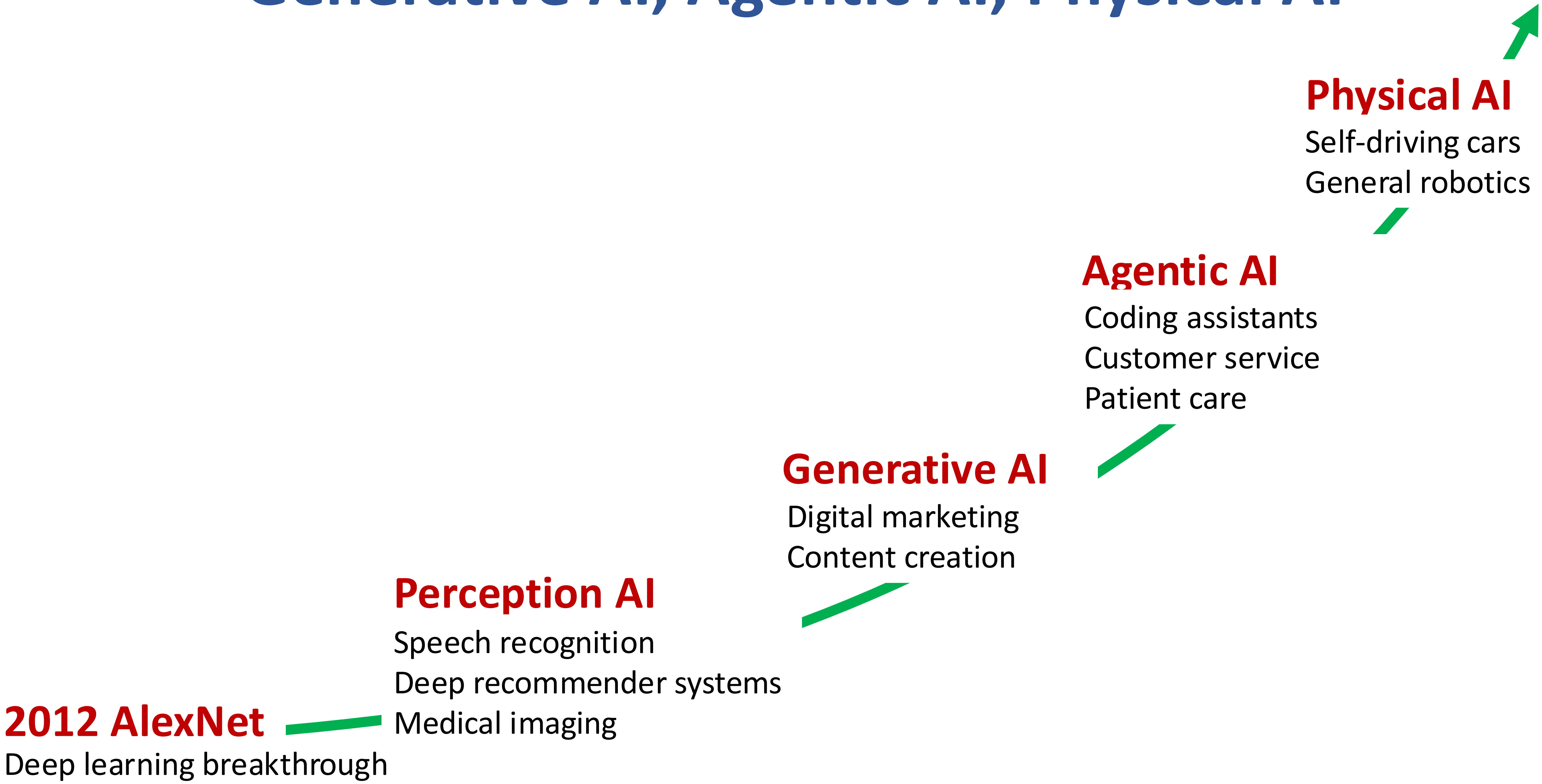
● Google	● xAI
● OpenAI	● Anthropic
● Meta	● NVIDIA
● Mistral AI	● BigScience
● Alibaba	

Key Milestones

2017:	Transformer architecture
2018:	First-gen GPT, BERT
2020:	GPT-3 (175B parameters)
2022:	Emergent abilities, instruction tuning
2023:	GPT-4, multimodal models
2024:	Open-weights race, Mamba2
2025:	DeepSeek-R1, Grok 3
	Claude 3.7, Gemma 3

Source: https://github.com/Hannibal046/Awesome-LLM

Generative AI, Agentic AI, Physical AI



2012 AlexNet

Deep learning breakthrough

Perception AI

Speech recognition
Deep recommender systems
Medical imaging

Generative AI

Digital marketing
Content creation

Agentic AI

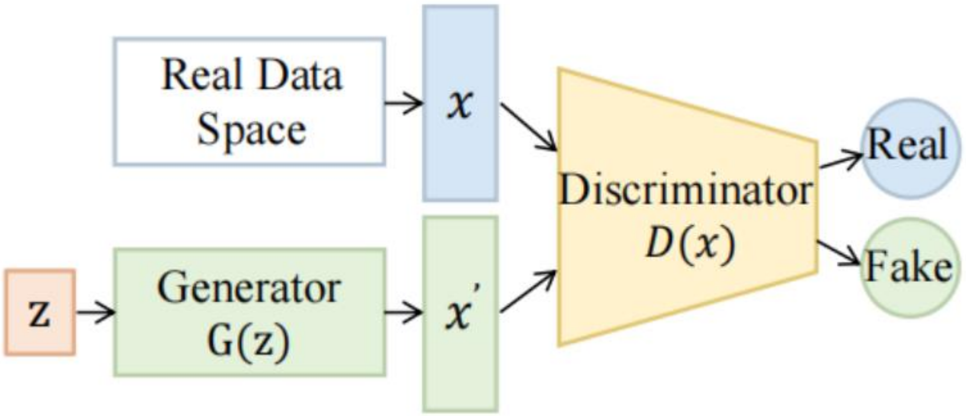
Coding assistants
Customer service
Patient care

Physical AI

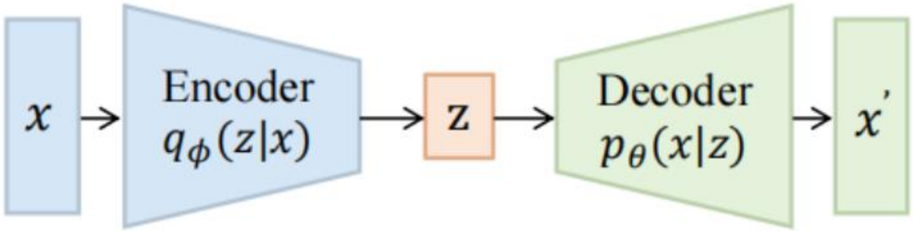
Self-driving cars
General robotics

Source: NVIDIA (2025), GTC March 2025 Keynote with NVIDIA CEO Jensen Huang, https://www.youtube.com/watch?v=_waPvOwL9Z8

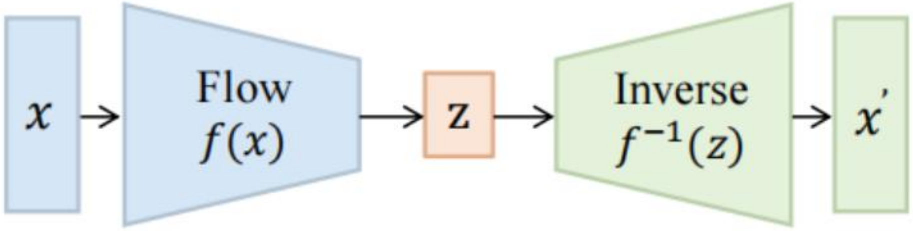
Categories of Vision Generative Models



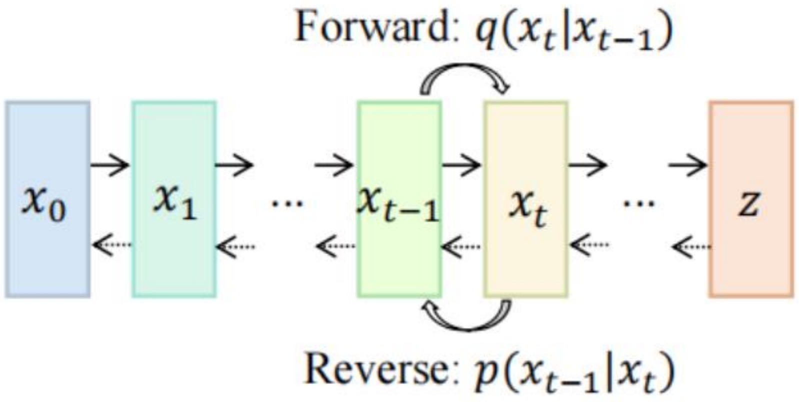
(1) Generative adversarial networks



(2) Variational autoencoders



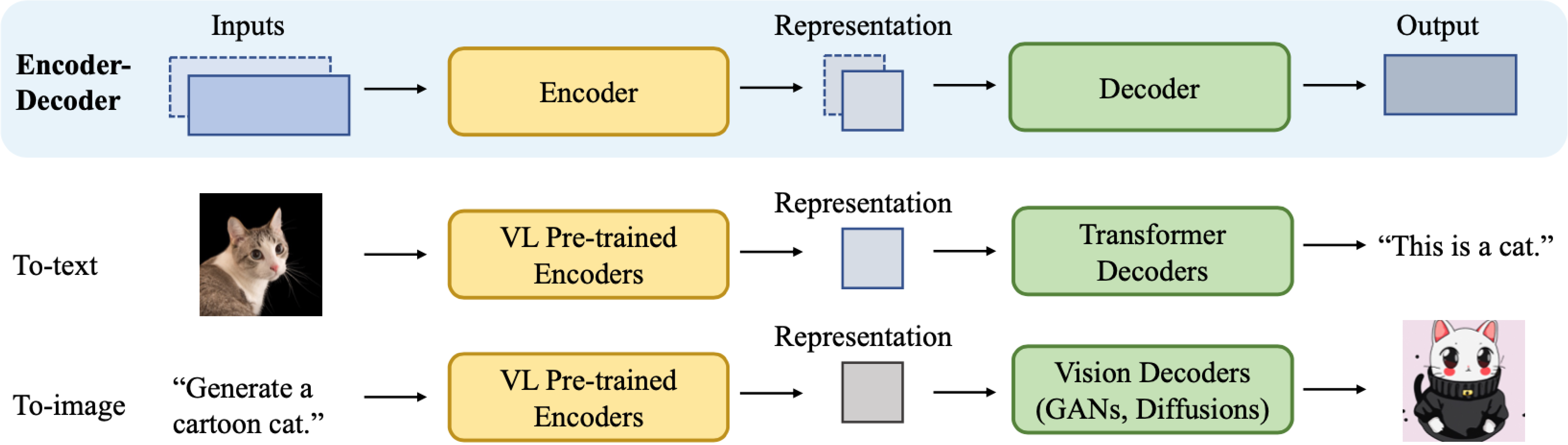
(3) Normalizing flows



(4) Diffusion models

Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

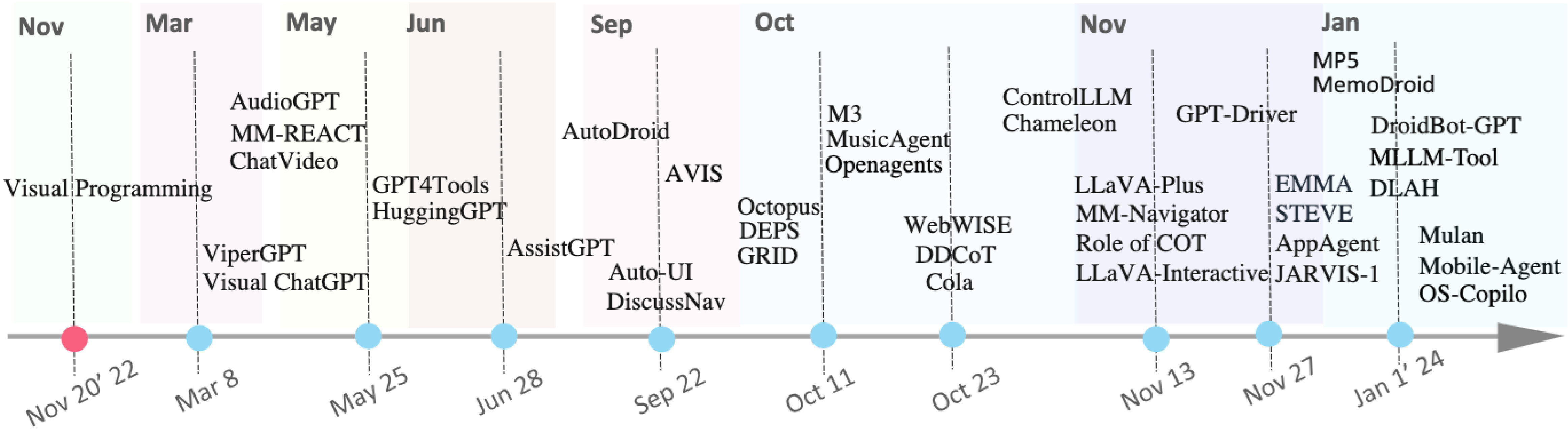
The General Structure of Generative Vision Language



Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

LLM-powered Multimodal Agents

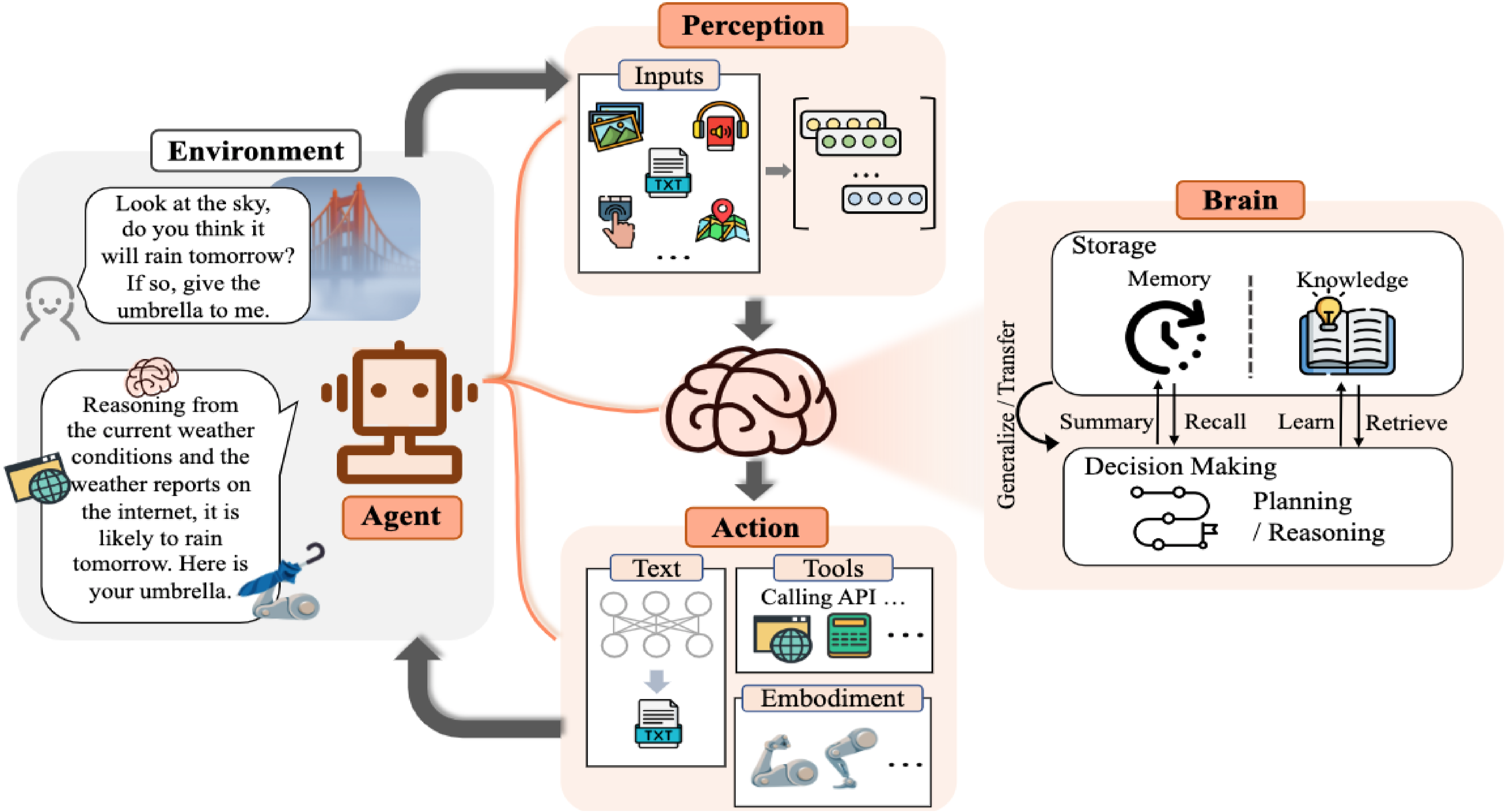
Large Multimodal Agents (LMAs)



2022 2023 2024

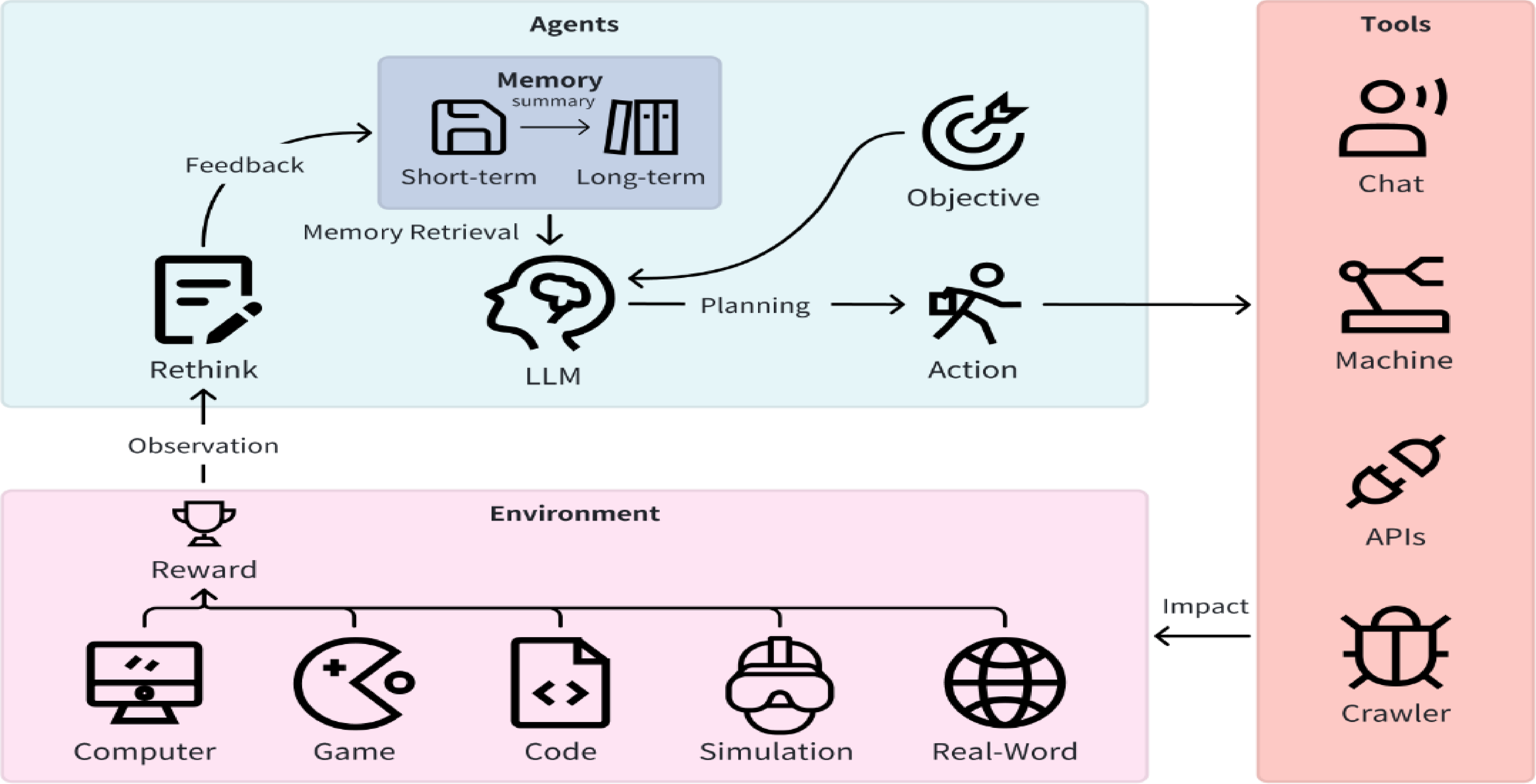
Source: Xie, Junlin, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. "Large Multimodal Agents: A Survey." arXiv preprint arXiv:2402.15116 (2024).

Large Language Model (LLM) based Agents


Perception
Inputs
TXT
...
Environment
Look at the sky, do you think it will rain tomorrow? If so, give the umbrella to me.
Reasoning from the current weather conditions and the weather reports on the internet, it is likely to rain tomorrow. Here is your umbrella.
Agent
Brain
Storage
Memory
Knowledge
Generalize / Transfer
Summary
Recall
Learn
Retrieve
Decision Making
Planning / Reasoning
Action
Text
Tools
Calling API ...
Embodiment
...

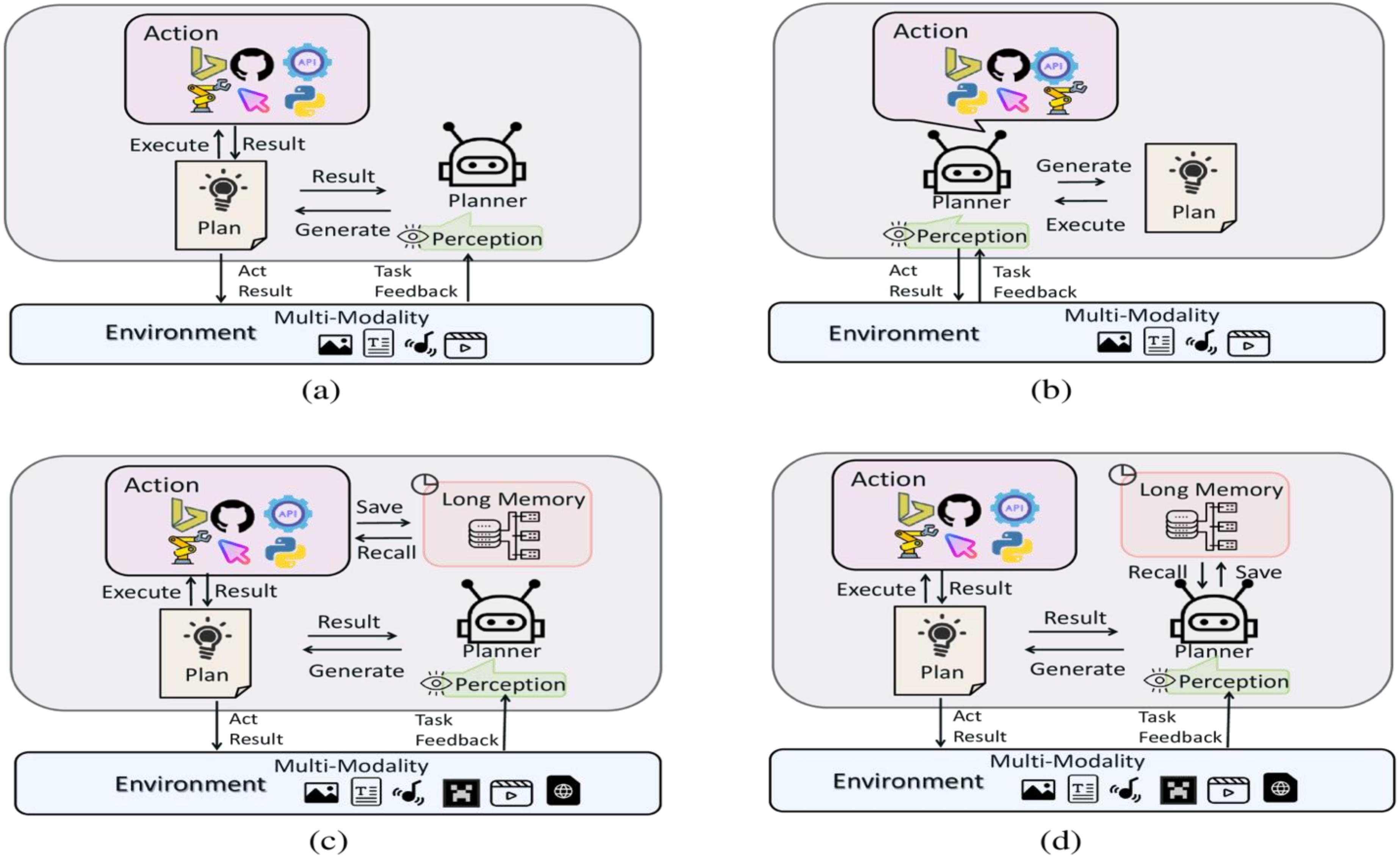
Source: Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.

LLM-based Agents



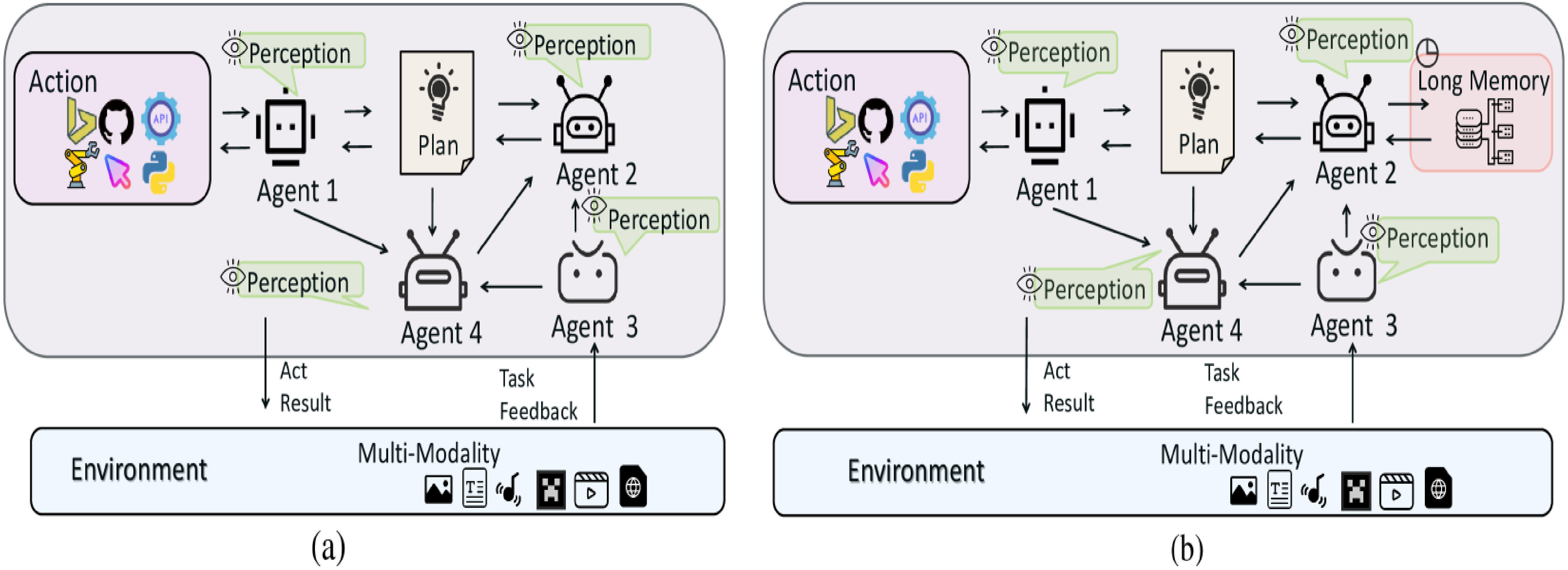
Source: Cheng, Yuheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang et al. "Exploring large language model based intelligent agents: Definitions, methods, and prospects." arXiv preprint arXiv:2401.03428 (2024).

Large Multimodal Agents (LMA)



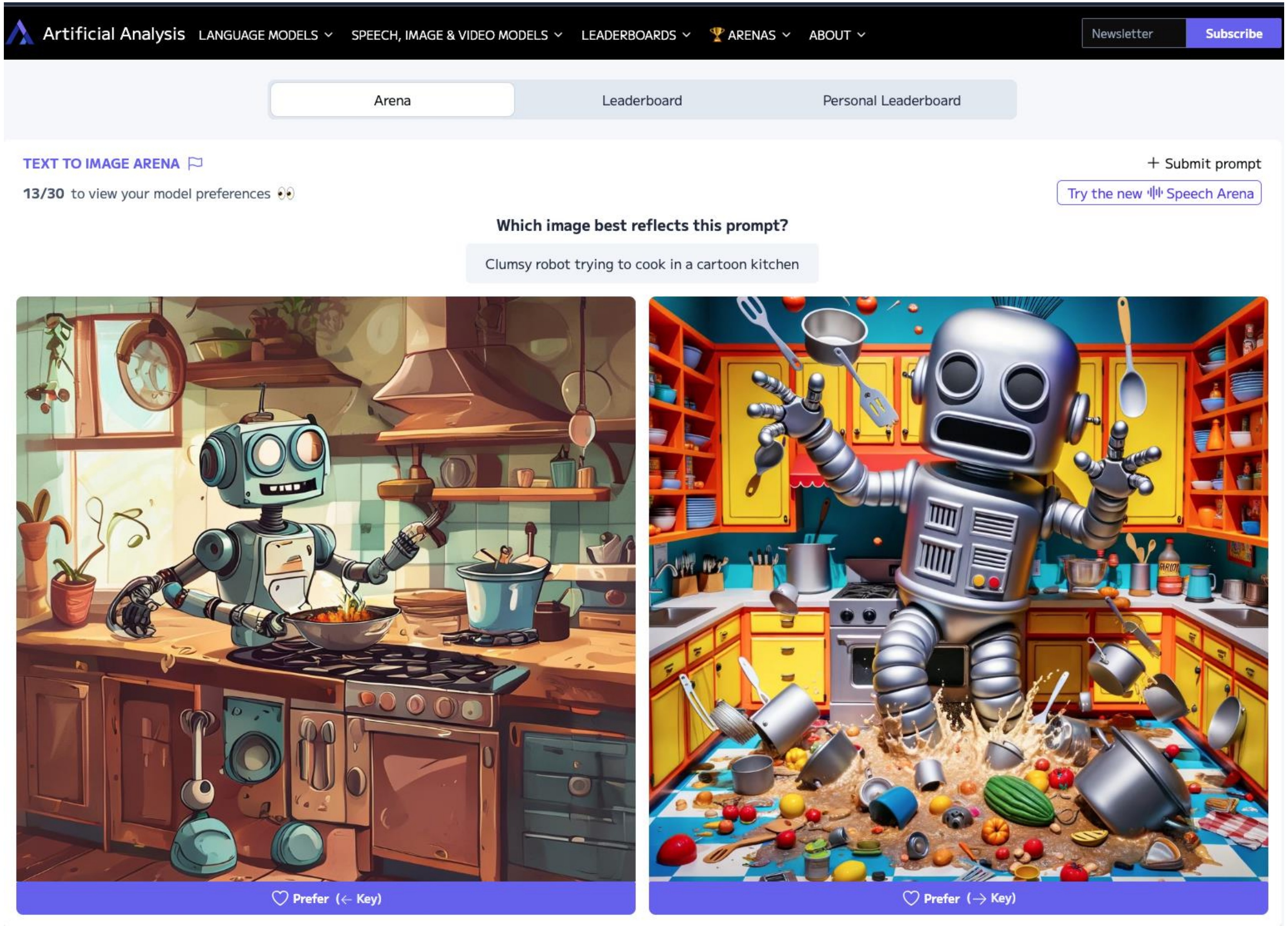
Source: Xie, J., Chen, Z., Zhang, R., Wan, X., & Li, G. (2024). Large Multimodal Agents: A Survey. ArXiv, abs/2402.15116.

Large Multimodal Agents (LMA)



Source: Xie, J., Chen, Z., Zhang, R., Wan, X., & Li, G. (2024). Large Multimodal Agents: A Survey. ArXiv, abs/2402.15116.

Artificial Analysis Text to Image Arena



TEXT TO IMAGE ARENA

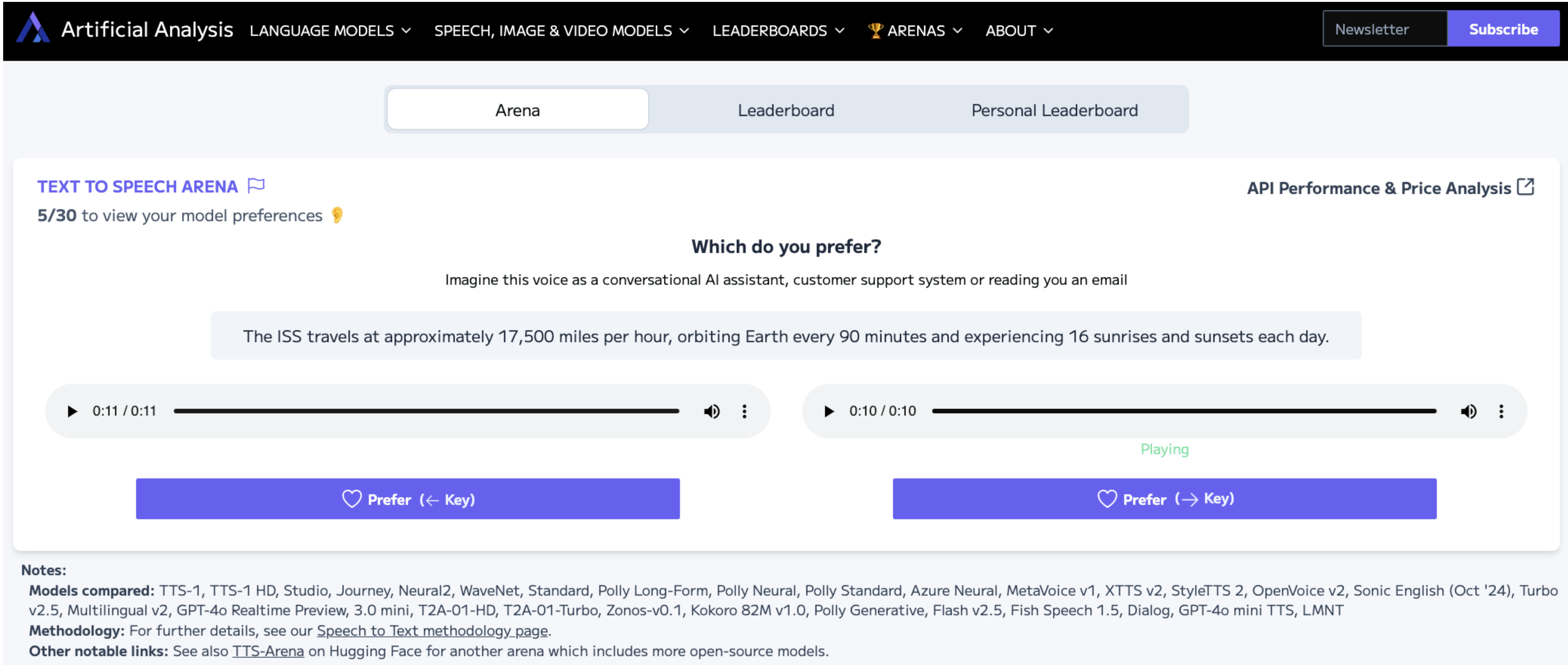
13/30 to view your model preferences 👀

Which image best reflects this prompt?

Clumsy robot trying to cook in a cartoon kitchen

Source: https://artificialanalysis.ai/text-to-image/arena

Artificial Analysis Text to Speech Arena



TEXT TO SPEECH ARENA

5/30 to view your model preferences 👂

API Performance & Price Analysis

Which do you prefer?

Imagine this voice as a conversational AI assistant, customer support system or reading you an email

The ISS travels at approximately 17,500 miles per hour, orbiting Earth every 90 minutes and experiencing 16 sunrises and sunsets each day.

Notes:

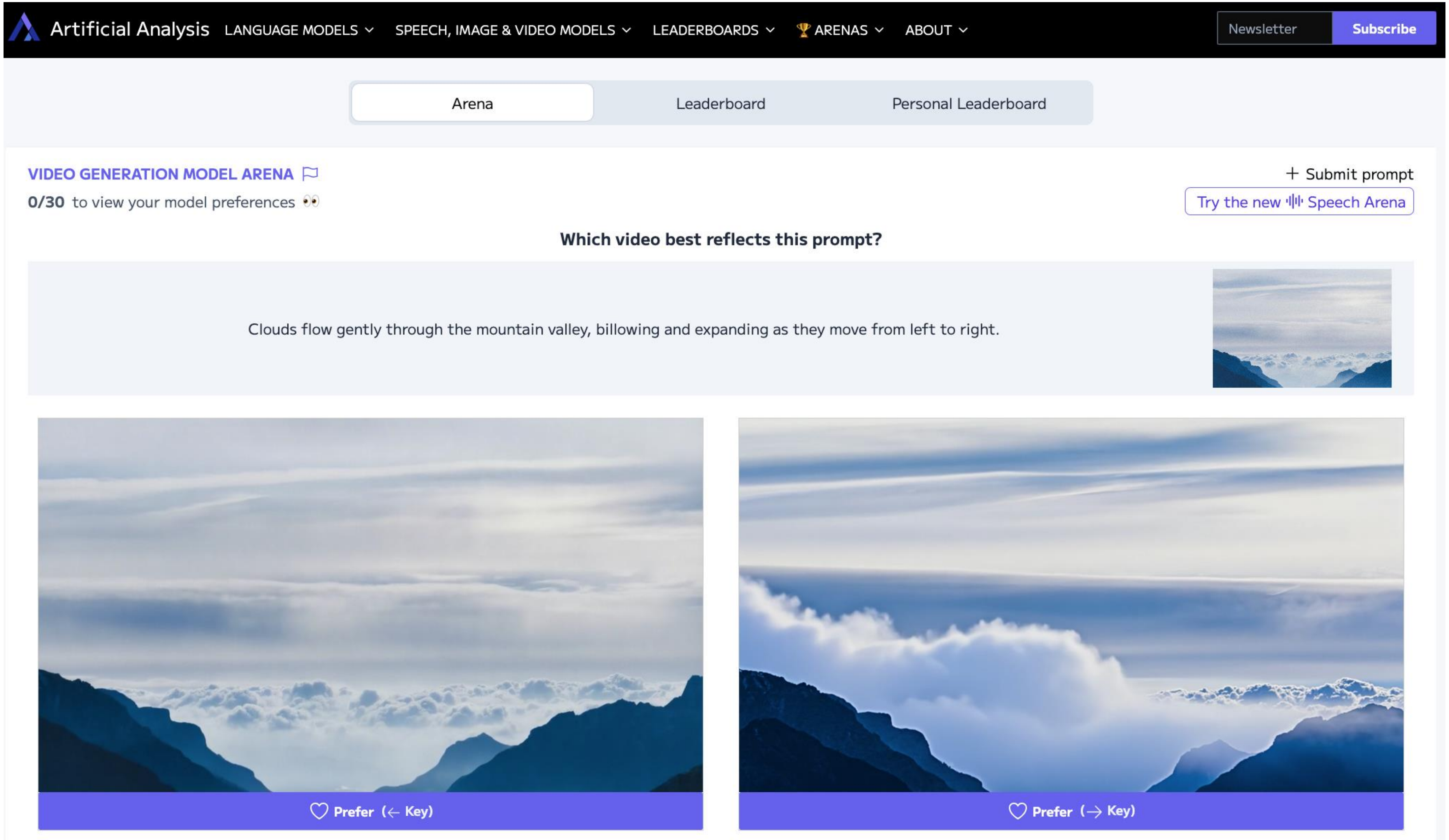
Models compared: TTS-1, TTS-1 HD, Studio, Journey, Neural2, WaveNet, Standard, Polly Long-Form, Polly Neural, Polly Standard, Azure Neural, MetaVoice v1, XTTS v2, StyleTTS 2, OpenVoice v2, Sonic English (Oct '24), Turbo v2.5, Multilingual v2, GPT-4o Realtime Preview, 3.0 mini, T2A-01-HD, T2A-01-Turbo, Zonos-v0.1, Kokoro 82M v1.0, Polly Generative, Flash v2.5, Fish Speech 1.5, Dialog, GPT-4o mini TTS, LMNT

Methodology: For further details, see our Speech to Text methodology page.

Other notable links: See also TTS-Arena on Hugging Face for another arena which includes more open-source models.

Source: https://artificialanalysis.ai/text-to-speech/arena

Artificial Analysis Video Generation Model Arena



VIDEO GENERATION MODEL ARENA

0/30 to view your model preferences 👀

+ Submit prompt

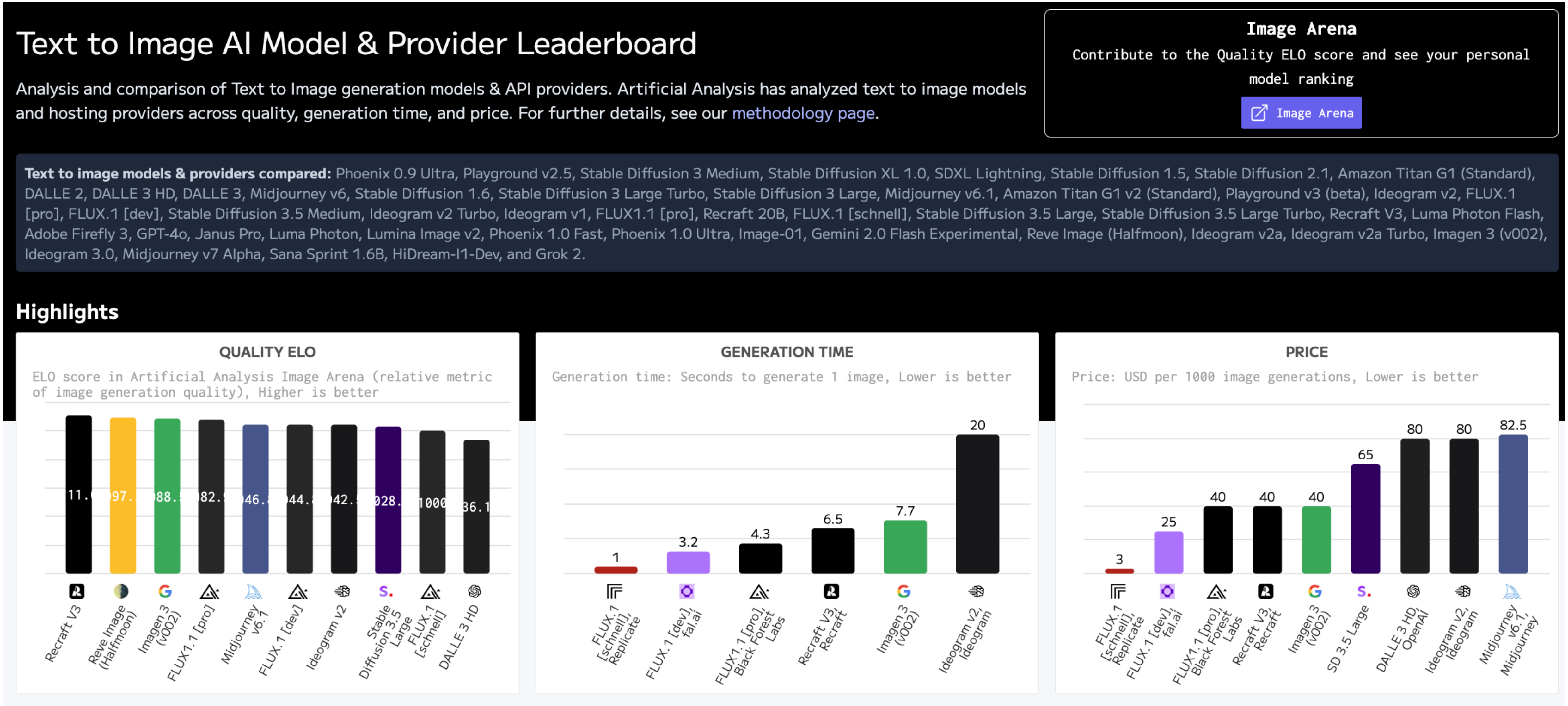
Try the new Speech Arena

Which video best reflects this prompt?

Clouds flow gently through the mountain valley, billowing and expanding as they move from left to right.

Source: https://artificialanalysis.ai/text-to-video/arena

Artificial Analysis Text to Image Leaderboard



Text to Image AI Model & Provider Leaderboard

Analysis and comparison of Text to Image generation models & API providers. Artificial Analysis has analyzed text to image models and hosting providers across quality, generation time, and price. For further details, see our methodology page.

Image Arena

Contribute to the Quality ELO score and see your personal model ranking

Image Arena

Text to image models & providers compared: Phoenix 0.9 Ultra, Playground v2.5, Stable Diffusion 3 Medium, Stable Diffusion XL 1.0, SDXL Lightning, Stable Diffusion 1.5, Stable Diffusion 2.1, Amazon Titan G1 (Standard), DALLE 2, DALLE 3 HD, DALLE 3, Midjourney v6, Stable Diffusion 1.6, Stable Diffusion 3 Large Turbo, Stable Diffusion 3 Large, Midjourney v6.1, Amazon Titan G1 v2 (Standard), Playground v3 (beta), Ideogram v2, FLUX.1 [pro], FLUX.1 [dev], Stable Diffusion 3.5 Medium, Ideogram v2 Turbo, Ideogram v1, FLUX1.1 [pro], Recraft 20B, FLUX.1 [schnell], Stable Diffusion 3.5 Large, Stable Diffusion 3.5 Large Turbo, Recraft V3, Luma Photon Flash, Adobe Firefly 3, GPT-4o, Janus Pro, Luma Photon, Lumina Image v2, Phoenix 1.0 Fast, Phoenix 1.0 Ultra, Image-01, Gemini 2.0 Flash Experimental, Reve Image (Halfmoon), Ideogram v2a, Ideogram v2a Turbo, Imagen 3 (v002), Ideogram 3.0, Midjourney v7 Alpha, Sana Sprint 1.6B, HiDream-I1-Dev, and Grok 2.

Highlights

QUALITY ELO

ELO score in Artificial Analysis Image Arena (relative metric of image generation quality), Higher is better

GENERATION TIME

Generation time: Seconds to generate 1 image, Lower is better

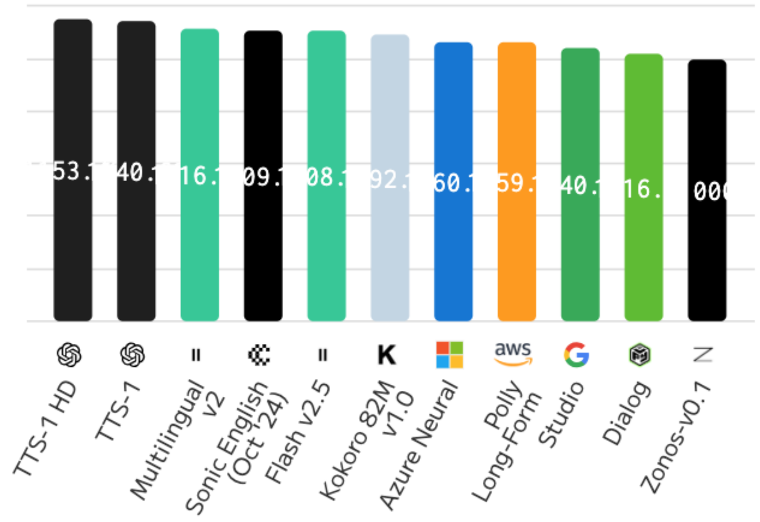
PRICE

Price: USD per 1000 image generations, Lower is better

Source: https://artificialanalysis.ai/text-to-image

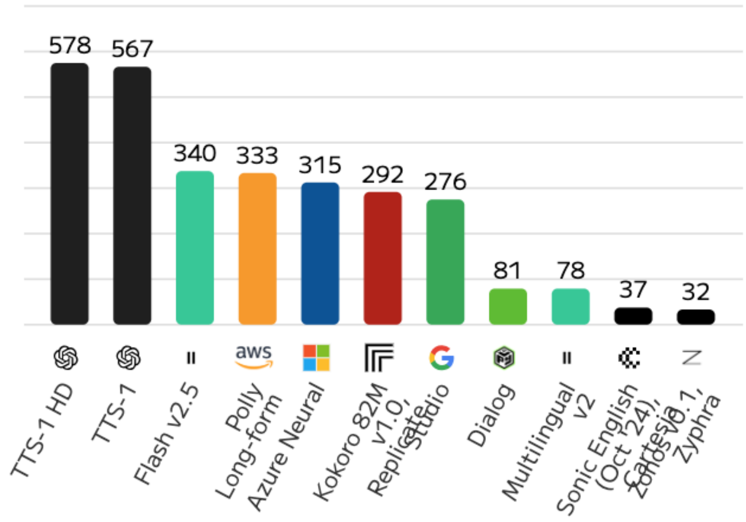
Text to Speech (TTS) AI Model & Provider Leaderboard

Text to speech models & providers compared: TTS-1, TTS-1 HD, Studio, Journey, Neural2, WaveNet, Standard, Polly Long-Form, Polly Neural, Polly Standard, Azure Neural, MetaVoice v1, XTTS v2, StyleTTS 2, OpenVoice v2, Sonic English (Oct '24), 3.0 mini, Turbo v2.5, Multilingual v2, T2A-01-HD, T2A-01-Turbo, Zonos-v0.1, Kokoro 82M v1.0, Polly Generative, Flash v2.5, Dialog, Murf Speech Gen 2, and Step TTS Mini.



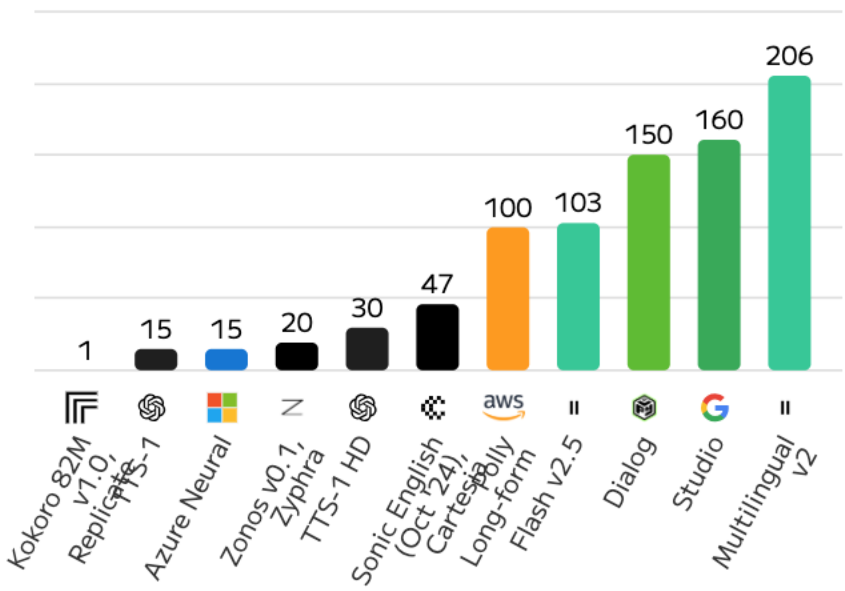
QUALITY ELO

Arena ELO: Average ELO rating of the model, Higher is better



CHARACTERS PER SECOND

Characters processed per second: # of characters per second of generation time, Higher is better



PRICE

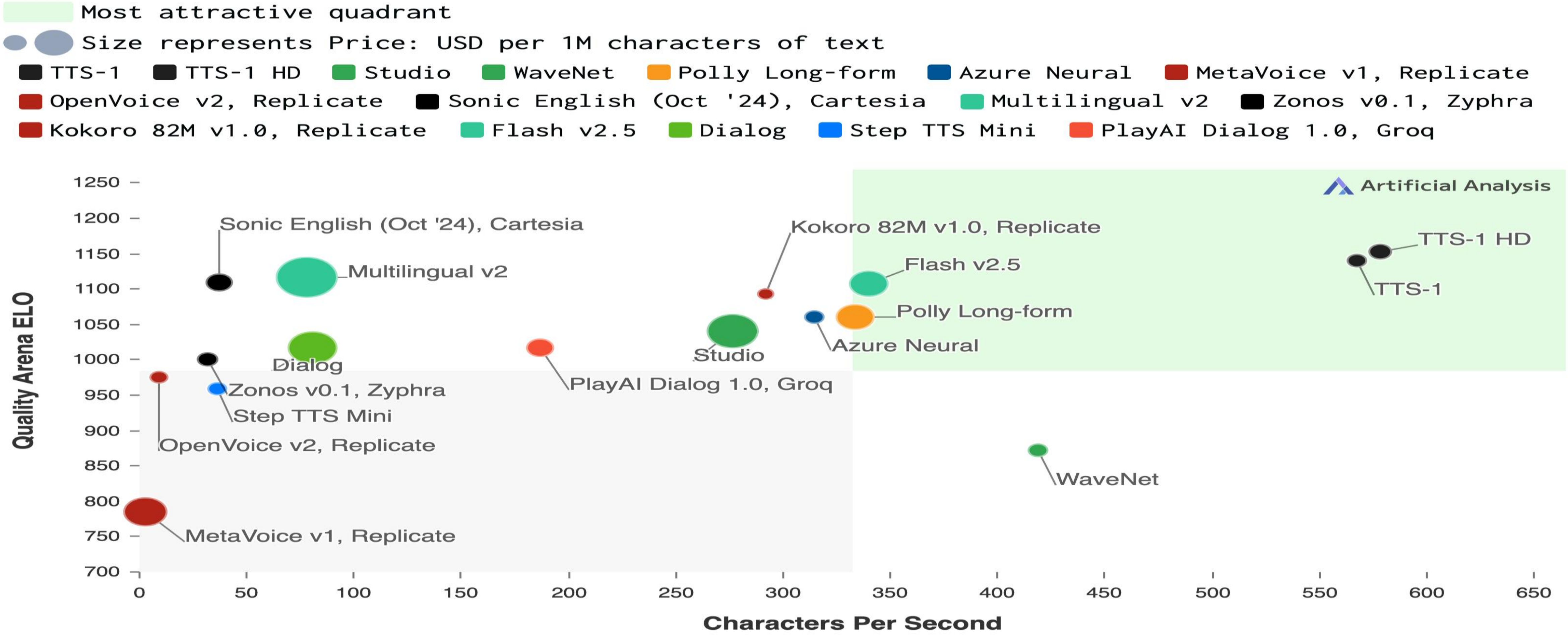
Price: USD per 1M characters of text, Lower is better

Source: https://artificialanalysis.ai/text-to-speech

Text to Speech (TTS) AI Model & Provider Leaderboard

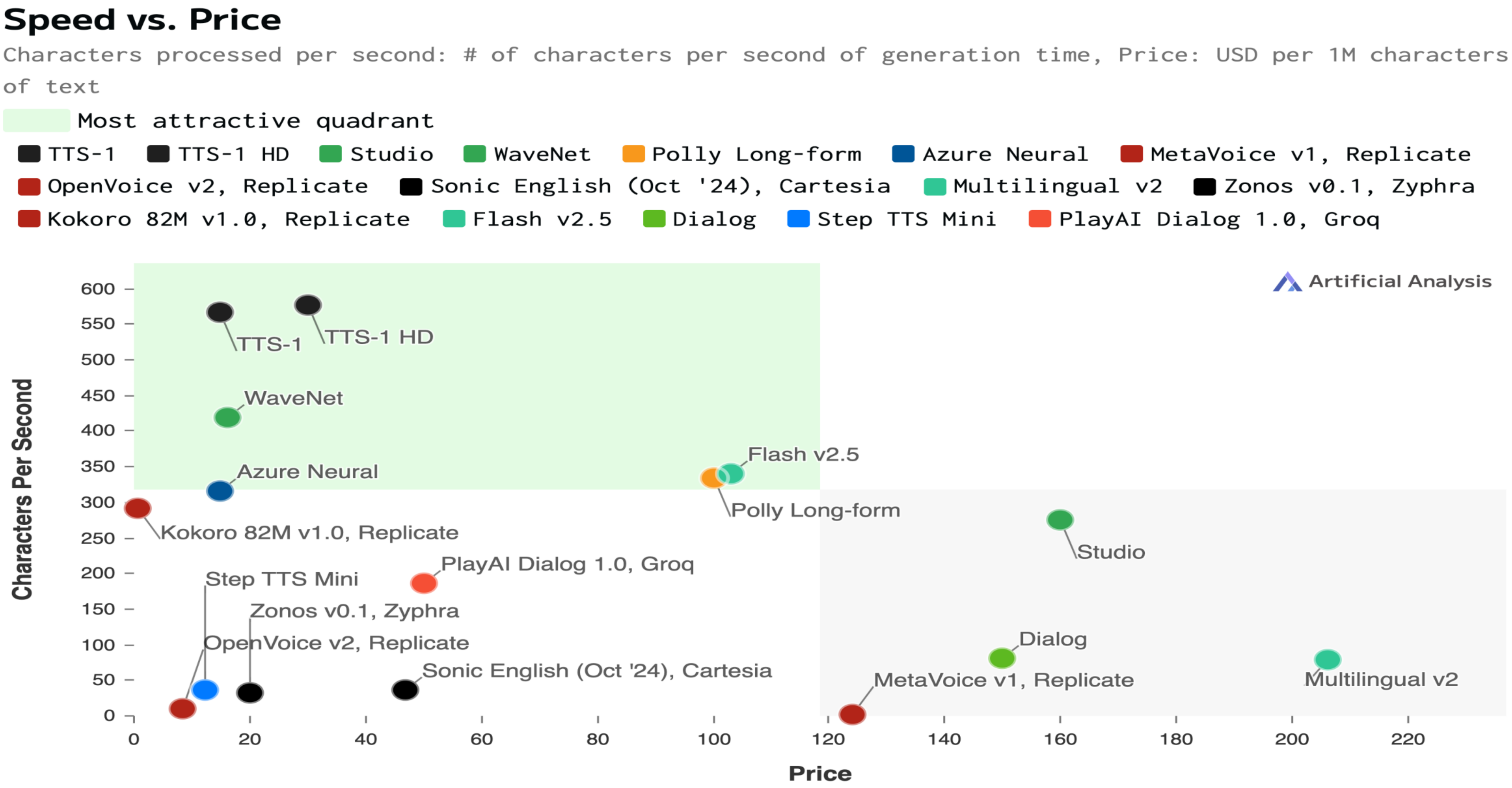
Quality vs. Speed

Arena ELO: Average ELO rating of the model, Characters processed per second: # of characters per second of generation time



Source: https://artificialanalysis.ai/text-to-speech

Text to Speech (TTS) AI Model & Provider Leaderboard



Speed vs. Price

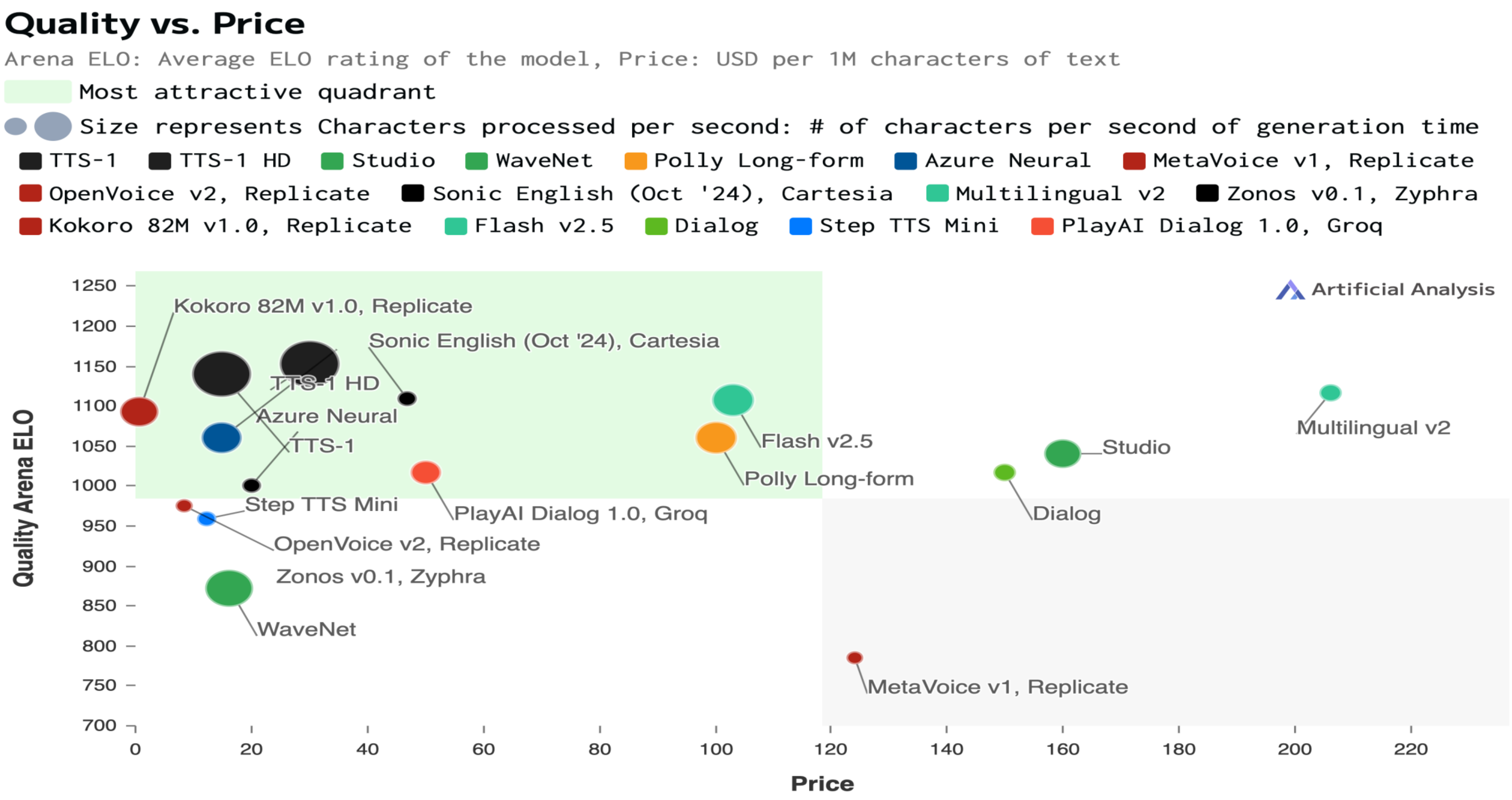
Characters processed per second: # of characters per second of generation time, Price: USD per 1M characters of text

🟩 Most attractive quadrant

⬛ TTS-1 ⬛ TTS-1 HD 🟩 Studio 🟩 WaveNet 🟧 Polly Long-form 🟦 Azure Neural 🟥 MetaVoice v1, Replicate
🟥 OpenVoice v2, Replicate ⬛ Sonic English (Oct '24), Cartesia 🟩 Multilingual v2 ⬛ Zonos v0.1, Zyphra
🟥 Kokoro 82M v1.0, Replicate 🟩 Flash v2.5 🟩 Dialog 🟦 Step TTS Mini 🟥 PlayAI Dialog 1.0, Groq

Source: https://artificialanalysis.ai/text-to-speech

Text to Speech (TTS) AI Model & Provider Leaderboard



Quality vs. Price

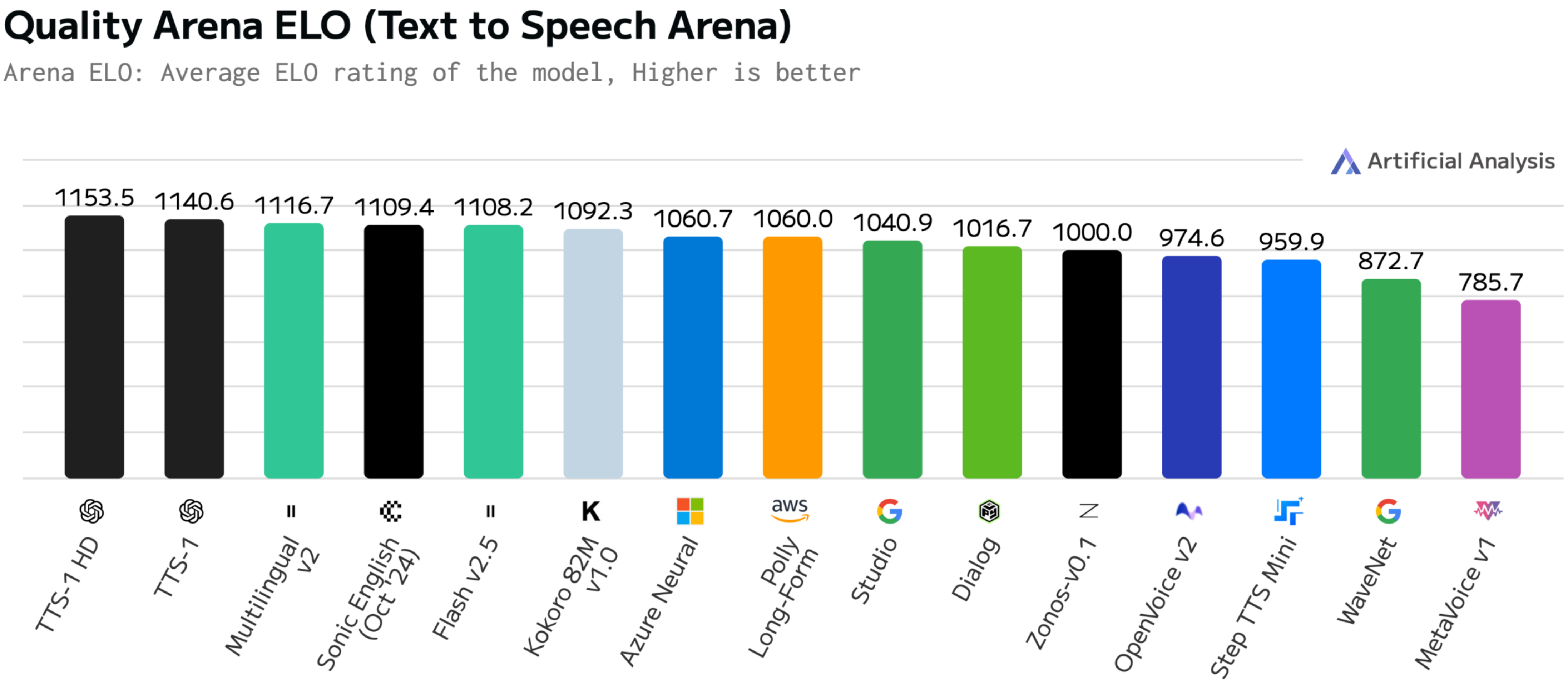
Arena ELO: Average ELO rating of the model, Price: USD per 1M characters of text

🟩 Most attractive quadrant

● ● Size represents Characters processed per second: # of characters per second of generation time

Source: https://artificialanalysis.ai/text-to-speech

Text to Speech (TTS) AI Model & Provider Leaderboard



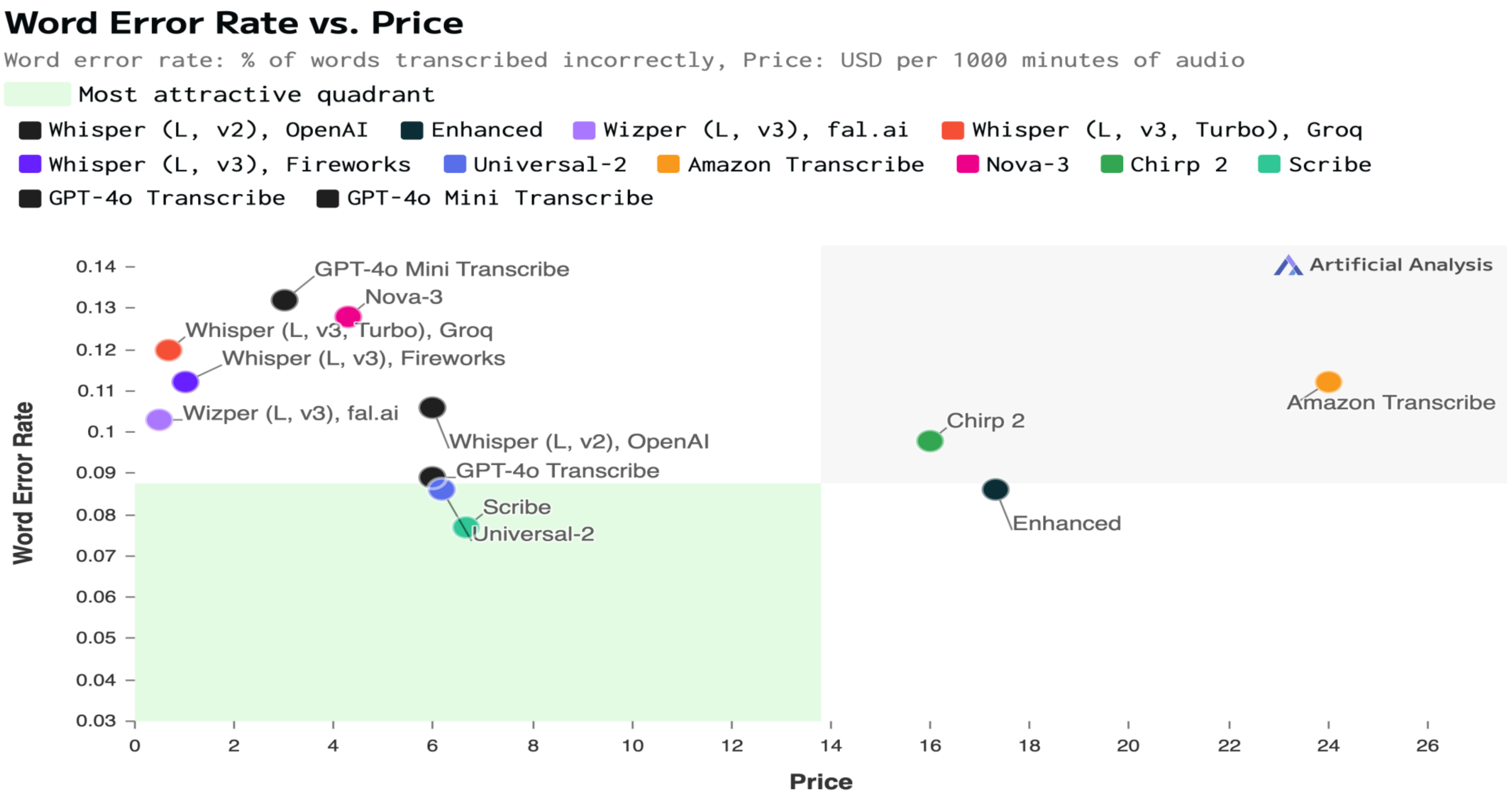
Quality Arena ELO (Text to Speech Arena)

Arena ELO: Average ELO rating of the model, Higher is better

Source: https://artificialanalysis.ai/text-to-speech

Speech to Text (STT) AI Model & Provider Leaderboard

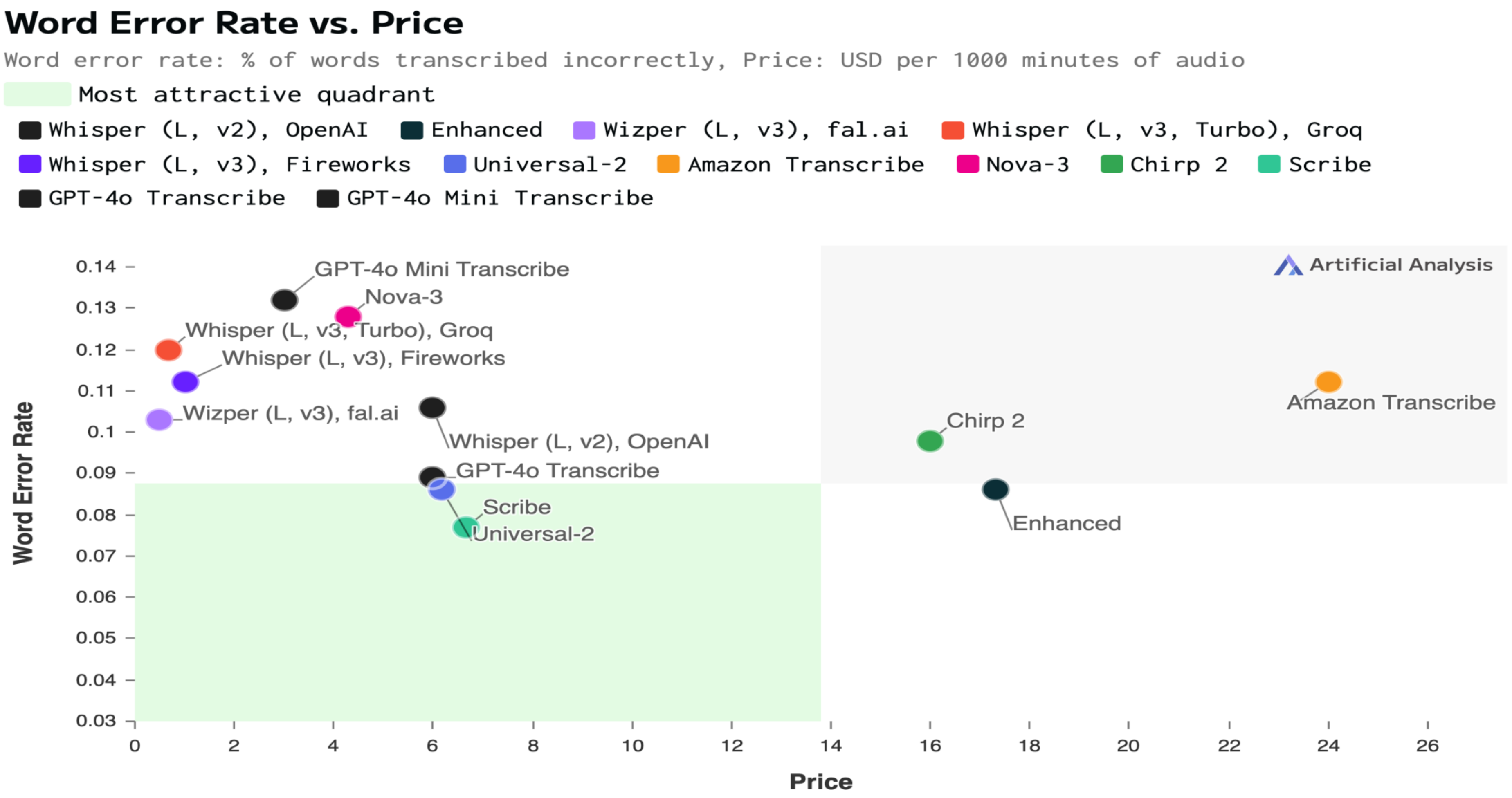
Speech-to-text models & providers compared: Whisper (L, v2), OpenAI, Universal-1, Standard, Whisper (L, v2), Azure, Enhanced, Nano, Wizper (L, v3), fal.ai, Incredibly Fast Whisper, Replicate, Nova-2, Whisper (L, v2), Replicate, Whisper (L, v3), Replicate, Base, WhisperX, Replicate, Whisper (L v2), Deepgram, Gladia, Whisper (L, v3), Groq, Distil-Whisper, Groq, Whisper (L, v3), fal.ai, Whisper (L, v3), Deepinfra, Whisper (L, v3, Turbo), Groq, Whisper (L, v3), Fireworks, Whisper (L, v3, Turbo), Fireworks, Universal-2, Amazon Transcribe, Fish Speech to Text, Nova-3, Chirp, Chirp 2, Scribe, GPT-4o Transcribe, and GPT-4o Mini Transcribe.



Source: https://artificialanalysis.ai/speech-to-text

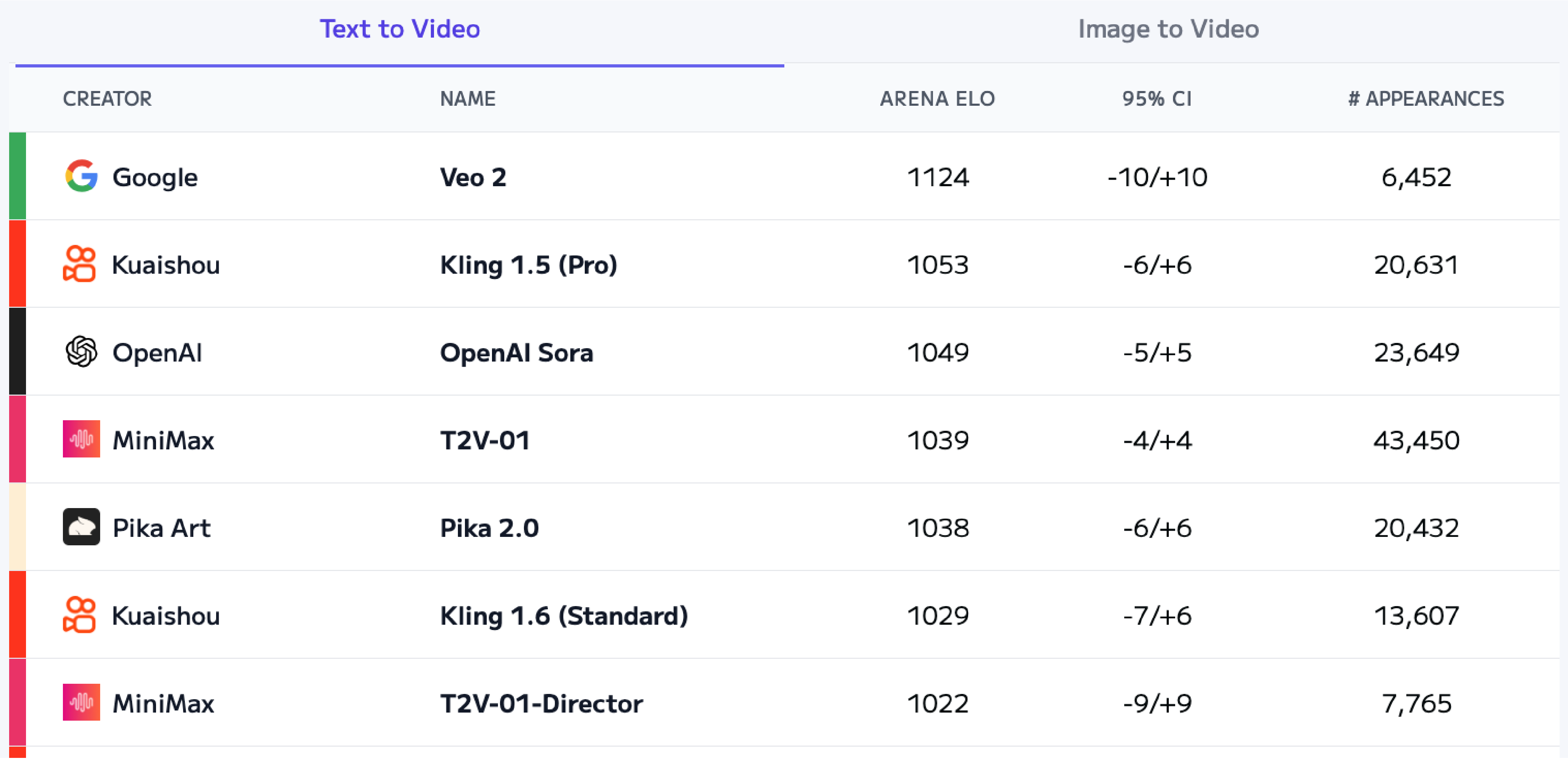
Speech to Text (STT) AI Model & Provider Leaderboard

Speech-to-text models & providers compared: Whisper (L, v2), OpenAI, Universal-1, Standard, Whisper (L, v2), Azure, Enhanced, Nano, Wizper (L, v3), fal.ai, Incredibly Fast Whisper, Replicate, Nova-2, Whisper (L, v2), Replicate, Whisper (L, v3), Replicate, Base, WhisperX, Replicate, Whisper (L v2), Deepgram, Gladia, Whisper (L, v3), Groq, Distil-Whisper, Groq, Whisper (L, v3), fal.ai, Whisper (L, v3), Deepinfra, Whisper (L, v3, Turbo), Groq, Whisper (L, v3), Fireworks, Whisper (L, v3, Turbo), Fireworks, Universal-2, Amazon Transcribe, Fish Speech to Text, Nova-3, Chirp, Chirp 2, Scribe, GPT-4o Transcribe, and GPT-4o Mini Transcribe.



Source: https://artificialanalysis.ai/speech-to-text

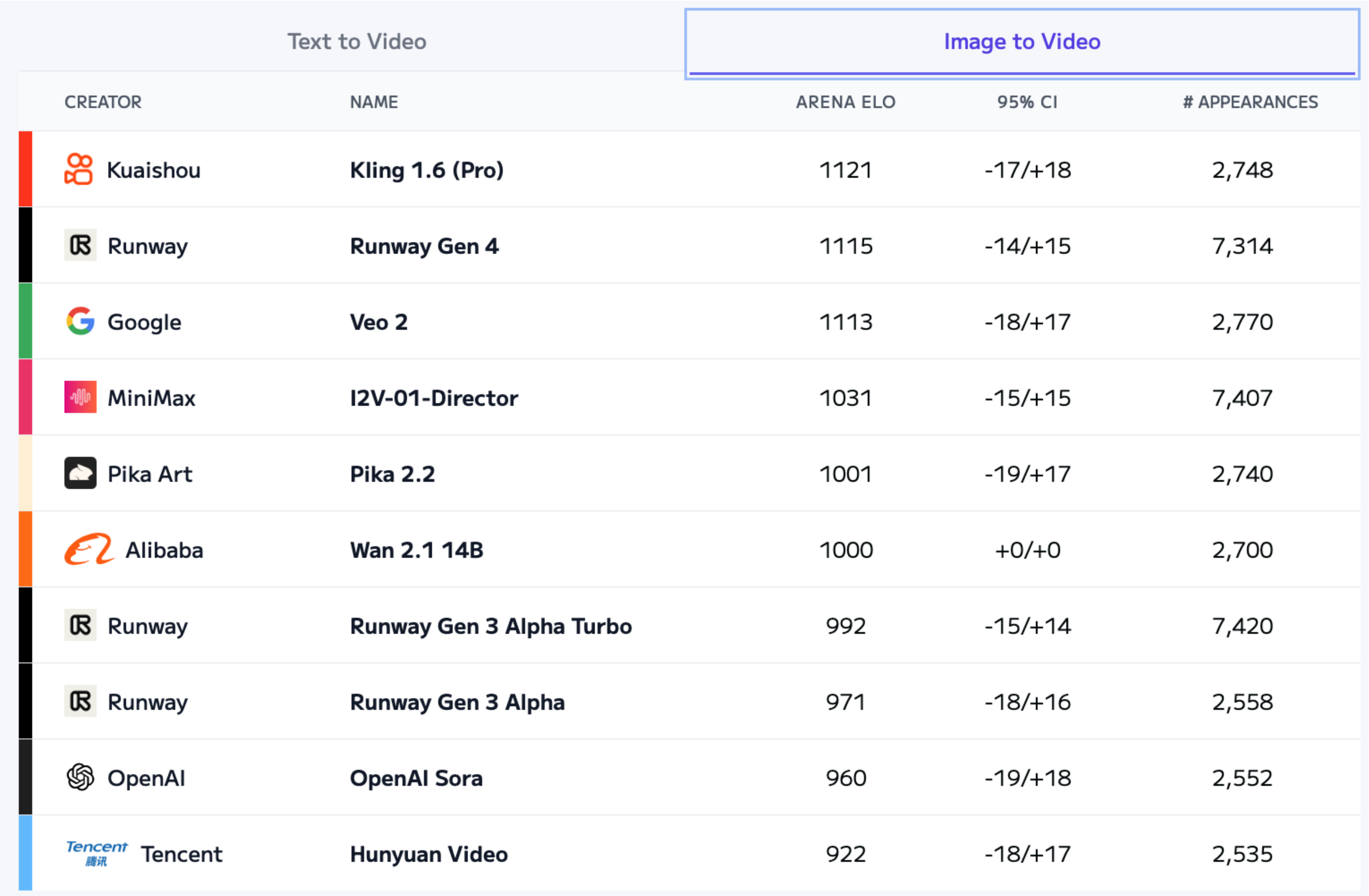
Artificial Analysis Text to Video Leaderboard



CREATOR	NAME	ARENA ELO	95% CI	# APPEARANCES
Google	**Veo 2**	1124	-10/+10	6,452
Kuaishou	**Kling 1.5 (Pro)**	1053	-6/+6	20,631
OpenAI	**OpenAI Sora**	1049	-5/+5	23,649
MiniMax	**T2V-01**	1039	-4/+4	43,450
Pika Art	**Pika 2.0**	1038	-6/+6	20,432
Kuaishou	**Kling 1.6 (Standard)**	1029	-7/+6	13,607
MiniMax	**T2V-01-Director**	1022	-9/+9	7,765

Source: https://artificialanalysis.ai/text-to-video/arena?tab=Leaderboard&leaderboard_tab=t2v

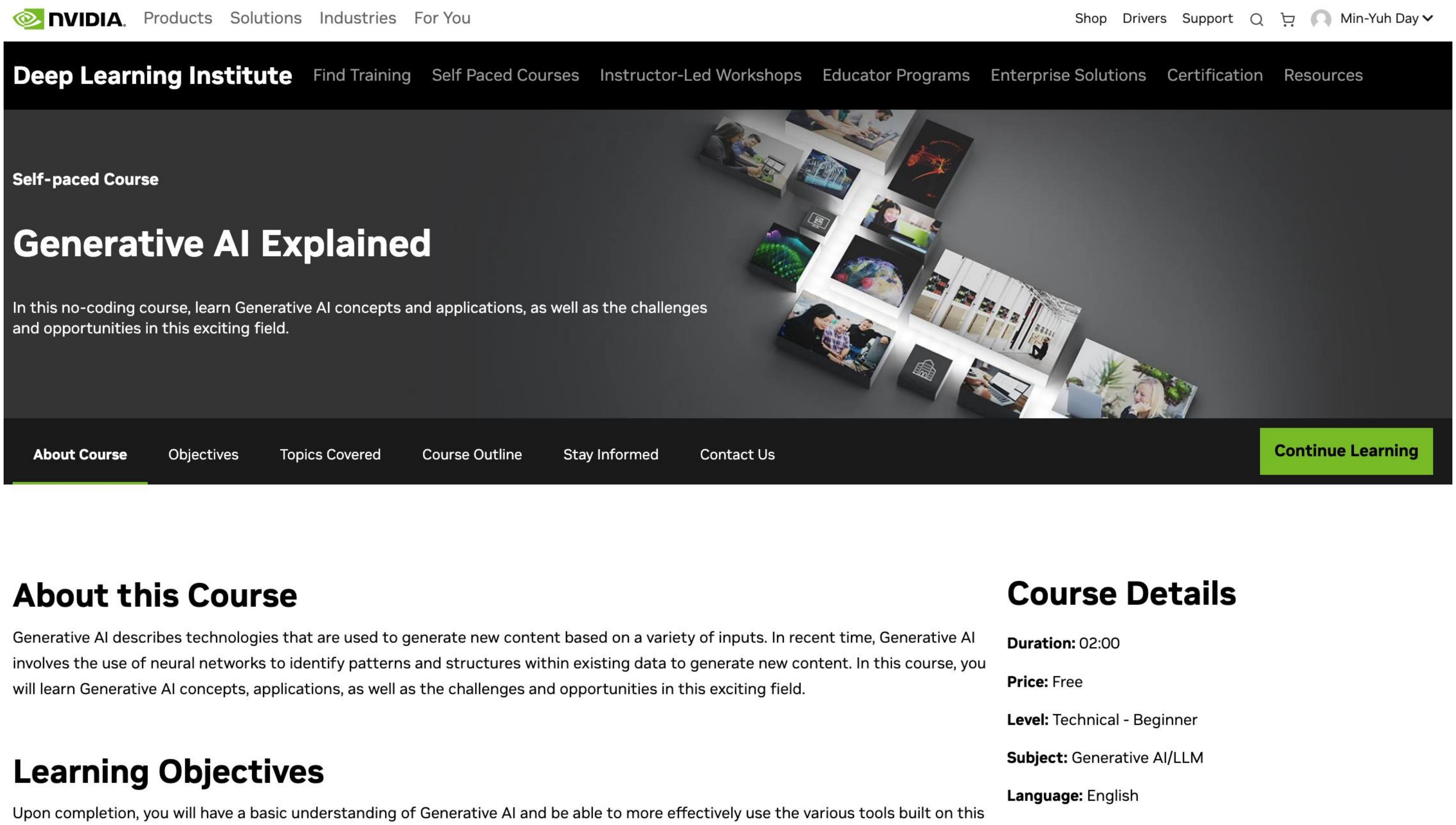
Artificial Analysis Image to Video Leaderboard



CREATOR	NAME	ARENA ELO	95% CI	# APPEARANCES
Kuaishou	Kling 1.6 (Pro)	1121	-17/+18	2,748
Runway	Runway Gen 4	1115	-14/+15	7,314
Google	Veo 2	1113	-18/+17	2,770
MiniMax	I2V-01-Director	1031	-15/+15	7,407
Pika Art	Pika 2.2	1001	-19/+17	2,740
Alibaba	Wan 2.1 14B	1000	+0/+0	2,700
Runway	Runway Gen 3 Alpha Turbo	992	-15/+14	7,420
Runway	Runway Gen 3 Alpha	971	-18/+16	2,558
OpenAI	OpenAI Sora	960	-19/+18	2,552
Tencent	Hunyuan Video	922	-18/+17	2,535

Source: https://artificialanalysis.ai/text-to-video/arena?tab=Leaderboard&leaderboard_tab=t2v

Generative AI Explained


NVIDIA. Products Solutions Industries For You
Shop Drivers Support Min-Yuh Day
Deep Learning Institute Find Training Self Paced Courses Instructor-Led Workshops Educator Programs Enterprise Solutions Certification Resources
About Course Objectives Topics Covered Course Outline Stay Informed Contact Us
Continue Learning

Self-paced Course

Generative AI Explained

In this no-coding course, learn Generative AI concepts and applications, as well as the challenges and opportunities in this exciting field.

About this Course

Generative AI describes technologies that are used to generate new content based on a variety of inputs. In recent time, Generative AI involves the use of neural networks to identify patterns and structures within existing data to generate new content. In this course, you will learn Generative AI concepts, applications, as well as the challenges and opportunities in this exciting field.

Learning Objectives

Upon completion, you will have a basic understanding of Generative AI and be able to more effectively use the various tools built on this

Course Details

Duration: 02:00

Price: Free

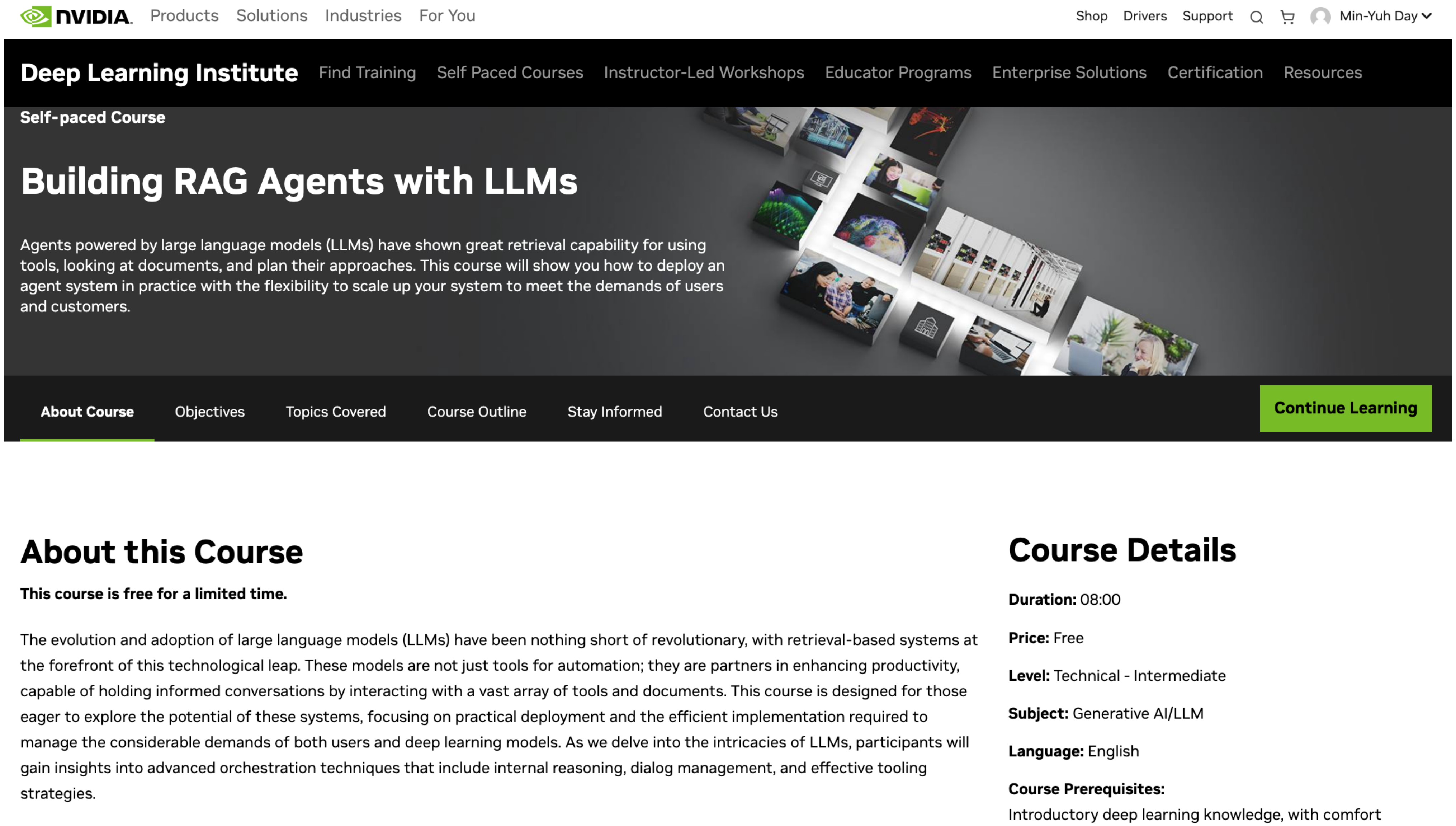
Level: Technical - Beginner

Subject: Generative AI/LLM

Language: English

https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Building RAG Agents with LLMs



Self-paced Course

Building RAG Agents with LLMs

Agents powered by large language models (LLMs) have shown great retrieval capability for using tools, looking at documents, and plan their approaches. This course will show you how to deploy an agent system in practice with the flexibility to scale up your system to meet the demands of users and customers.

About this Course

This course is free for a limited time.

The evolution and adoption of large language models (LLMs) have been nothing short of revolutionary, with retrieval-based systems at the forefront of this technological leap. These models are not just tools for automation; they are partners in enhancing productivity, capable of holding informed conversations by interacting with a vast array of tools and documents. This course is designed for those eager to explore the potential of these systems, focusing on practical deployment and the efficient implementation required to manage the considerable demands of both users and deep learning models. As we delve into the intricacies of LLMs, participants will gain insights into advanced orchestration techniques that include internal reasoning, dialog management, and effective tooling strategies.

Course Details

Duration: 08:00

Price: Free

Level: Technical - Intermediate

Subject: Generative AI/LLM

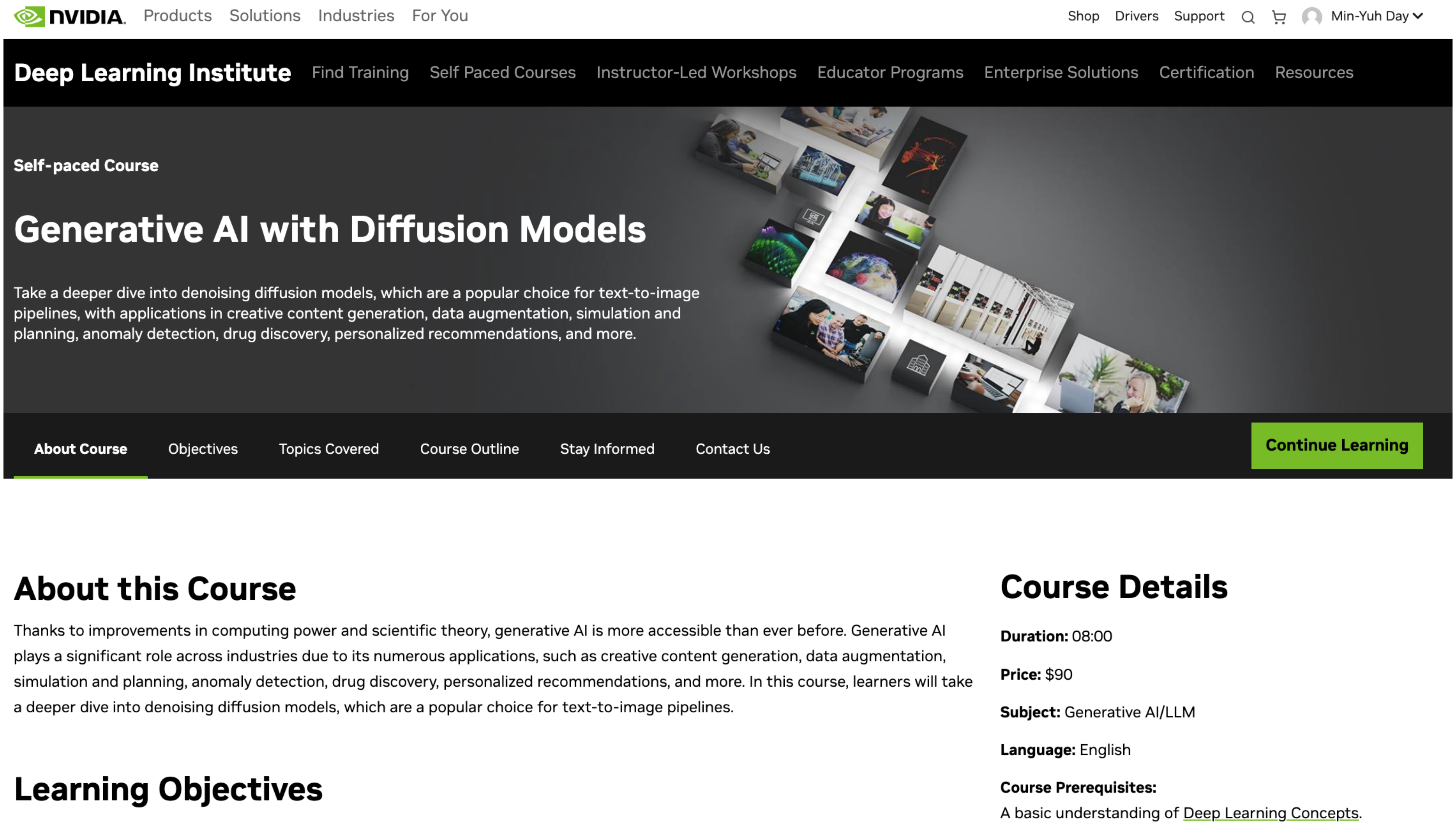
Language: English

Course Prerequisites:

Introductory deep learning knowledge, with comfort

https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Generative AI with Diffusion Models



Self-paced Course

Generative AI with Diffusion Models

Take a deeper dive into denoising diffusion models, which are a popular choice for text-to-image pipelines, with applications in creative content generation, data augmentation, simulation and planning, anomaly detection, drug discovery, personalized recommendations, and more.

About this Course

Thanks to improvements in computing power and scientific theory, generative AI is more accessible than ever before. Generative AI plays a significant role across industries due to its numerous applications, such as creative content generation, data augmentation, simulation and planning, anomaly detection, drug discovery, personalized recommendations, and more. In this course, learners will take a deeper dive into denoising diffusion models, which are a popular choice for text-to-image pipelines.

Learning Objectives

Course Details

Duration: 08:00

Price: $90

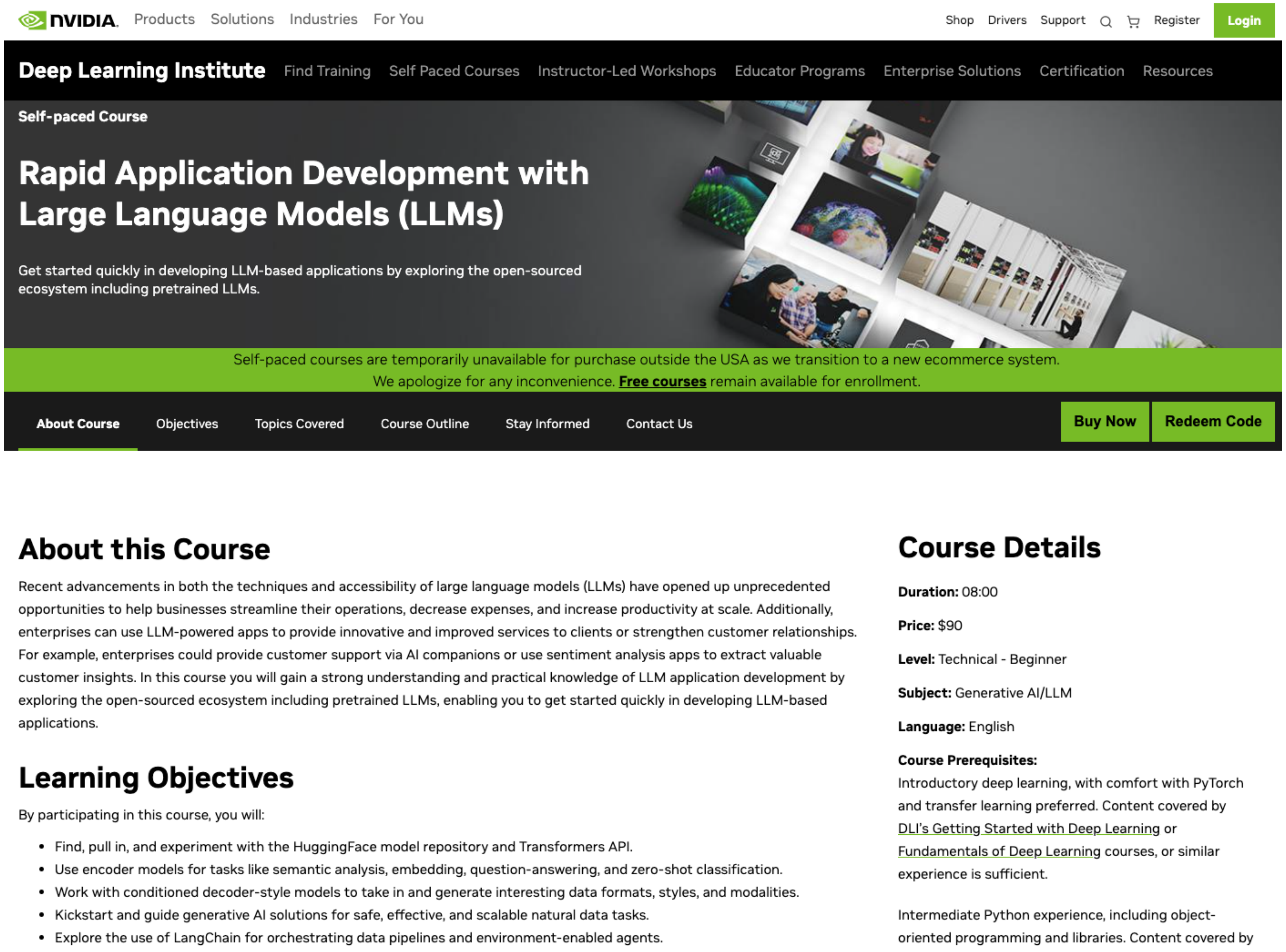
Subject: Generative AI/LLM

Language: English

Course Prerequisites:
A basic understanding of Deep Learning Concepts.

https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-14+V1

Rapid Application Development with Large Language Models (LLMs)



Self-paced Course

Rapid Application Development with Large Language Models (LLMs)

Get started quickly in developing LLM-based applications by exploring the open-sourced ecosystem including pretrained LLMs.

Self-paced courses are temporarily unavailable for purchase outside the USA as we transition to a new ecommerce system. We apologize for any inconvenience. **Free courses** remain available for enrollment.

About this Course

Recent advancements in both the techniques and accessibility of large language models (LLMs) have opened up unprecedented opportunities to help businesses streamline their operations, decrease expenses, and increase productivity at scale. Additionally, enterprises can use LLM-powered apps to provide innovative and improved services to clients or strengthen customer relationships. For example, enterprises could provide customer support via AI companions or use sentiment analysis apps to extract valuable customer insights. In this course you will gain a strong understanding and practical knowledge of LLM application development by exploring the open-sourced ecosystem including pretrained LLMs, enabling you to get started quickly in developing LLM-based applications.

Learning Objectives

By participating in this course, you will:

- Find, pull in, and experiment with the HuggingFace model repository and Transformers API.
- Use encoder models for tasks like semantic analysis, embedding, question-answering, and zero-shot classification.
- Work with conditioned decoder-style models to take in and generate interesting data formats, styles, and modalities.
- Kickstart and guide generative AI solutions for safe, effective, and scalable natural data tasks.
- Explore the use of LangChain for orchestrating data pipelines and environment-enabled agents.

Course Details

Duration: 08:00

Price: $90

Level: Technical - Beginner

Subject: Generative AI/LLM

Language: English

Course Prerequisites:
Introductory deep learning, with comfort with PyTorch and transfer learning preferred. Content covered by DLI's Getting Started with Deep Learning or Fundamentals of Deep Learning courses, or similar experience is sufficient.

Intermediate Python experience, including object-oriented programming and libraries. Content covered by

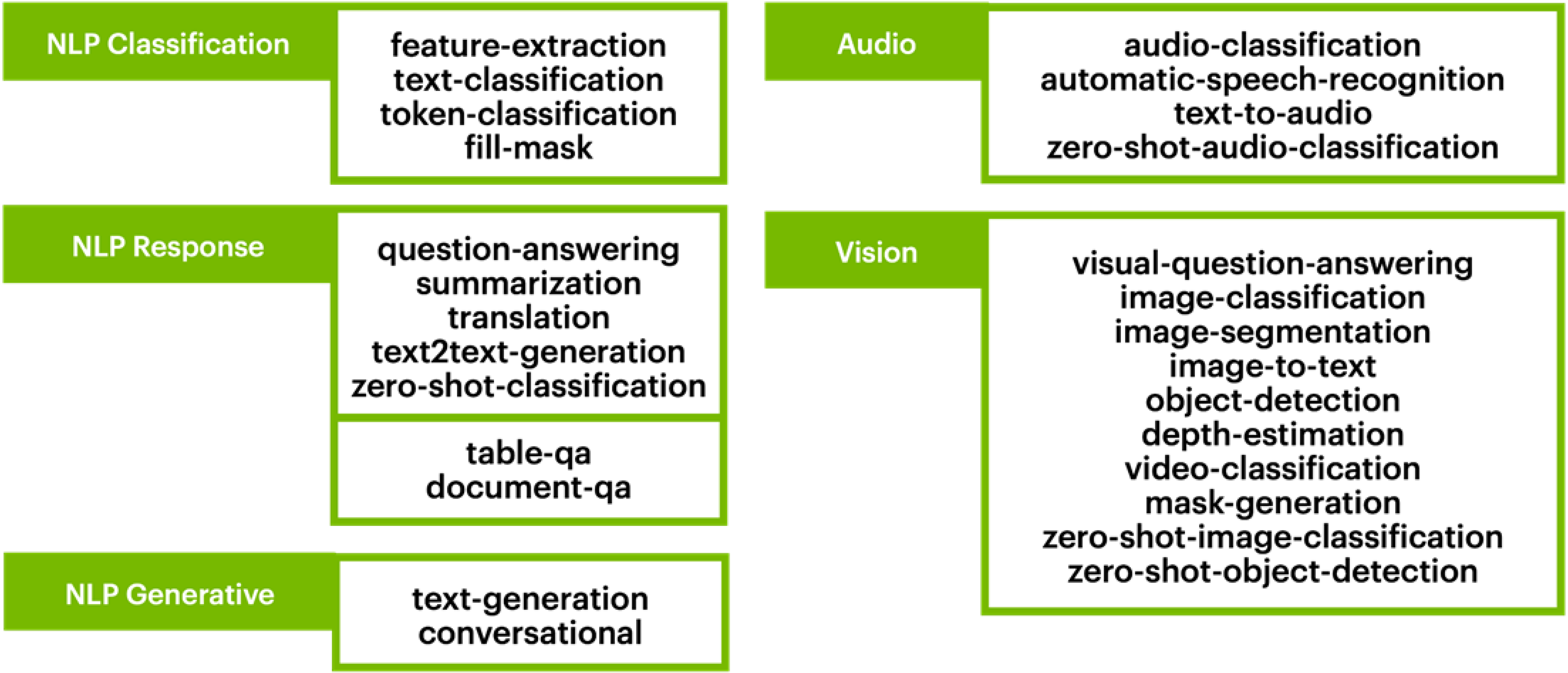
https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Rapid Application Development using Large Language Models

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

HF Pipeline Options

Automatic End-to-End Pipelines



NLP Classification

feature-extraction
text-classification
token-classification
fill-mask

NLP Response

question-answering
summarization
translation
text2text-generation
zero-shot-classification

table-qa
document-qa

NLP Generative

text-generation
conversational

Audio

audio-classification
automatic-speech-recognition
text-to-audio
zero-shot-audio-classification

Vision

visual-question-answering
image-classification
image-segmentation
image-to-text
object-detection
depth-estimation
video-classification
mask-generation
zero-shot-image-classification
zero-shot-object-detection

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

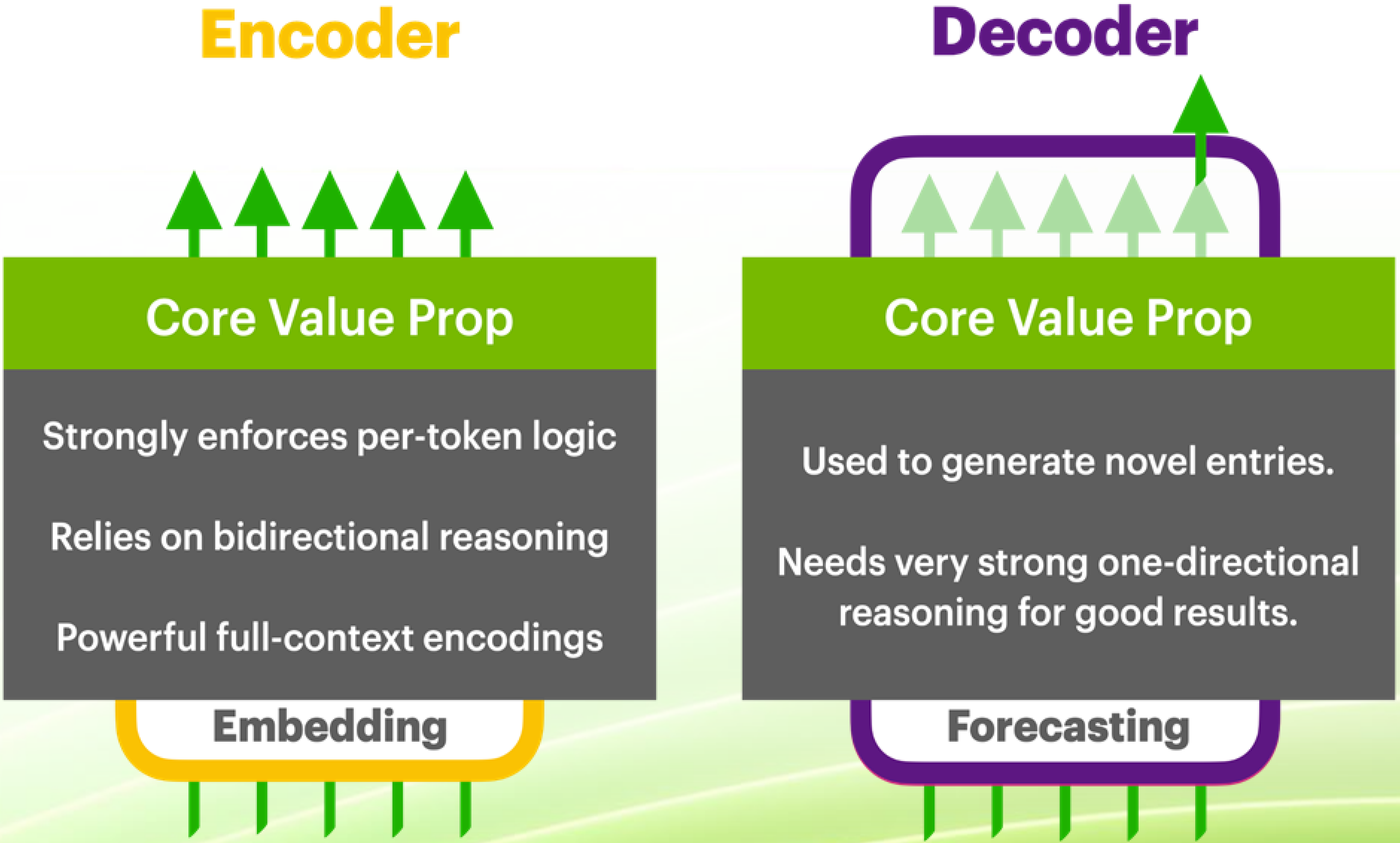
Large Language Models

Backbones for Language Understanding

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Encoders vs Decoders

Both Options Have Pros and Cons



Encoder – Core Value Prop

Strongly enforces per-token logic

Relies on bidirectional reasoning

Powerful full-context encodings

Decoder – Core Value Prop

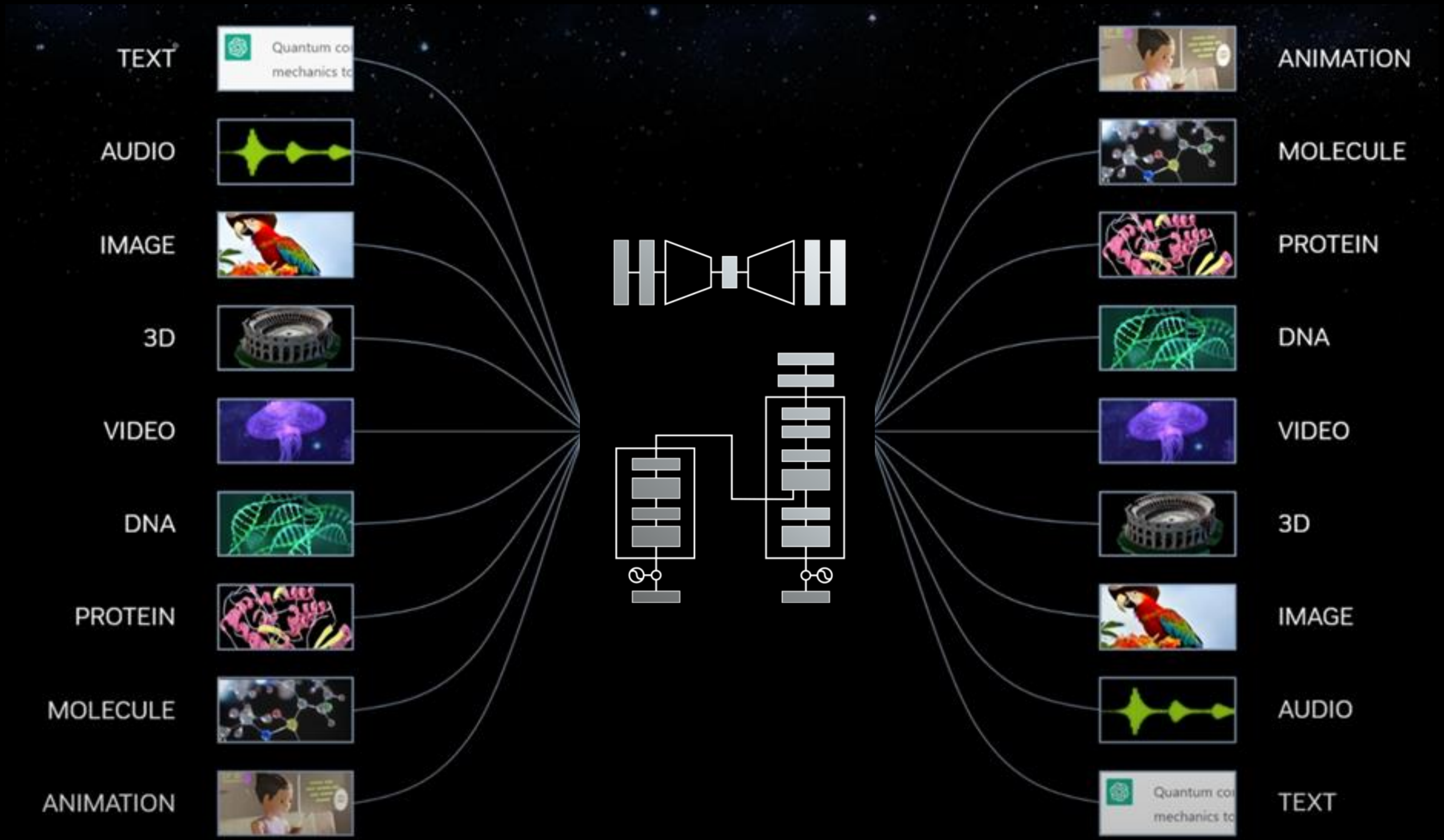
Used to generate novel entries.

Needs very strong one-directional reasoning for good results.

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Modular Modalities

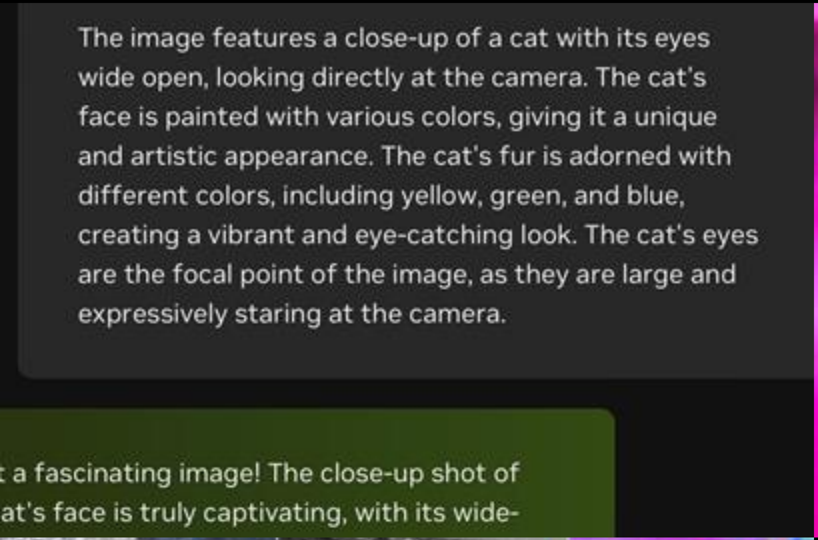
Where Can The Transformer Fit?



Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

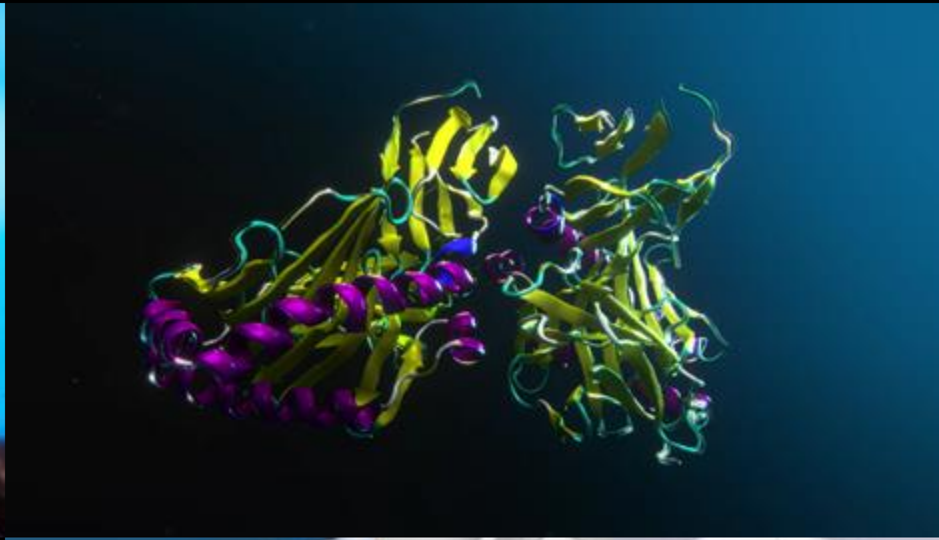
Information Modalities

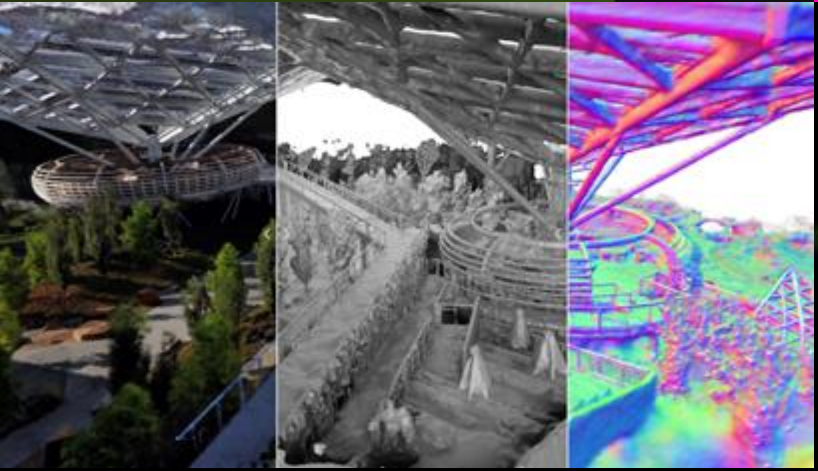
Core Component of Generative AI



The image features a close-up of a cat with its eyes wide open, looking directly at the camera. The cat's face is painted with various colors, giving it a unique and artistic appearance. The cat's fur is adorned with different colors, including yellow, green, and blue, creating a vibrant and eye-catching look. The cat's eyes are the focal point of the image, as they are large and expressively staring at the camera.

t a fascinating image! The close-up shot of at's face is truly captivating, with its wide-

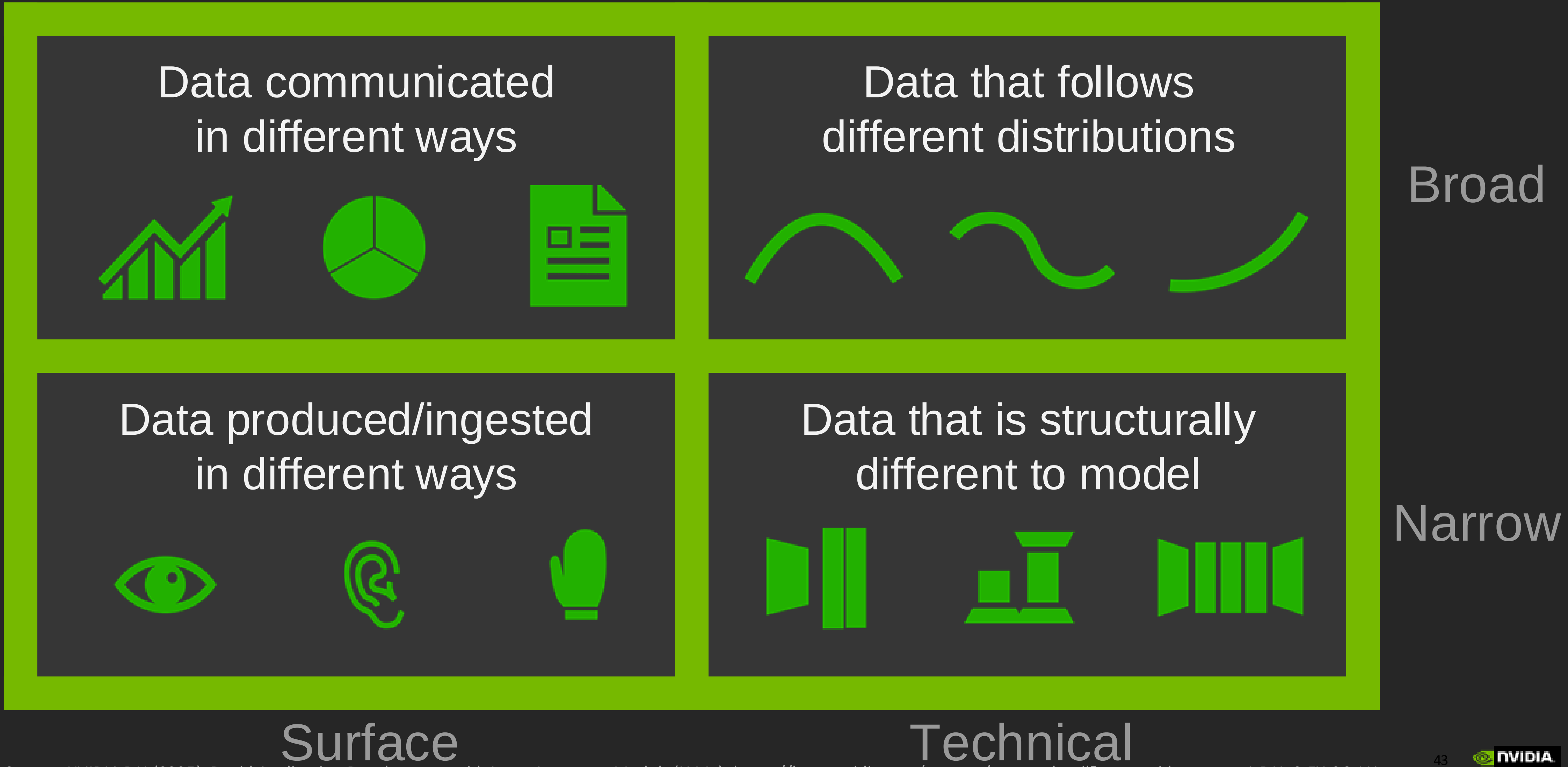




Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

What Makes Modalities Different?

Core Component of Generative AI



Data communicated in different ways

Data that follows different distributions

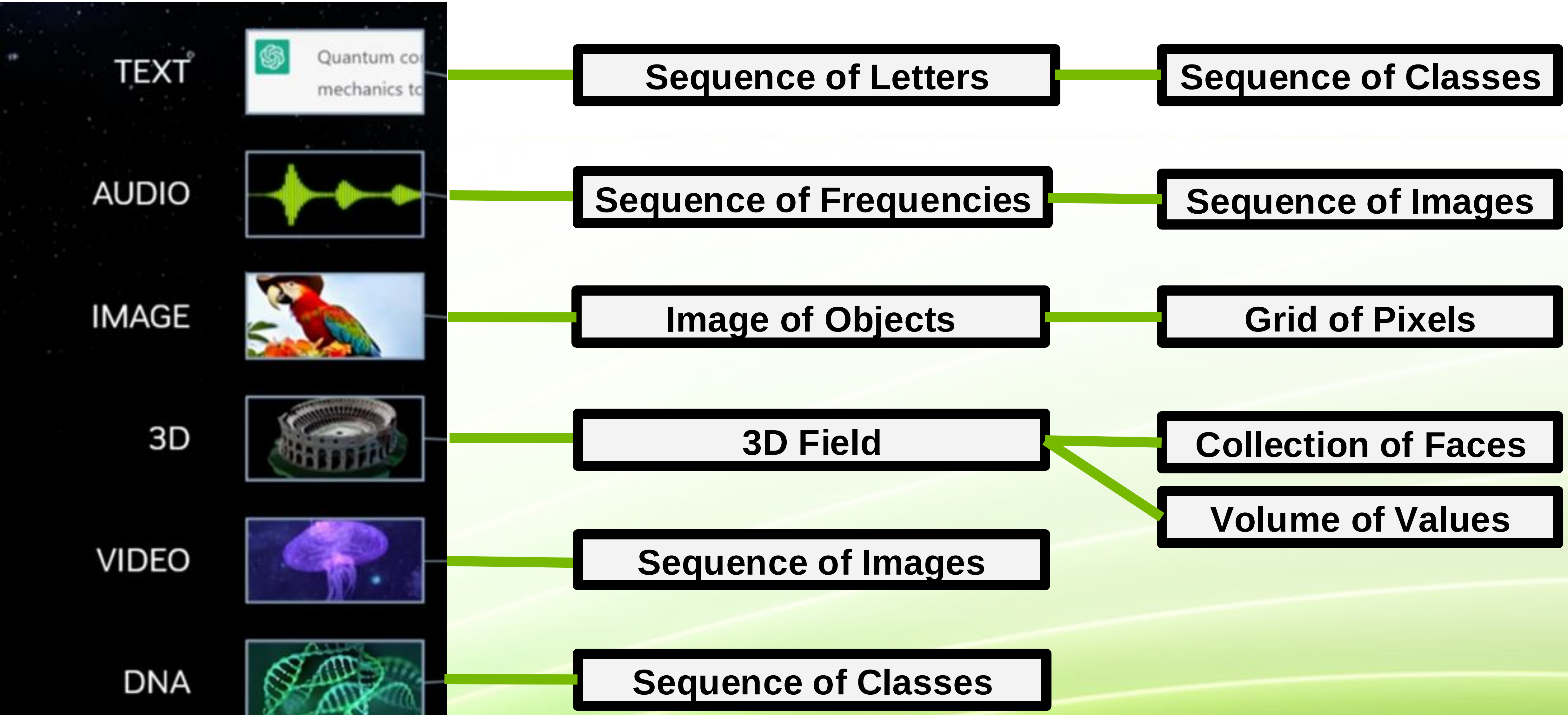
Data produced/ingested in different ways

Data that is structurally different to model

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Multimodal Connections

Core Components of Generative AI



Sequence of Letters

Sequence of Classes

Sequence of Frequencies

Sequence of Images

Image of Objects

Grid of Pixels

3D Field

Collection of Faces

Volume of Values

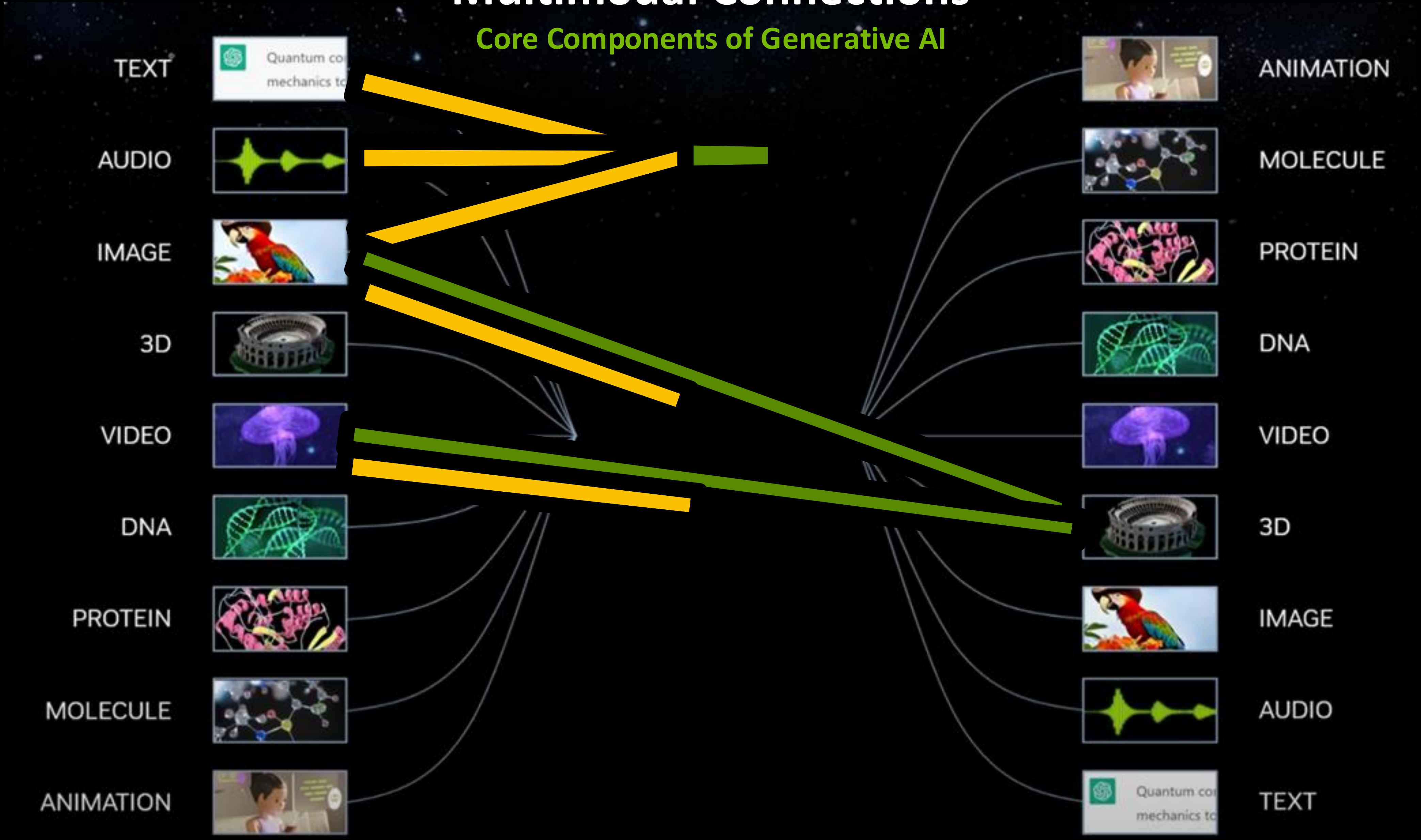
Sequence of Images

Sequence of Classes

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Multimodal Connections

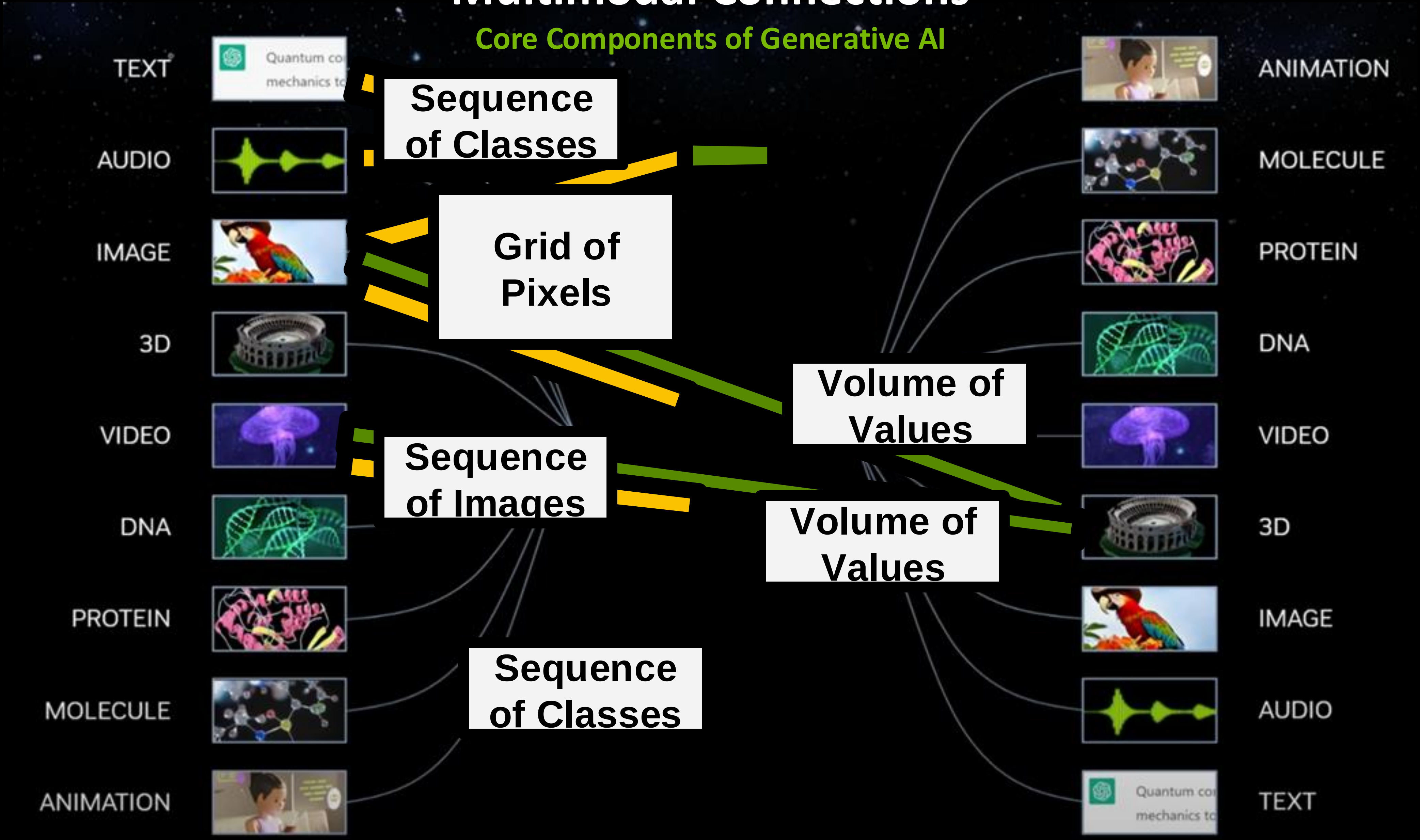
Core Components of Generative AI



Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Multimodal Connections

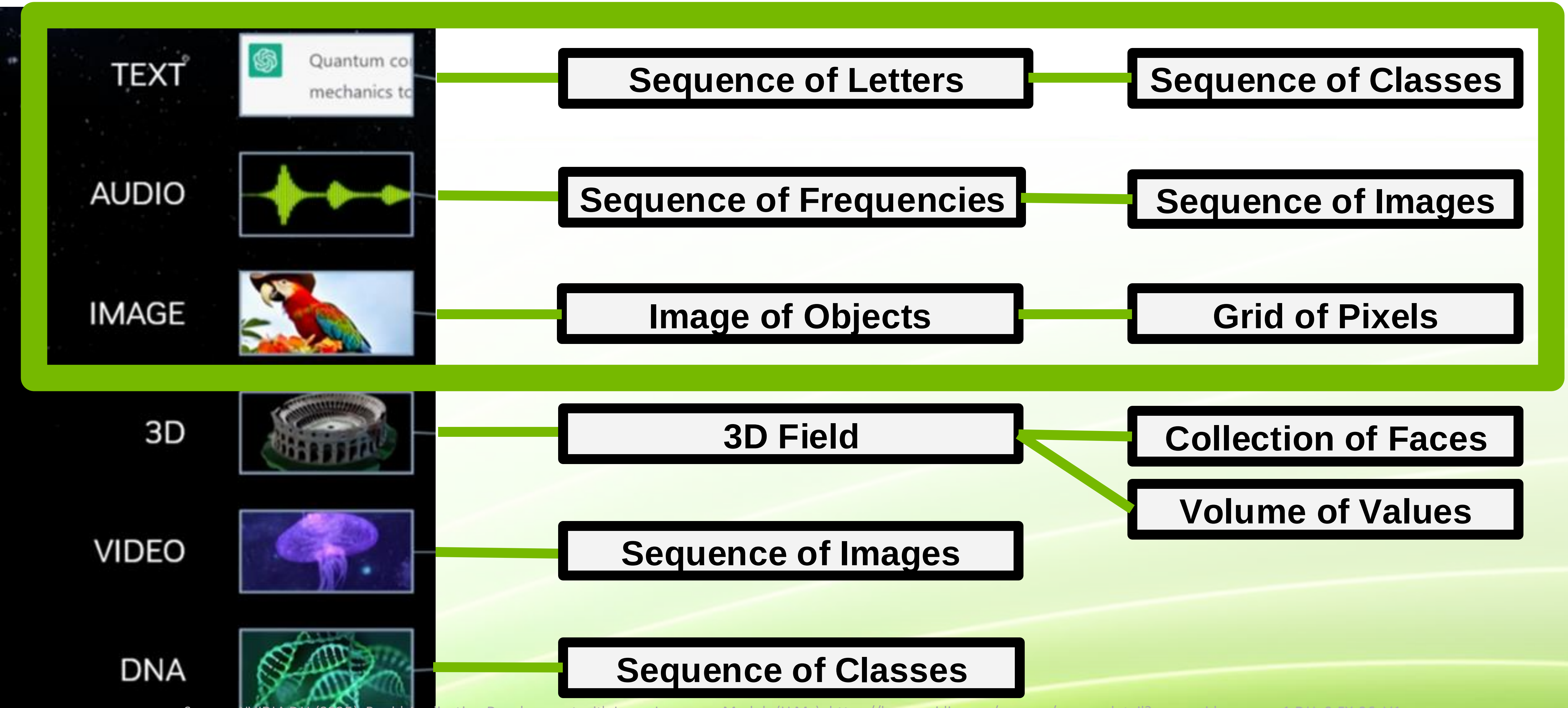
Core Components of Generative AI



Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Multimodal Connections

Core Components of Generative AI



TEXT — Sequence of Letters — Sequence of Classes

AUDIO — Sequence of Frequencies — Sequence of Images

IMAGE — Image of Objects — Grid of Pixels

3D — 3D Field — Collection of Faces

Volume of Values

VIDEO — Sequence of Images

DNA — Sequence of Classes

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Transformer Benefits

What Are They Good For

Misconception:

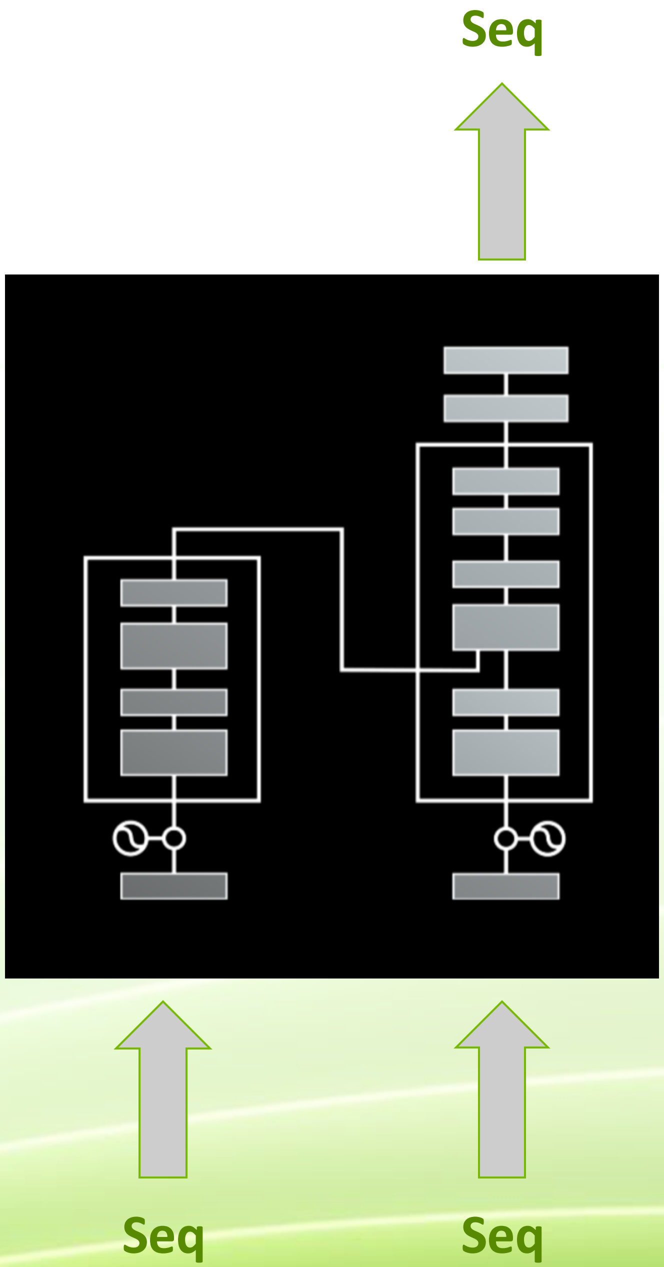
Transformers are only good for language

Recall:

Language Text = Ordered Sequence of Classes

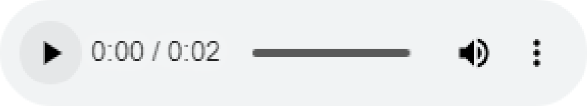
Resolution:

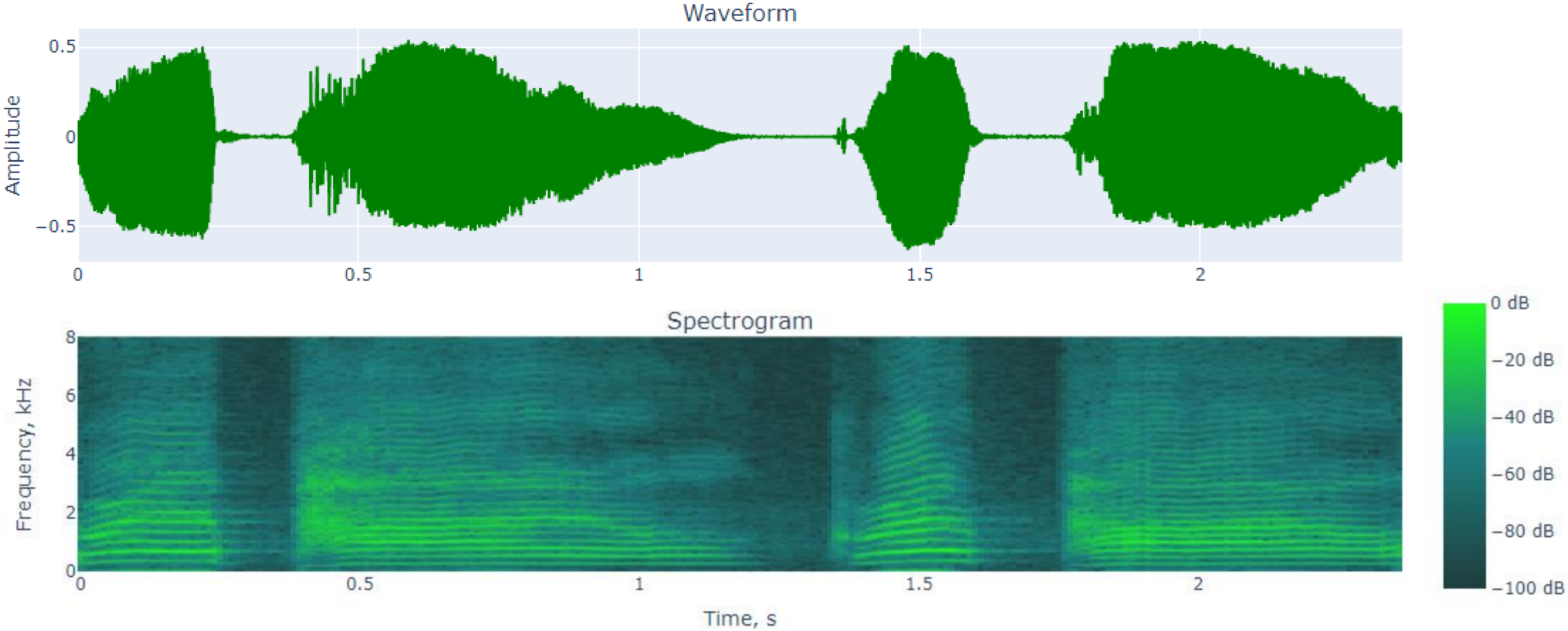
Transformers are good for ordered sequences



Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Speech-Guided Encoders

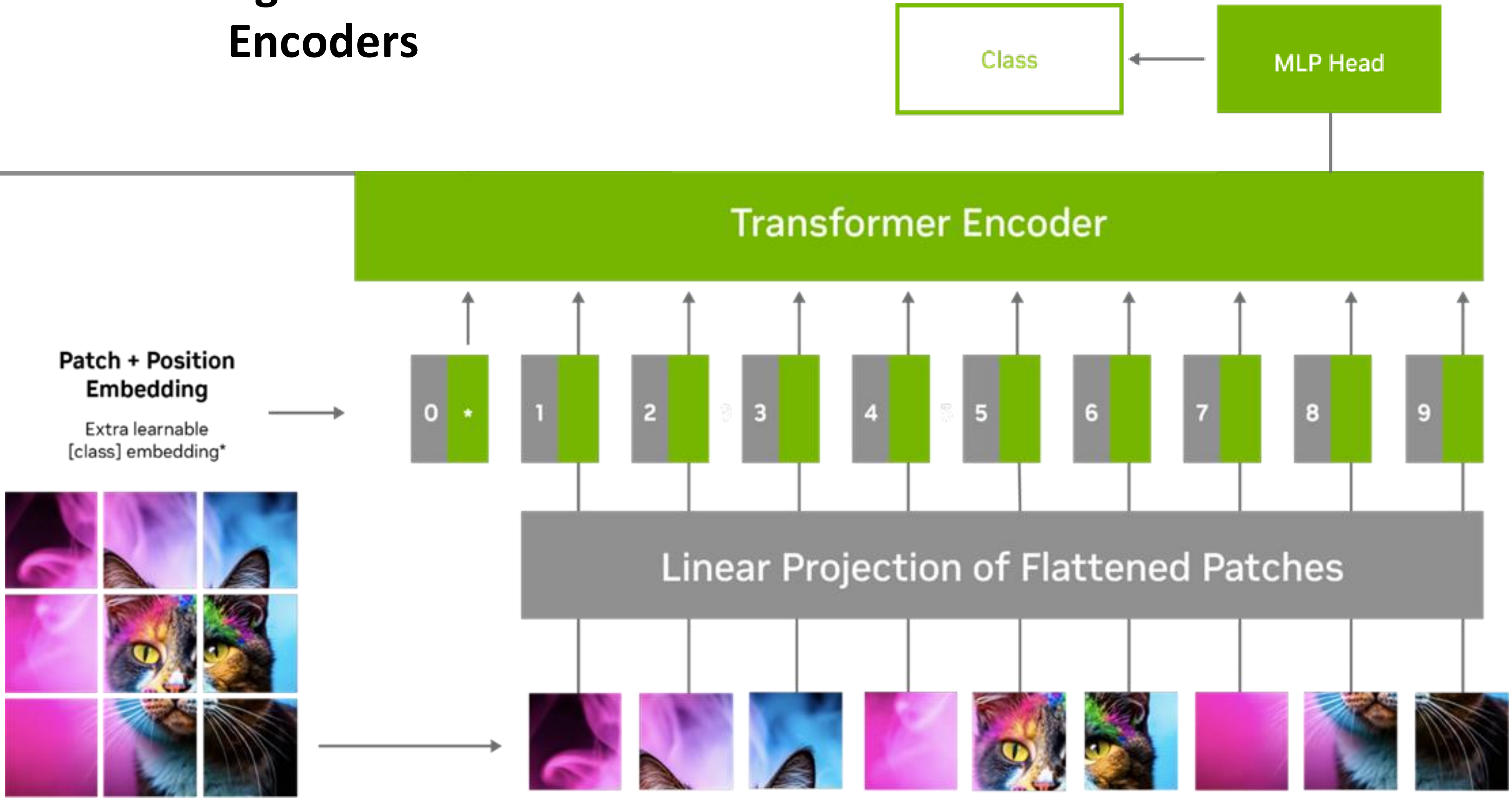




Speech Data Explorer | NeMo Framework User Guide

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Image-Guided Encoders

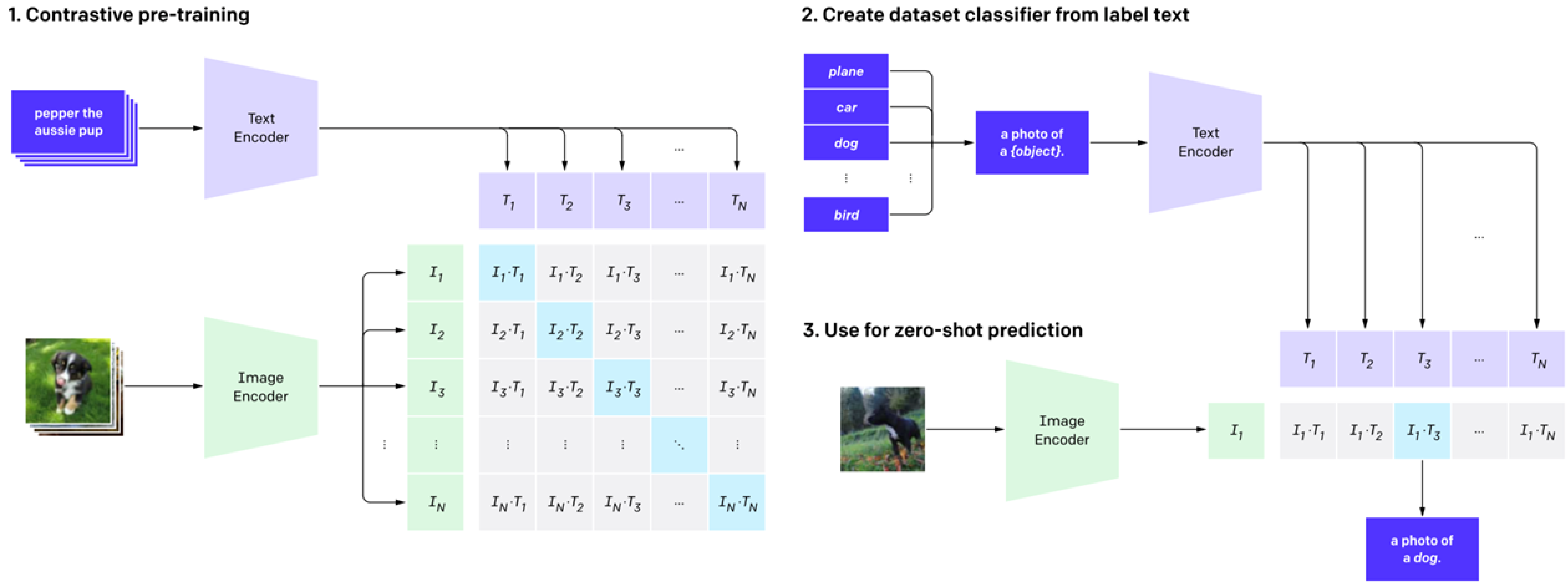

Class
MLP Head
Transformer Encoder
Patch + Position Embedding
Extra learnable [class] embedding*
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Improve Accuracy and Robustness of Vision AI Apps with Vision Transformers and NVIDIA TAO | NVIDIA Blog

Source: NVIDIA DLI (2025) Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Synergizing Encoders

CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset. We then use this behavior to turn CLIP into a zero-shot classifier. We convert all of a dataset's classes into captions such as "a photo of a dog" and predict the class of the caption CLIP estimates best pairs with a given image.



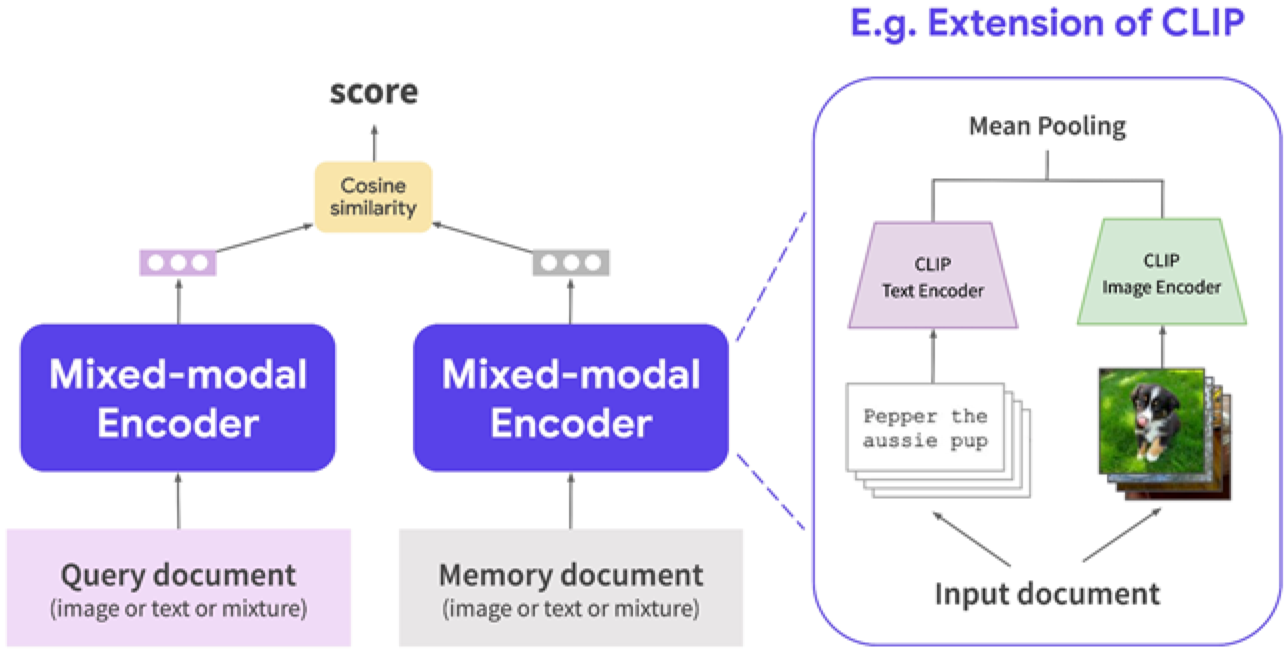
1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ ... $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ ... $I_3{\cdot}T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ ... $I_N{\cdot}T_N$
2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
a photo of a dog.

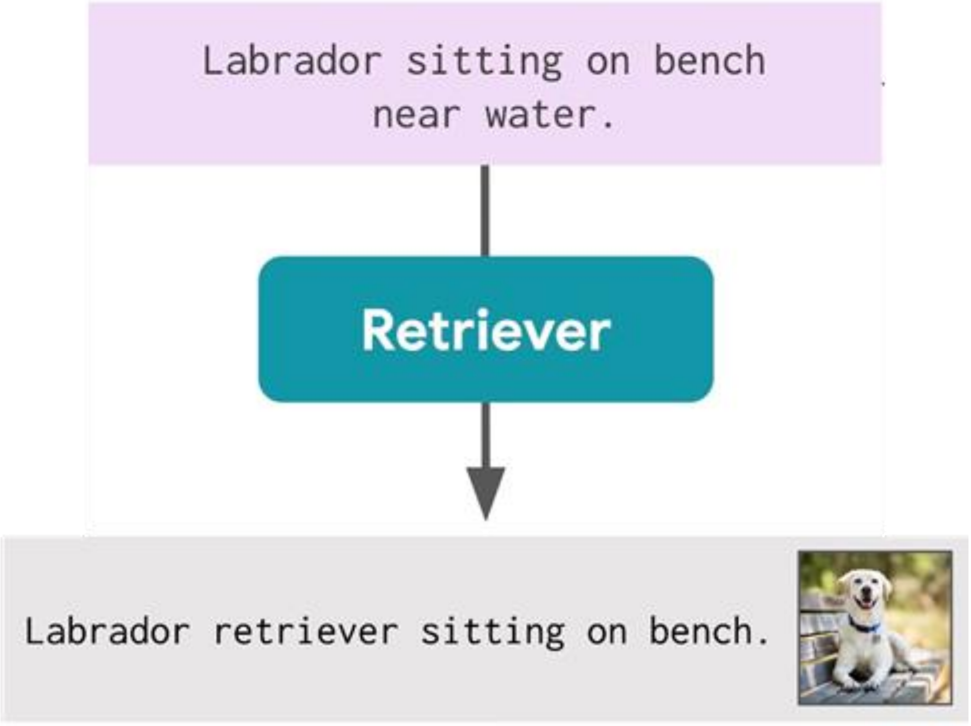
CLIP: Connecting text and images | OpenAI

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Multimodal Retrievers

f(query, memory) → score




Labrador sitting on bench near water.
Retriever
Labrador retriever sitting on bench.

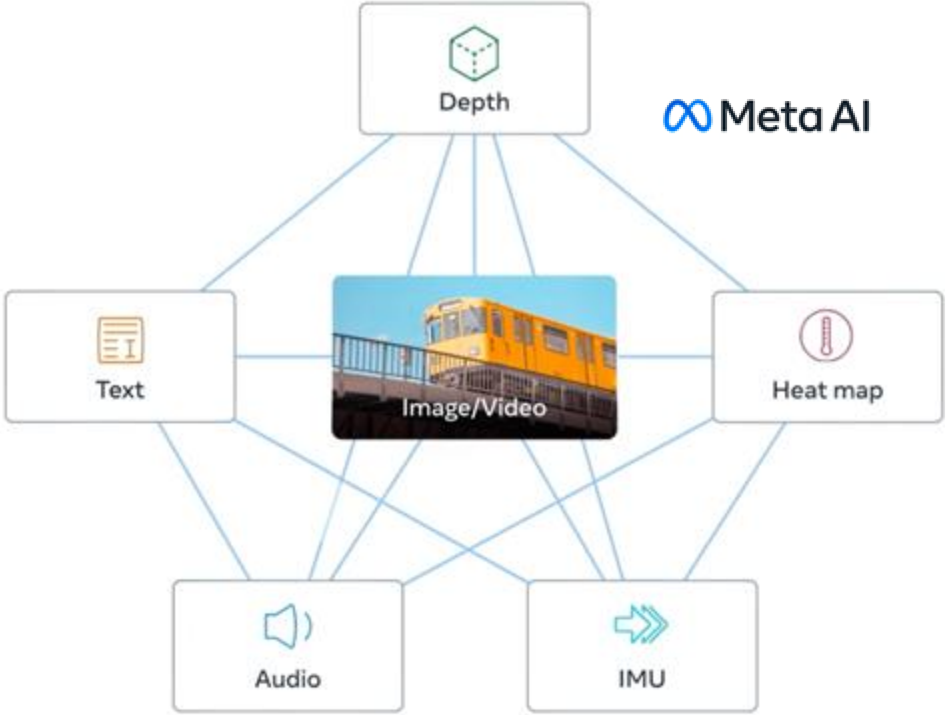
Multimodal document: image, text, or mixture of them

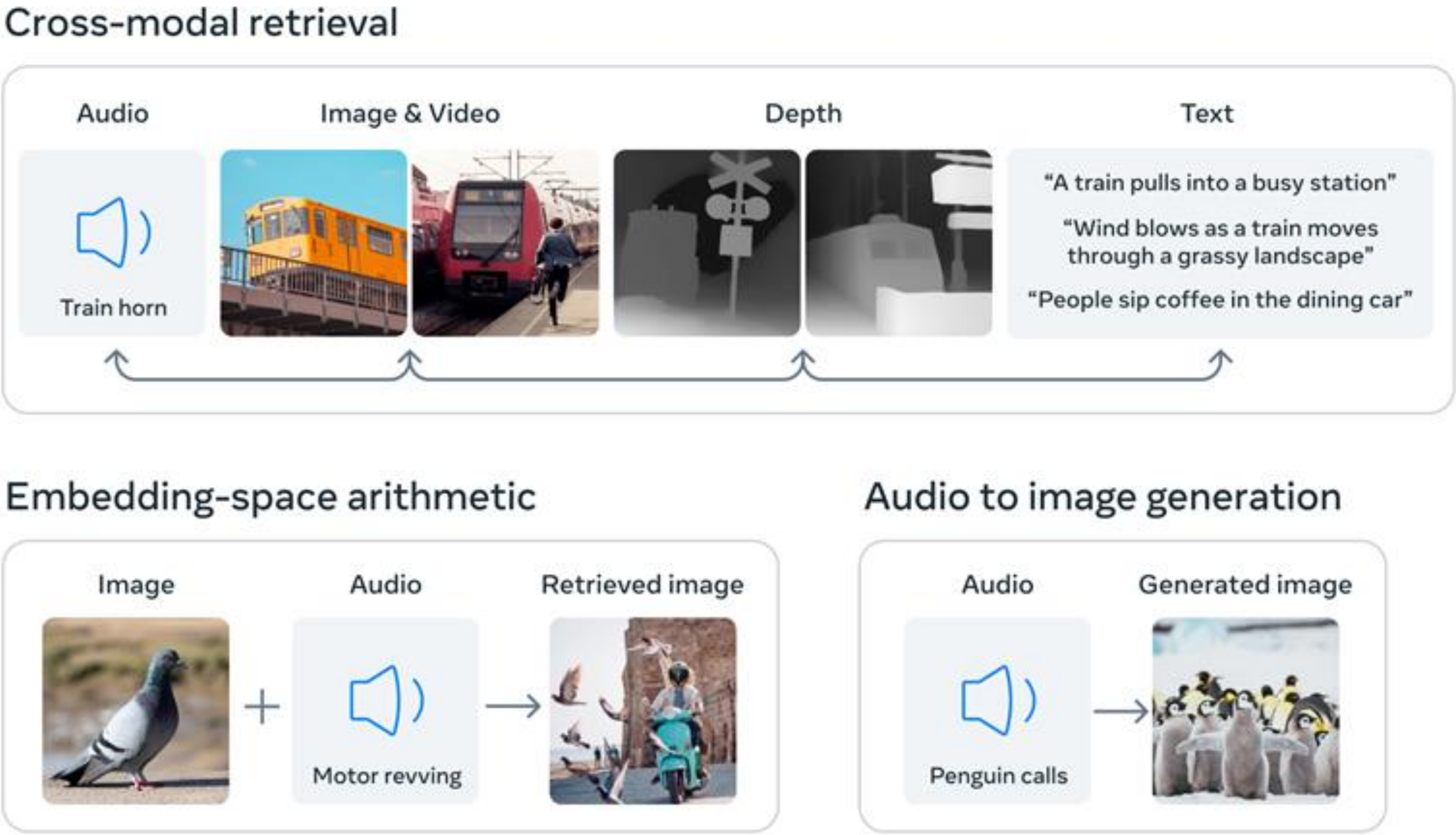
RA-CM3: Retrieval-Augmented Multimodal Modeling (2023)

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Synergizing Encoders

By aligning six modalities' embedding into a common space, ImageBind enables cross-modal retrieval of different types of content that aren't observed together, the addition of embeddings from different modalities to naturally compose their semantics, and audio-to-image generation by using our audio embeddings with a pretrained DALLE-2 decoder to work with CLIP text embeddings.

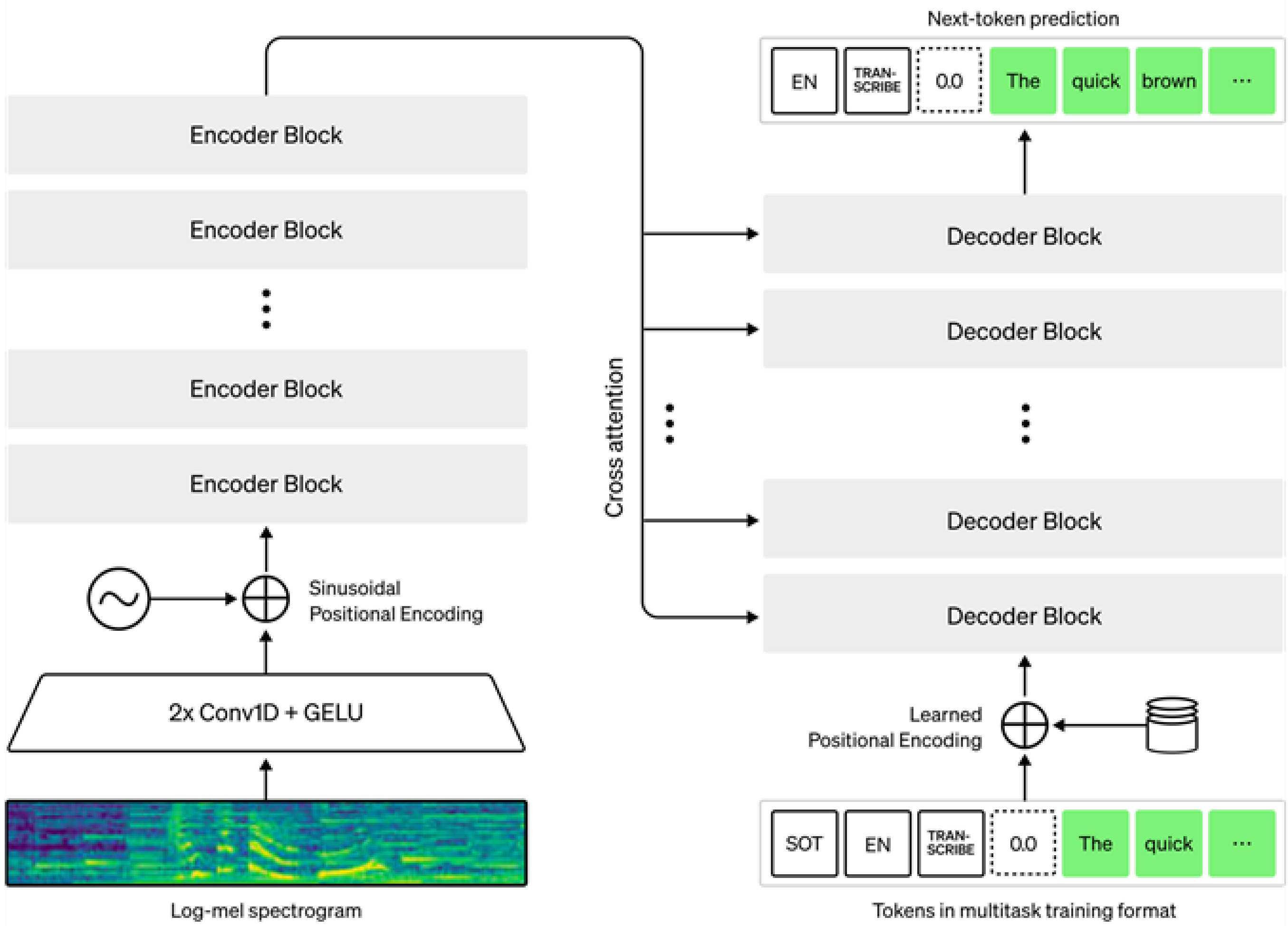



Cross-modal retrieval
Audio
Image & Video
Depth
Text
Train horn
"A train pulls into a busy station"
"Wind blows as a train moves through a grassy landscape"
"People sip coffee in the dining car"
Embedding-space arithmetic
Image
Audio
Retrieved image
Motor revving
Audio to image generation
Audio
Generated image
Penguin calls

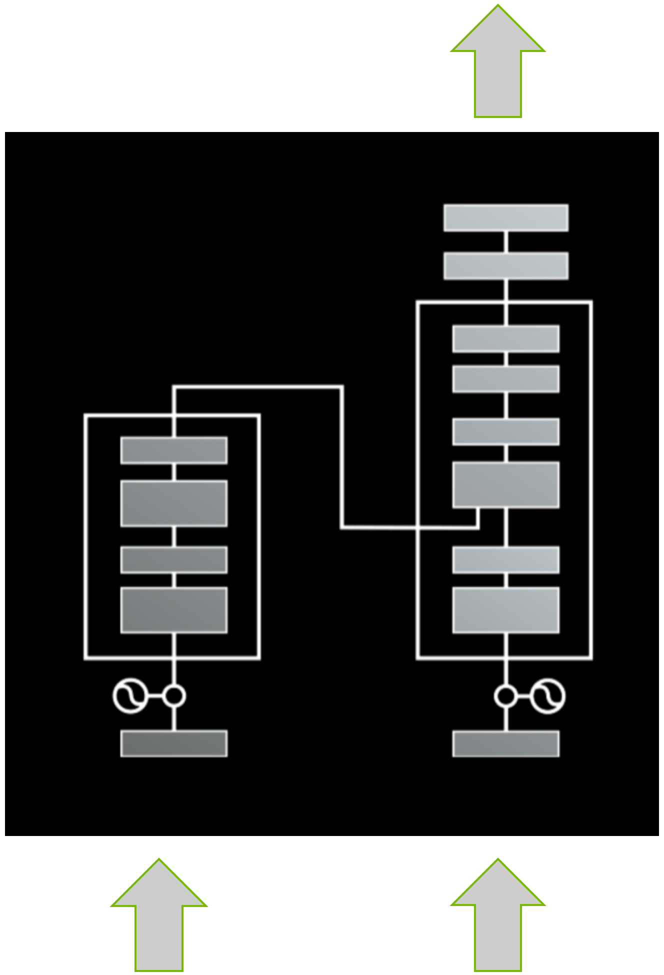
[ImageBind: Holistic AI learning across six modalities (2023)](https://ai.meta.com/blog/imagebind-six-modalities-binding-ai/)

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Speech-Guided Text Generation



Text



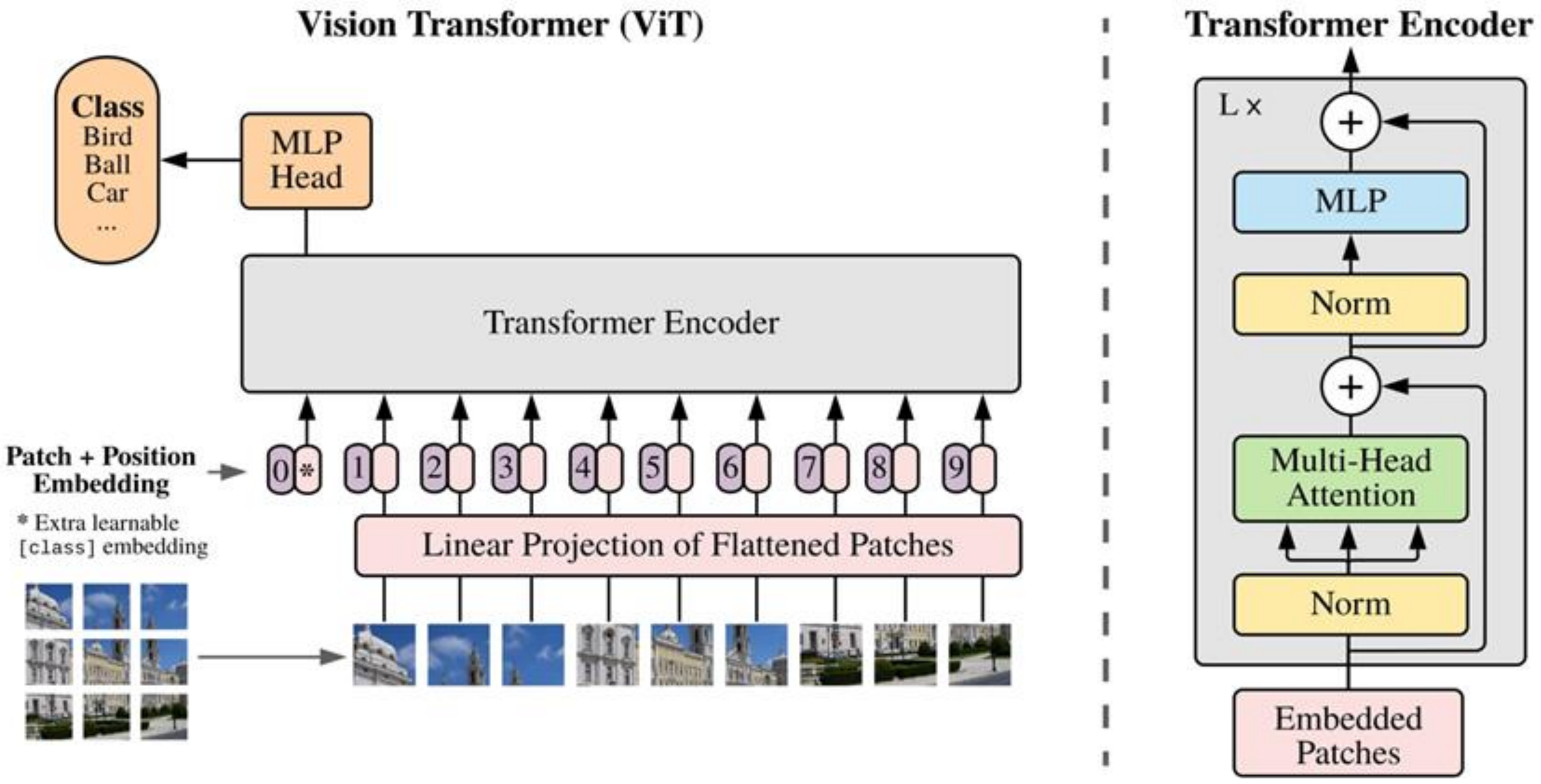
Audio

Text

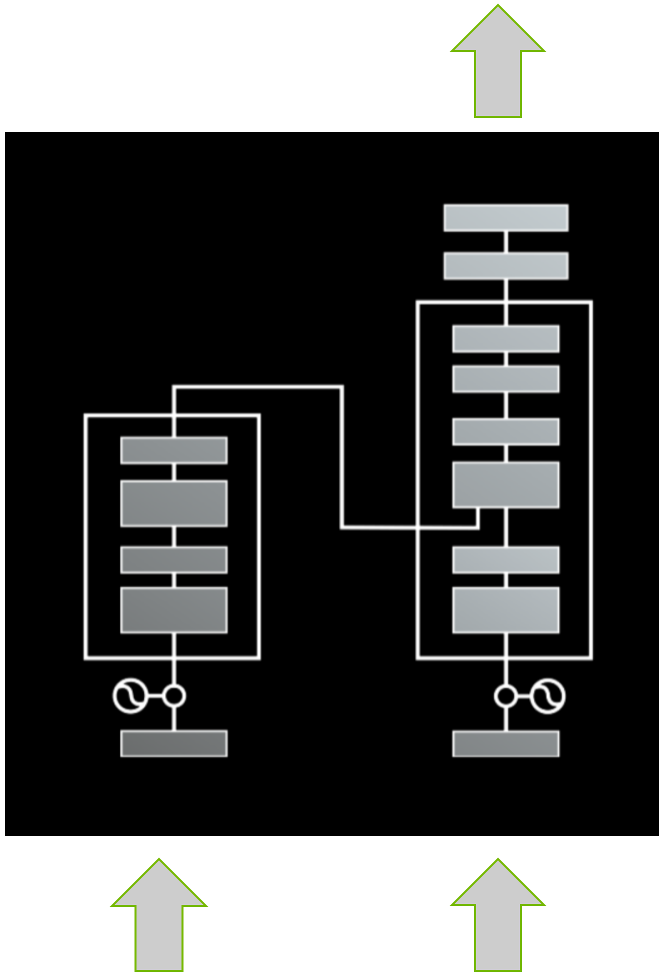
[Introducing Whisper | OpenAI](https://openai.com/index/whisper/)

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Image-Guided Text Generation



Text



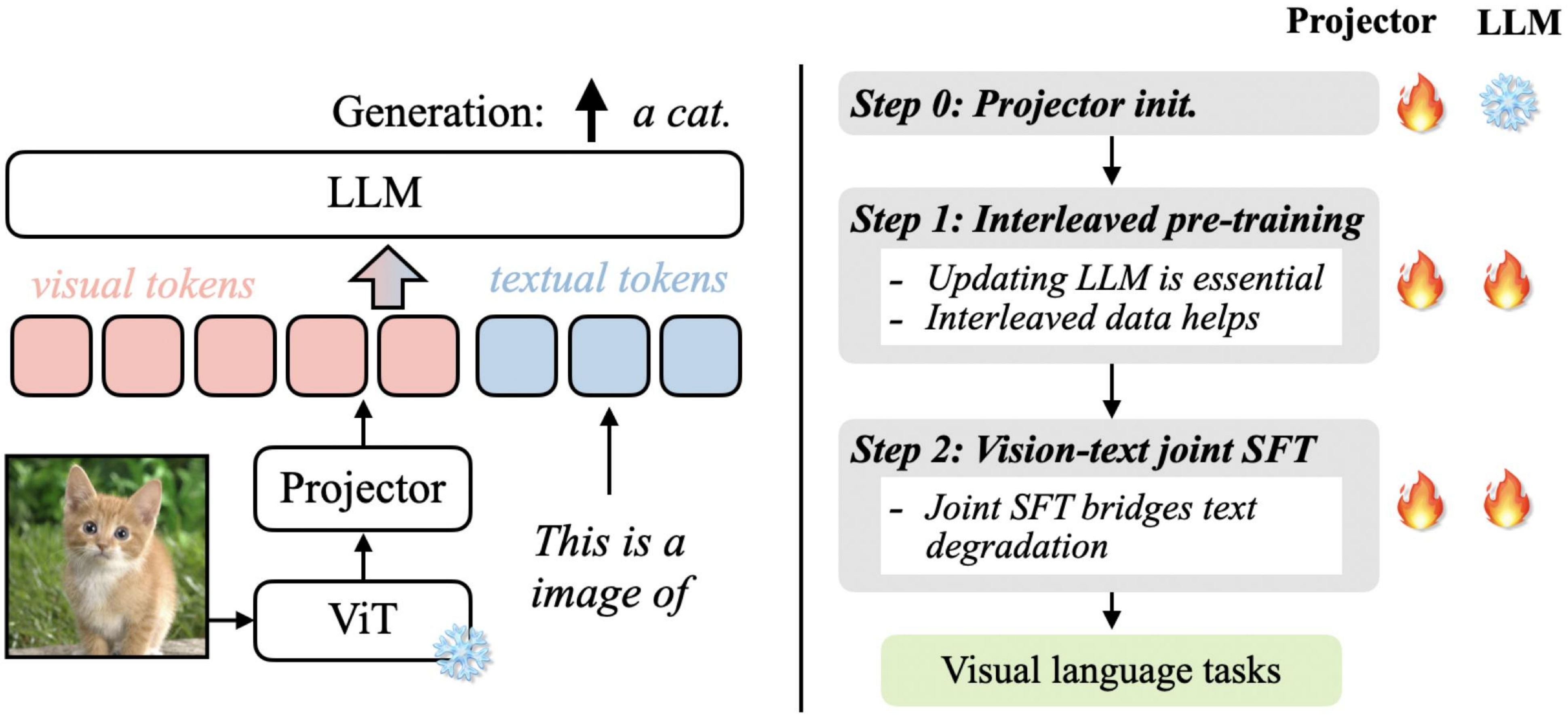
Image

Text

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021)

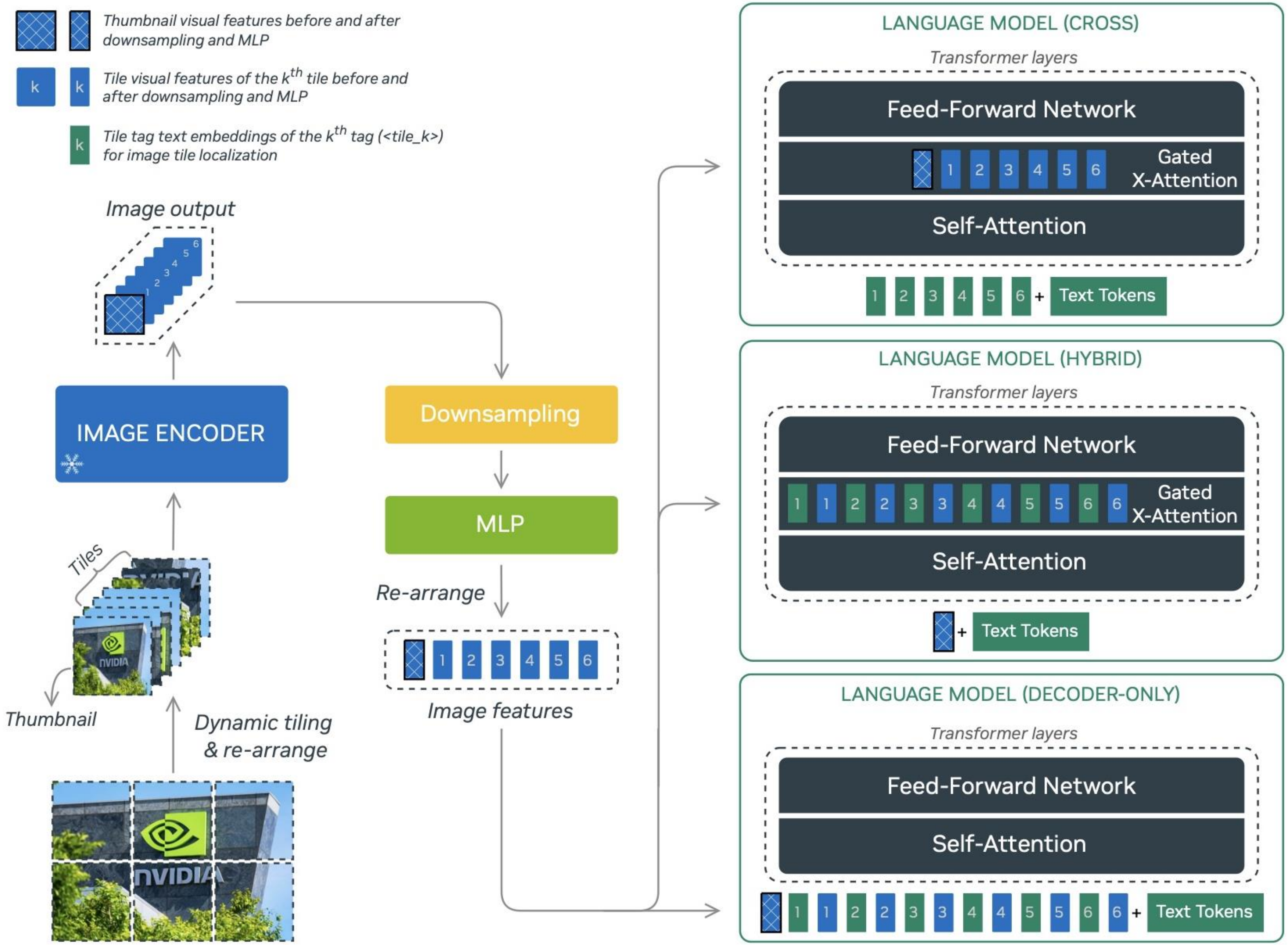
Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Multi-Image Guided Text Generation



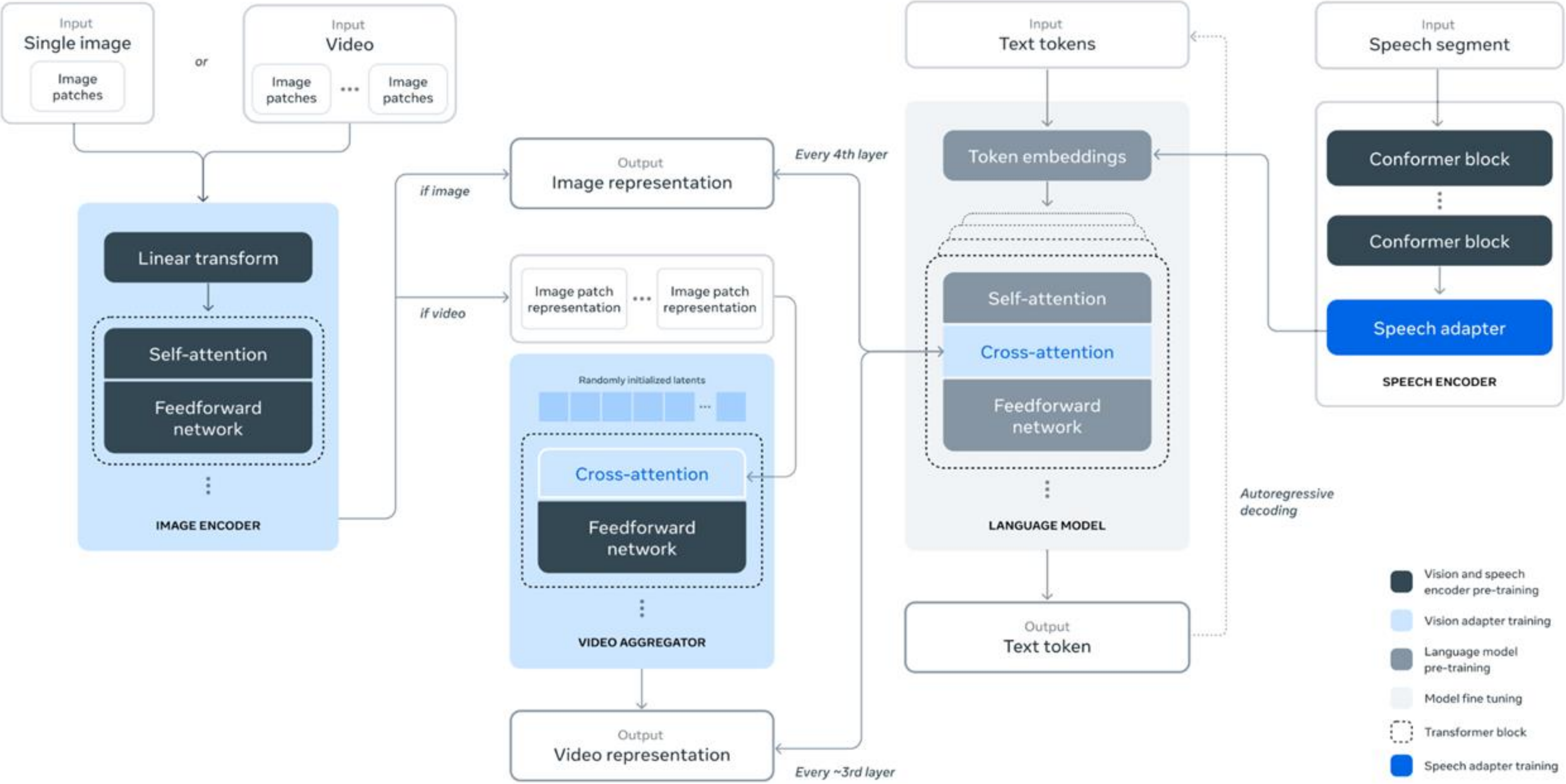
VILA: On Pre-training for Visual Language Models

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1



NVLM: Open Frontier-Class Multimodal LLMS (2024)

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1



[Introducing Llama 3.1: Our most capable models to date | Meta (2024)]

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Core Transformer Benefits

When Does It Shine

Delivers Semantically-Dense Embeddings

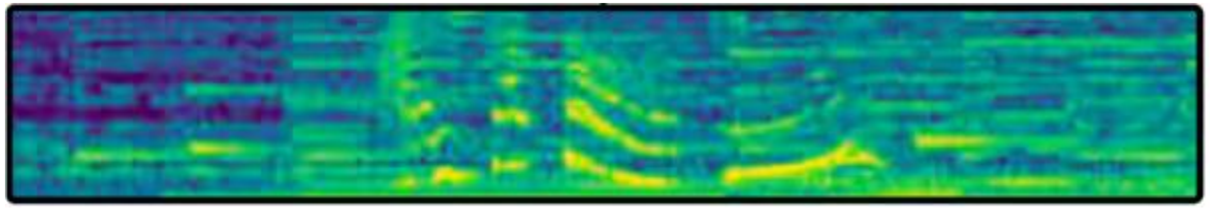
- Easy to reuse. Easy to transfer
- Sequence logic and token logic

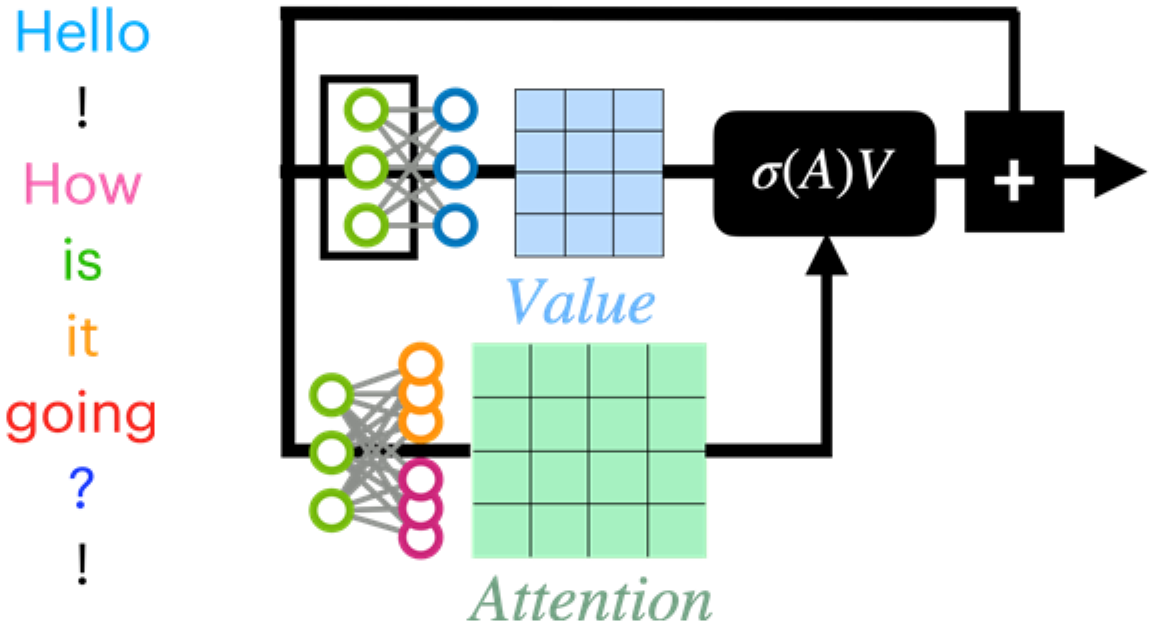
Enforce Attention-Oriented Logic

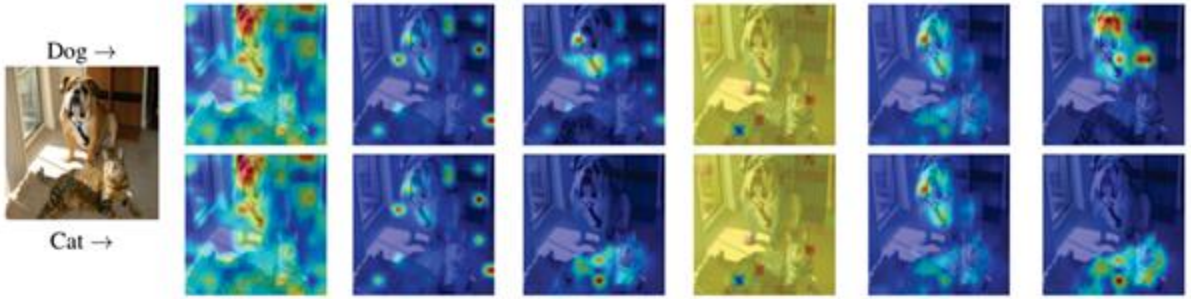
- Lightweight supervisory interface
- Doubles as interpretable “I care about X”

Synergize Semantically-Dense Embeddings

- Link two modalities for powerful results
- Good for zero-shot and training enforcement





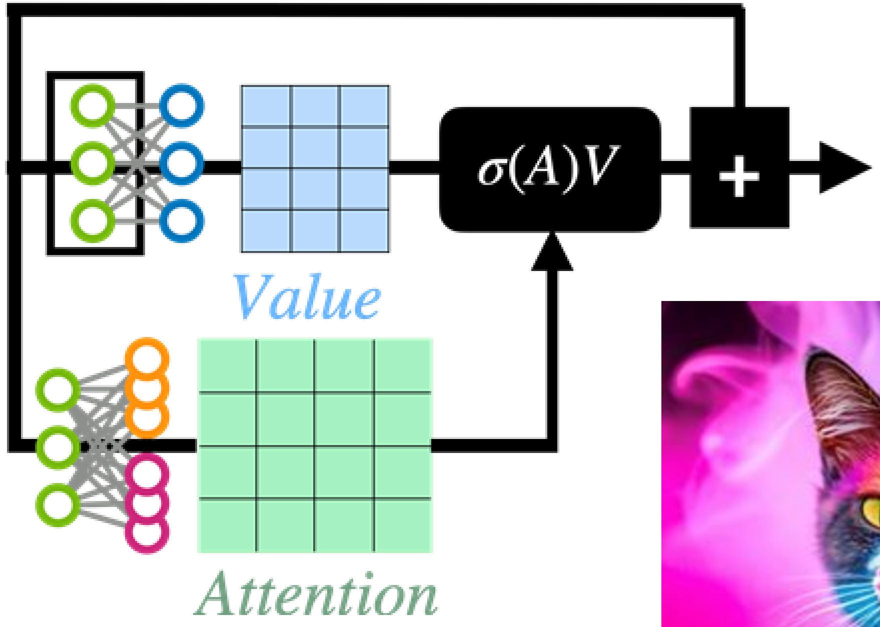


Transformer Interpretability Beyond Attention Visualization (2021)

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Image Generation?

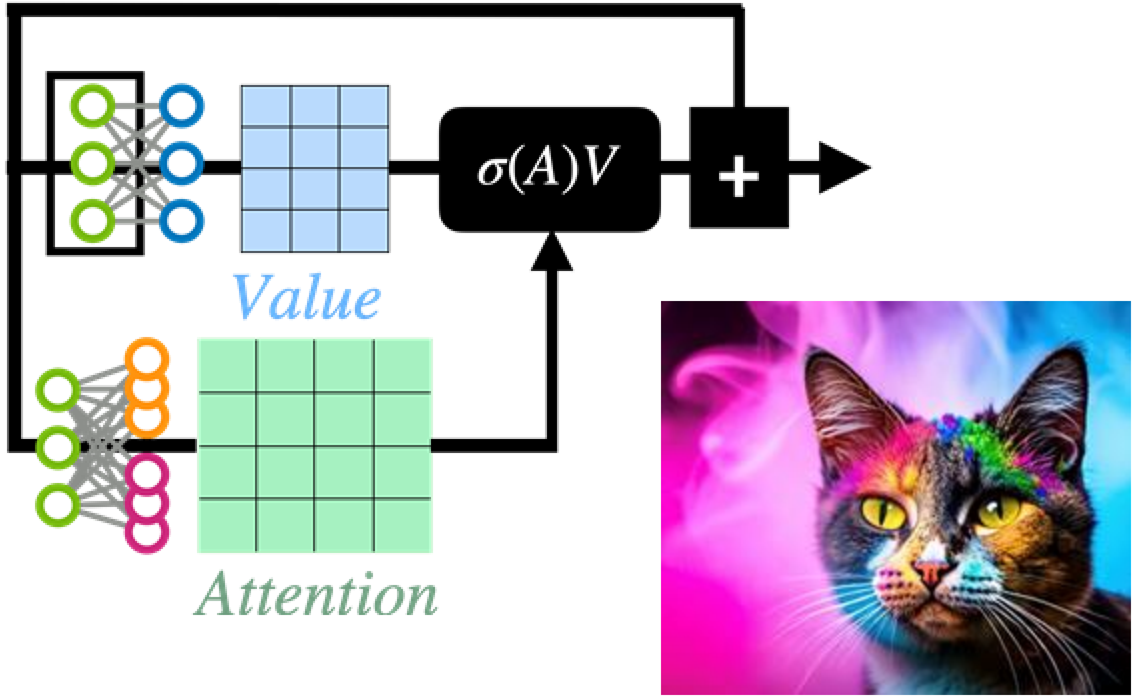
Can Transformers Generate Images?

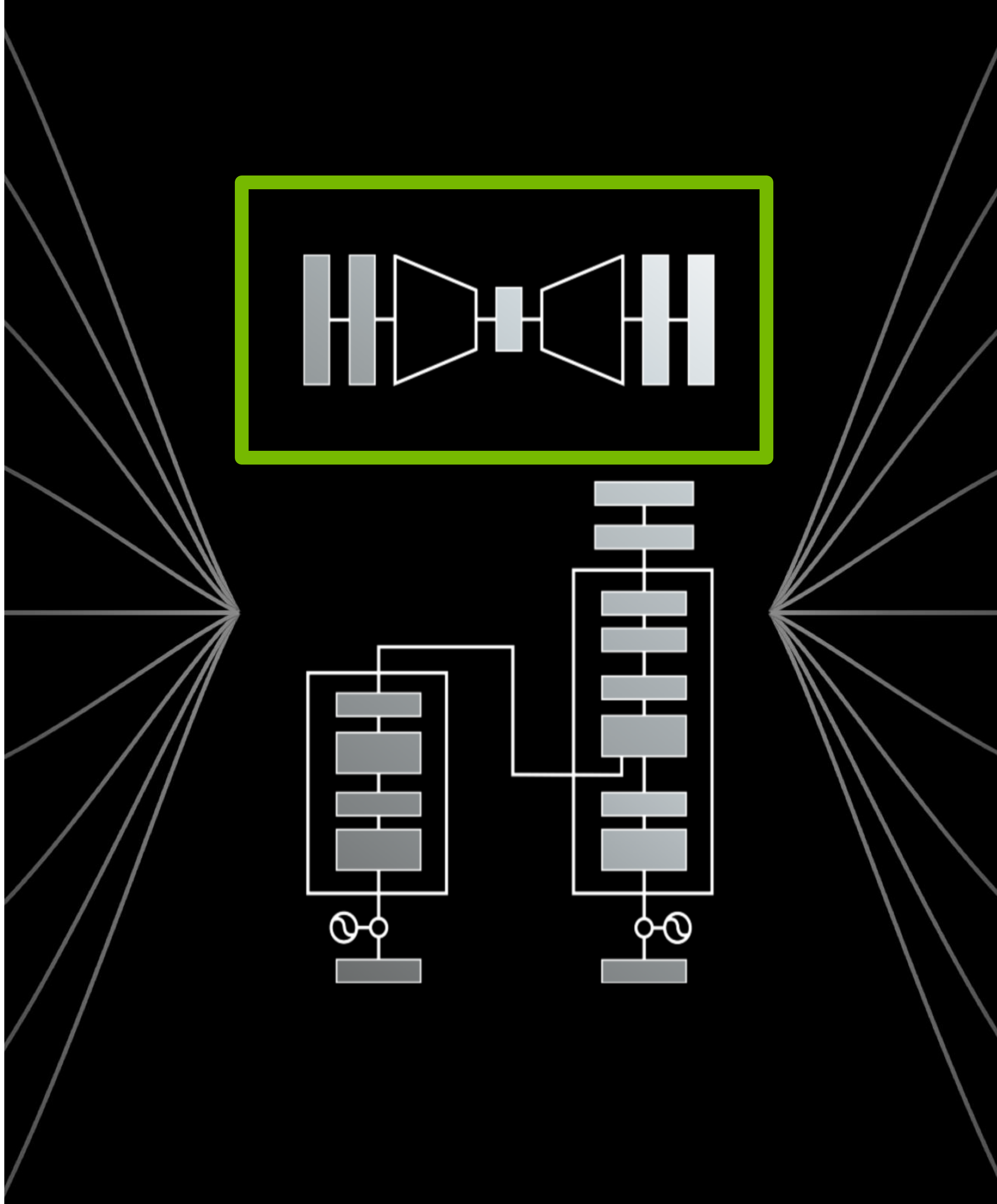


Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Image Generation?

Can Transformers Generate Images?

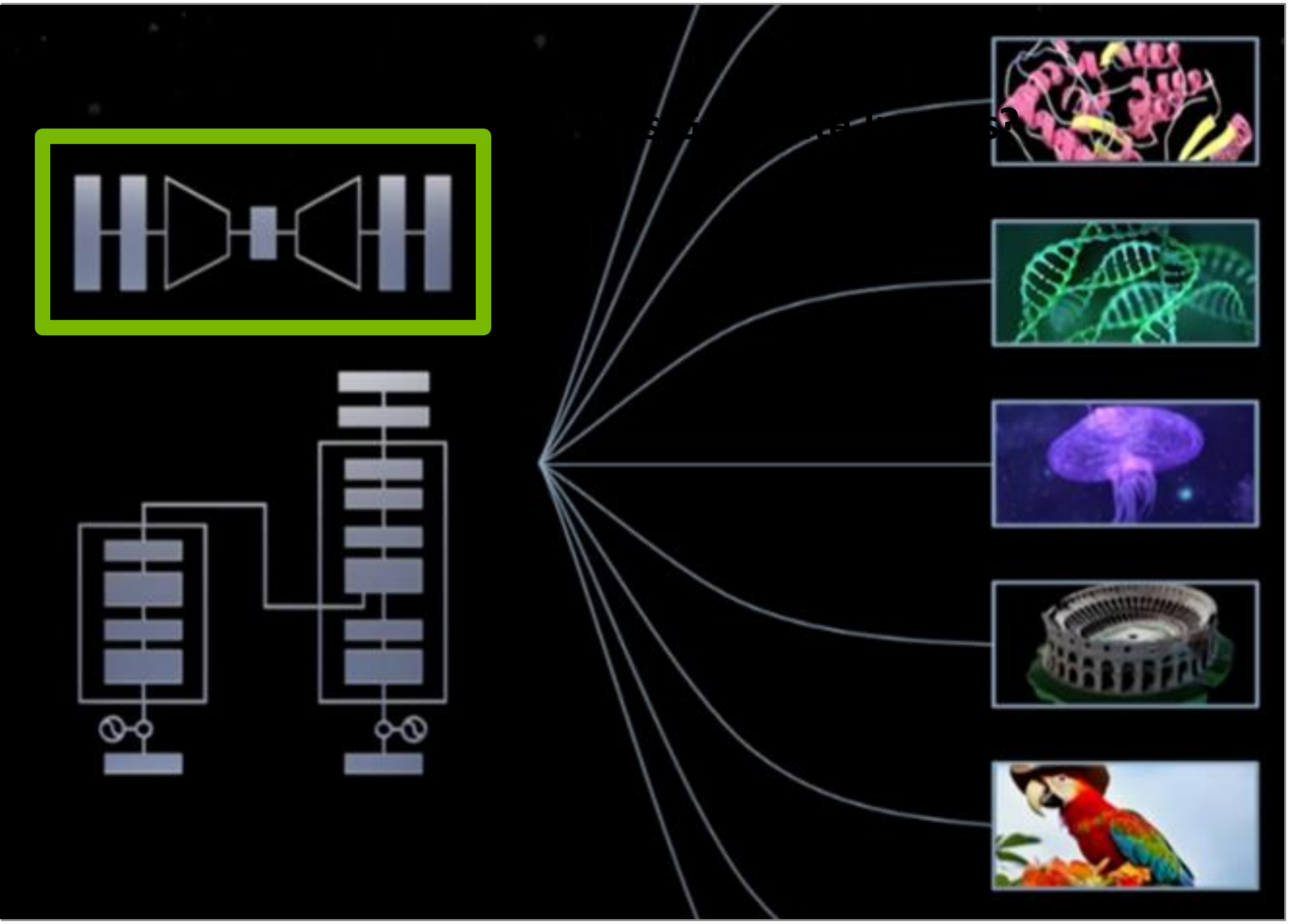


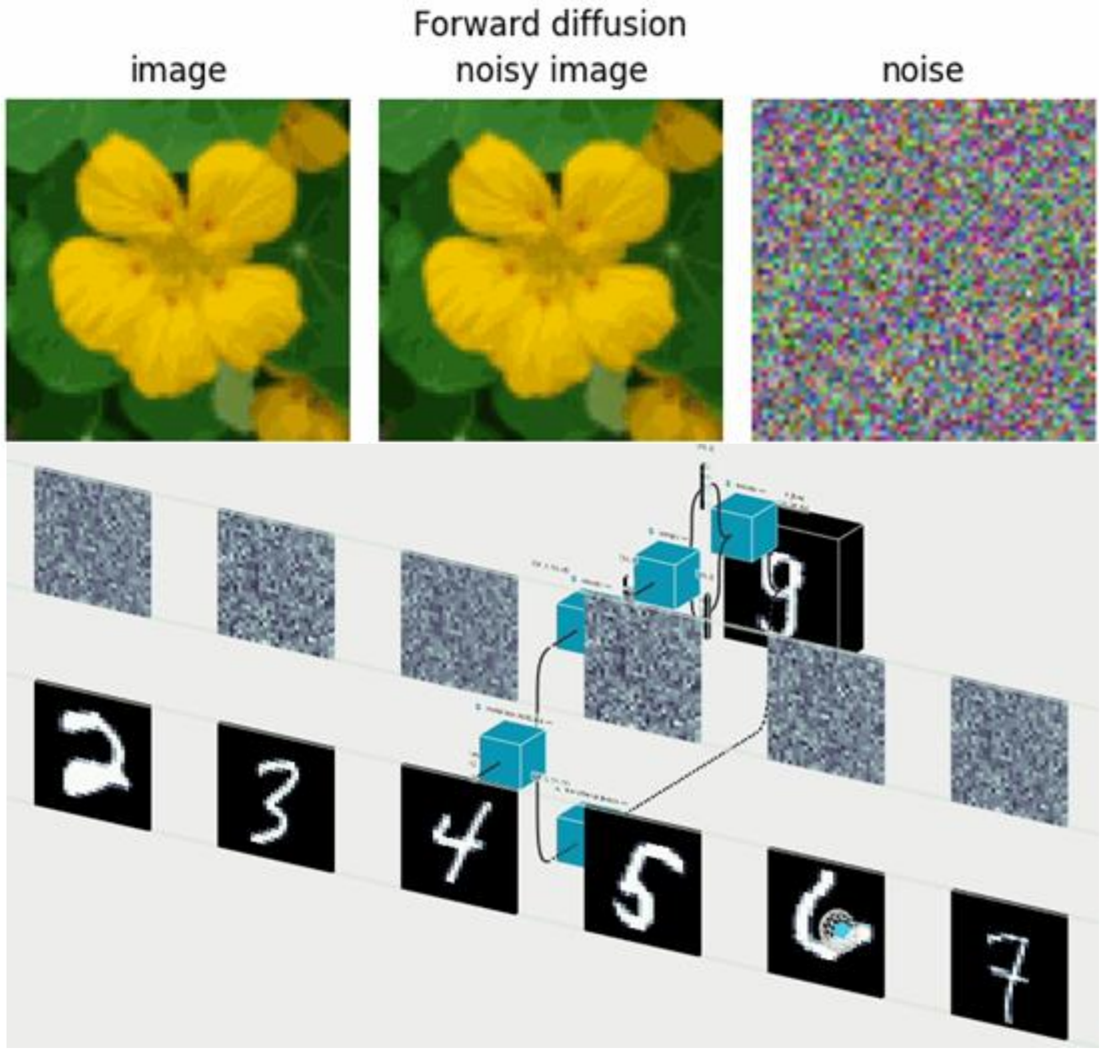


Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Image Generation?

Can Transformers Generate Images?





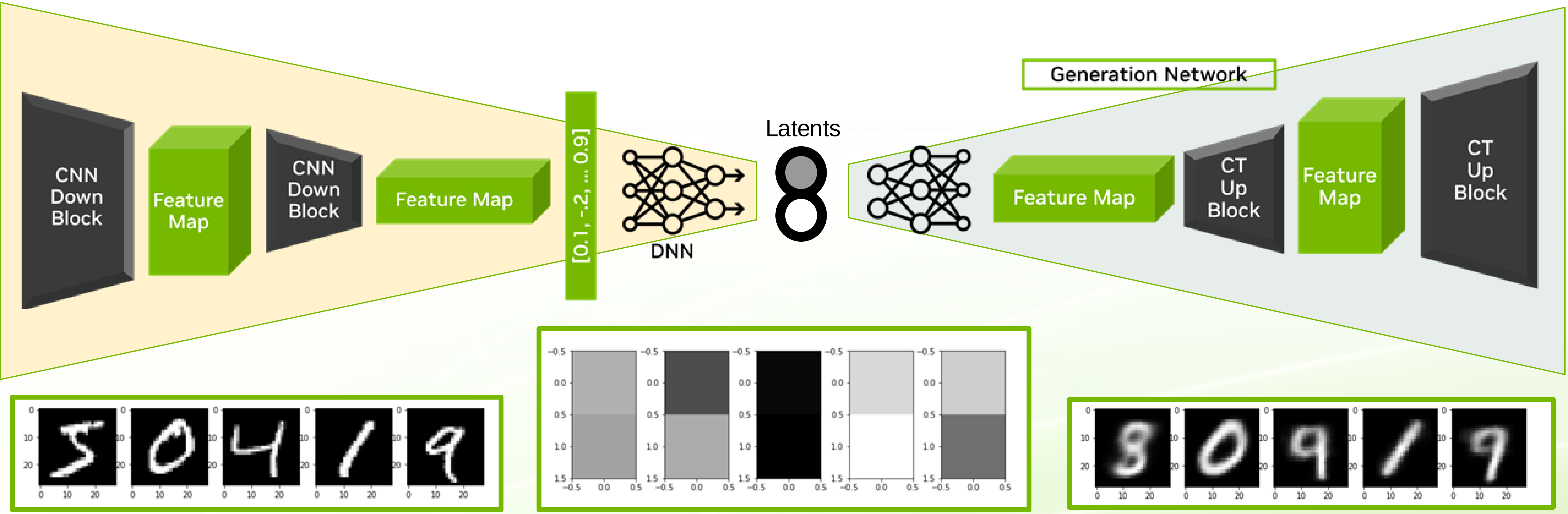
VAE in 3D | Differentiable Blog (2020)

Denoising Diffusion Implicit Models | Keras (2022)

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Making An Autoencoder

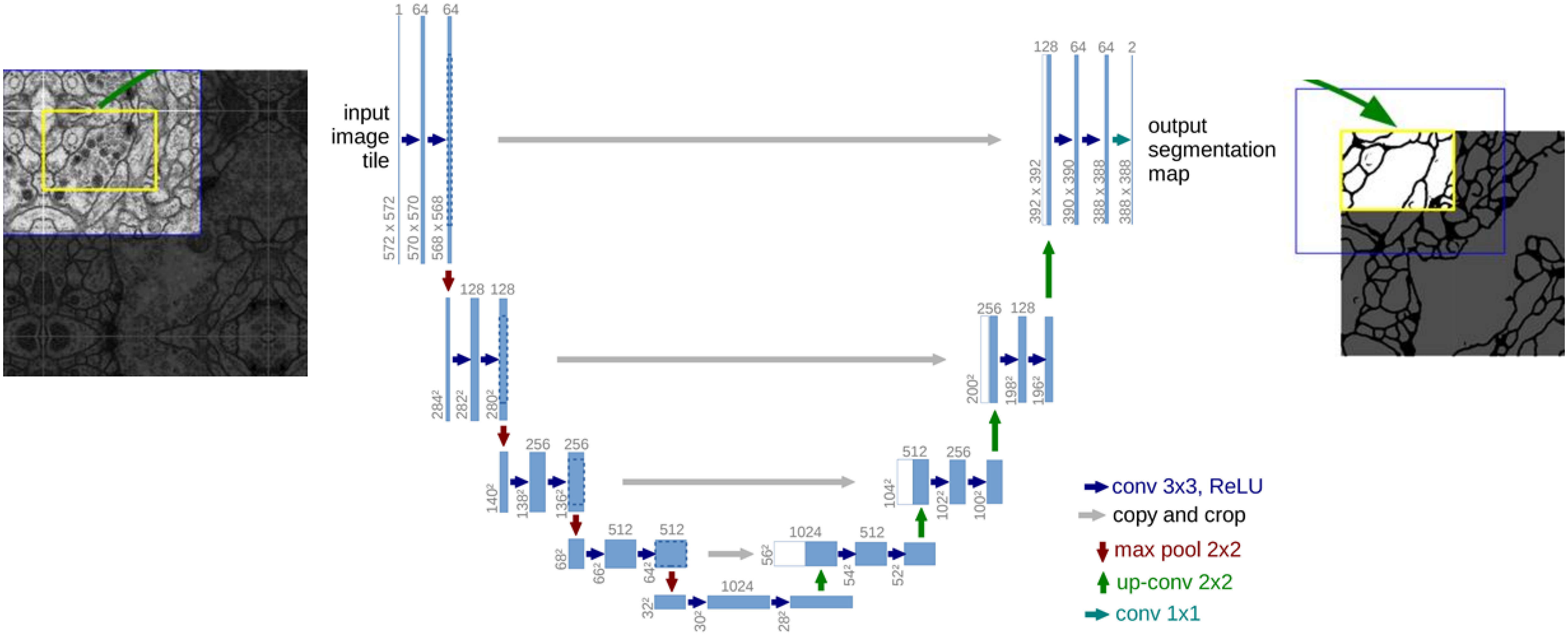
Decoding Your Encoding



Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Making A U-Net

Restyle An Image

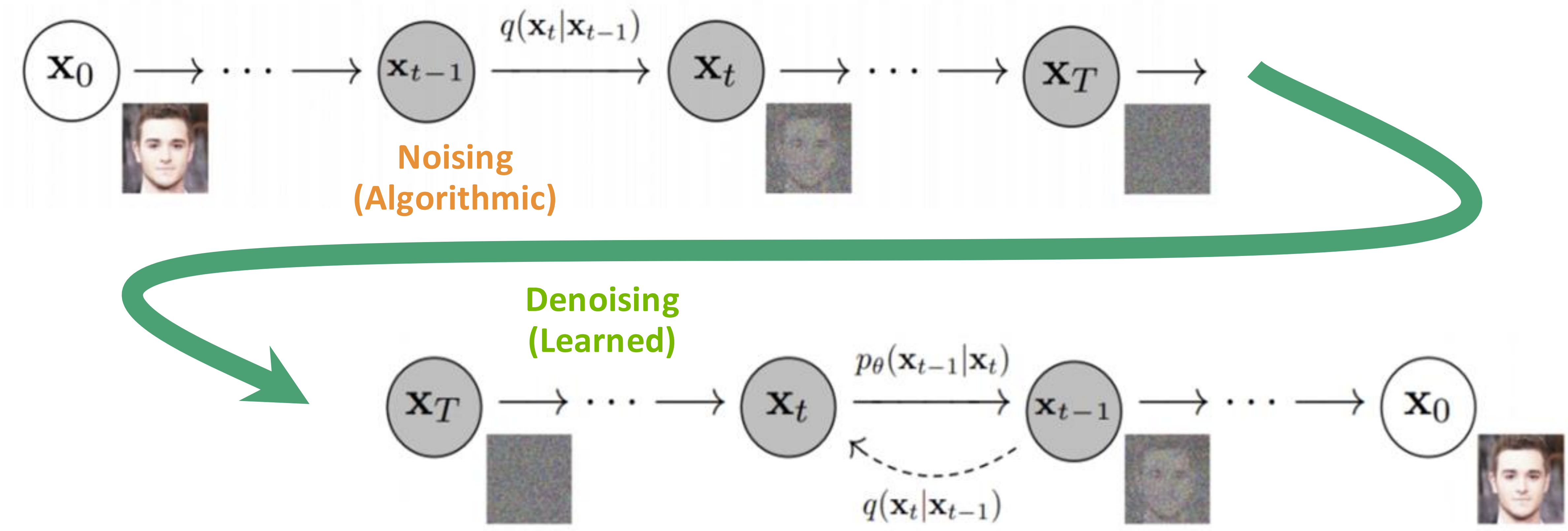


[U-Net: Convolutional Networks for Biomedical Image Segmentation (2015)]

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Making A Denoiser

Data by Noising, Train/Inferending by Denoising

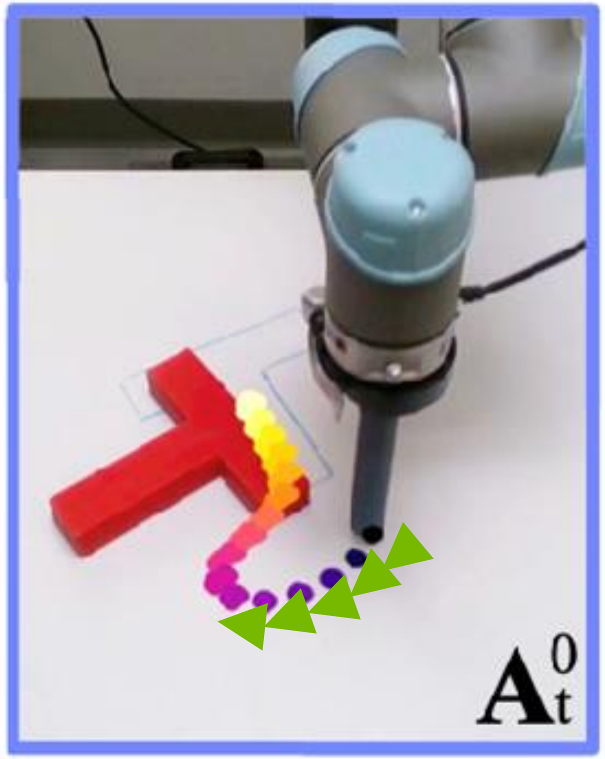

$\mathbf{x}_0 \longrightarrow \cdots \longrightarrow \mathbf{x}_{t-1} \xrightarrow{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \mathbf{x}_t \longrightarrow \cdots \longrightarrow \mathbf{x}_T \longrightarrow$
Noising
(Algorithmic)
Denoising
(Learned)
$\mathbf{x}_T \longrightarrow \cdots \longrightarrow \mathbf{x}_t \xrightarrow{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \mathbf{x}_{t-1} \longrightarrow \cdots \longrightarrow \mathbf{x}_0$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$

Denoising Diffusion Probabilistic Models (2020)

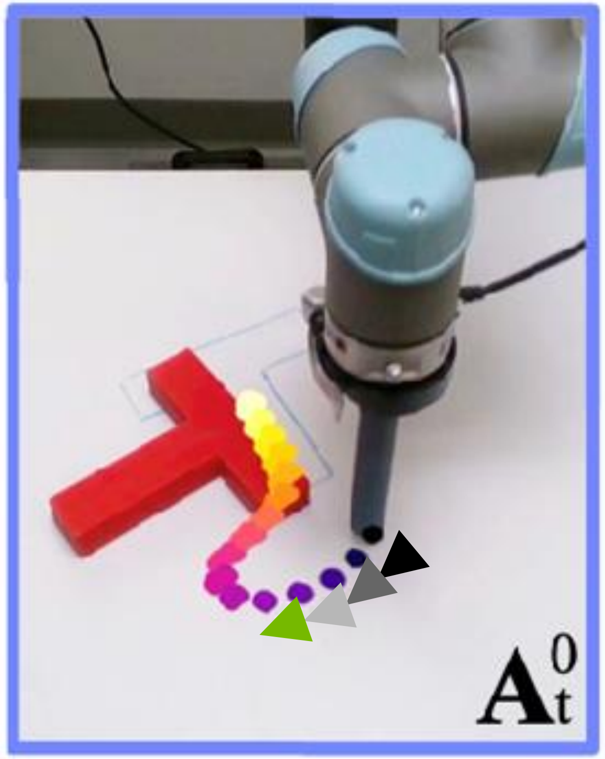
Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Policy Networks

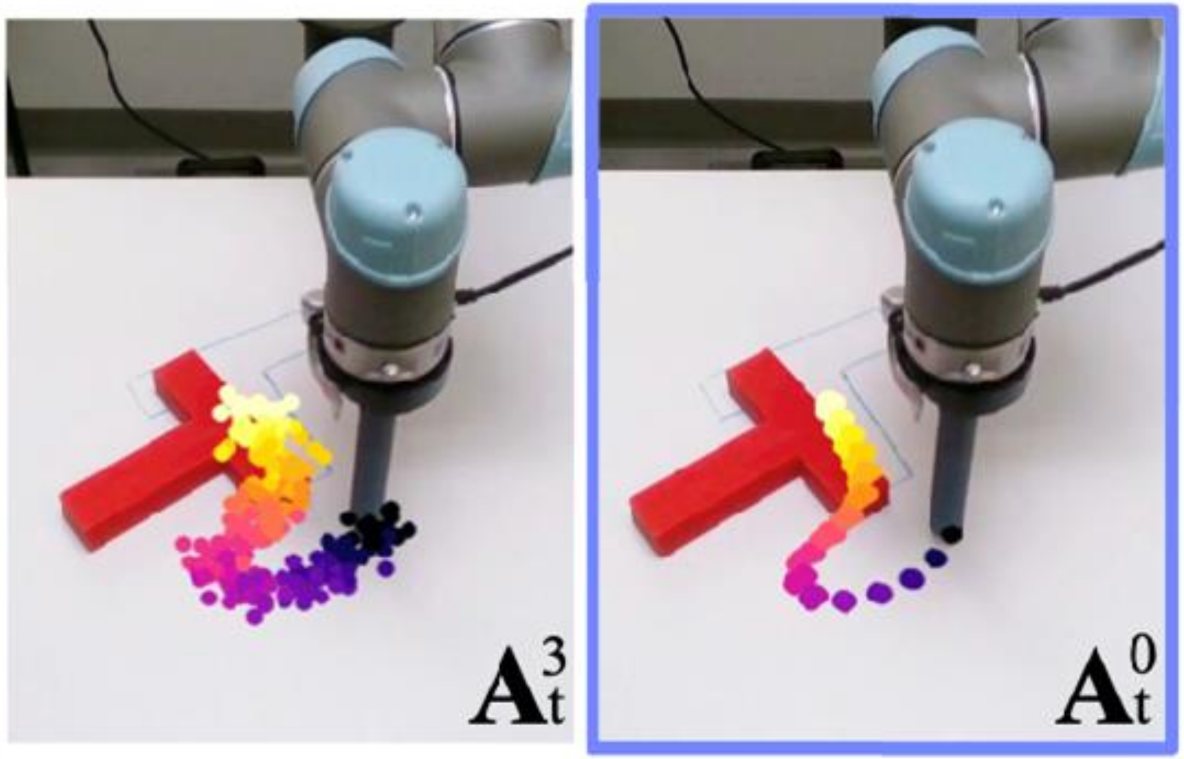
Where Direct, Autoregression, or Diffusion Can Work



[Spatial] Delta Prediction From Previous Position/Delta



[Sequential] Prediction From Previous Position(s)



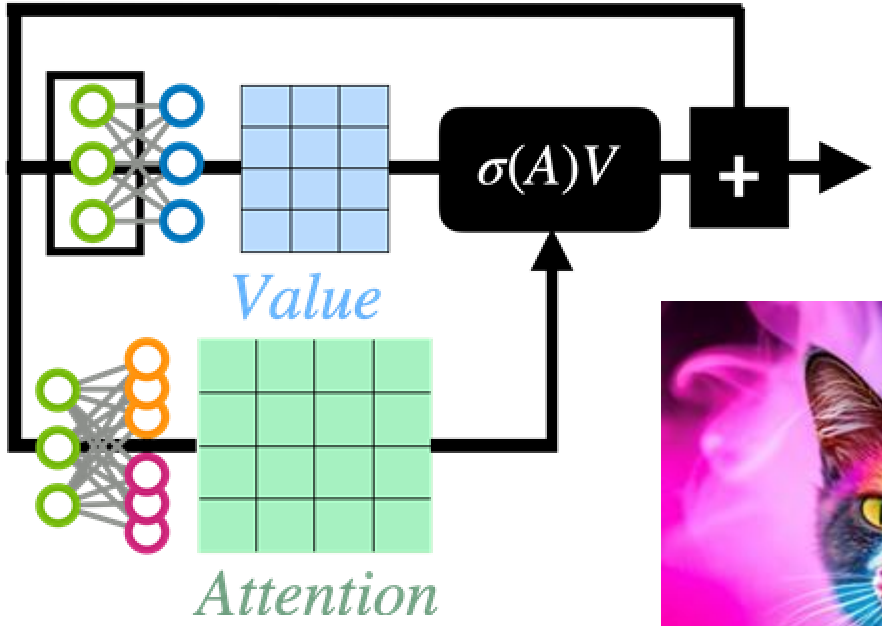
[Diffusion] Progression of points into final structured alignment

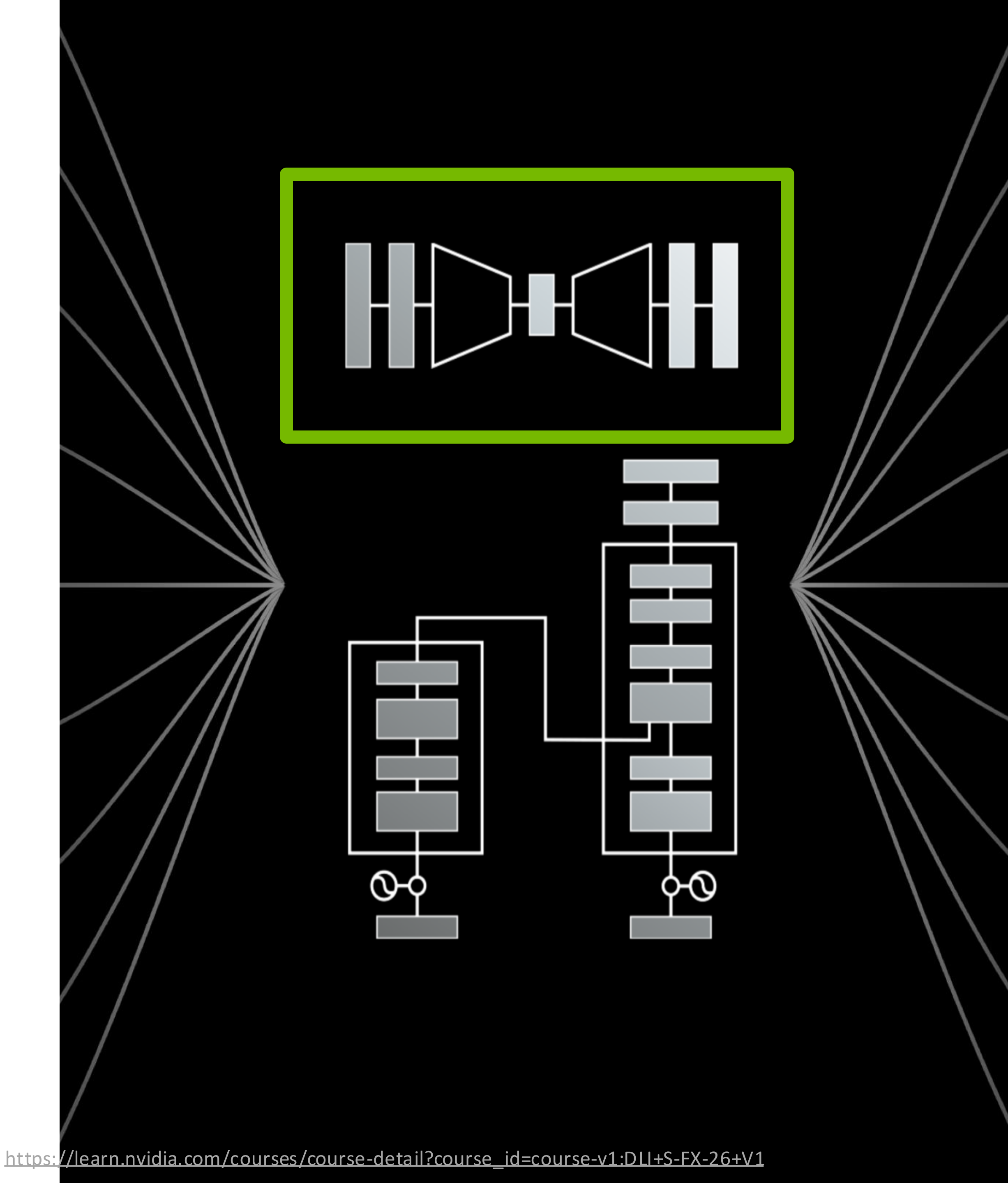
[Diffusion Policy: Visuomotor Policy Learning via Action Diffusion (2024)]

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Helping Image Generation?

Can Transformers HELP Generate Images?

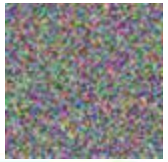


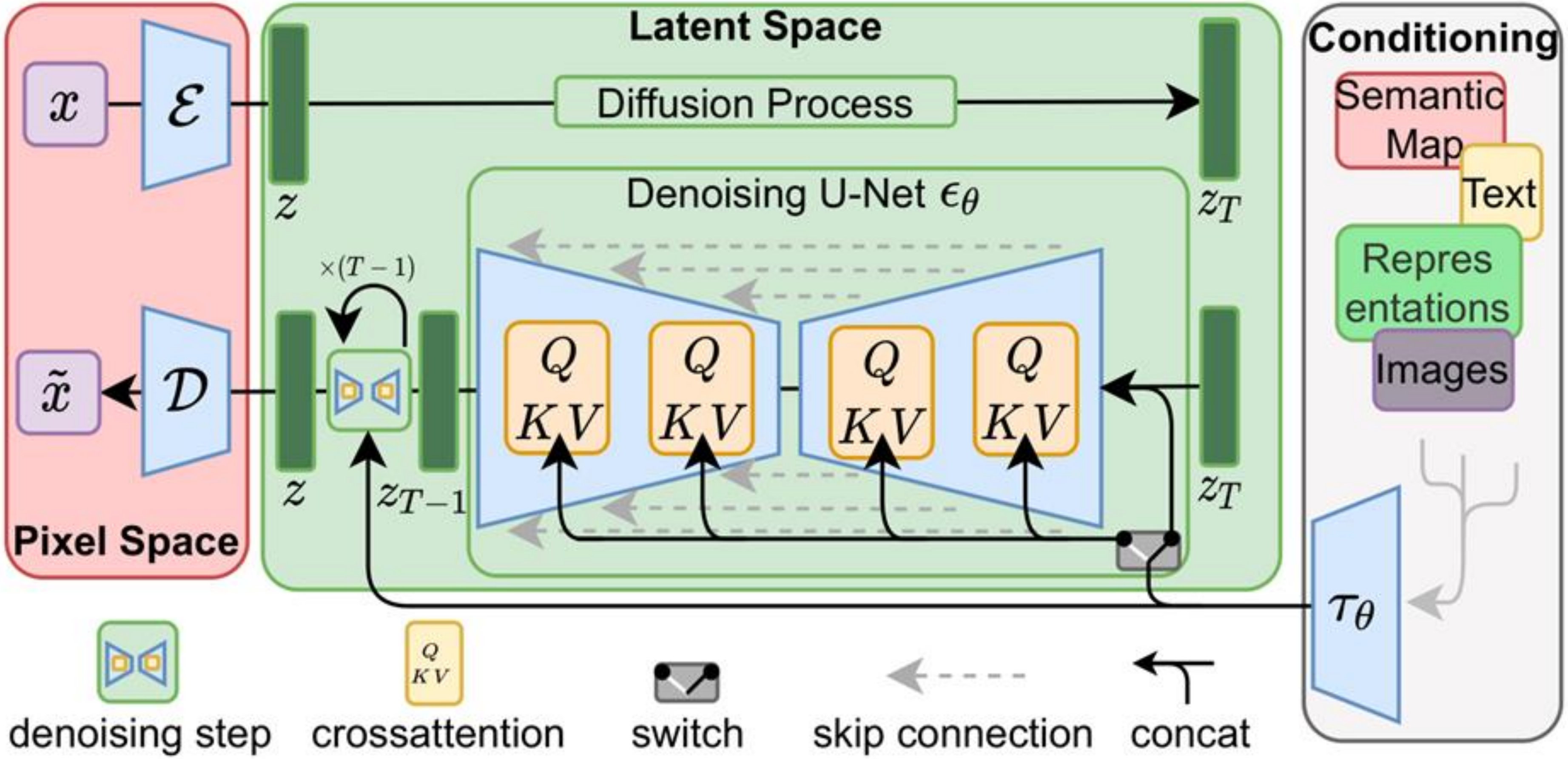


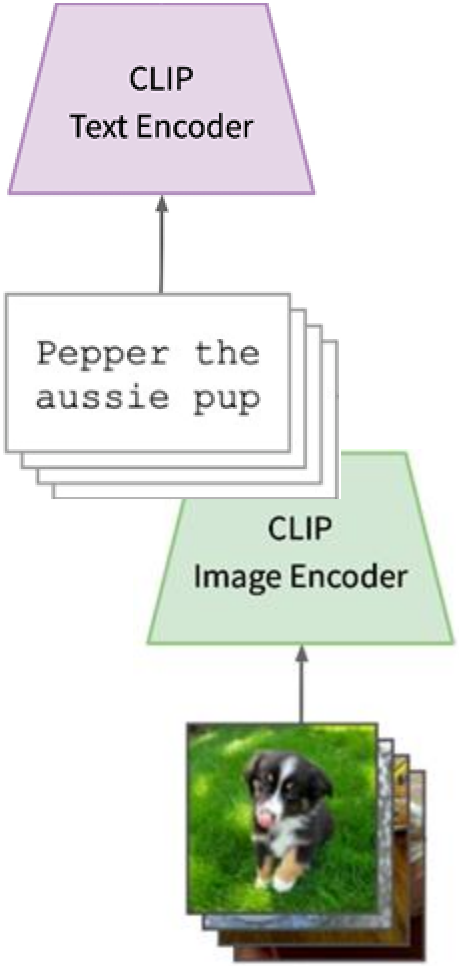
Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Image Generation?

Can Transformers **Help** Generate Images?



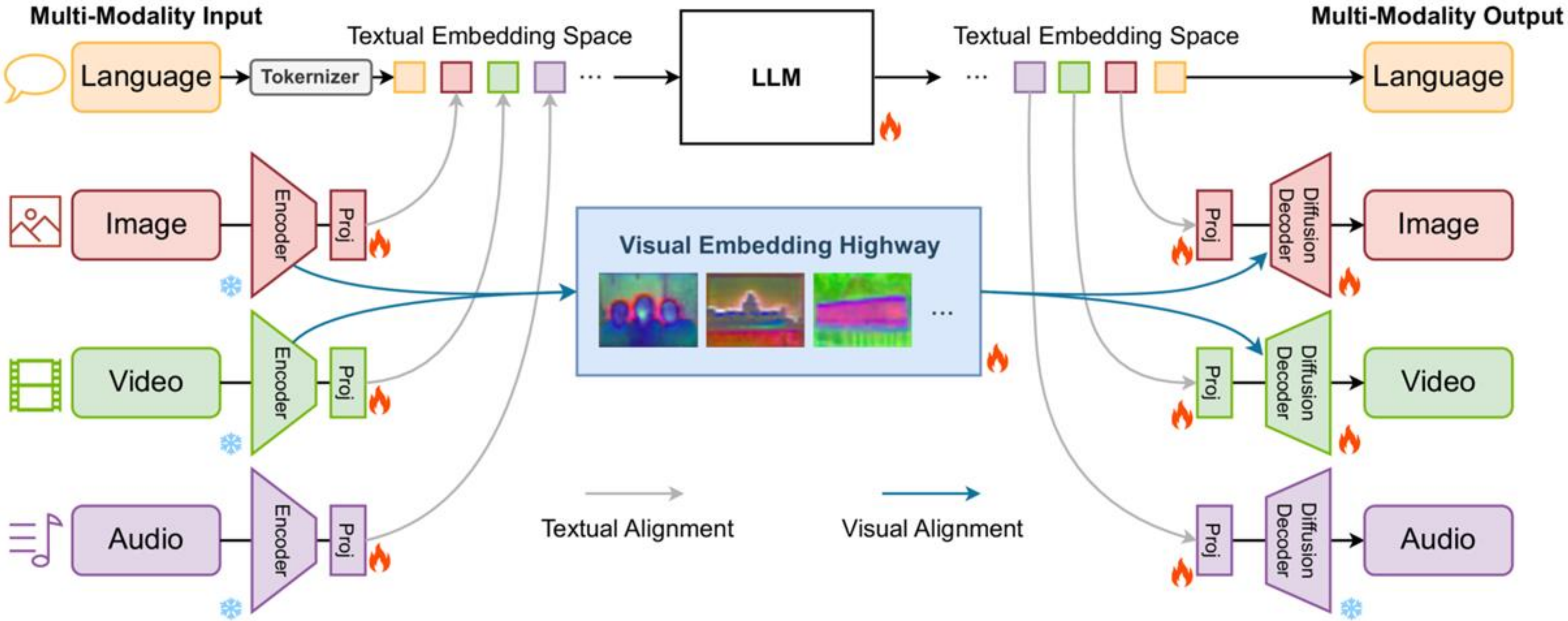




High-Resolution Image Synthesis with Latent Diffusion Models

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Synergized Multi-Domain Encoders

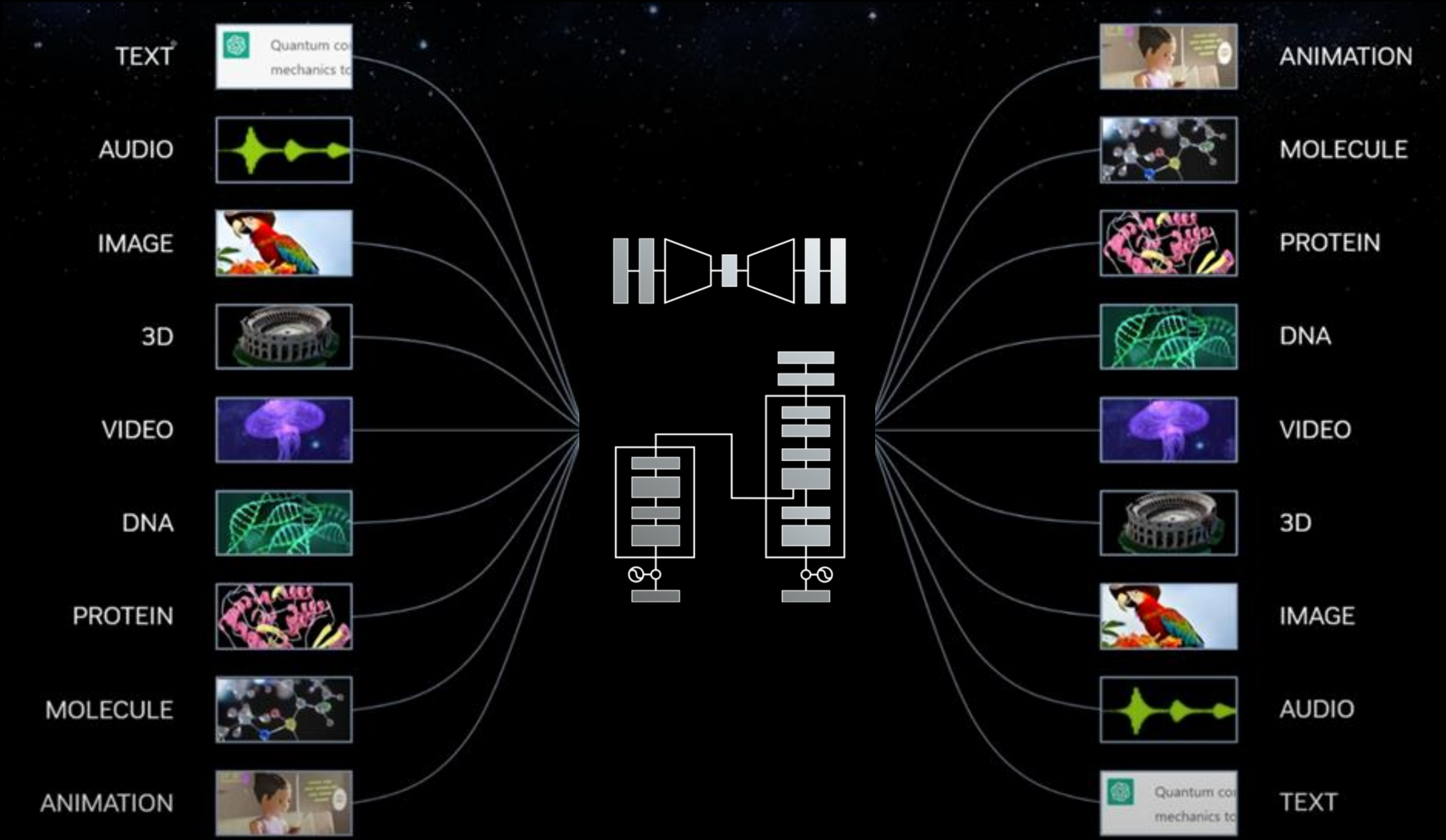


X-VILA: Cross-Modality Alignment for Large Language Model (2024)

Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Modular Modalities

Where Can The Transformer Fit?



Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

References

- Numa Dhamani and Maggie Engler (2024), Introduction to Generative AI, Manning
- Denis Rothman (2024), Transformers for Natural Language Processing and Computer Vision: Explore Generative AI and Large Language Models with Hugging Face, ChatGPT, GPT-4V, and DALL-E 3, 3rd Edition, Packt Publishing
- NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1
- NVIDIA DLI (2024), Building RAG Agents with LLMs, https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1
- NVIDIA DLI (2024), Generative AI with Diffusion Models, https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-14+V1