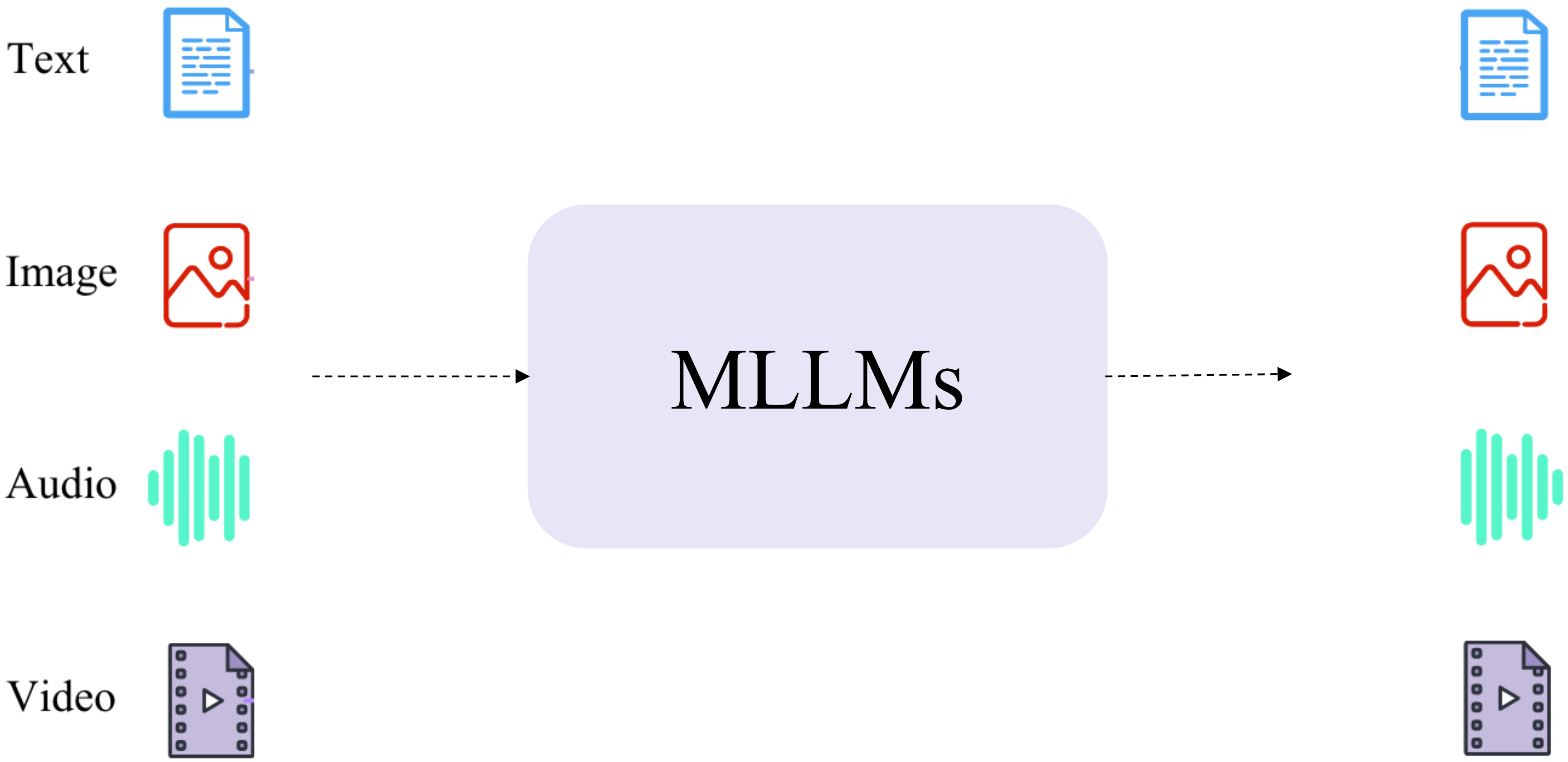
## Module 10 - Extra

# Multimodal
# Large Language Models

AI VIET NAM
Nguyen Quoc Thai

# Objectives

## Multimodal Large Language Models

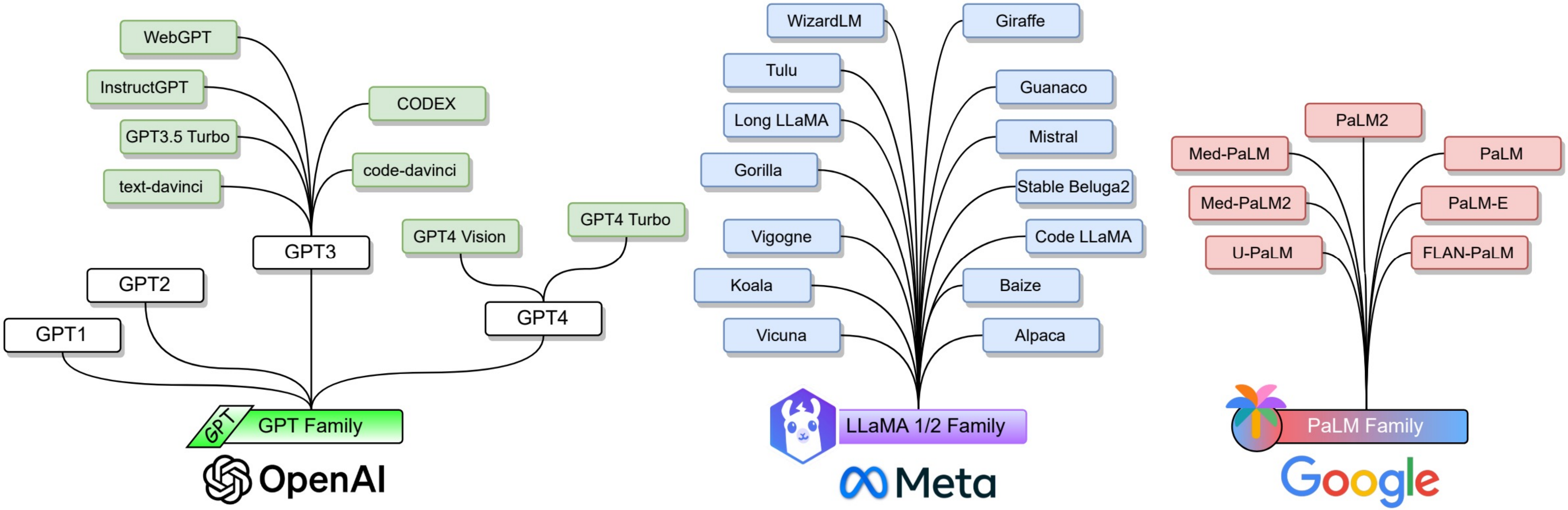


# Outline

- Introduction
- Multimodal Large Language Models
- BLIP-2
- NExT-GPT: Any-to-Any MLLM

# Introduction

## Large Language Models



# Introduction

## ! Large Language Models

- “Very” large LMs: models of 100+ billion parameters

  GPT3 (175B), BLOOM (176B), PaLM (540B), GLaM (1200B)…

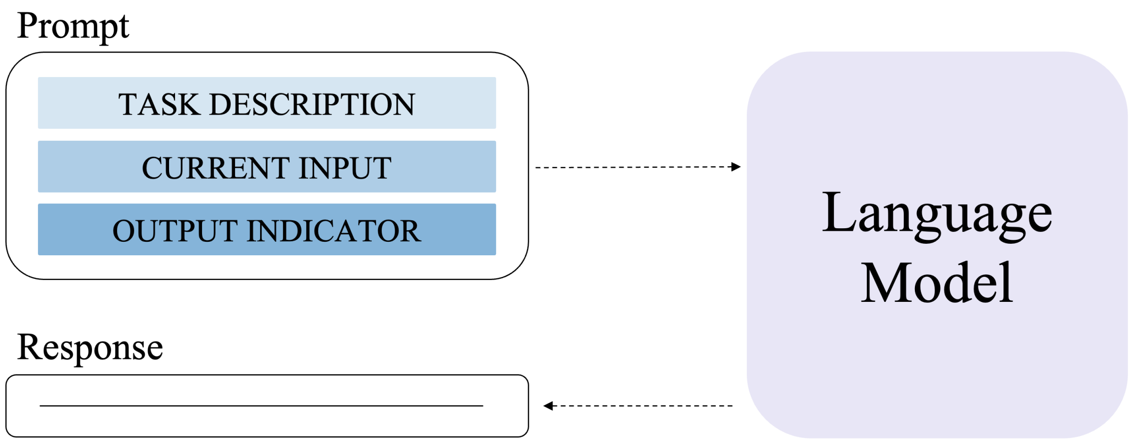
- Data scale: usually in the order of trillions of tokens

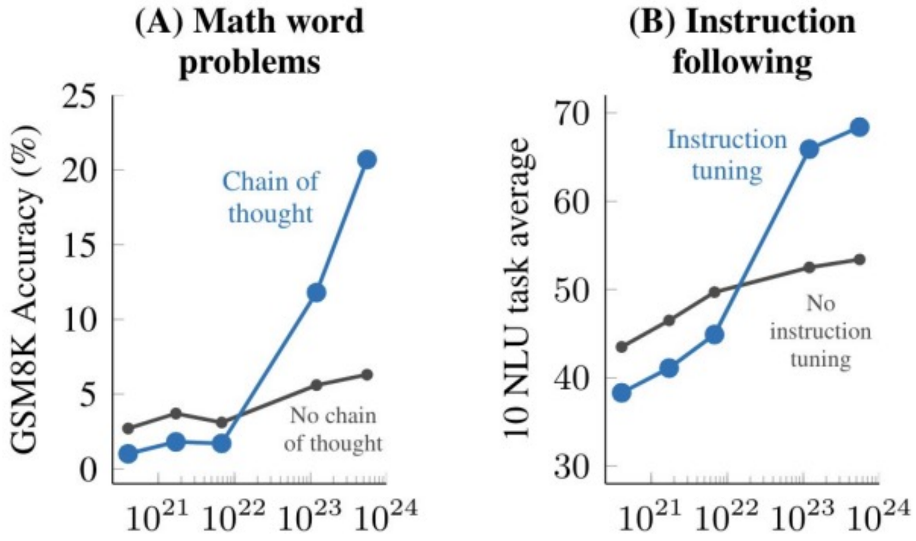
  GPT3 (0.5 trillion tokens), LLaMa (1.4 trillion tokens)

# Introduction

## Large Language Models

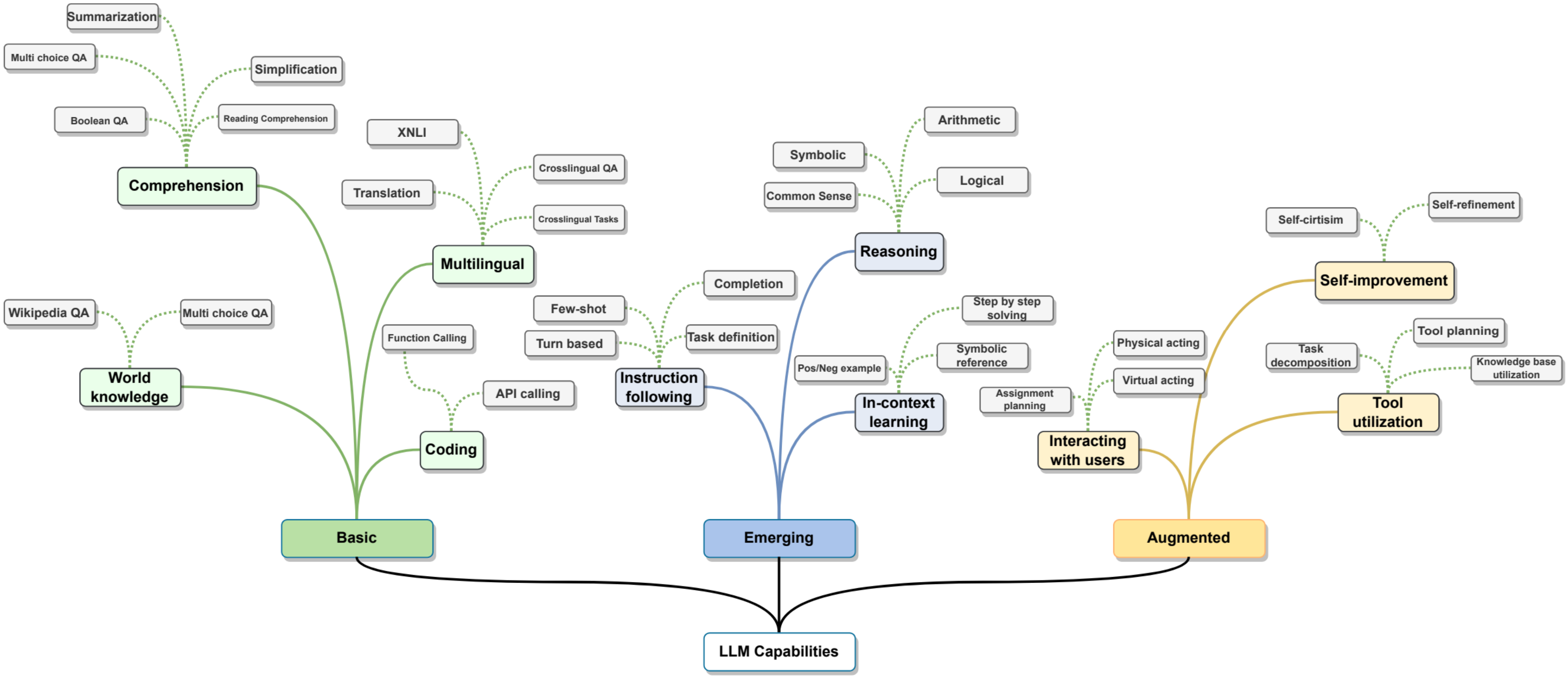
- The promise: one single model to solve many NLP tasks
- Emergent properties in LLMs





# Introduction

## Large Language Models



LLM Capabilities

Basic

Comprehension
Summarization
Multi choice QA
Simplification
Boolean QA
Reading Comprehension

Multilingual
XNLI
Crosslingual QA
Translation
Crosslingual Tasks

World knowledge
Wikipedia QA
Multi choice QA

Coding
Function Calling
API calling

Emerging

Reasoning
Arithmetic
Symbolic
Logical
Common Sense

Instruction following
Completion
Few-shot
Task definition
Turn based

In-context learning
Step by step solving
Symbolic reference
Pos/Neg example

Augmented

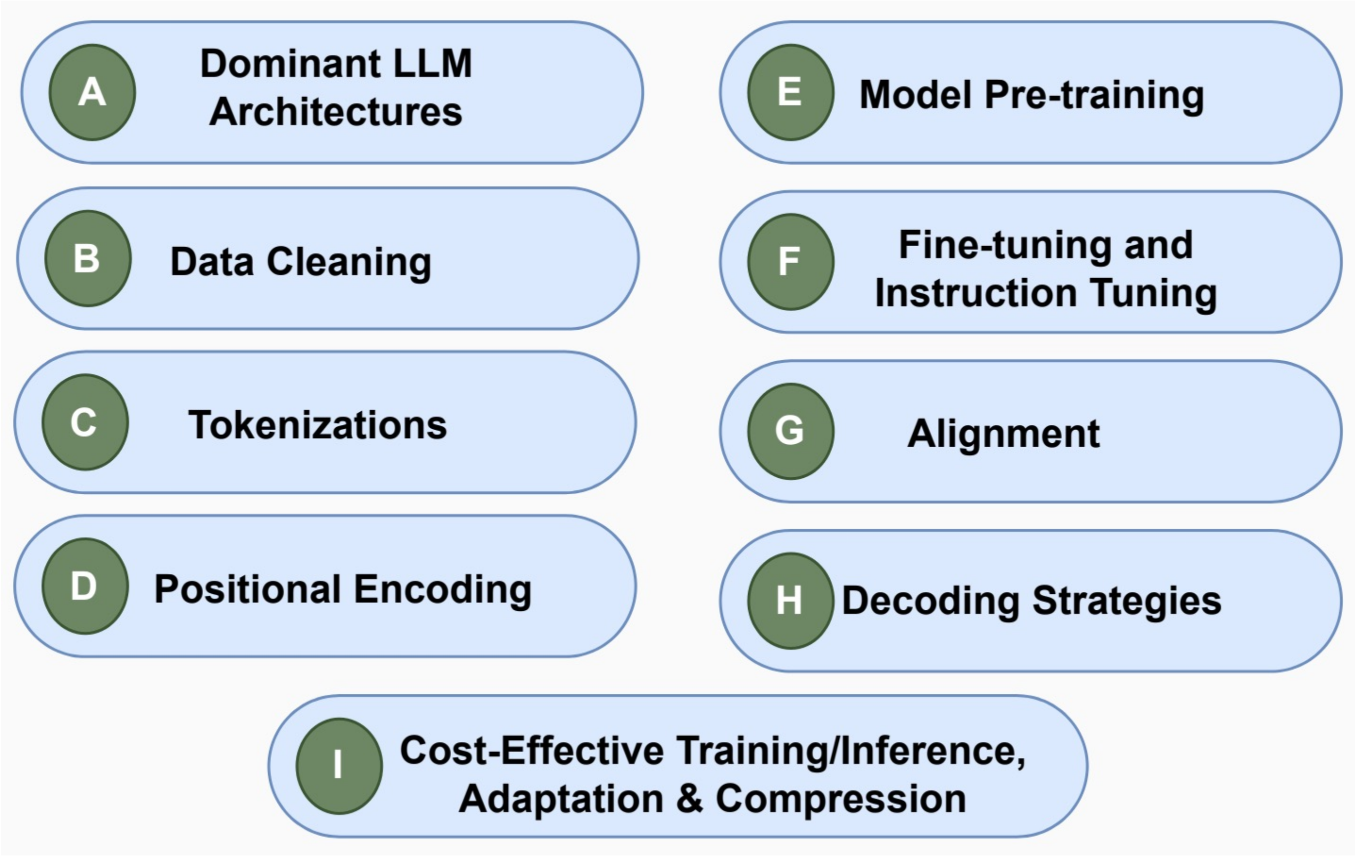
Self-improvement
Self-cirtisim
Self-refinement

Tool utilization
Tool planning
Task decomposition
Knowledge base utilization

Interacting with users
Physical acting
Virtual acting
Assignment planning

# Introduction

## Large Language Models



A Dominant LLM Architectures

B Data Cleaning

C Tokenizations

D Positional Encoding

E Model Pre-training

F Fine-tuning and Instruction Tuning

G Alignment

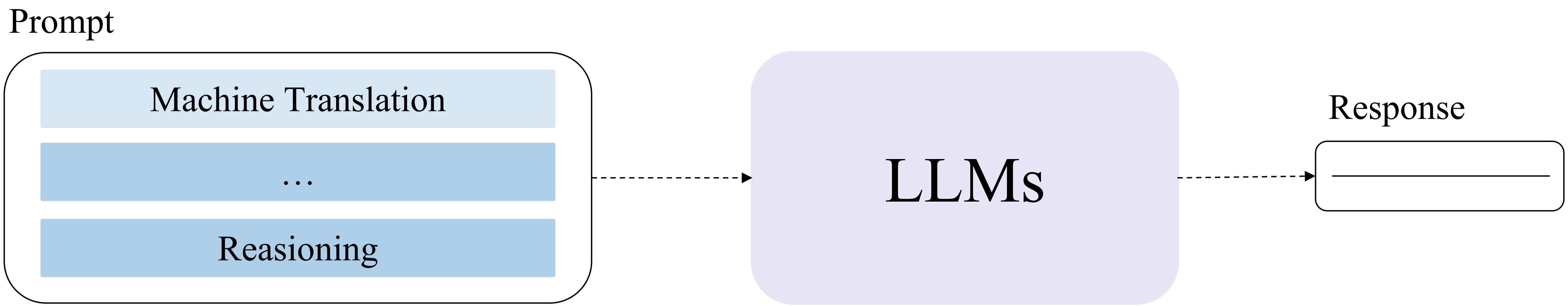
H Decoding Strategies

I Cost-Effective Training/Inference, Adaptation & Compression

# Introduction

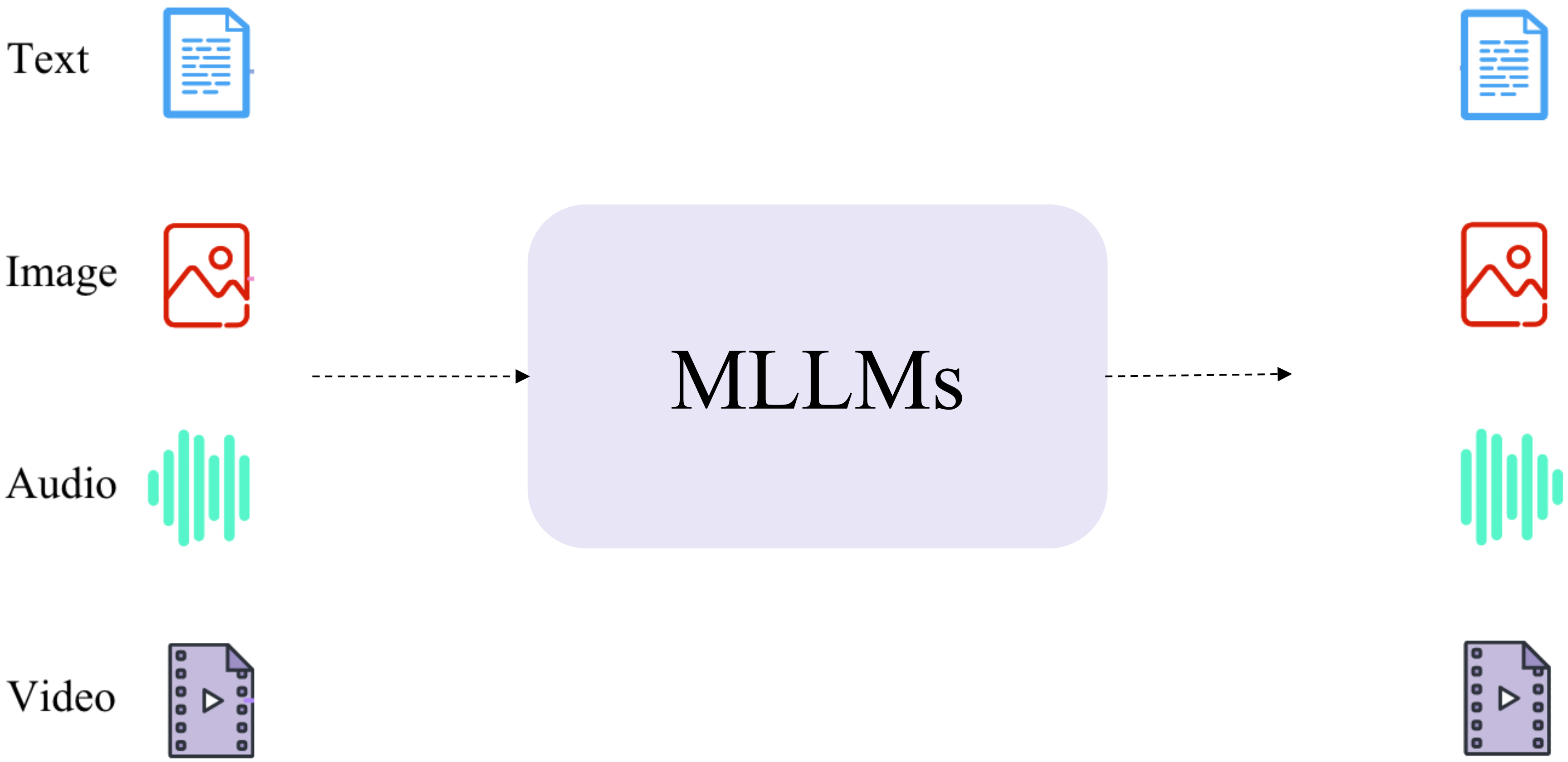
## Large Language Models

- Solve many NLP tasks



# Introduction

## Multimodal Large Language Models

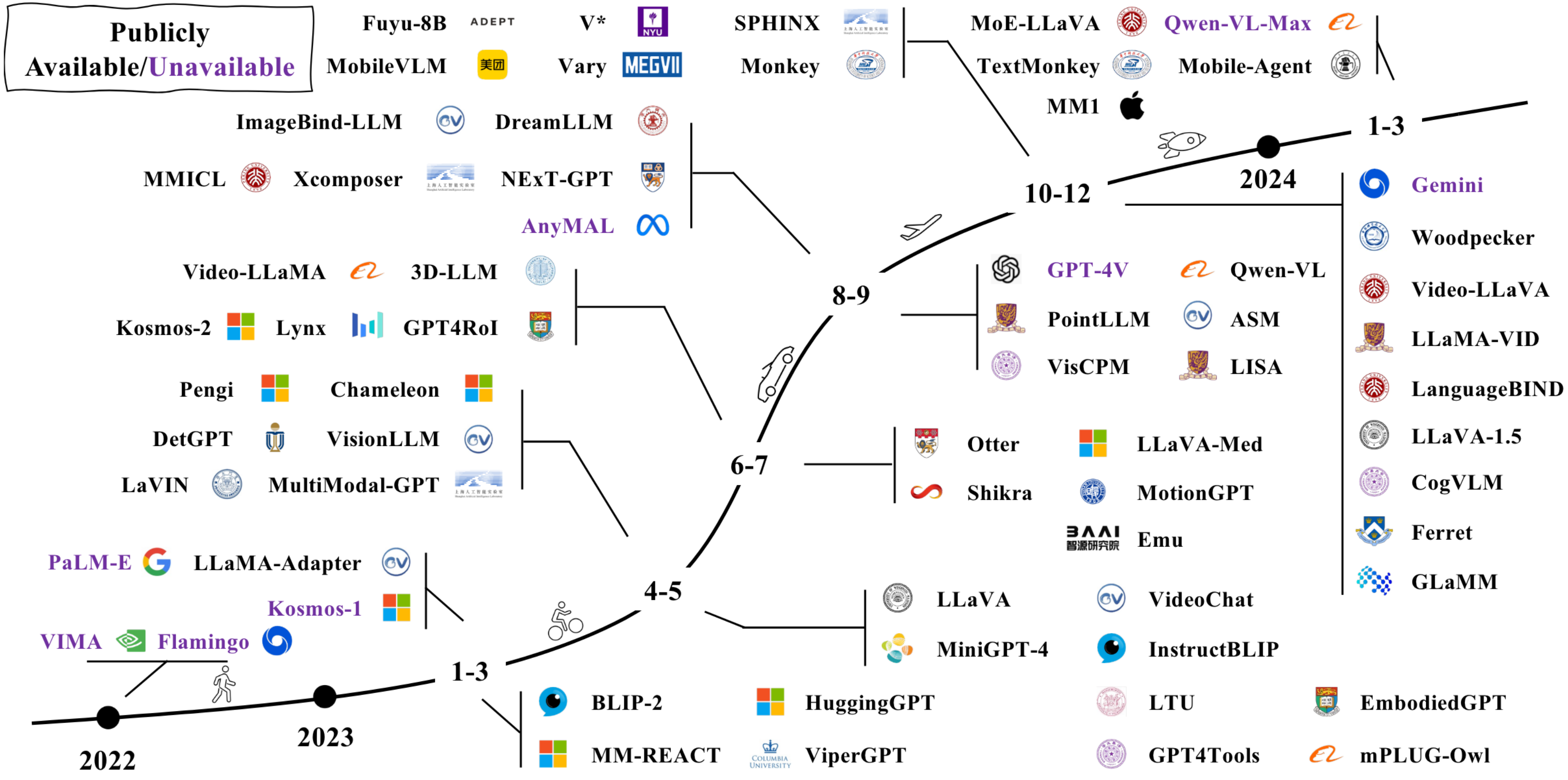


# Outline

- Introduction
- Multimodal Large Language Models
- BLIP-2
- NExT-GPT: Any-to-Any MLLM

# Multimodal LLMs

## The milestones of Multimodal LLMs



Publicly Available/Unavailable

2022

VIMA

2023

1-3: PaLM-E, LLaMA-Adapter, Kosmos-1, Flamingo, BLIP-2, MM-REACT, HuggingGPT, ViperGPT, LTU, GPT4Tools, EmbodiedGPT, mPLUG-Owl

4-5: Pengi, Chameleon, DetGPT, VisionLLM, LaVIN, MultiModal-GPT, LLaVA, VideoChat, MiniGPT-4, InstructBLIP

6-7: Video-LLaMA, 3D-LLM, Kosmos-2, Lynx, GPT4RoI, Otter, LLaVA-Med, Shikra, MotionGPT, Emu

8-9: ImageBind-LLM, DreamLLM, MMICL, Xcomposer, NExT-GPT, AnyMAL, GPT-4V, Qwen-VL, PointLLM, ASM, VisCPM, LISA

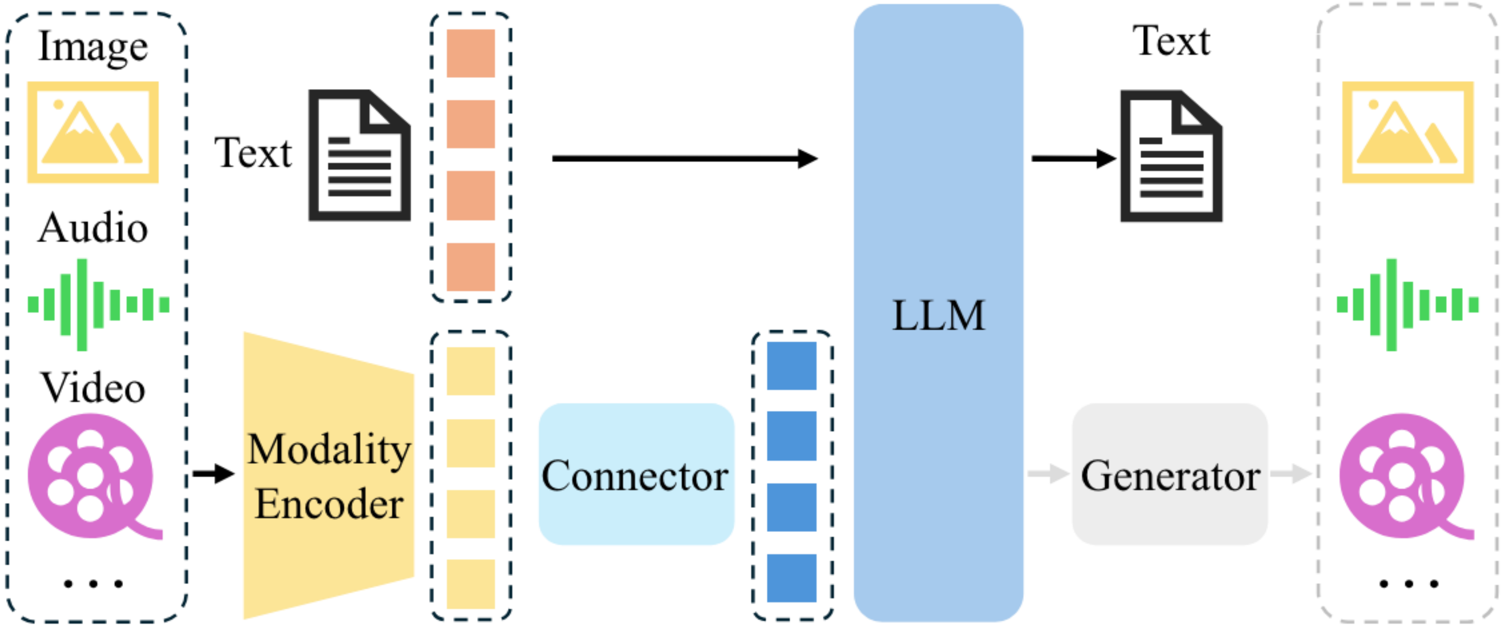
10-12: Fuyu-8B, V*, SPHINX, MobileVLM, Vary, Monkey, Gemini, Woodpecker, Video-LLaVA, LLaMA-VID, LanguageBIND, LLaVA-1.5, CogVLM, Ferret, GLaMM

2024

1-3: MoE-LLaVA, Qwen-VL-Max, TextMonkey, Mobile-Agent, MM1

# Multimodal LLMs

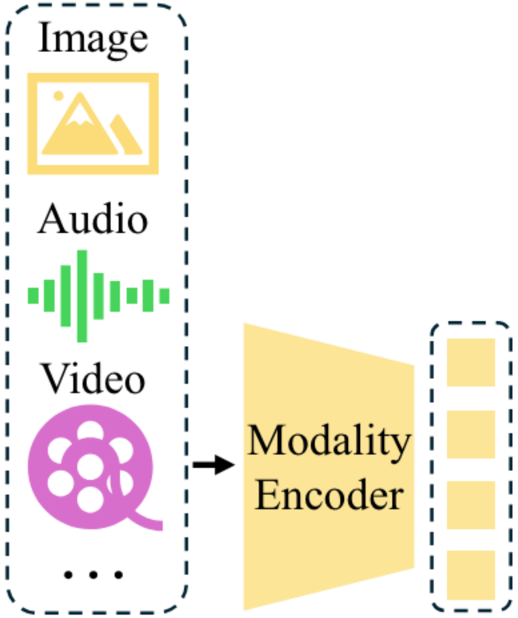
## Architecture



# Multimodal LLMs

## Architecture – Modility Encoder

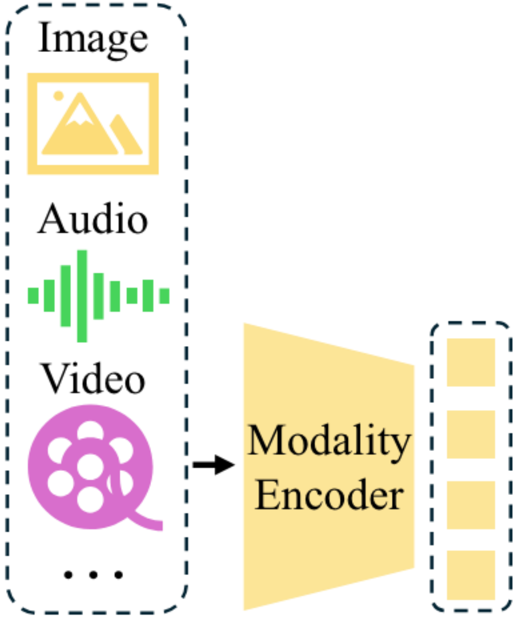
- Encode inputs from diverse modalities to obtain corresponding features

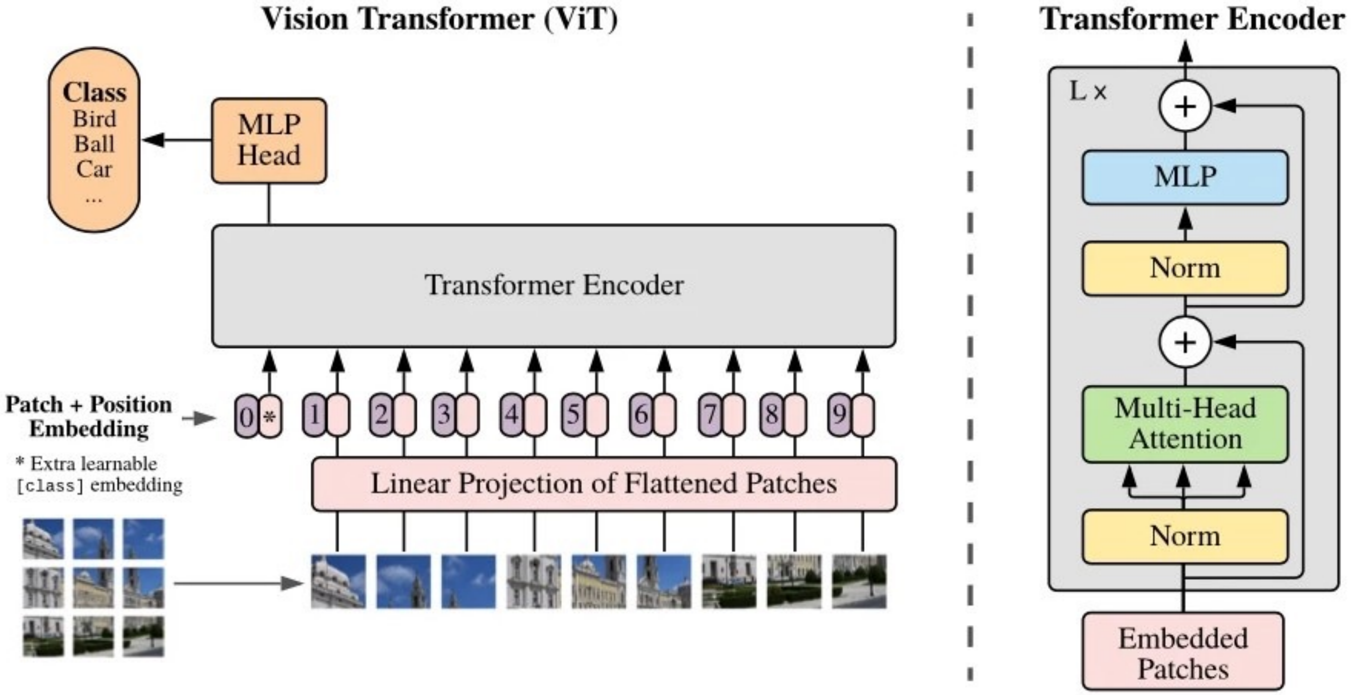


# Multimodal LLMs

## Architecture – Modility Encoder

- Image/Video Encoder: ViT

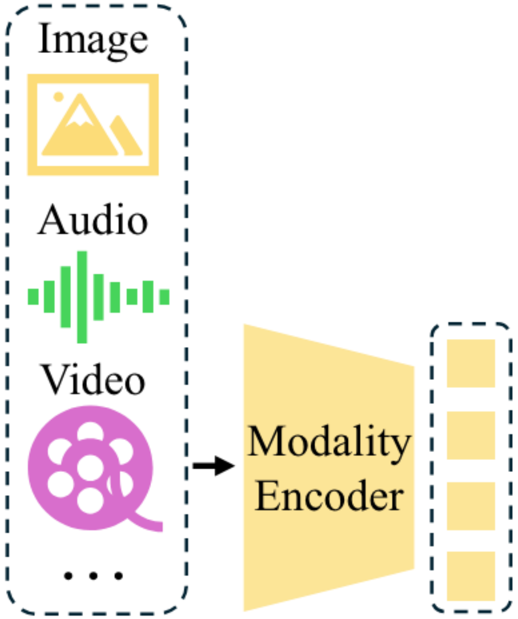


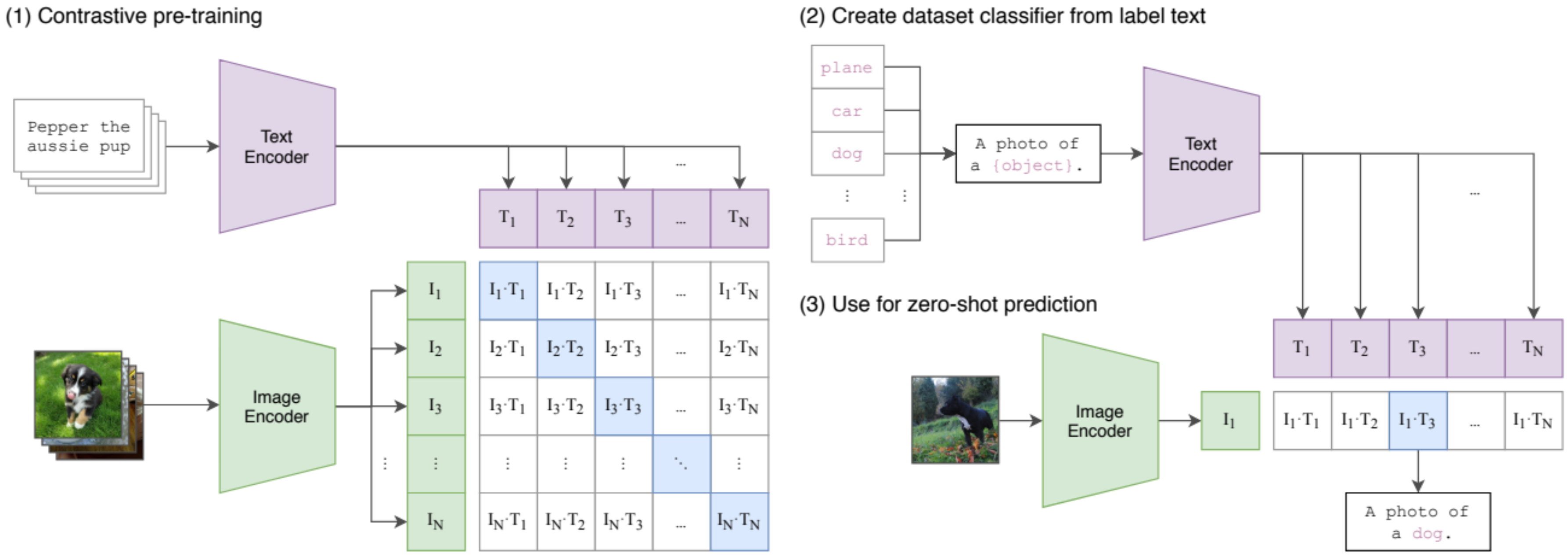


# Multimodal LLMs

## Architecture – Modility Encoder

- Image/Video Encoder: ViT/ CLIP ViT/ Eva-CLIP ViT





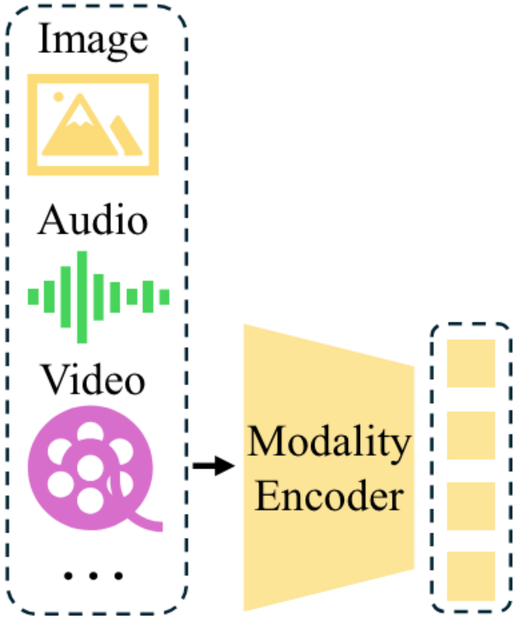
|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

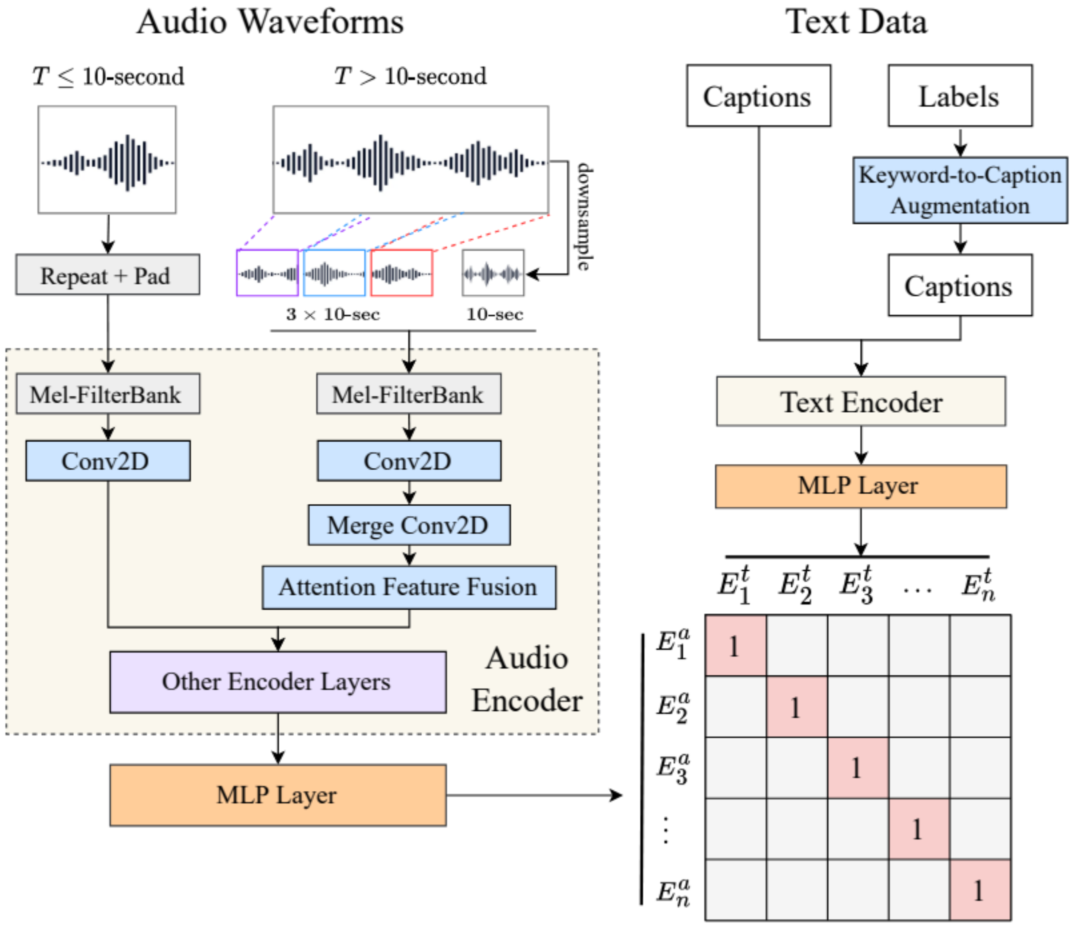
|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |

# Multimodal LLMs

## Architecture – Modility Encoder

- Audio Encoder: C-Former / HuBERT / BEATs / Whisper / **CLAP**

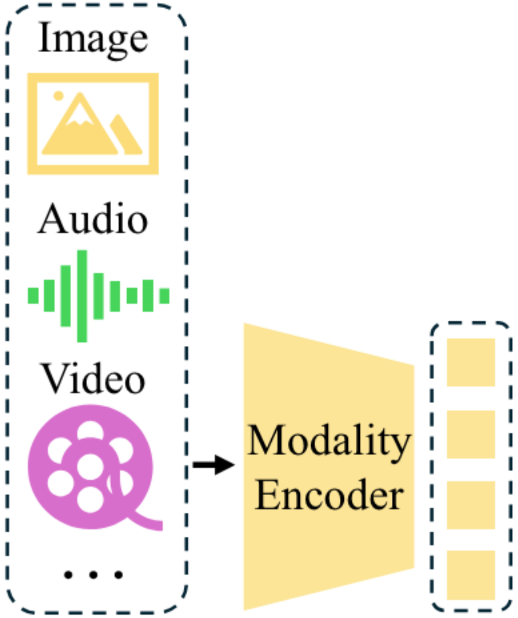


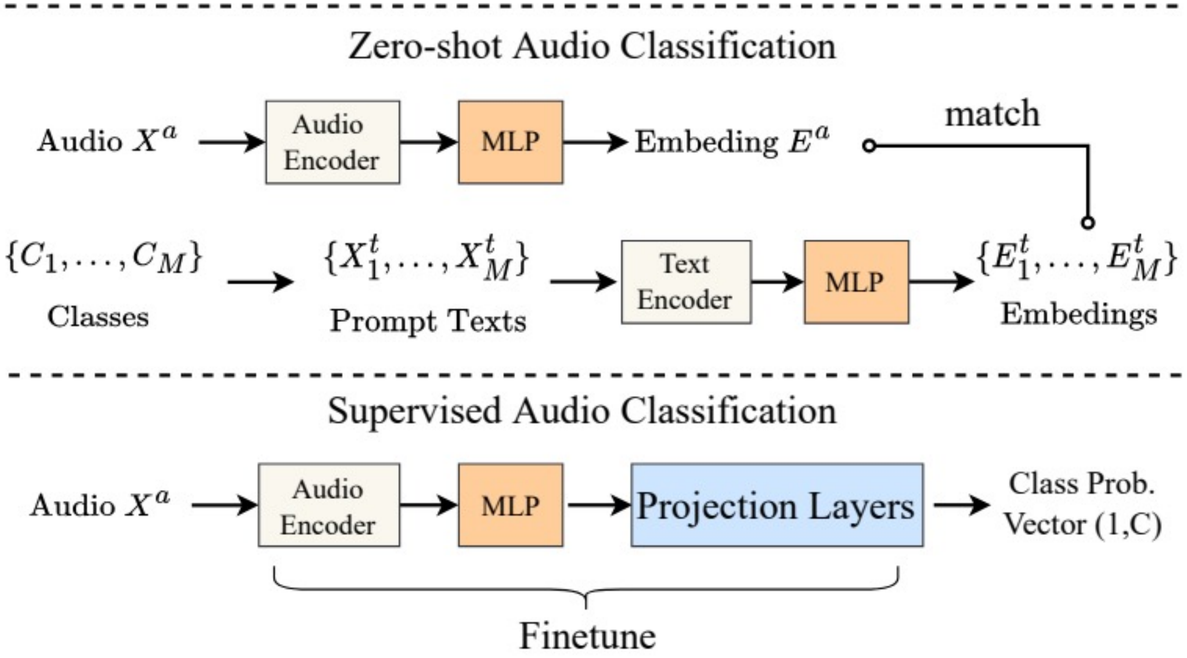


# Multimodal LLMs

## Architecture – Modility Encoder

- Audio Encoder: C-Former / HuBERT / BEATs / Whisper / **CLAP**

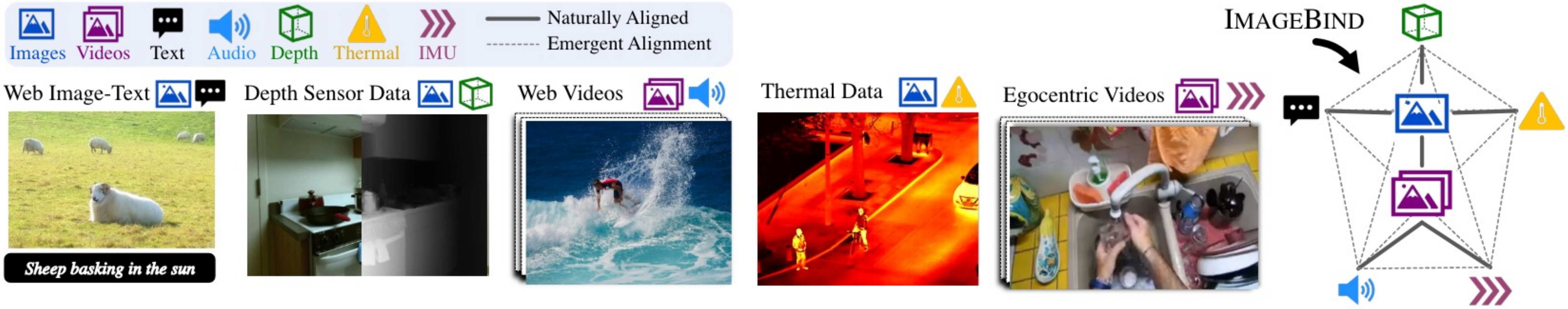




# Multimodal LLMs

## Architecture – Modility Encoder

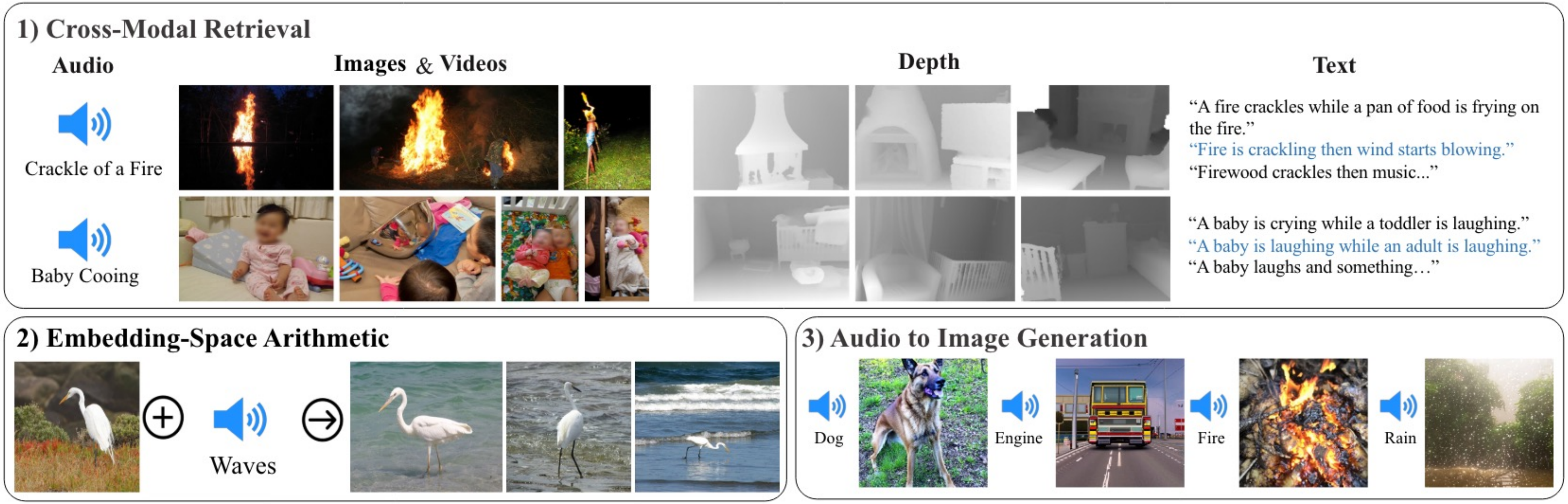
- IMAGEBLIND: One Embedding Space To Bind Them All
- Join embedding space enables novel multimodal capabilities



# Multimodal LLMs

## Architecture – Modility Encoder

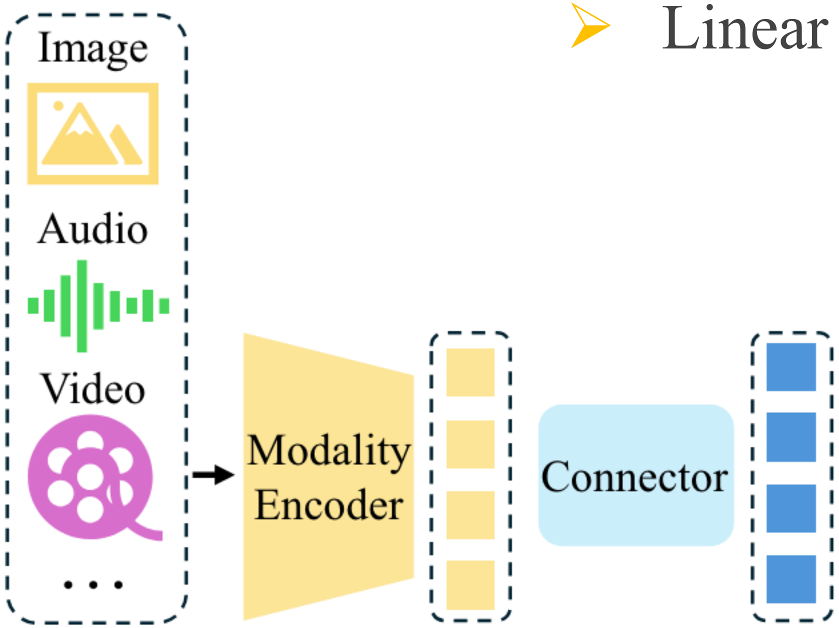
- IMAGEBLIND: One Embedding Space To Bind Them All
- Join embedding space enables novel multimodal capabilities


1) Cross-Modal Retrieval
Audio
Images & Videos
Depth
Text
Crackle of a Fire
"A fire crackles while a pan of food is frying on the fire."
"Fire is crackling then wind starts blowing."
"Firewood crackles then music..."
Baby Cooing
"A baby is crying while a toddler is laughing."
"A baby is laughing while an adult is laughing."
"A baby laughs and something…"
2) Embedding-Space Arithmetic
Waves
3) Audio to Image Generation
Dog
Engine
Fire
Rain

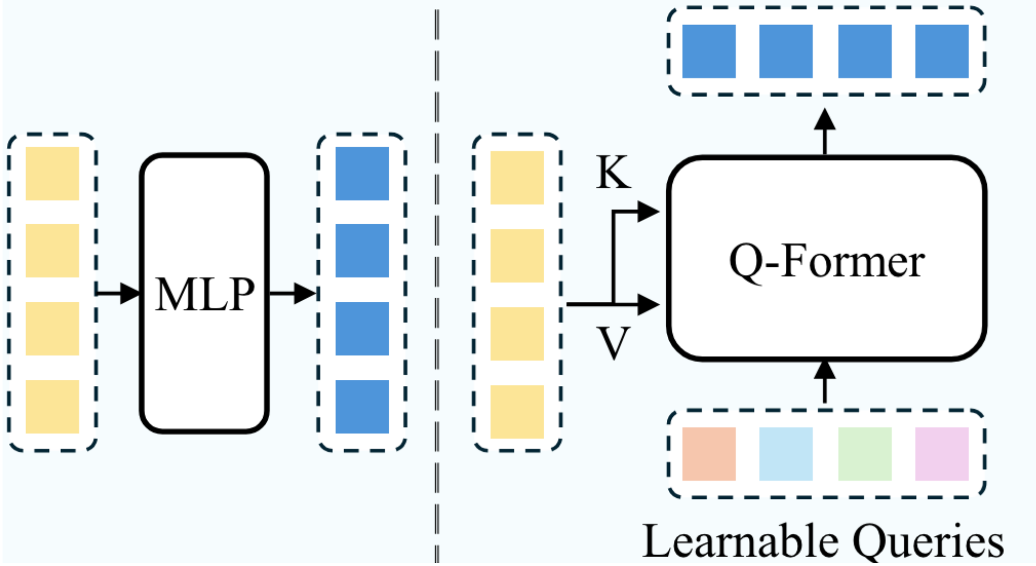
# Multimodal LLMs

## Architecture – Connecter (Input Projector)

- Align the encoded features of other modalities with the text feature



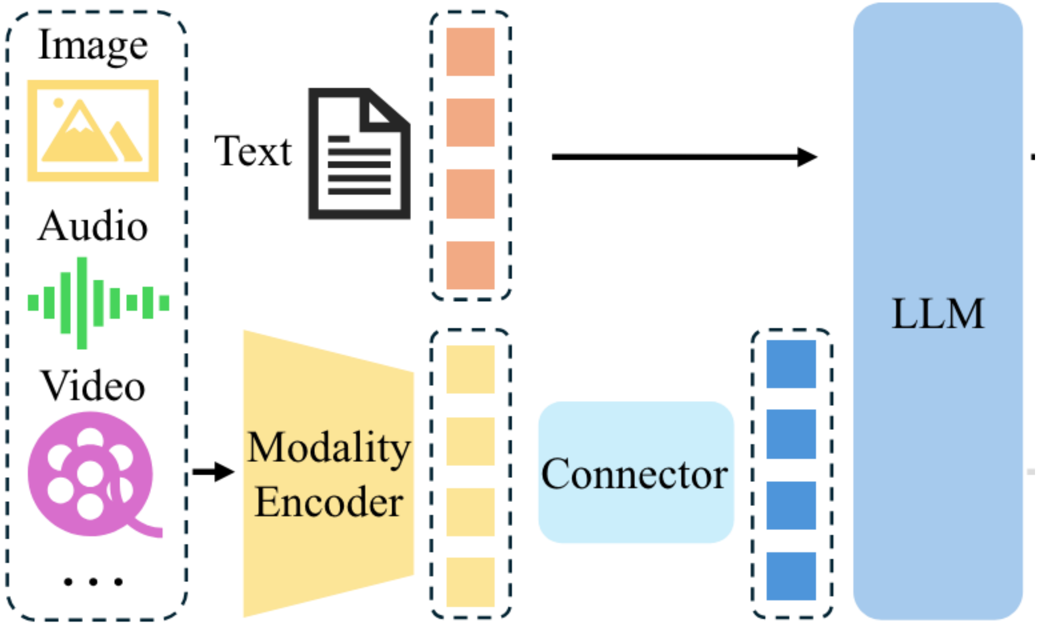
- Linear Projector / MLP / Cross-attention / Q-Former / P-Former

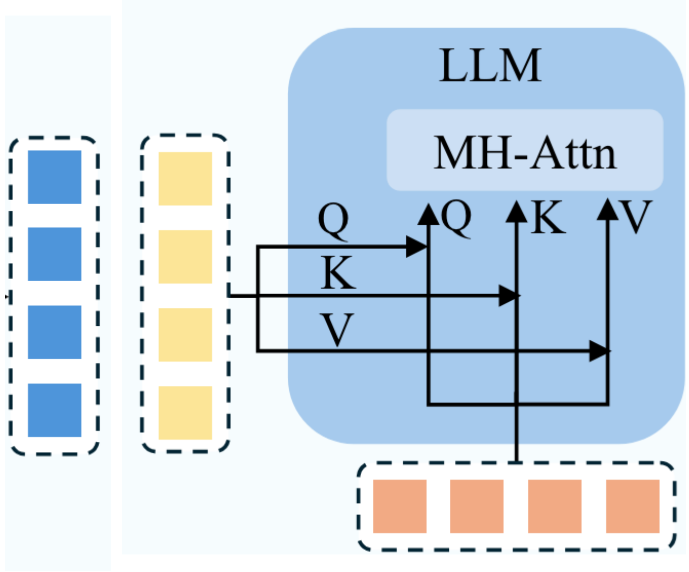


# Multimodal LLMs

## Architecture – LLMs

- LLMs: PaLM, LLaMA, Vicuna, Qwen,…

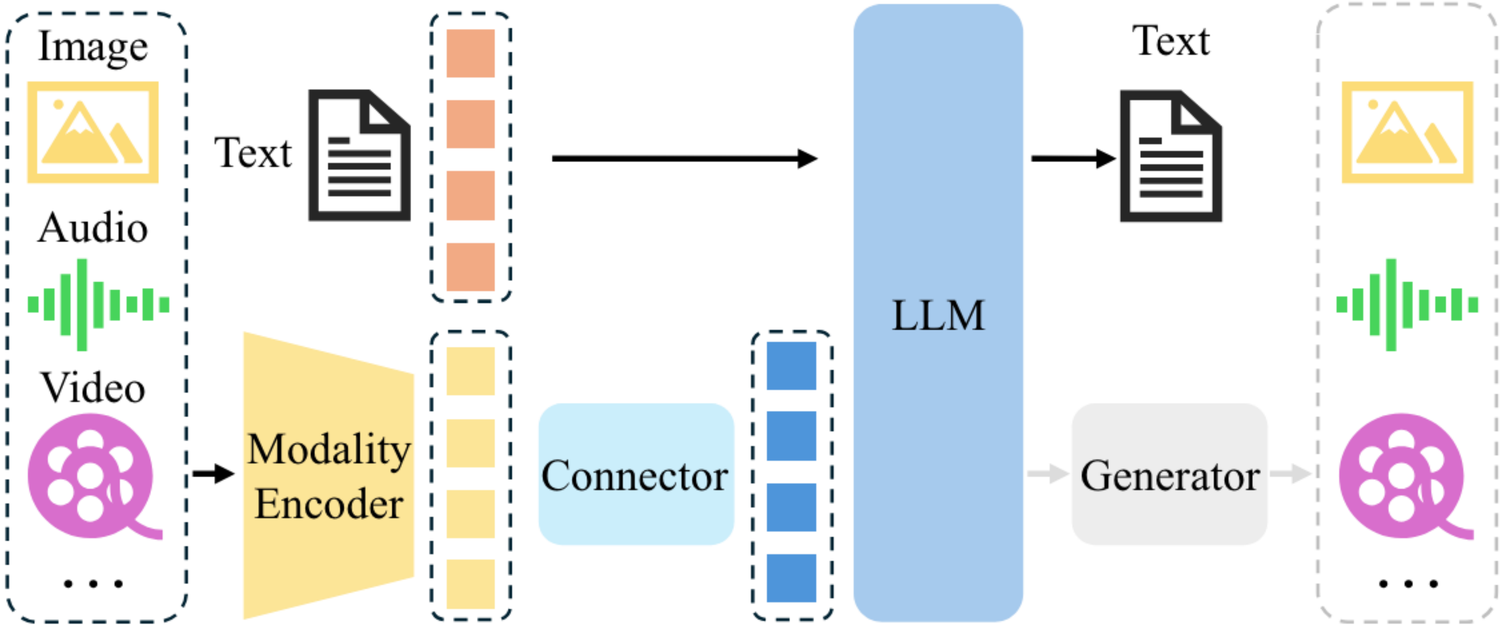




# Multimodal LLMs

## Architecture – Output Projector

- Output Projector: maps the signal token representation from LLM into features

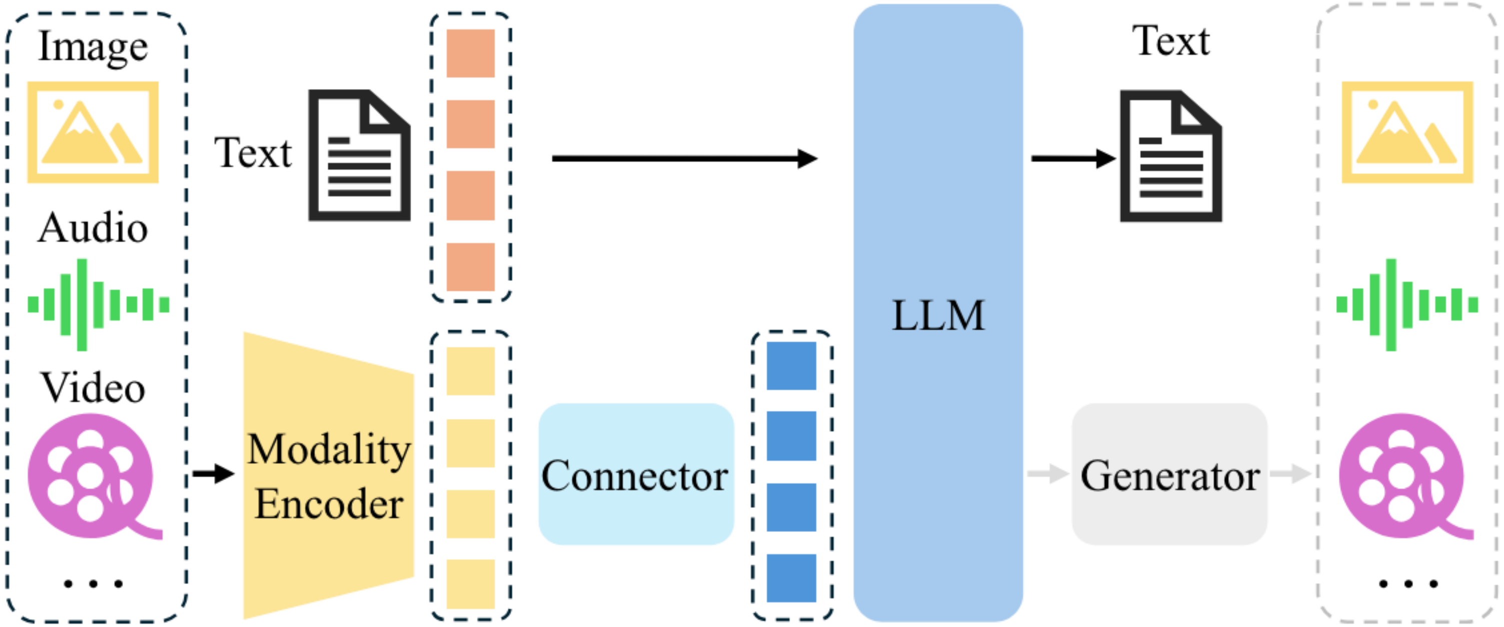


- MLP / Tiny Transformer

# Multimodal LLMs

## Architecture – Modality Generator

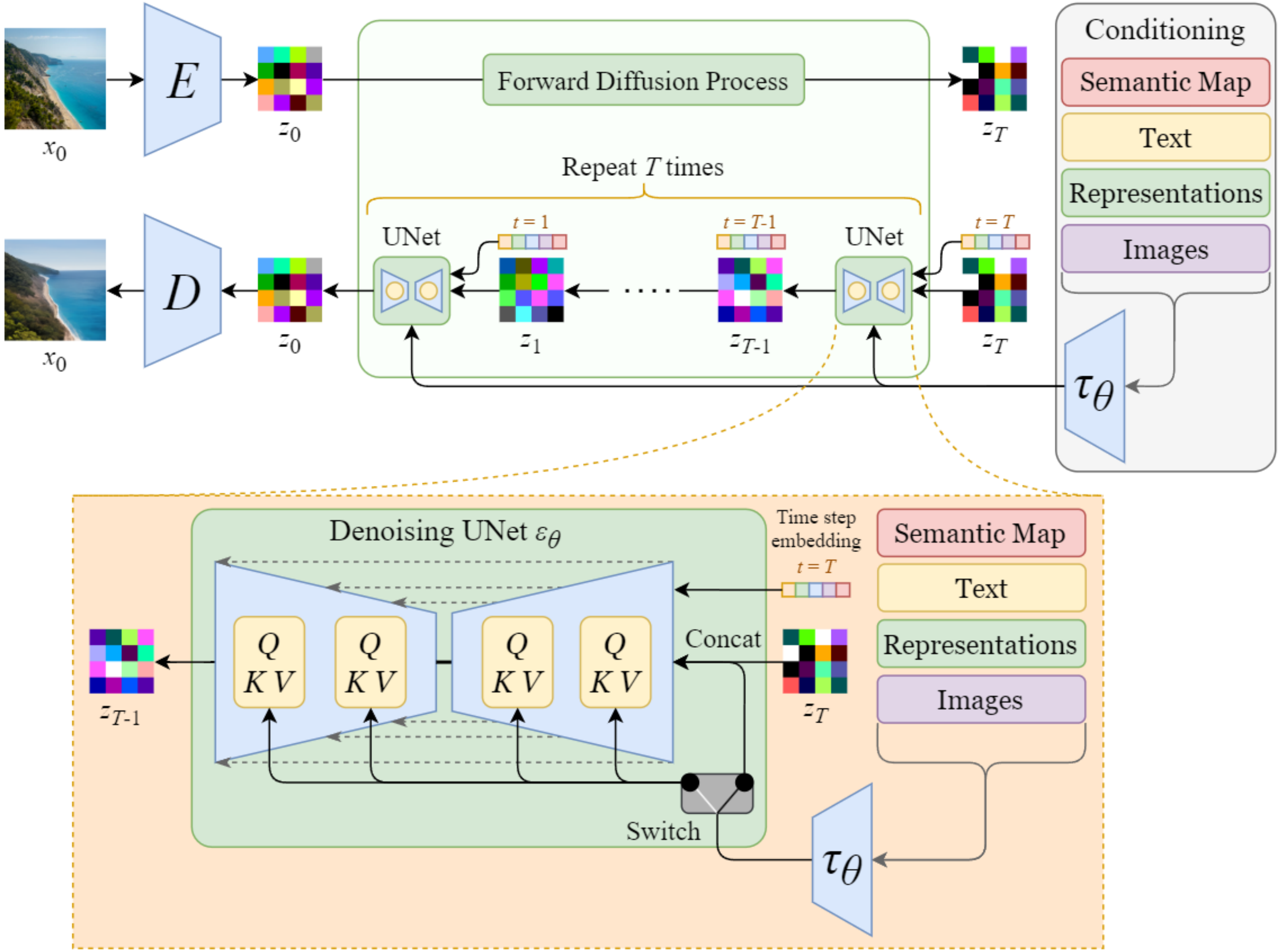
- Product outputs in distinct modalities
- **Stable Diffusion Model**



# Multimodal LLMs

## Architecture – Modality Generator

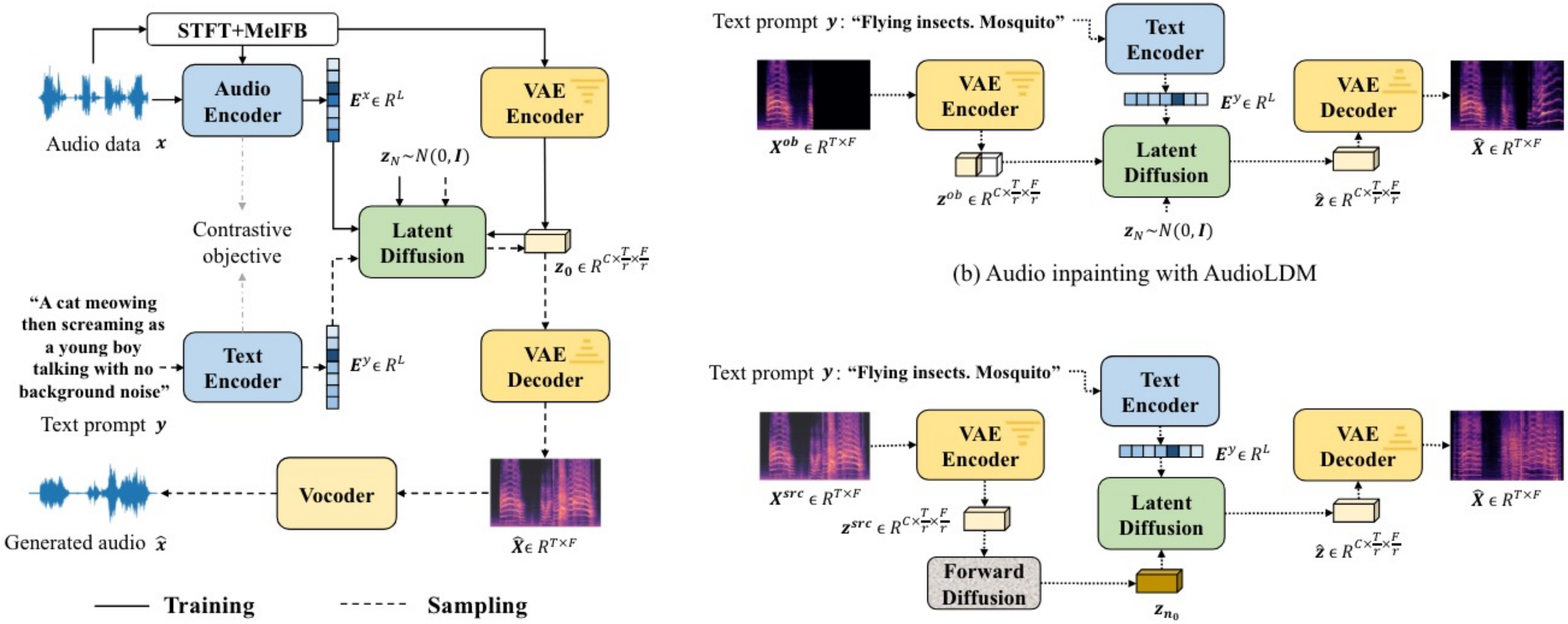
- Stable Diffusion Model for Image



# Multimodal LLMs

## Architecture – Modality Generator

- ➢ **Stable Diffusion Model for Audio (AudioLDM)**


STFT+MelFB
Audio Encoder
$E^x \in R^L$
VAE Encoder
Audio data $x$
$z_N \sim N(0, I)$
Latent Diffusion
Contrastive objective
$z_0 \in R^{C \times \frac{T}{r} \times \frac{F}{r}}$
"A cat meowing then screaming as a young boy talking with no background noise"
Text Encoder
$E^y \in R^L$
VAE Decoder
Text prompt $y$
Vocoder
Generated audio $\hat{x}$
$\hat{X} \in R^{T \times F}$
Training
Sampling
Text prompt $y$: "Flying insects. Mosquito"
Text Encoder
VAE Encoder
$E^y \in R^L$
VAE Decoder
$X^{ob} \in R^{T \times F}$
Latent Diffusion
$\hat{X} \in R^{T \times F}$
$z^{ob} \in R^{C \times \frac{T}{r} \times \frac{F}{r}}$
$\hat{z} \in R^{C \times \frac{T}{r} \times \frac{F}{r}}$
$z_N \sim N(0, I)$
Text prompt $y$: "Flying insects. Mosquito"
Text Encoder
VAE Encoder
$E^y \in R^L$
VAE Decoder
$X^{src} \in R^{T \times F}$
$\hat{X} \in R^{T \times F}$
$z^{src} \in R^{C \times \frac{T}{r} \times \frac{F}{r}}$
Latent Diffusion
$\hat{z} \in R^{C \times \frac{T}{r} \times \frac{F}{r}}$
Forward Diffusion
$z_{n_0}$

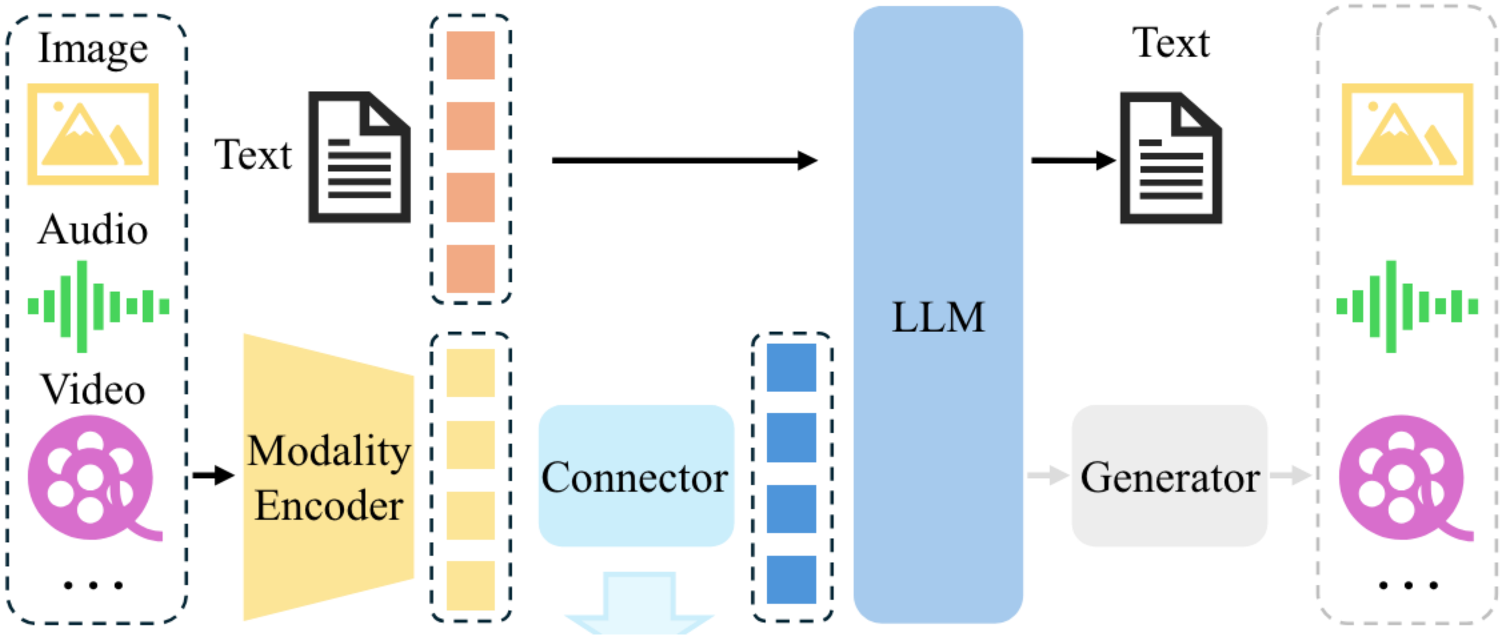
(b) Audio inpainting with AudioLDM

(a) Training and sampling process of AudioLDM

(c) Audio style transfer with AudioLDM

# Multimodal LLMs

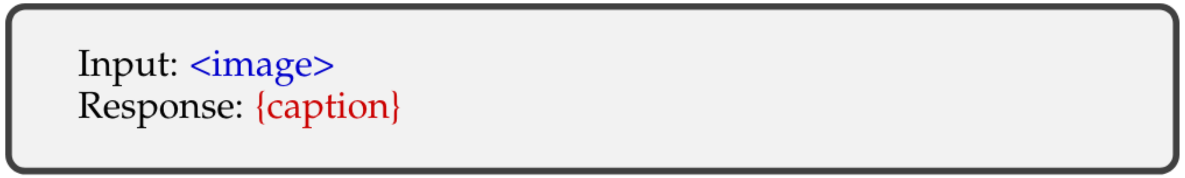
## Architecture



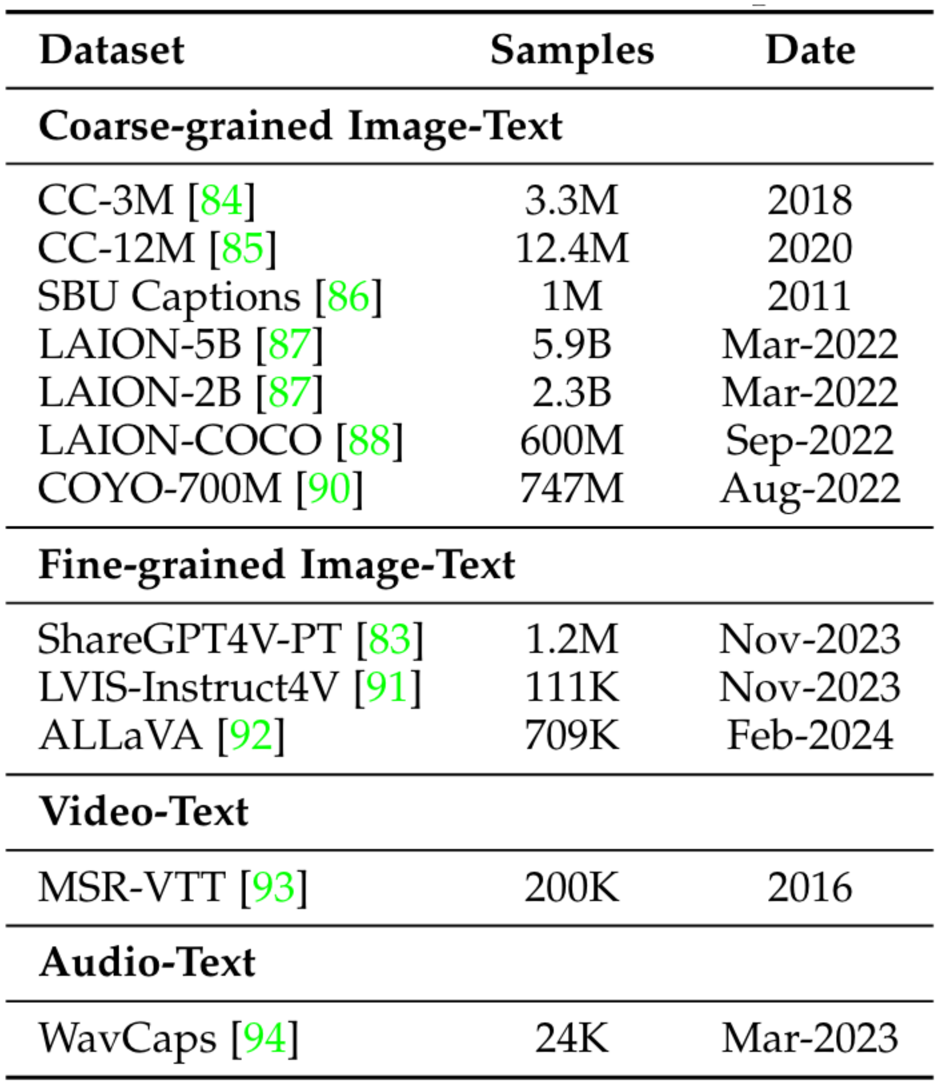
# Multimodal LLMs

## Training Strategy – Pre-Training

- Align different modalities and learn multimodal world knowledge
- Entails large-scale text-paired data



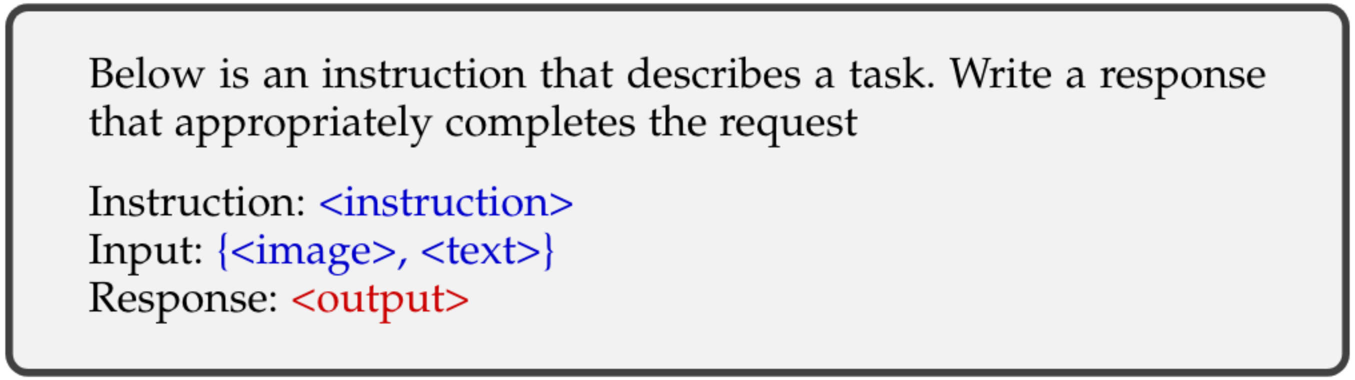
Input: <image>
Response: {caption}



| Dataset | Samples | Date |
|---|---|---|
| **Coarse-grained Image-Text** | | |
| CC-3M [84] | 3.3M | 2018 |
| CC-12M [85] | 12.4M | 2020 |
| SBU Captions [86] | 1M | 2011 |
| LAION-5B [87] | 5.9B | Mar-2022 |
| LAION-2B [87] | 2.3B | Mar-2022 |
| LAION-COCO [88] | 600M | Sep-2022 |
| COYO-700M [90] | 747M | Aug-2022 |
| **Fine-grained Image-Text** | | |
| ShareGPT4V-PT [83] | 1.2M | Nov-2023 |
| LVIS-Instruct4V [91] | 111K | Nov-2023 |
| ALLaVA [92] | 709K | Feb-2024 |
| **Video-Text** | | |
| MSR-VTT [93] | 200K | 2016 |
| **Audio-Text** | | |
| WavCaps [94] | 24K | Mar-2023 |

# Multimodal LLMs

## Training Strategy – Instruction-Tuning

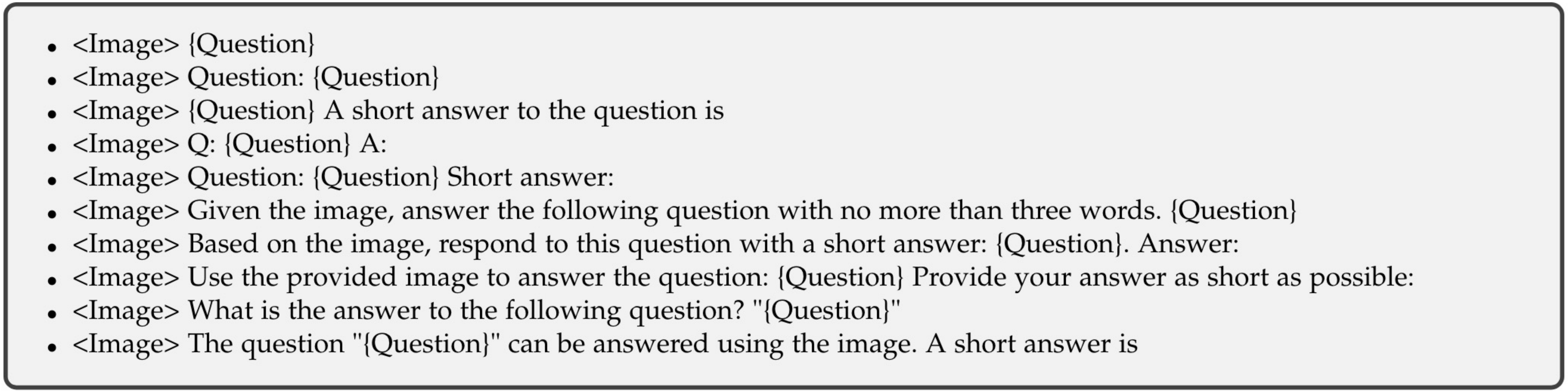


Below is an instruction that describes a task. Write a response that appropriately completes the request

Instruction: <instruction>
Input: {<image>, <text>}
Response: <output>

# Multimodal LLMs

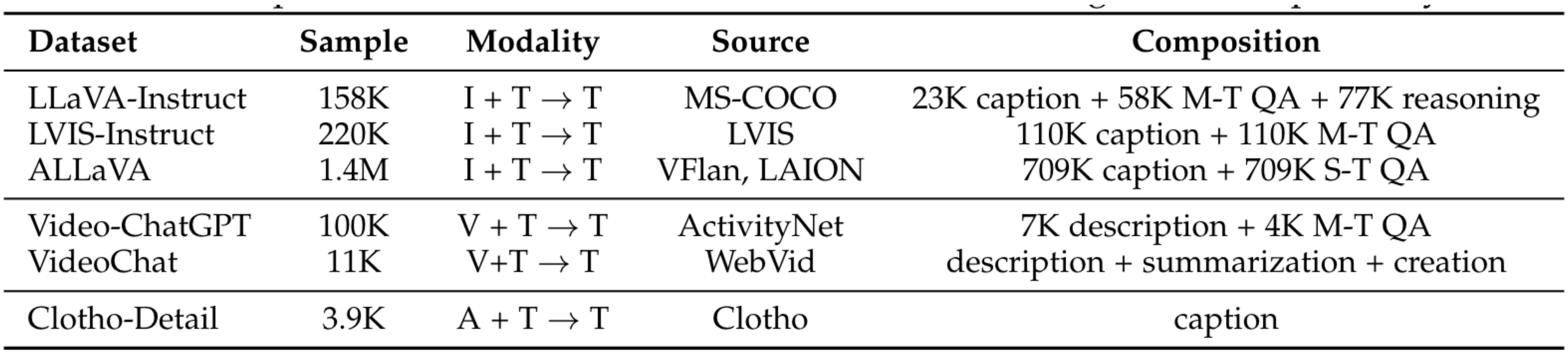
## Training Strategy – Instruction-Tuning



- <Image> {Question}
- <Image> Question: {Question}
- <Image> {Question} A short answer to the question is
- <Image> Q: {Question} A:
- <Image> Question: {Question} Short answer:
- <Image> Given the image, answer the following question with no more than three words. {Question}
- <Image> Based on the image, respond to this question with a short answer: {Question}. Answer:
- <Image> Use the provided image to answer the question: {Question} Provide your answer as short as possible:
- <Image> What is the answer to the following question? "{Question}"
- <Image> The question "{Question}" can be answered using the image. A short answer is

# Multimodal LLMs

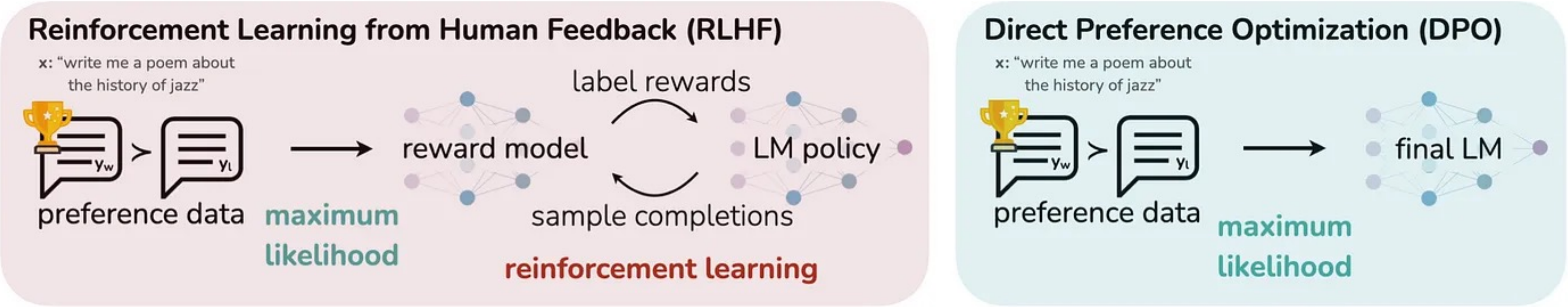
## Training Strategy – Instruction-Tuning



| Dataset | Sample | Modality | Source | Composition |
|---|---|---|---|---|
| LLaVA-Instruct | 158K | I + T → T | MS-COCO | 23K caption + 58K M-T QA + 77K reasoning |
| LVIS-Instruct | 220K | I + T → T | LVIS | 110K caption + 110K M-T QA |
| ALLaVA | 1.4M | I + T → T | VFlan, LAION | 709K caption + 709K S-T QA |
| Video-ChatGPT | 100K | V + T → T | ActivityNet | 7K description + 4K M-T QA |
| VideoChat | 11K | V+T → T | WebVid | description + summarization + creation |
| Clotho-Detail | 3.9K | A + T → T | Clotho | caption |

# Multimodal LLMs

## Training Strategy – Alignment Tuning



Reinforcement Learning from Human Feedback (RLHF)
x: "write me a poem about the history of jazz"
$y_w$
>
$y_l$
preference data
maximum likelihood
reward model
label rewards
LM policy
sample completions
reinforcement learning
Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
$y_w$
>
$y_l$
preference data
maximum likelihood
final LM

# Multimodal LLMs

## SOTA MLLMs

| Model | I/O | Modality Encoder | Input Projector | LLM | Output Projector | Modality Generator |
|---|---|---|---|---|---|---|
| BLIP-2 | IT => T | CLIP ViT | Q-Former<br>Linear | Flan-T5<br>OPT | - | - |
| LLaVA | IT => T | CLIP ViT | Linear | Vicuna | - | - |
| miniGPT-4 | IT => T | Eva-CLIP ViT | Q-Former<br>Linear | Vicuna | - | - |
| InstructBLIP | IVT => T | ViT | Q-Former<br>Linear | Flan-T5<br>Vicuna | - | - |
| Next-GPT | IVAT => IVAT | ImageBlind | Linear | Vicuna | Tiny Transformer | Stable Diffusion Model |
| ModaVerse | IVAT => IVAT | ImageBlind | Linear | LLaMA2 | MLP | Stable Diffusion Model |

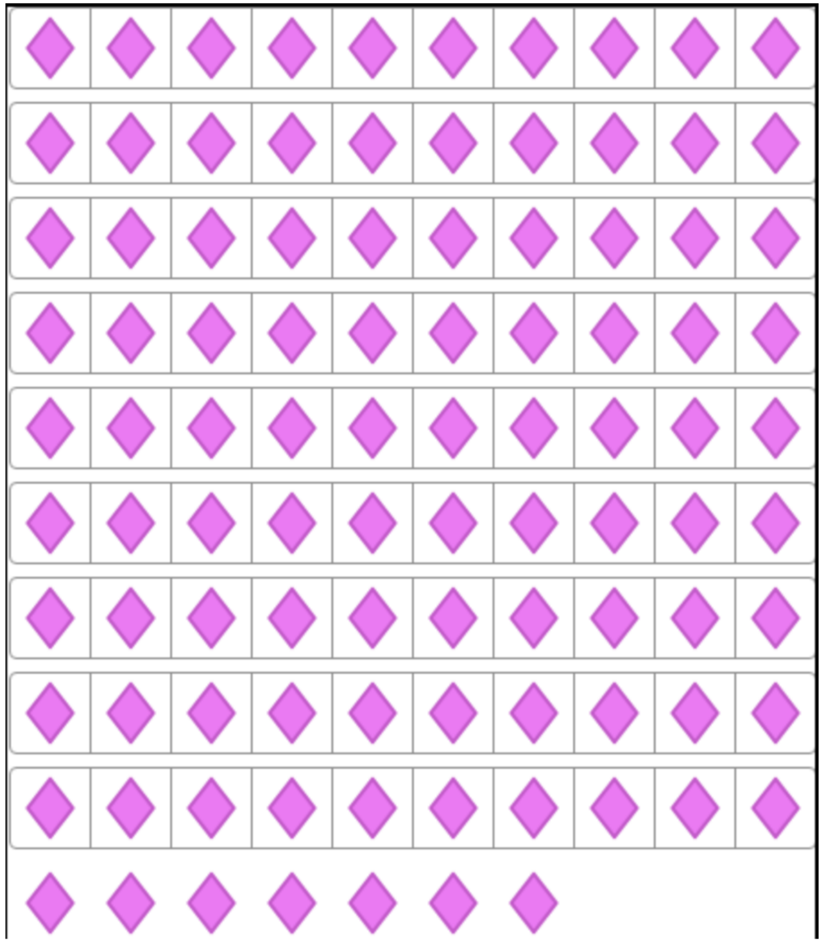
# Outline

- ➢ Introduction
- ➢ Multimodal Large Language Models
- ➢ BLIP-2 for Visual Question Answering
- ➢ NExT-GPT: Any-to-Any MLLM

# BLIP-2 for VQA

## VQA Dataset

Question:
How many diamonds are there?

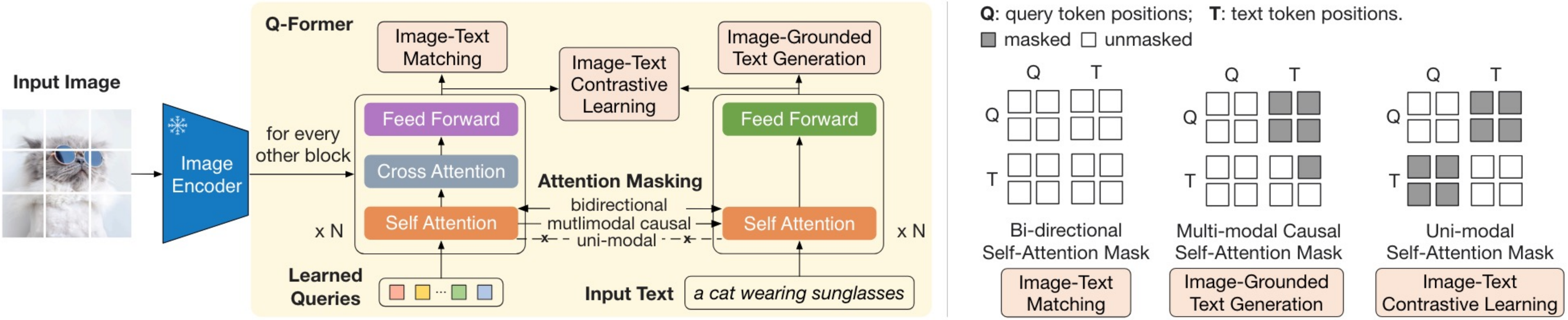


Response:
97

# BLIP-2 for VQA

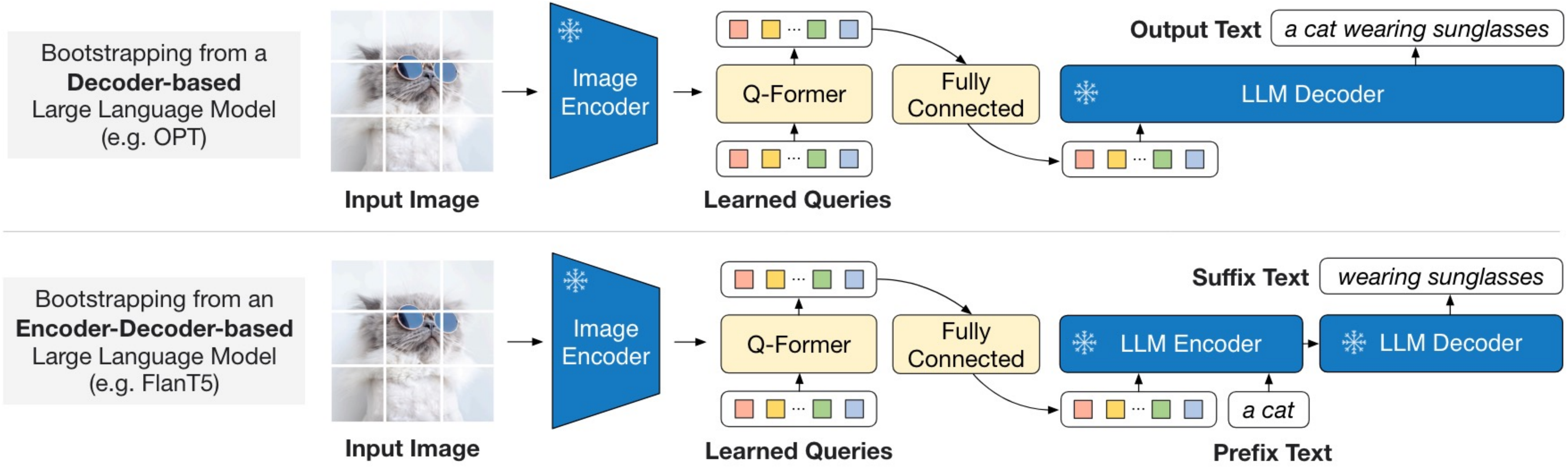
## BLIP-2 - Training

| Model | I/O | Modality Encoder | Input Projector | LLM | Output Projector | Modality Generator |
|---|---|---|---|---|---|---|
| BLIP-2 | IT => T | CLIP ViT | Q-Former<br>Linear | Flan-T5<br>OPT | - | - |



# Multimodal LLMs

## BLIP-2 - Inference



# Multimodal LLMs

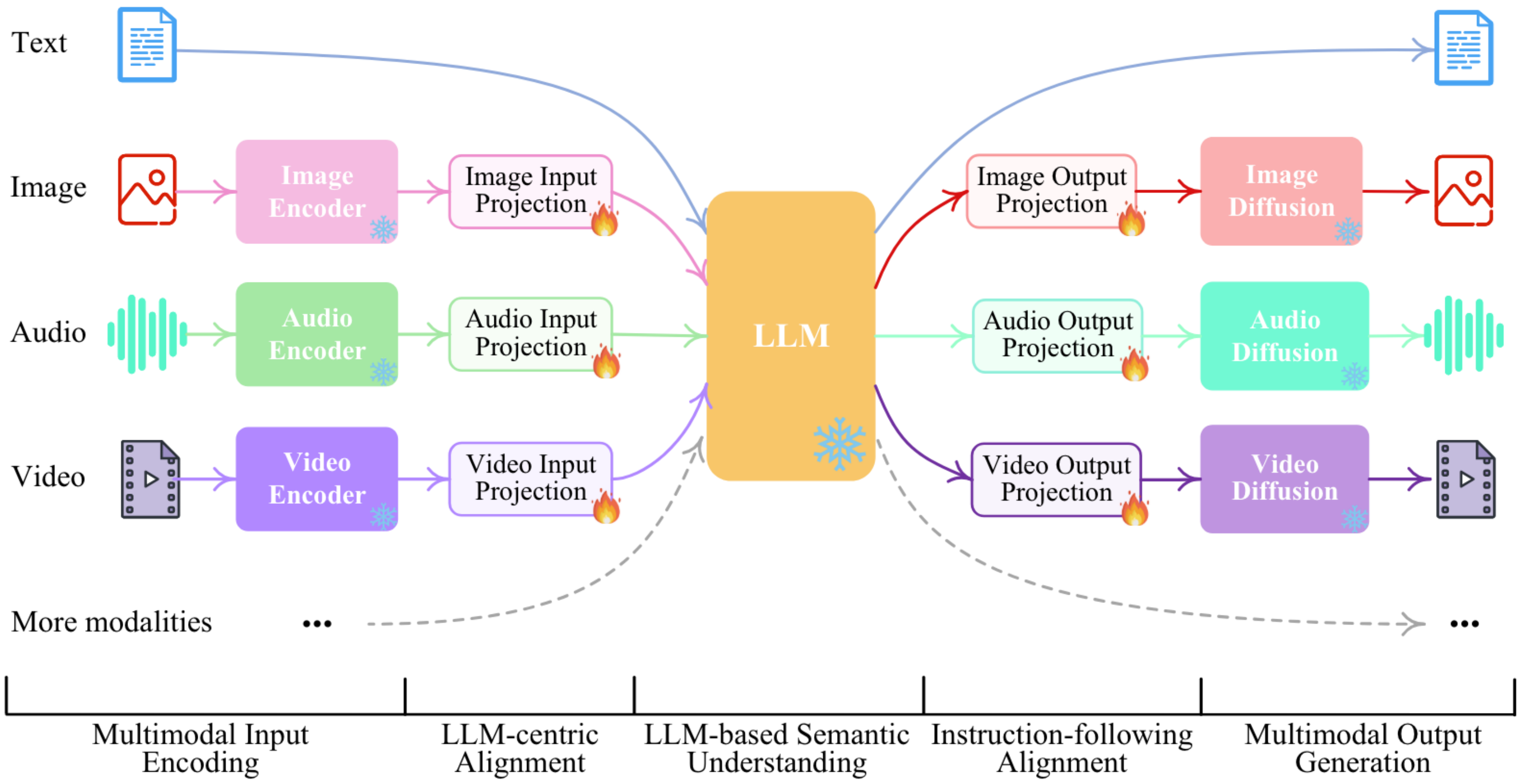
## BLIP-2 - Demo

# Outline

- ➢ Introduction
- ➢ Multimodal Large Language Models
- ➢ BLIP-2 for Visual Question Answering
- ➢ NExT-GPT: Any-to-Any MLLM

# NExT-GPT

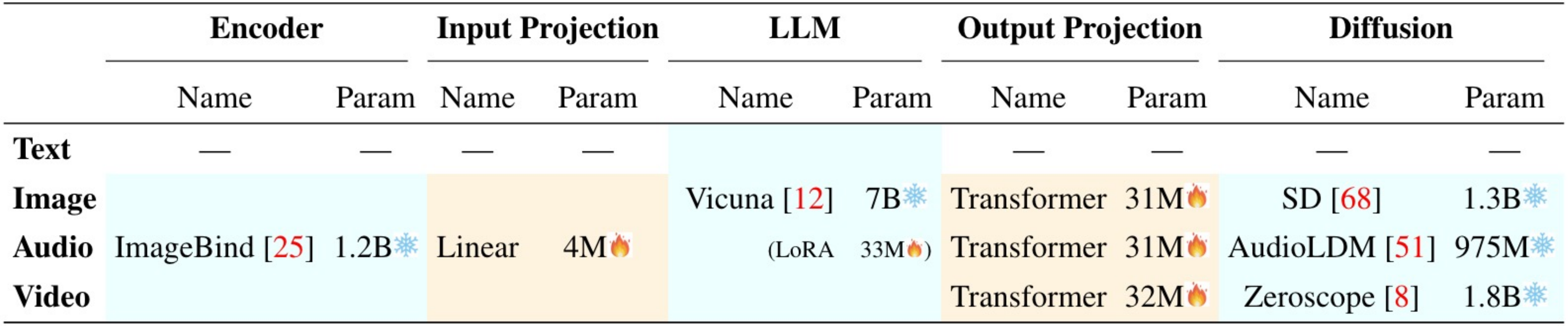
## NExT-GPT



# NExT-GPT

## NExT-GPT

| Model | I/O | Modality Encoder | Input Projector | LLM | Output Projector | Modality Generator |
|---|---|---|---|---|---|---|
| Next-GPT | IVAT => IVAT | ImageBlind | Linear | Vicuna | Tiny Transformer | Stable Diffusion Model |

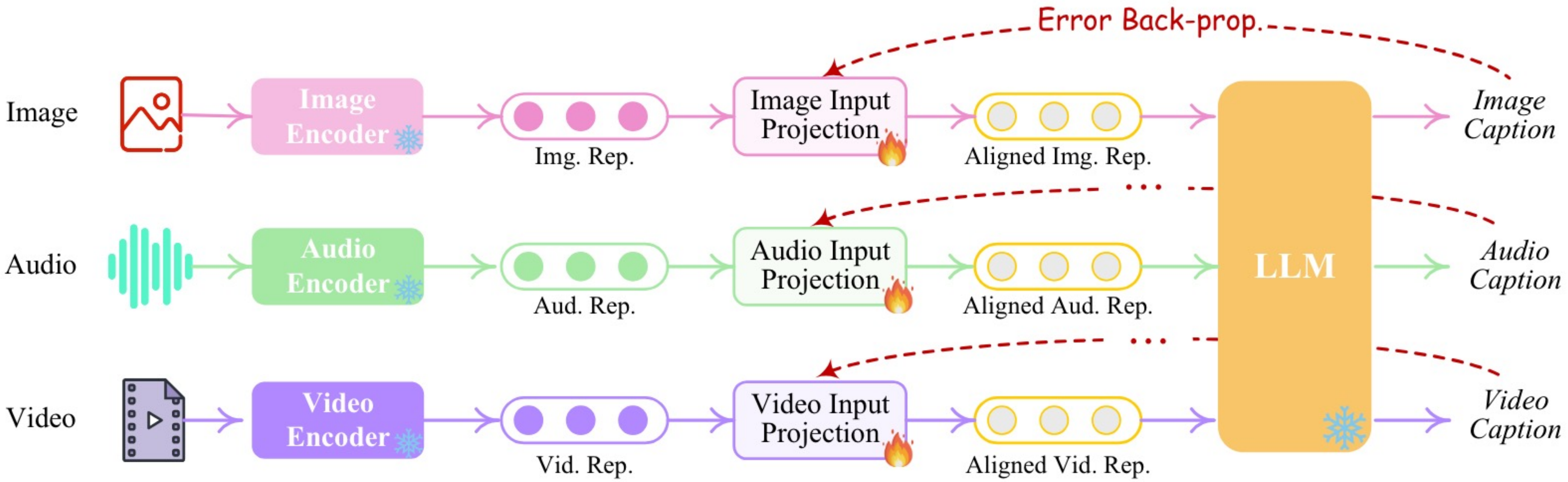


| | Encoder | | Input Projection | | LLM | | Output Projection | | Diffusion | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Name | Param | Name | Param | Name | Param | Name | Param | Name | Param |
| **Text** | — | — | — | — | | | — | — | — | — |
| **Image** | | | | | Vicuna [12] | 7B❄ | Transformer | 31M🔥 | SD [68] | 1.3B❄ |
| **Audio** | ImageBind [25] | 1.2B❄ | Linear | 4M🔥 | (LoRA | 33M🔥) | Transformer | 31M🔥 | AudioLDM [51] | 975M❄ |
| **Video** | | | | | | | Transformer | 32M🔥 | Zeroscope [8] | 1.8B❄ |

# NExT-GPT

## NExT-GPT: Lightweight Multimodal Alignment Learning

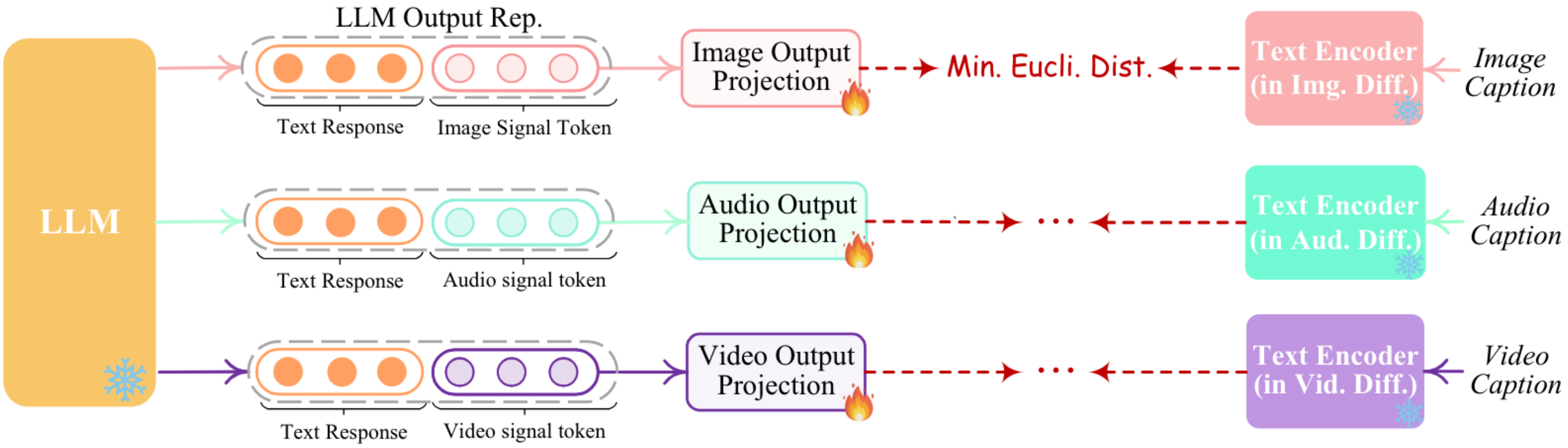
- Encoding-side LLM-centric Multimodal Alignment



# NExT-GPT

## NExT-GPT: Lightweight Multimodal Alignment Learning

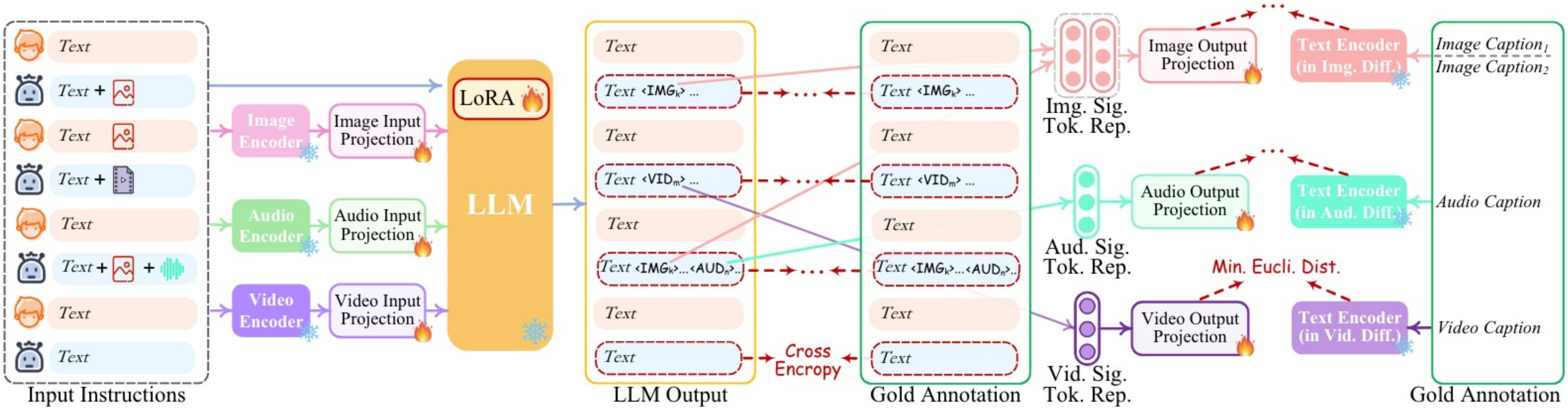
- Decoding-side Instruction-following Alignment



# NExT-GPT

## NExT-GPT: Lightweight Multimodal Alignment Learning

- Modality-switching Instruction Tuning

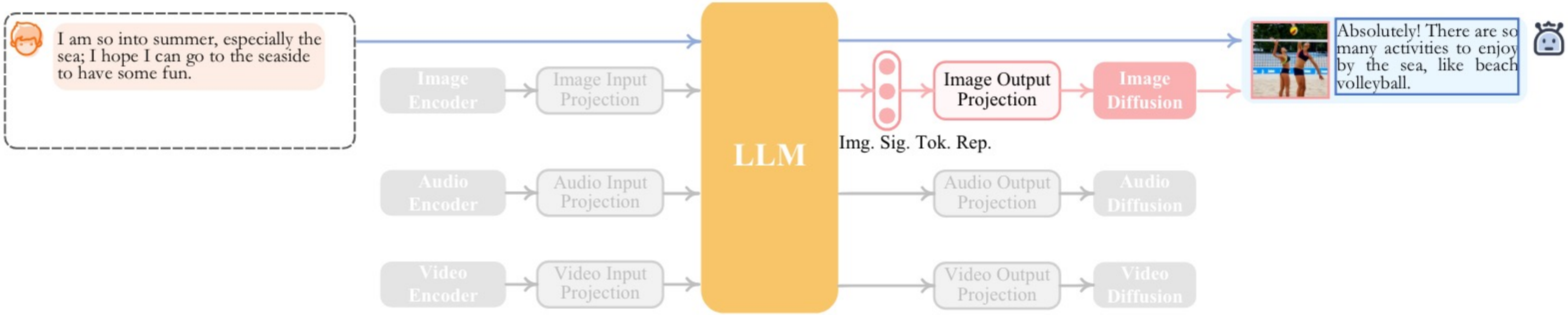


# NExT-GPT

## NExT-GPT - Demo

# NExT-GPT

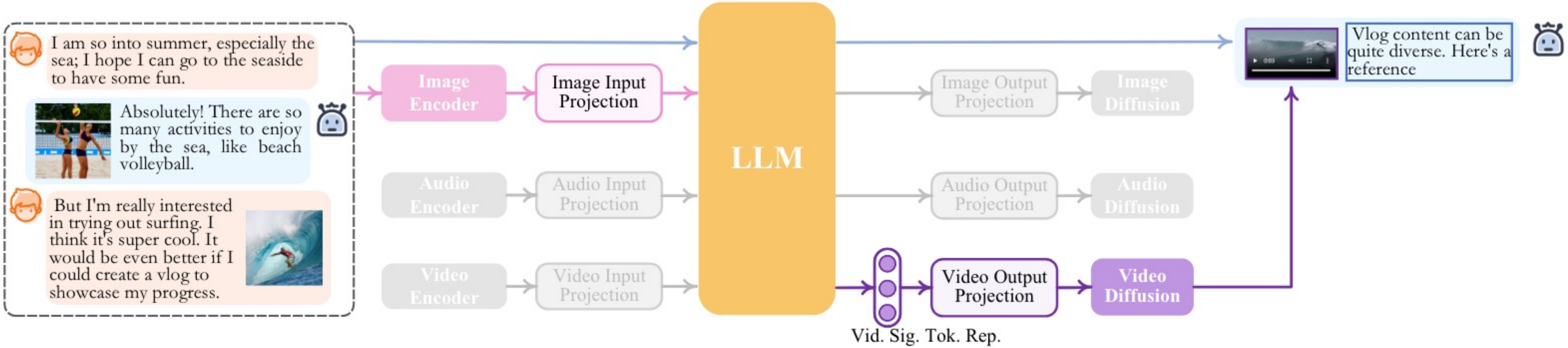
## NExT-GPT - Demo


I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.

Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.

# NExT-GPT

## NExT-GPT - Demo



I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.

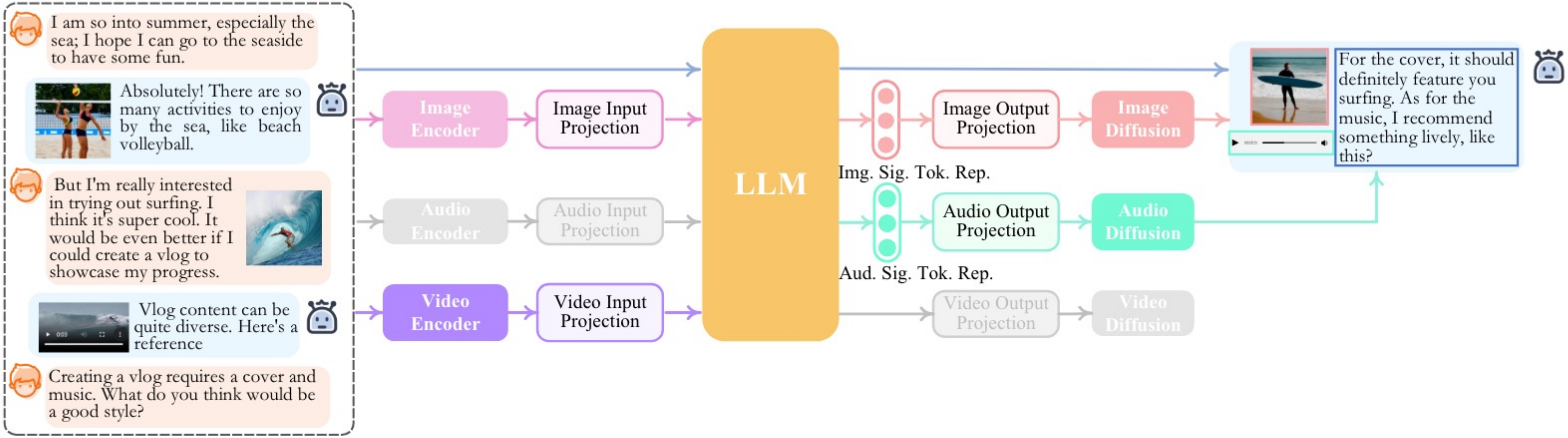
Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.

But I'm really interested in trying out surfing. I think it's super cool. It would be even better if I could create a vlog to showcase my progress.

Vlog content can be quite diverse. Here's a reference

# NExT-GPT

## NExT-GPT - Demo


I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.

Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.

But I'm really interested in trying out surfing. I think it's super cool. It would be even better if I could create a vlog to showcase my progress.

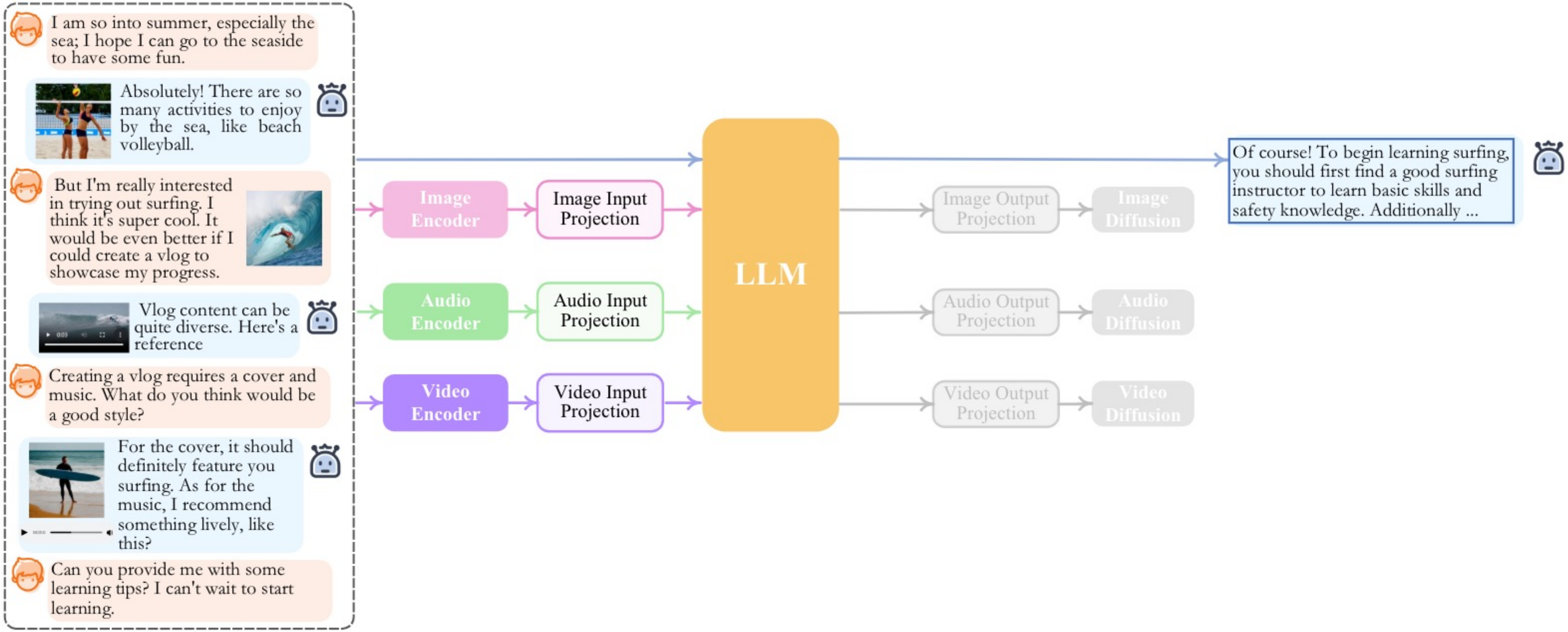
Vlog content can be quite diverse. Here's a reference

Creating a vlog requires a cover and music. What do you think would be a good style?

For the cover, it should definitely feature you surfing. As for the music, I recommend something lively, like this?

# NExT-GPT

## NExT-GPT - Demo



I am so into summer, especially the sea; I hope I can go to the seaside to have some fun.

Absolutely! There are so many activities to enjoy by the sea, like beach volleyball.

But I'm really interested in trying out surfing. I think it's super cool. It would be even better if I could create a vlog to showcase my progress.

Vlog content can be quite diverse. Here's a reference

Creating a vlog requires a cover and music. What do you think would be a good style?

For the cover, it should definitely feature you surfing. As for the music, I recommend something lively, like this?

Can you provide me with some learning tips? I can't wait to start learning.

Of course! To begin learning surfing, you should first find a good surfing instructor to learn basic skills and safety knowledge. Additionally ...

# Summary

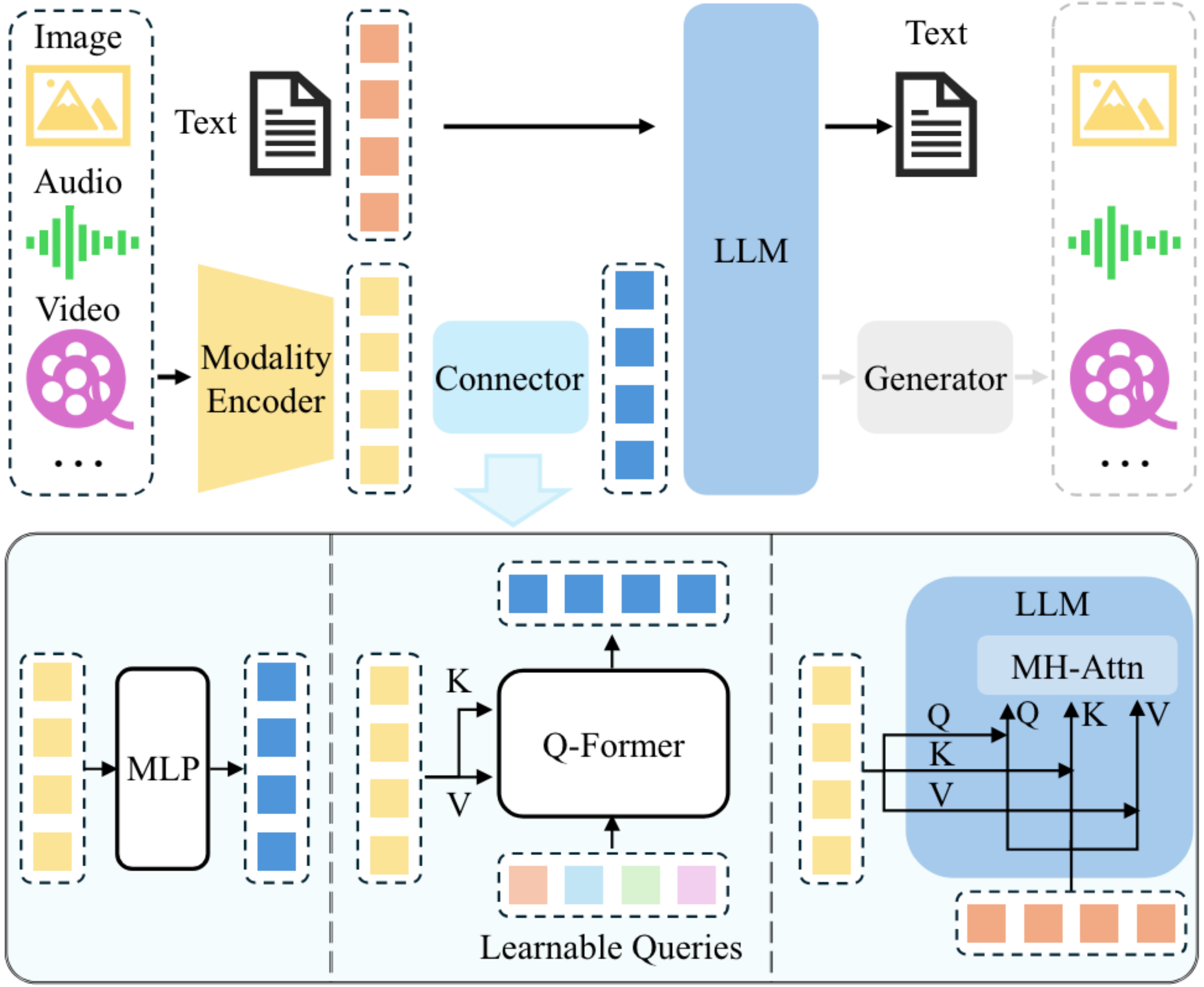


Image
Audio
Video
...
Text
Modality Encoder
Connector
LLM
Text
Generator
...
MLP
K
V
Q-Former
Learnable Queries
LLM
MH-Attn
Q
K
V
Q
K
V

# Thanks!

## Any questions?