Multimodality & Large Language Models

November 29th, 2024

Gaspard Michel
LORIA & Deezer Research

Institut des sciences du Digital Management & Cognition
UNIVERSITÉ DE LORRAINE
Loria Laboratoire lorrain de recherche en informatique et ses applications
DEEZER

Contents

You
Generate a fun meme about multimodal LLMs like yourself

DALL·E

Institut des sciences du Digital Management & Cognition
UNIVERSITÉ DE LORRAINE
Loria Laboratoire lorrain de recherche en informatique et ses applications
DEEZER

Introduction

- Real World Environment inherently multimodal
- Interaction through diverse channel:
  - Speech
  - Sound
  - Vision
  - Language
  - Touch

Introduction - Examples

Pdf description

(2) Prompt:
Describe the pointed region in the image.

Method	Validataion set								Test set							
	in.		near.		out.		overall		in.		near.		out.		overall	
	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S
OSCAR	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4	84.8	12.1	82.1	11.5	73.8	9.7	80.9	11.3
Human	84.4	14.3	85.0	14.3	95.7	14.0	87.1	14.2	80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6
VIVO	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	89.0	12.9	87.8	12.6	80.1	11.1	86.6	12.4
VinVL	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	98.0	13.6	95.2	13.4	78.0	11.5	92.5	13.1
UFO	103.9	14.5	95.5	13.8	83.5	12.3	94.3	13.6	98.9	14.3	94.7	13.9	77.9	12.1	92.3	13.6
mPLUG	-	-	-	-	-	-	114.8	14.8	-	-	-	-	-	-	-	-
SimVLM	113.7	-	110.9	-	115.2	-	115.2	-	113.7	-	110.9	-	115.2	-	115.2	-
LEMON	118.0	15.4	116.3	15.1	120.2	14.5	117.3	15.0	112.8	15.2	115.5	15.1	110.1	13.7	114.3	14.9
UniversalCap	123.2	15.0	121.5	15.3	123.4	14.4	122.1	15.0	118.9	15.4	120.6	15.3	114.3	14.1	119.3	15.1
CoCa	-	-	-	-	-	-	122.4	15.5	-	-	-	-	-	-	120.6	15.5
GIT_B	100.7	13.8	97.7	13.5	89.6	12.5	96.6	13.4	-	-	-	-	-	-	-	-
GIT_L	107.7	14.9	107.8	14.5	102.5	13.7	106.9	14.4	-	-	-	-	-	-	-	-
GIT	**129.8**	**16.3**	124.1	16.0	127.1	15.7	125.5	16.0	122.4	16.2	123.9	16.0	122.0	**15.7**	123.4	15.9
GIT2	126.9	16.1	**125.8**	**16.2**	**130.6**	**15.8**	**126.9**	**16.1**	**124.2**	**16.4**	**125.5**	**16.1**	**122.3**	15.6	**124.8**	**16.1**

GPT-4V:
The highlighted red region in the image is a table cell that contains a numerical value of 122.3. It is located in the "Test set" section of the table, under the "out" column and the "GIT2" row. This value represents the overall performance metric for the GIT2 method on the test set, in the "out" category.

Prompt:
Localize each person in the image using bounding box. What is the image size of the input image?

GPT-4V:
The image size of the input image is 1000 x 333 pixels.

Bounding boxes for each person:
- Person 1 (leftmost): (x1, y1) = (22, 108), (x2, y2) = (170, 300)
- Person 2: (x1, y1) = (200, 80), (x2, y2) = (290, 270)
- Person 3: (x1, y1) = (350, 70), (x2, y2) = (460, 280)
- Person 4: (x1, y1) = (540, 80), (x2, y2) = (640, 270)
- Person 5 (rightmost): (x1, y1) = (760, 100), (x2, y2) = (920, 290)

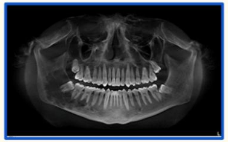
Note: These bounding box coordinates are approximate and may not be exact.

Visualization of Detected Bbox:

Image reasoning

Health care

Prompt:
Which wisdom tooth is hardest to remove?



Note: Same query image as shown above, resized to save space

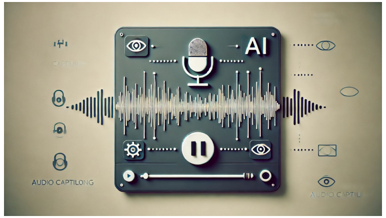
GPT-4V:
The lower wisdom teeth, especially those that are impacted, are typically considered the hardest to remove. This is because the lower jawbone is denser than the upper jawbone, and the impacted wisdom teeth may be more difficult to access. Additionally, the roots of the lower wisdom teeth are often longer and more curved than those of the upper wisdom teeth, making them more challenging to remove.

Prompt:
Count the number of apples in the image.

GPT-4V:
There are 7 apples in the image.

Introduction - Applications

- **Generation**
  - Image generation (DALL-E, MidJourney)
  - Text Generation (Image/Video/Audio captioning)



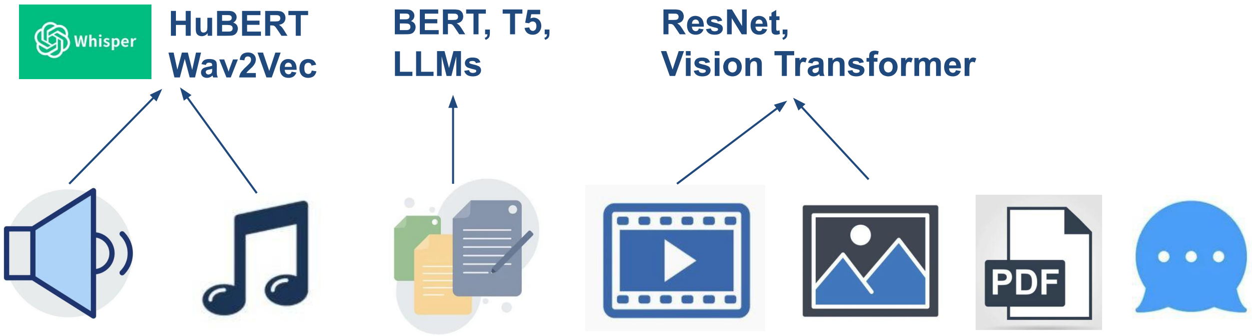
- **Vision/Audio-Language Understanding**
  - Various Understanding tasks (source separation, reasoning over images/video)
  - Text-based retrieval (image/audio/video search)

    → Spotify/Deezer, Youtube, Netflix…

Introduction

- Years of research on unimodal models

- High-quality representation derived from pretrained uni-modal
  Foundation models

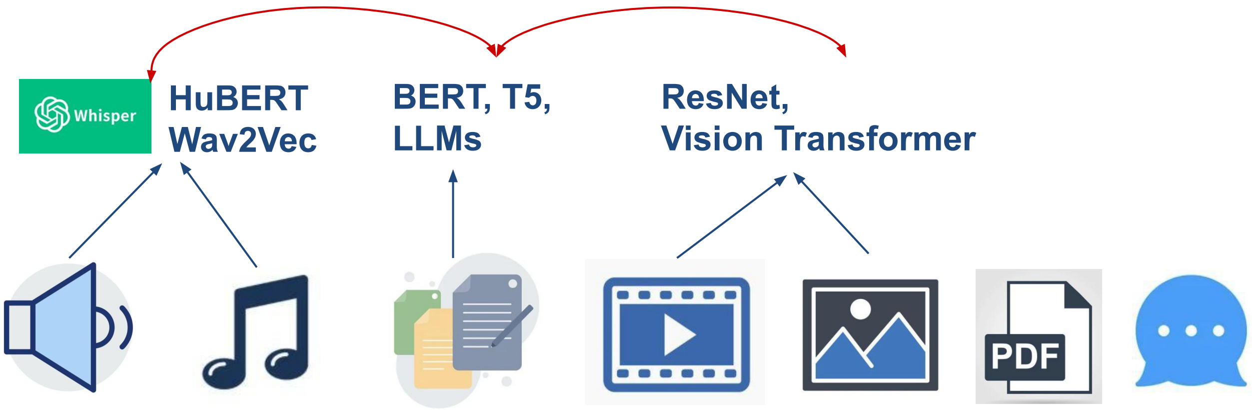


Introduction

- Recently, a core research idea came-up:

<u>**How to connect performant foundation unimodal models together?**</u>

  - Most efforts on connecting **text** with other modalities



A Brief History of MM Models

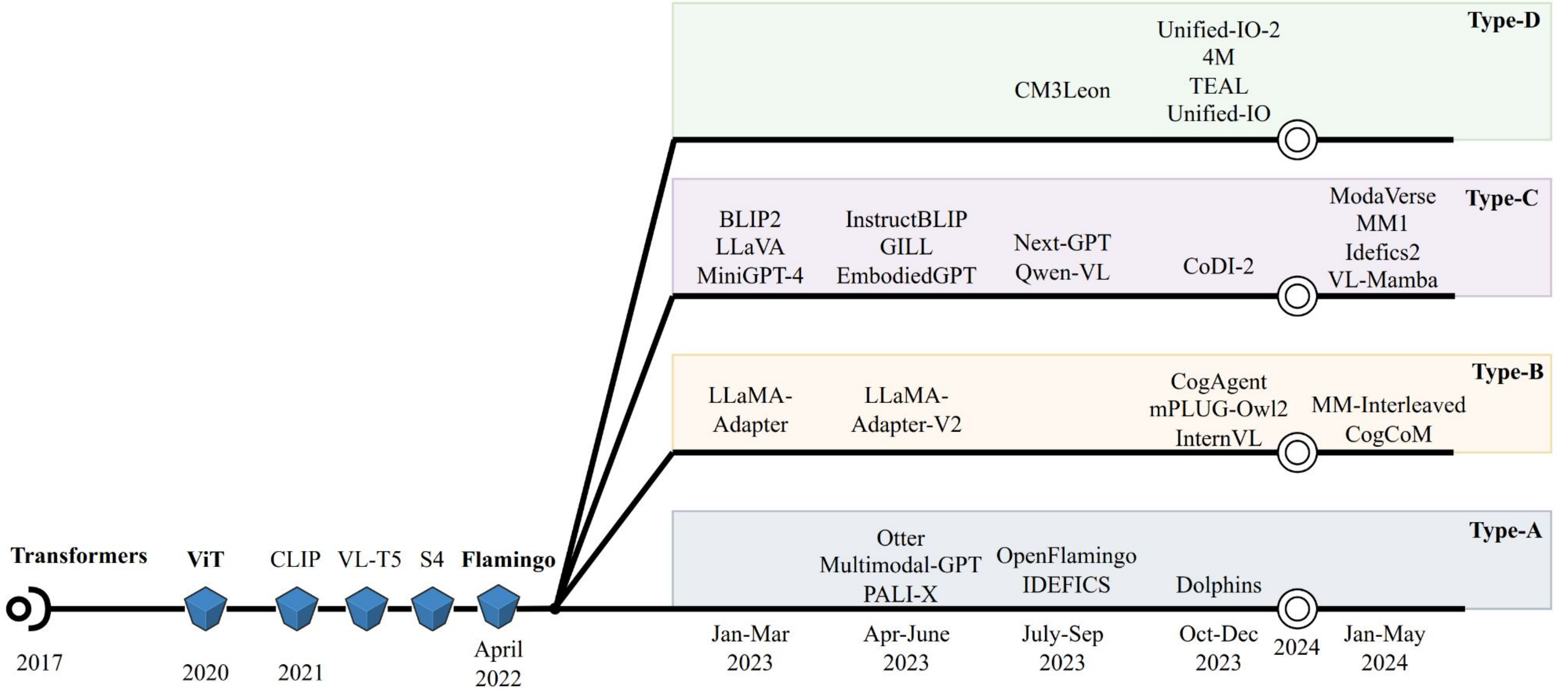
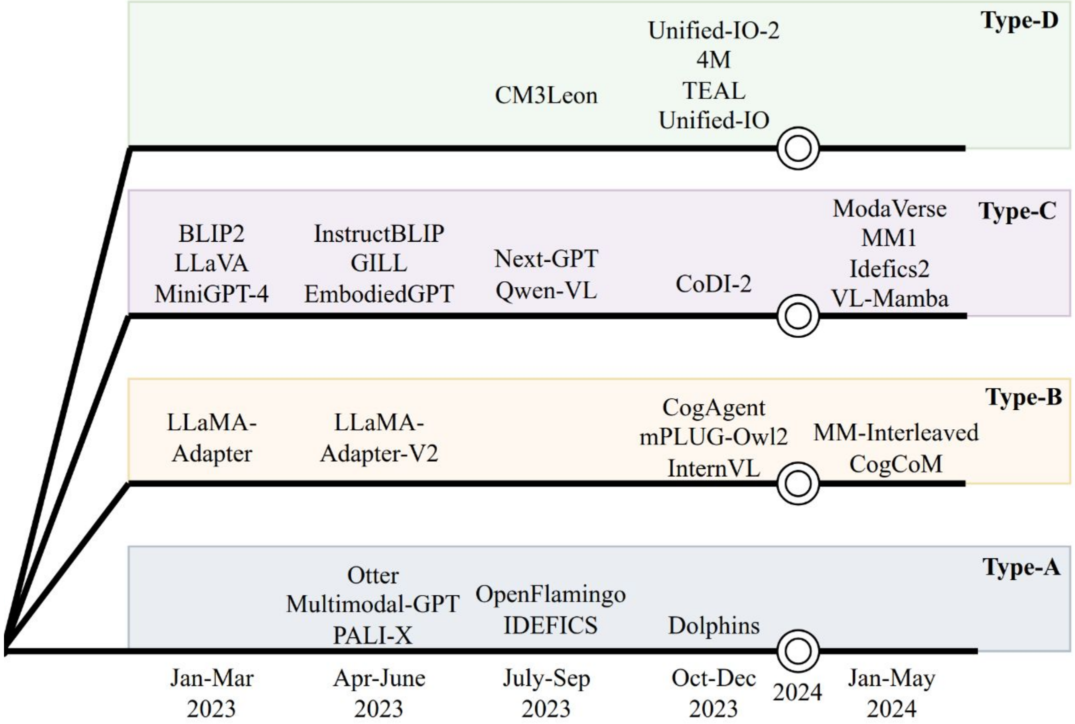


Figure 1: Development timeline of Multimodal models grouped in four proposed architecture types.

1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

A Brief History of MM Models



Modality Alignment

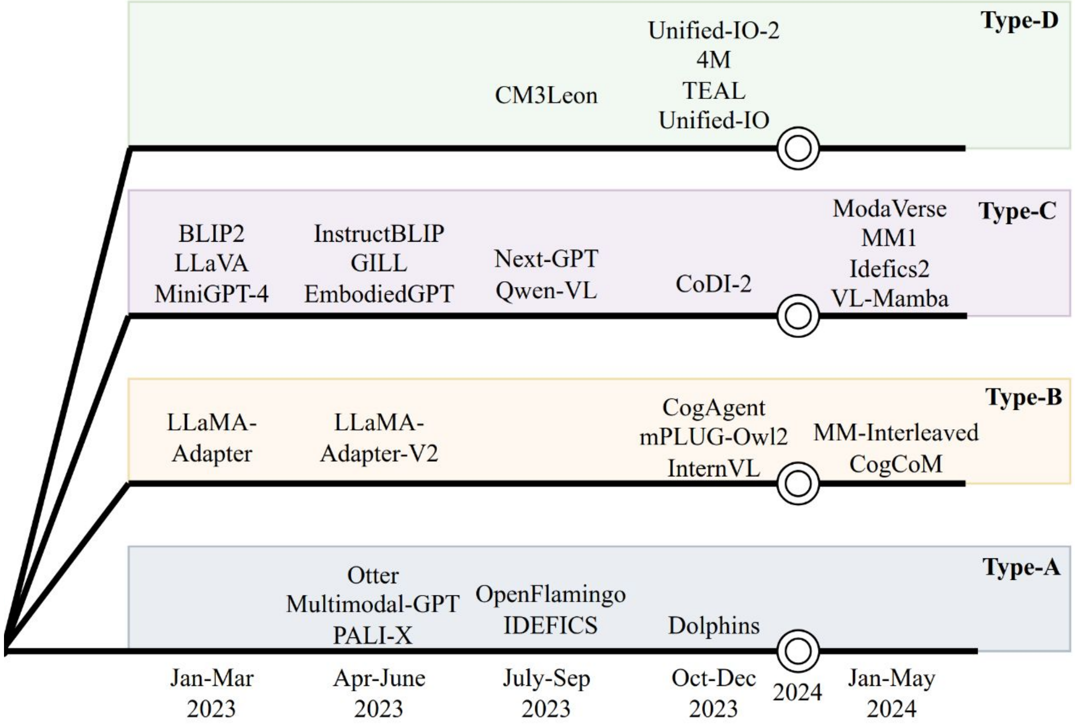
Modality are encoded/tokenized separately and then aligned

Fusion Models

Modality are fused at intermediate layers

1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

A Brief History of MM Models



Modality Encoders

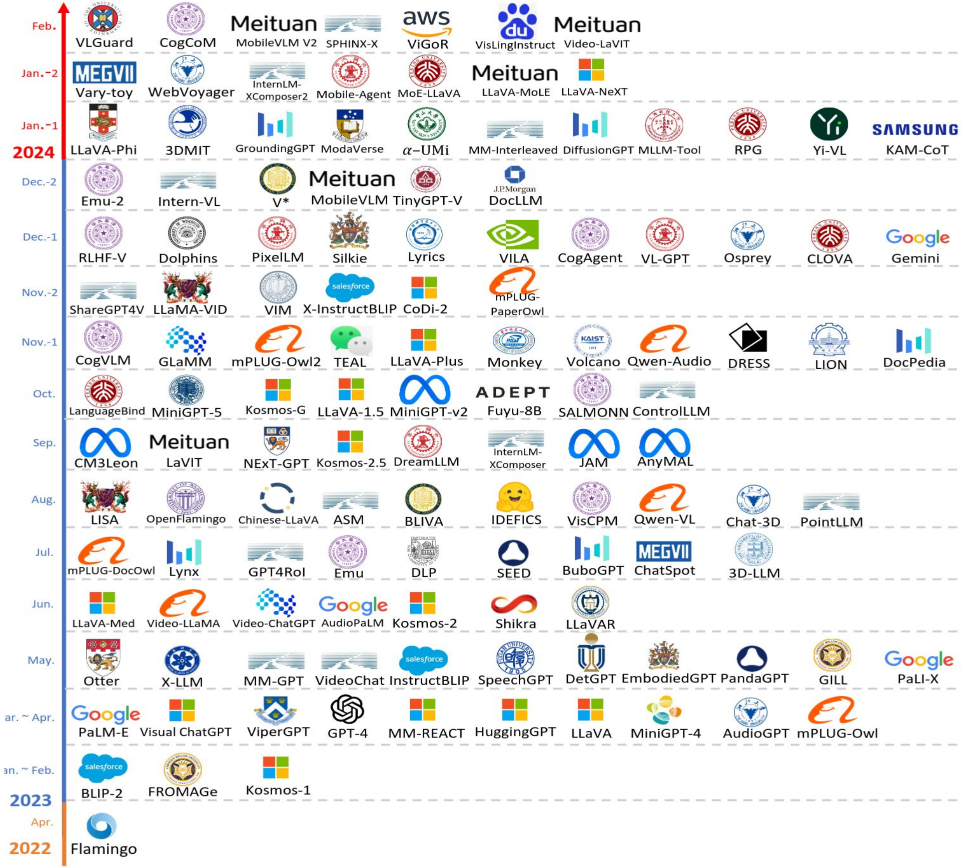
Modality are encoded/tokenized separately and then aligned

Fusion Models

Modality are fused at intermediate layers

1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

A Brief History of MM Models



Zhang, Duzhen, et al. "Mm-llms: Recent advances in multimodal large language models." arXiv preprint arXiv:2401.13601 (2024).

Contents

You
Generate a fun meme about multimodal LLMs like yourself

DALL·E

Multimodal Debuts: CLIP

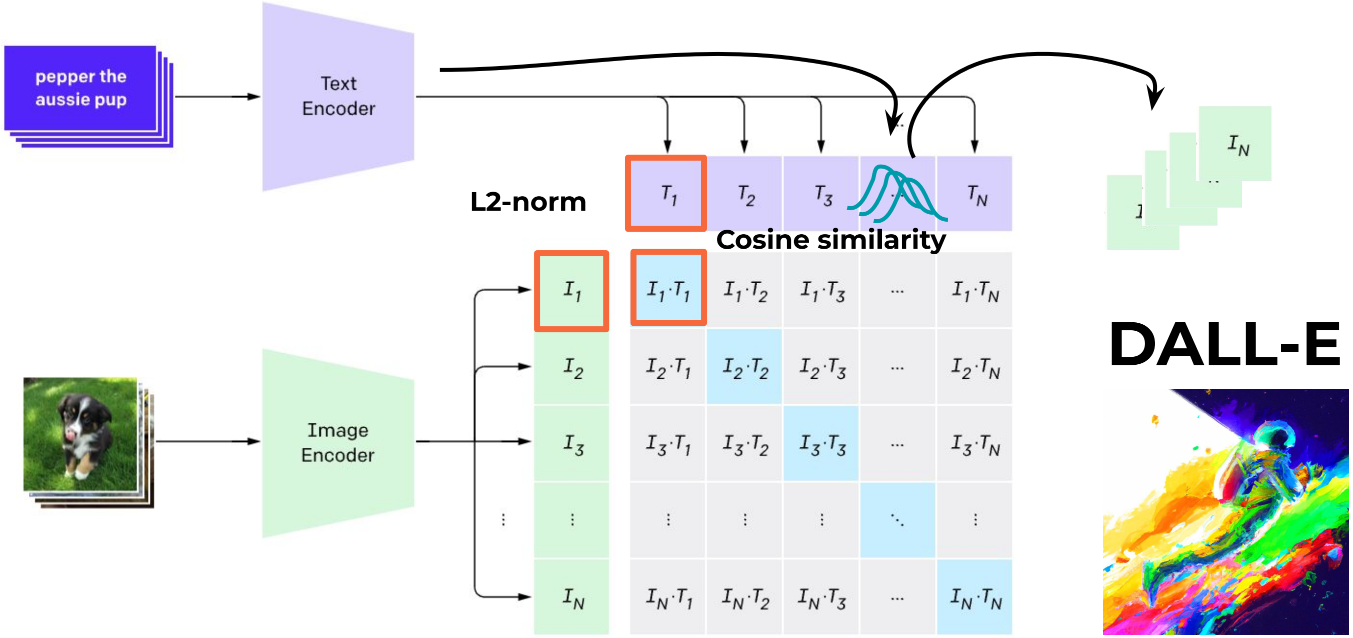
- One of the most prominent works that started a lot of research efforts on Multimodality

- Vision-Text Representation Learning: Image + Text

- Trained on large database of (*image, description*) pairs to learn a **multimodal embedding space**

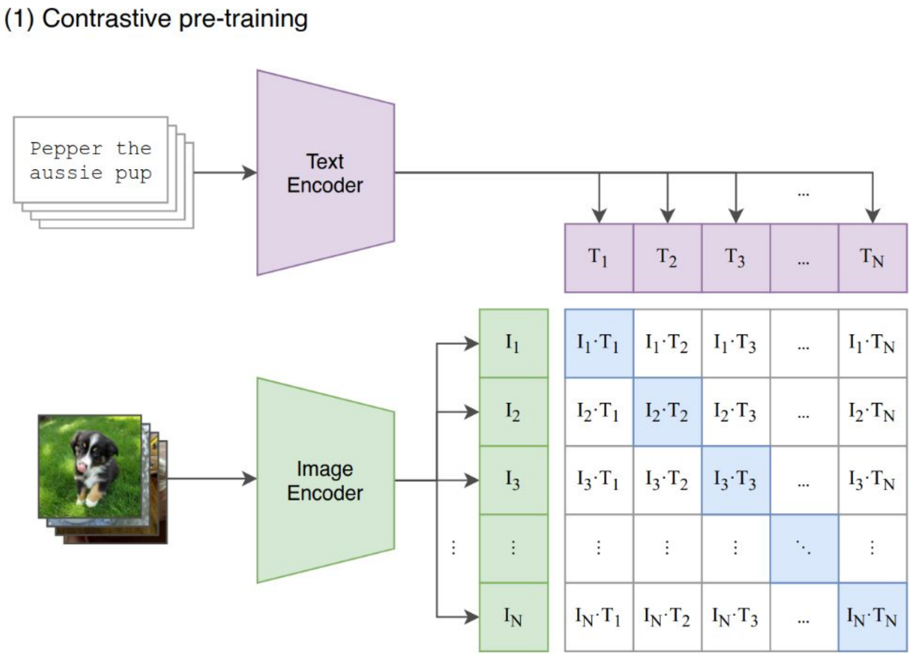
- Allows *zero-shot* transfer to most vision tasks using natural language supervision

Multimodal Debuts: CLIP



1. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Multimodal Debuts: CLIP



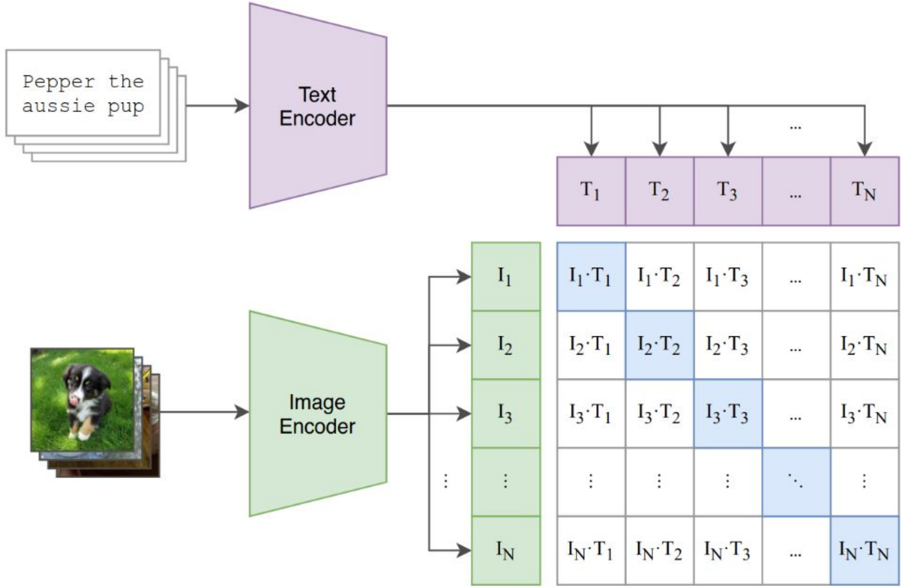
- Dataset: **WebImageText**
- 400 million (image, text) pair
- Various internet sources

1. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Multimodal Debuts: CLIP

(1) Contrastive pre-training

$$\sum_{i=1}^{B} -\log\left[\frac{h[f(\mathbf{x}^{(i)}), g(\mathbf{t}^{(i)})]}{\sum_{j\neq i} h[f(\mathbf{x}^{(i)}), g(\mathbf{t}^{(j)})] + h[f(\mathbf{x}^{(j)}), g(\mathbf{t}^{(i)})]}\right]$$

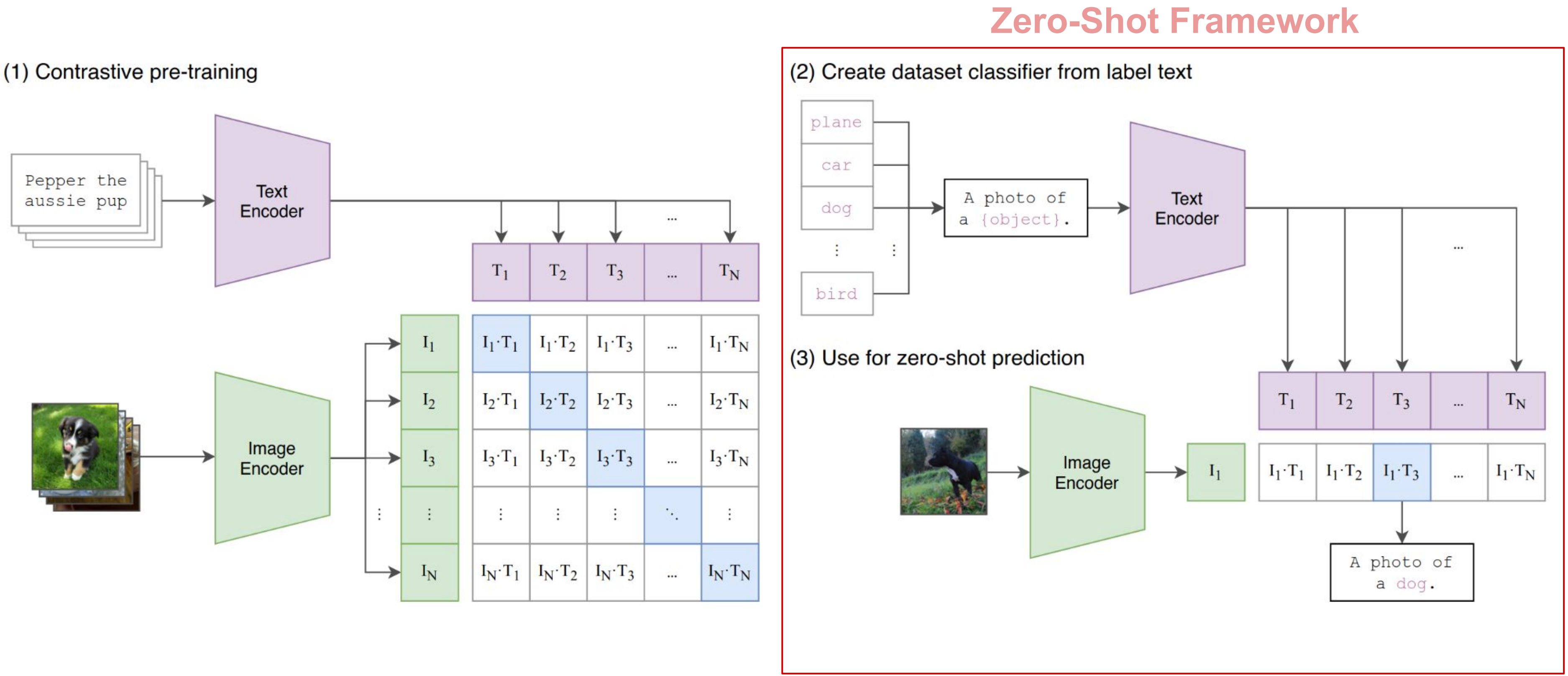



```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

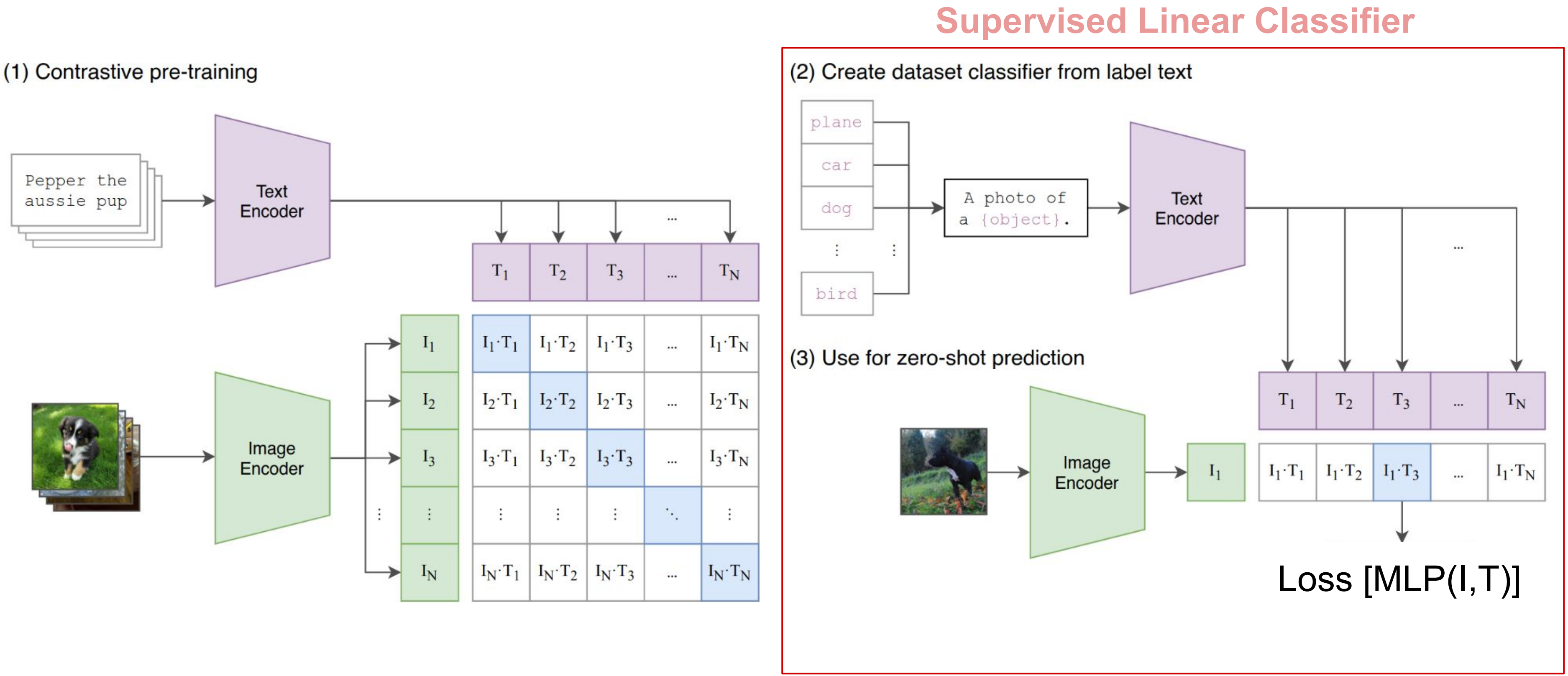
1. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Multimodal Debuts: CLIP



1. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Multimodal Debuts: CLIP



1. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

CLIP Outline

- Powerful **multimodal embedding space** using contrastive learning

- This space can be used as a **feature encoder**
    - Features can be used for downstream tasks (linear classifier)

- Can also be used for more **complex tasks**
    - Image Generation from text prompts (Dall-E)

From CLIP to Image Generation

- CLIP: Robust representations of images that capture both semantics and style

- Joint embedding space: language-guided image manipulations

- Goal: Generate Images from natural language description

a photo of a cat → an anime drawing of a super saiyan cat, artstation

vibrant portrait painting of Salvador Dalí with a robotic half face | a shiba inu wearing a beret and black turtleneck | a close up of a handpalm with leaves growing from it

1. Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.

From CLIP to Image Generation: unCLIP

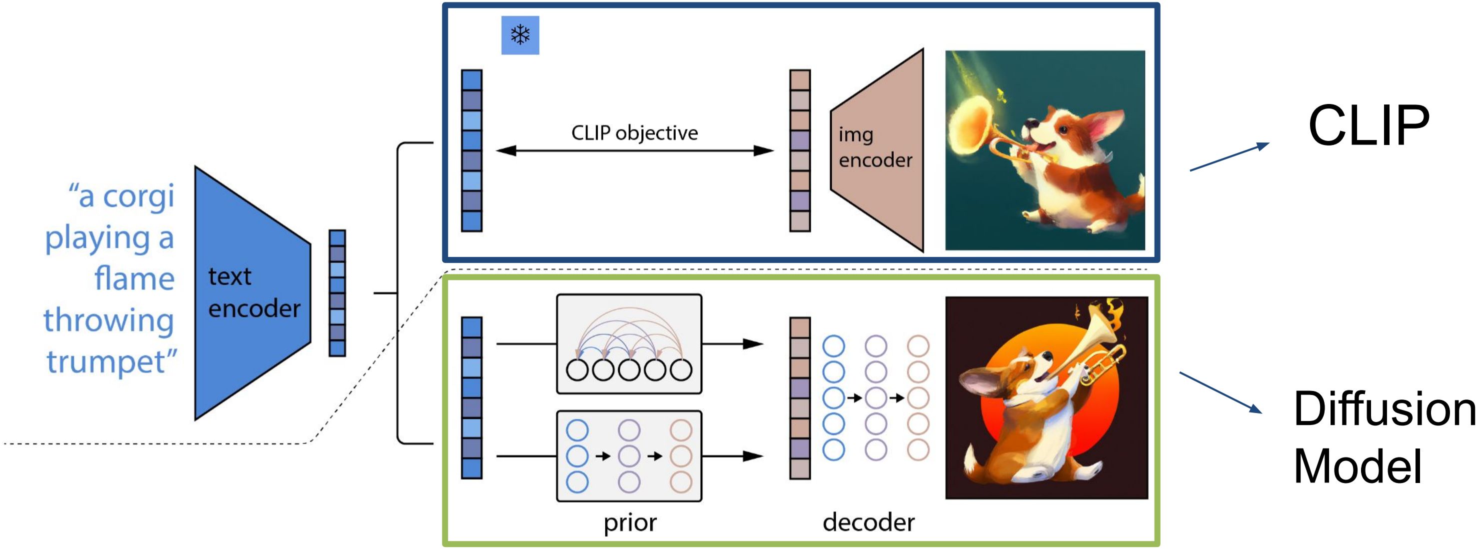
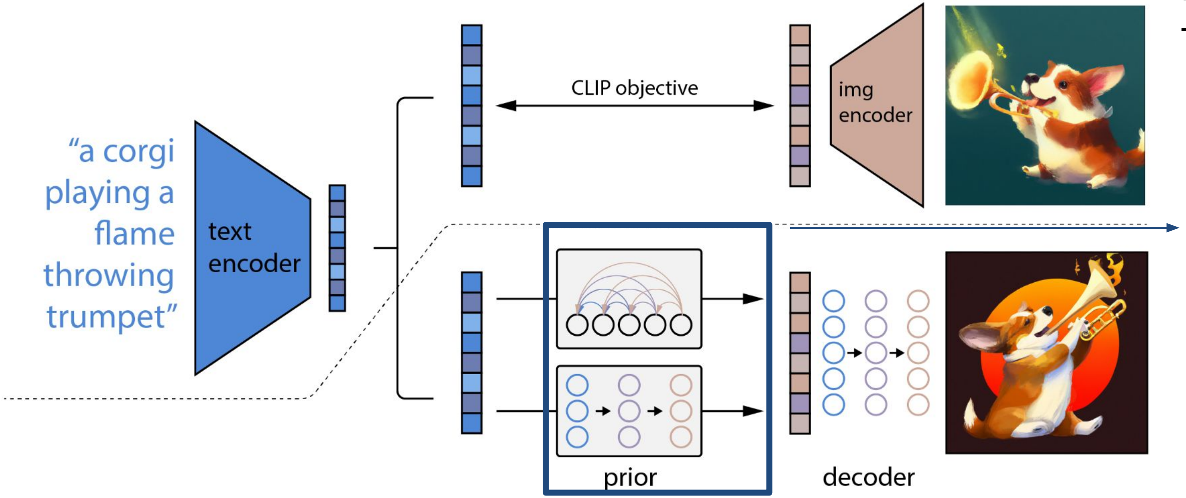


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

1. Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.

From CLIP to Image Generation: unCLIP

Goal: Produces CLIP Image embbedings z_i from captions y.



Decoder-only Transformer:
[text, *timestep*, CLIP(text), noised(z_i)]
—> z_i

Diffusion Prior: $P(z_i|y)$

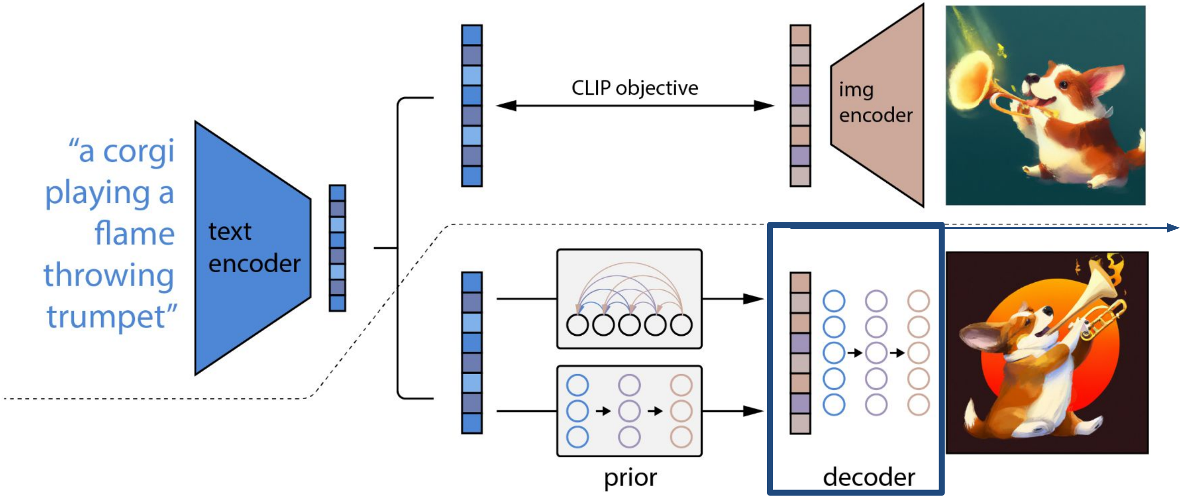
CLIP Image Embeddings z_i

Given caption y.

1. Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.

From CLIP to Image Generation: unCLIP

Goal: Invert CLIP Image embeddings z_i to produce images x



Decoder: $P(x|z_i, y)$

Generate image x
Given:

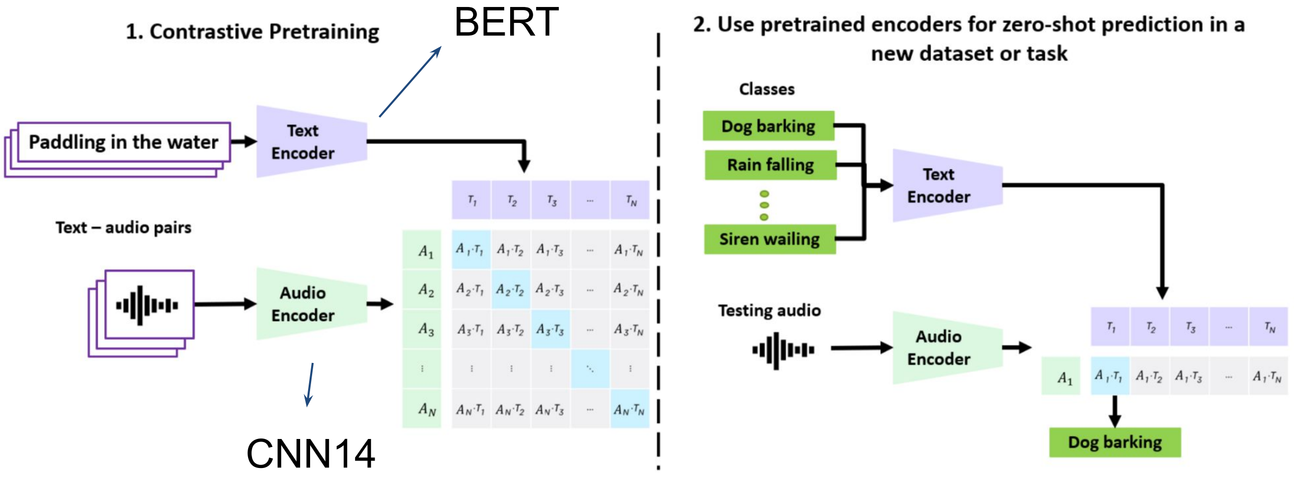
CLIP Image Embeddings z_i

(optional) caption y.

1. Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.

From Images to Audio

From Images to Audio: CLAP



128K (audio, text) pairs

Fig. 1. CLAP jointly trains an audio and a text encoder to learn the (dis)similarity of audio and text pairs in a batch using contrastive learning. At testing time, the pretrained encoders are used to extract audio embeddings from the testing audio and text embeddings from the class labels. Zero-Shot linear classification is achieved by computing cosine similarity between the embeddings.

1. Elizalde, Benjamin, et al. "Clap learning audio concepts from natural language supervision." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023

From Images to Audio: Mulan

44 million 30-second (music, description) pairs

Resnet-50
Or
Audio
Spectrogram
Transformer

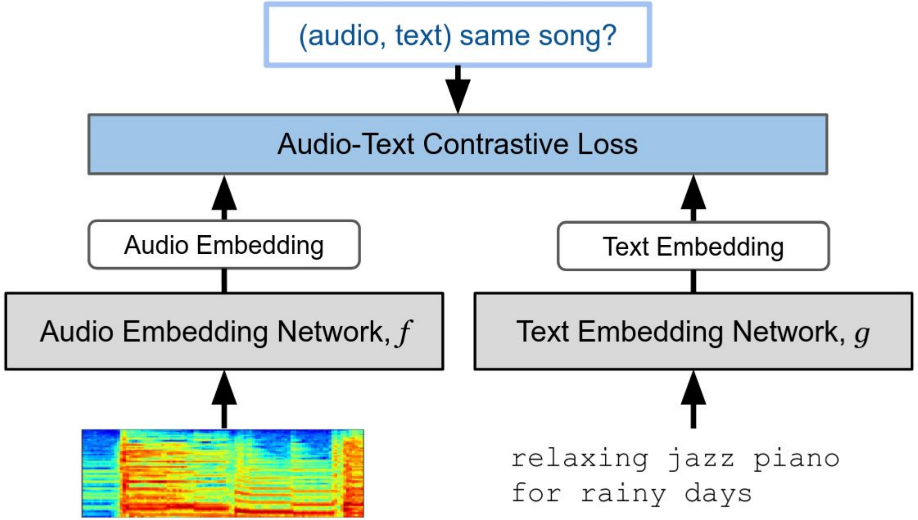


Figure 1. Learning framework diagram.

BERT

1. Huang, Qingqing, et al. "Mulan: A joint embedding of music audio and natural language." arXiv preprint arXiv:2208.12415 (2022)

From MULAN to Music Generation: MusicLM

- Audio synthesis as a language modeling task in a discrete representation space

- MULAN Tokens: joint audio-language embedding space

- Goal: Generate Music from natural language description

1. Agostinelli, Andrea, et al. "Musiclm: Generating music from text." arXiv preprint arXiv:2301.11325 (2023).

From MULAN to Music Generation: MusicLM

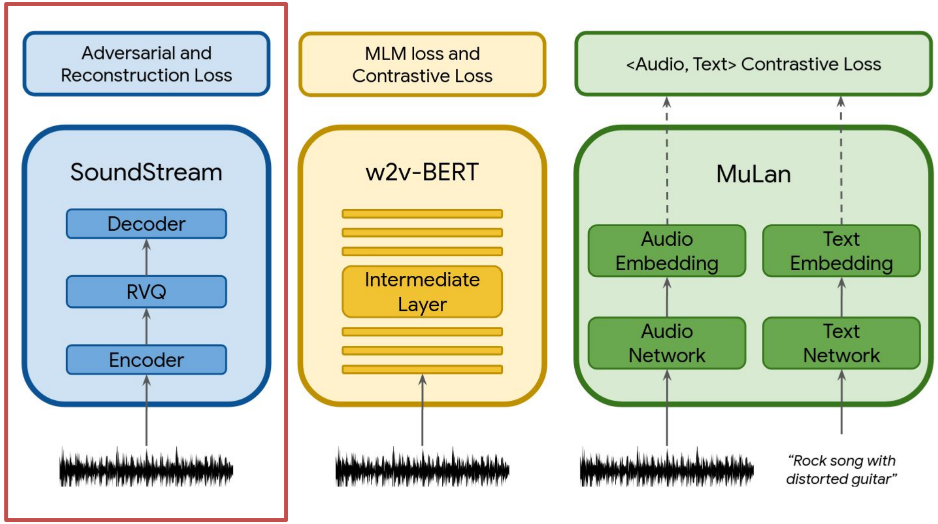


Figure 1. Independent pretraining of the models providing the audio and text representations for MusicLM: SoundStream (Zeghidour et al., 2022), w2v-BERT (Chung et al., 2021), and MuLan (Huang et al., 2022).

SoundStream: Self-Supervised Audio Encoder → **Acoustic Tokens**

1. Agostinelli, Andrea, et al. "Musiclm: Generating music from text." arXiv preprint arXiv:2301.11325 (2023).

From MULAN to Music Generation: MusicLM

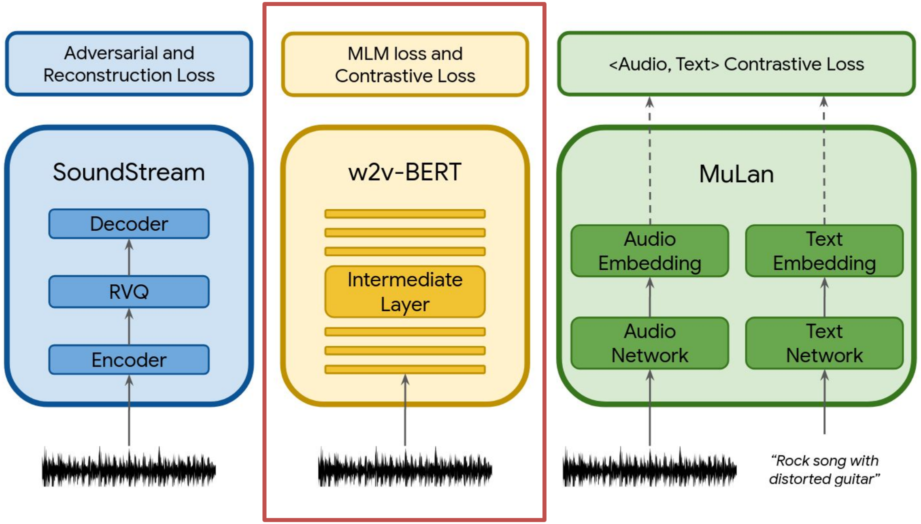


Figure 1. Independent pretraining of the models providing the audio and text representations for MusicLM: SoundStream (Zeghidour et al., 2022), w2v-BERT (Chung et al., 2021), and MuLan (Huang et al., 2022).

SoundStream: Self-Supervised Audio Encoder ⟶ **Acoustic Tokens**

w2v-BERT: Self-Supervised Speech Pretraining ⟶ **Semantic Tokens**

1. Agostinelli, Andrea, et al. "Musiclm: Generating music from text." arXiv preprint arXiv:2301.11325 (2023).

From MULAN to Music Generation: MusicLM

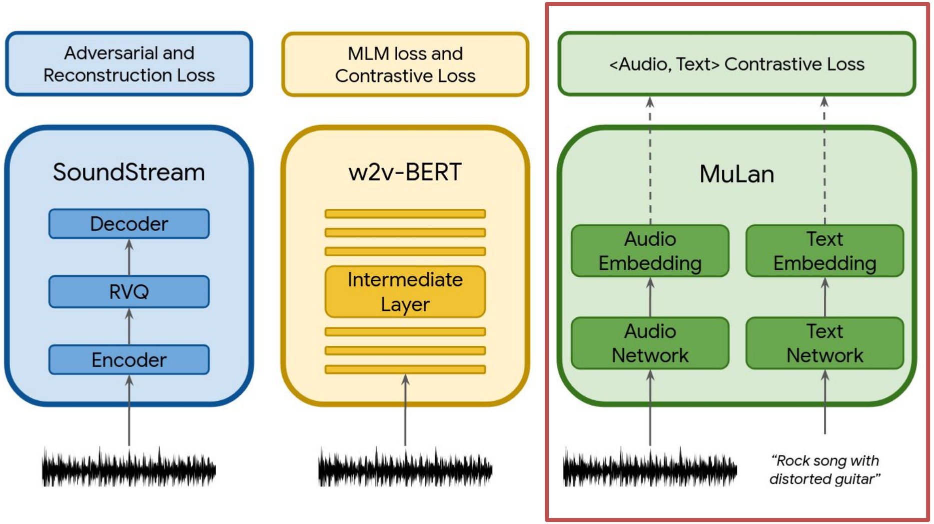


Figure 1. Independent pretraining of the models providing the audio and text representations for MusicLM: SoundStream (Zeghidour et al., 2022), w2v-BERT (Chung et al., 2021), and MuLan (Huang et al., 2022).

SoundStream: Self-Supervised Audio Encoder ⟶ **Acoustic Tokens**

w2v-BERT: Self-Supervised Speech Pretraining ⟶ **Semantic Tokens**

MuLAN: Audio-Text joint embedding space: **used to condition music generation**

1. Agostinelli, Andrea, et al. "Musiclm: Generating music from text." arXiv preprint arXiv:2301.11325 (2023).

From MULAN to Music Generation: MusicLM

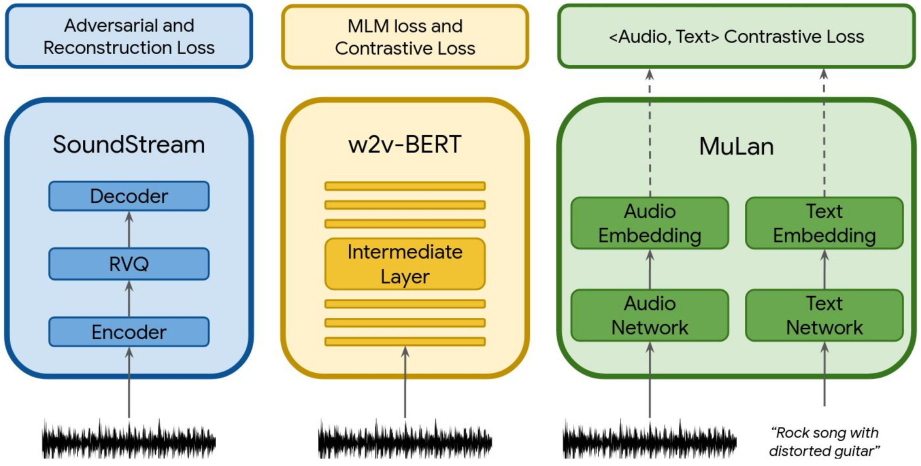


Figure 1. Independent pretraining of the models providing the audio and text representations for MusicLM: SoundStream (Zeghidour et al., 2022), w2v-BERT (Chung et al., 2021), and MuLan (Huang et al., 2022).

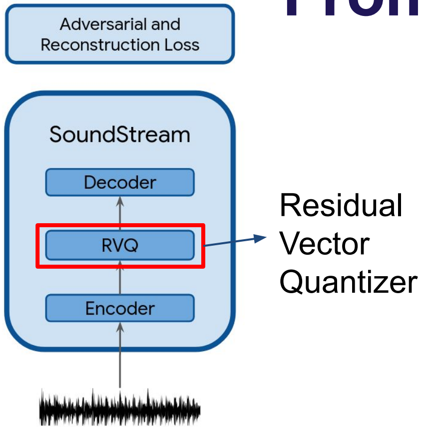
SoundStream: Self-Supervised Audio Encoder ⟶ **Acoustic Tokens**

w2v-BERT: Self-Supervised Speech Pretraining ⟶ **Semantic Tokens**

MuLAN: Audio-Text joint embedding space: **used to condition music generation**

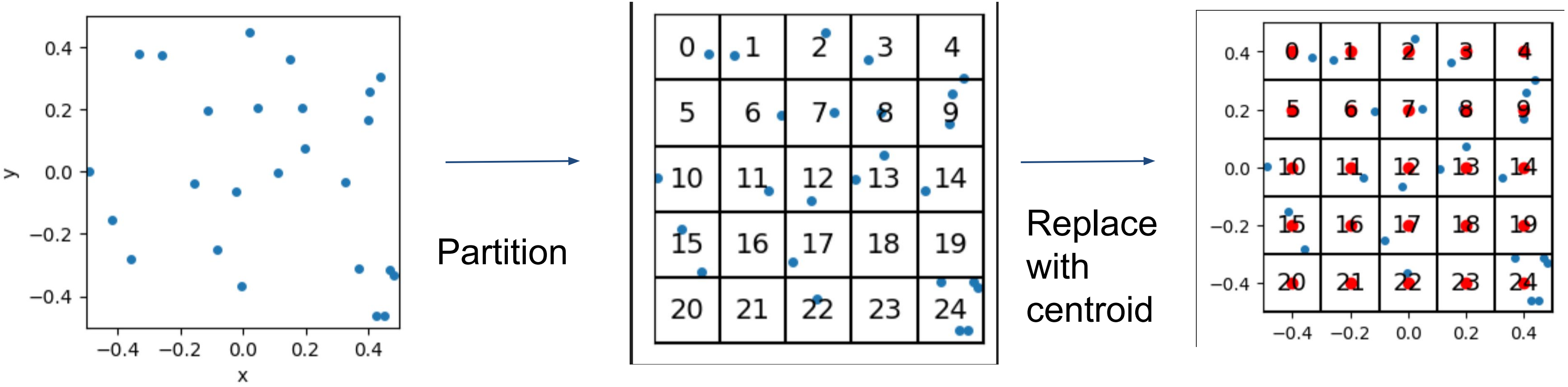
1. Agostinelli, Andrea, et al. "Musiclm: Generating music from text." arXiv preprint arXiv:2301.11325 (2023).

From MULAN to Music Generation: MusicLM



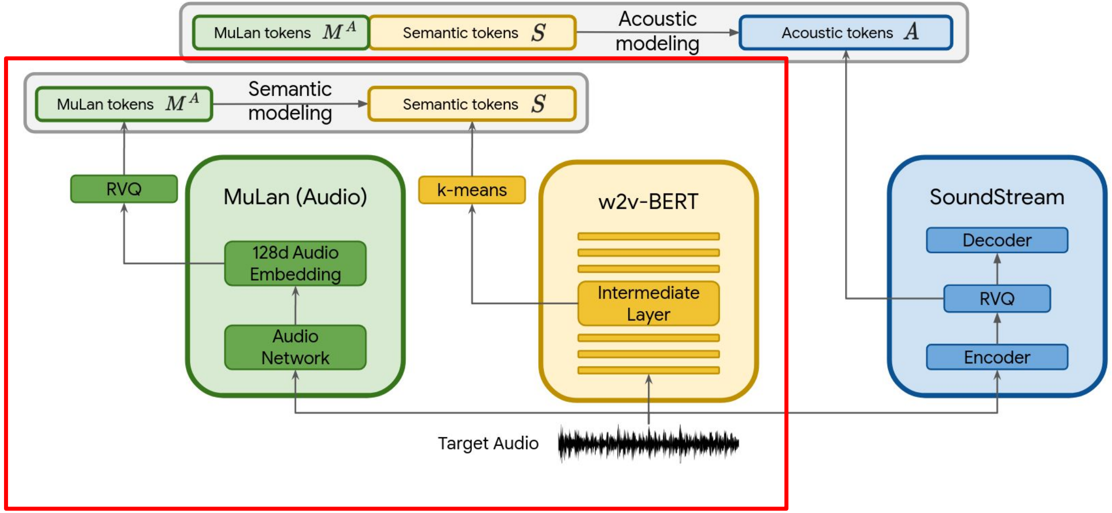
Quantization of continuous representation, also known as **audio-codecs**

Goal: Compress the continuous output to a set of one hot vectors → reduces the computational cost



Zeghidour, Neil, et al. "Soundstream: An end-to-end neural audio codec." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 495-507.

From MULAN to Music Generation: MusicLM



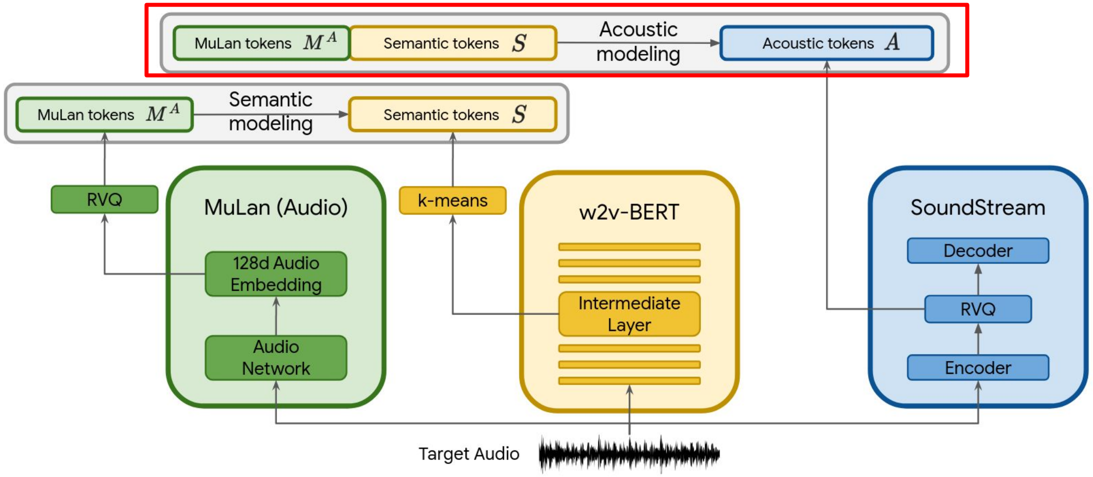
Training: only audio
Sequence-to-sequence modeling (Transformer Decoder)

Step 1. Semantic modeling

$$p(S_t | S_{<t}, M_A)$$

1. Agostinelli, Andrea, et al. "Musiclm: Generating music from text." arXiv preprint arXiv:2301.11325 (2023).

From MULAN to Music Generation: MusicLM



Training: only audio Sequence-to-sequence modeling (Transformer Decoder)

Step 1. Semantic Modeling

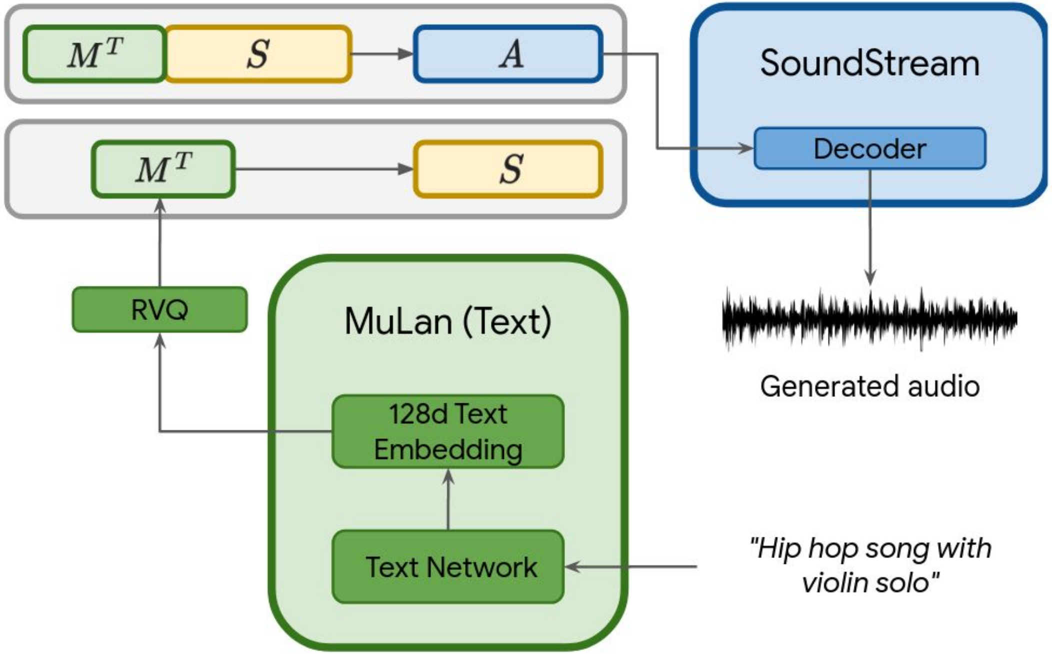
$$p(S_t|S_{<t}, M_A)$$

Step 2. Acoustic Modeling

$$p(A_t|A_{<t}, S, M_A).$$

1. Agostinelli, Andrea, et al. "Musiclm: Generating music from text." arXiv preprint arXiv:2301.11325 (2023).

From MULAN to Music Generation: MusicLM



Inference: only Text

→ Mulan tokens M^T used to predict Semantic Tokens S

→ Given M^T and S, predict Acoustic Tokens A

→ Use A to generate audio with SoundStream

[Website for examples]

1. Agostinelli, Andrea, et al. "Musiclm: Generating music from text." arXiv preprint arXiv:2301.11325 (2023).

MusicLM: Outline

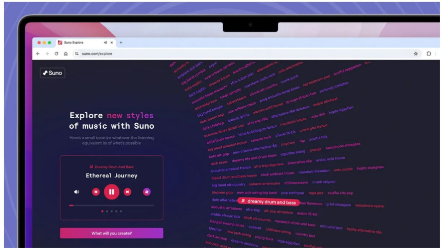
- Music generation from natural language description is feasible
- High audio quality thanks to powerful audio encoders using quantization
- MusicLM (google) was released in 2023. Since then, multiple tools created to generate music freely (link to Suno)

What is Suno? The viral AI song generator explained – and how to use it for free

Features By Mark Wilson | Contributions from Graham Barlow last updated July 23, 2024

It's time to stage-dive into AI song generators

When you purchase through links on our site, we may earn an affiliate commission. Here's how it works.

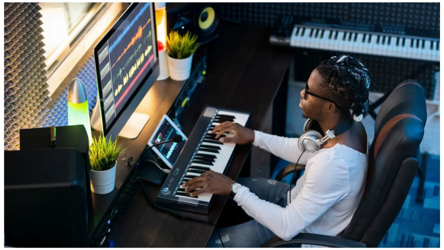


Adobe's new AI music tool could make you a text-to-musical genius

News By Cesar Cadenas published February 28, 2024

It can create musically complex tracks

When you purchase through links on our site, we may earn an affiliate commission. Here's how it works.

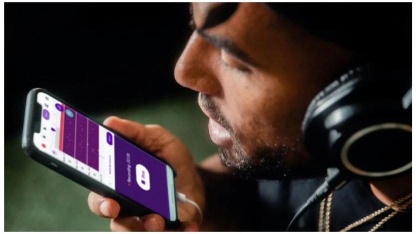


YouTube's new AI tool will let you create your dream song with a famous singer's voice

News By Cesar Cadenas published November 16, 2023

Dream Track can generate 30-second clips for YouTube Shorts

When you purchase through links on our site, we may earn an affiliate commission. Here's how it works.



Multimodal Debuts - Outline

- Multimodal architectures to generate images, audio, videos
  - Based on “multimodal encoders”: joint-modality embedding space
- So far, we have seen text-encoders based on “classic models”: BERT or related
- **What about Large Language Models?**

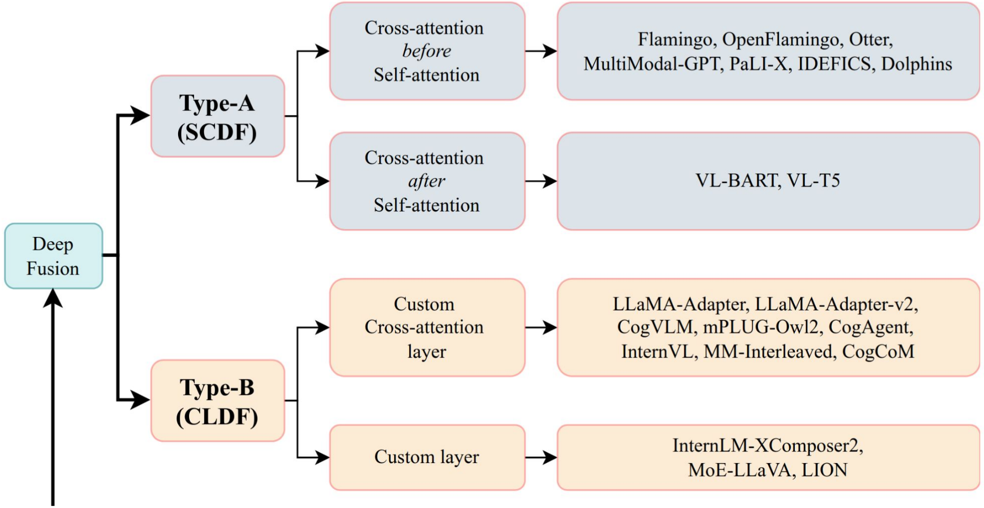
Contents

You
Generate a fun meme about multimodal LLMs like yourself

DALL·E

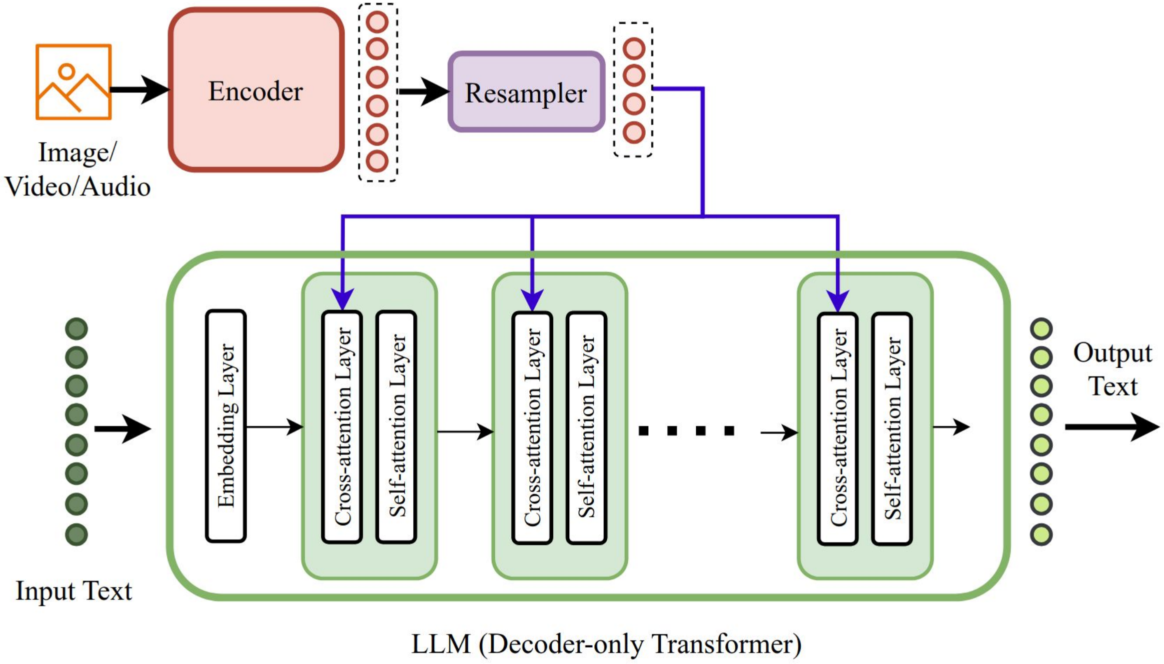
Deep Fusion Models

- Models that fuse modalities at intermediate layers
- Each model is different in how it fuses modalities → complex architectures



1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

Deep Fusion Models
Type A



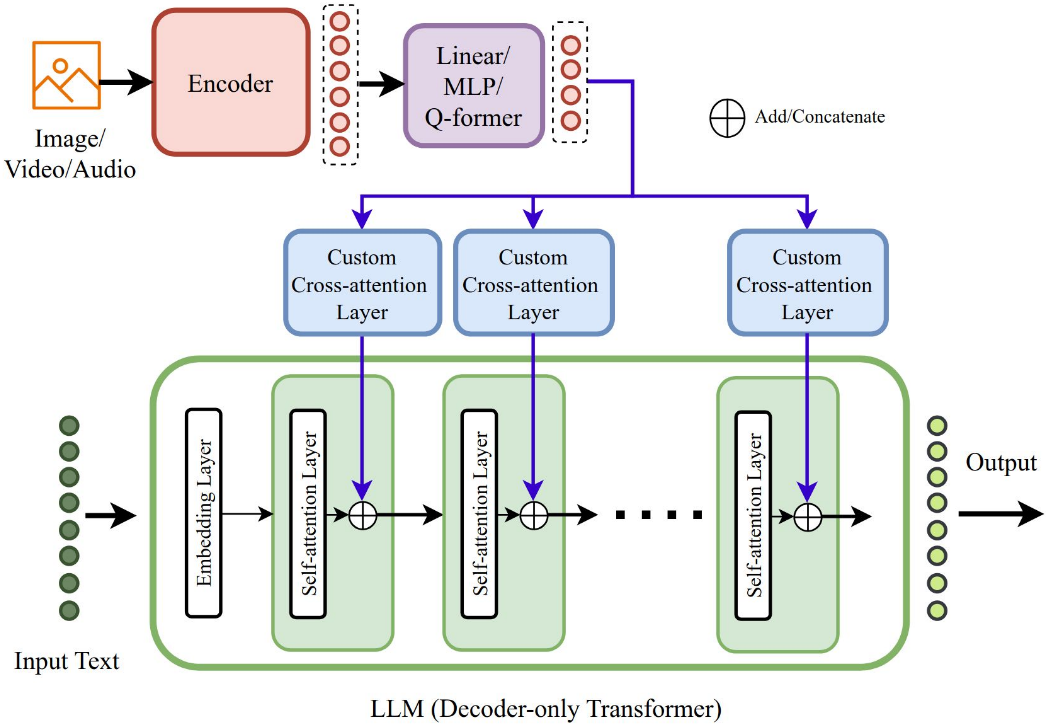
Cross-Attention **Before** Self-Attention

Early multimodal models

Hard to more that one other modalities

1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

Deep Fusion Models
Type B



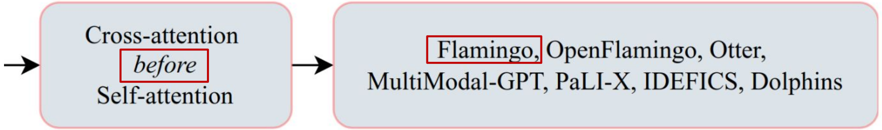
Cross-Attention **After** Self-Attention

More Scalable than Type A (because custom layers)

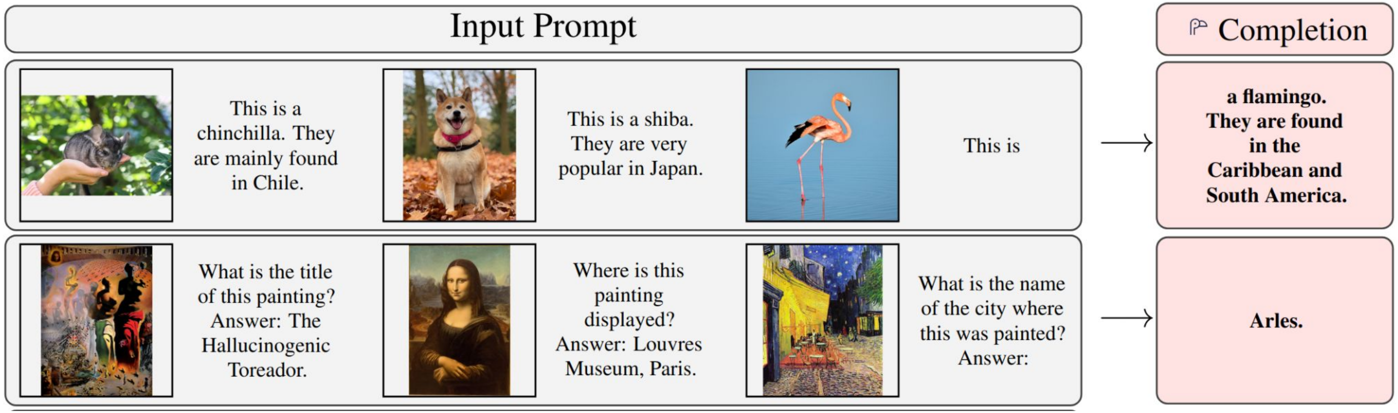
Easier to add input modalities

1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

Example - Flamingo (2022)



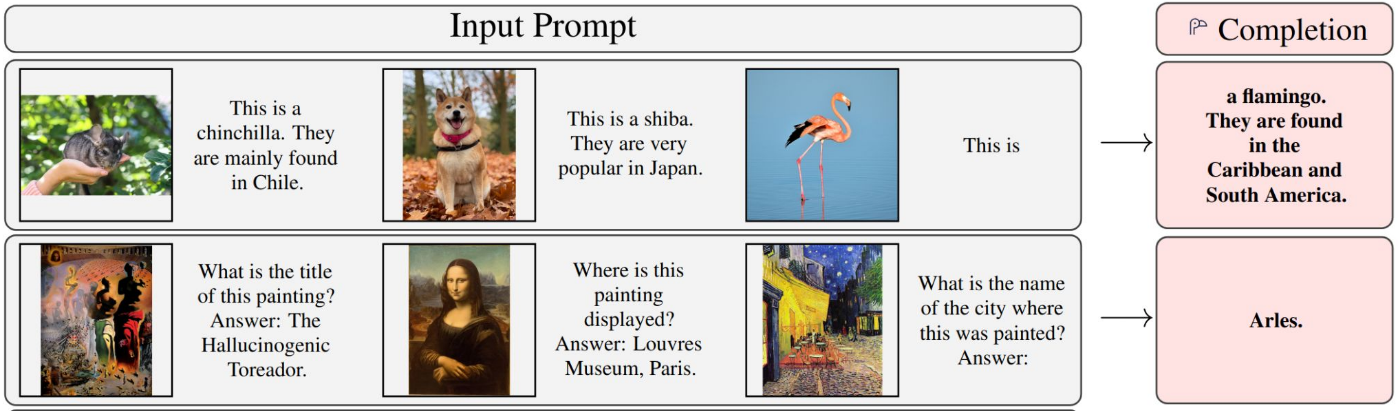
- Visual Language Model:
  - Vision + Text ⟶ Text



1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Flamingo

- Contributions:
  - Flexible: process sequences of text and images (video) of arbitrary length
  - Accurate: Few-shot Flamingo >= fine-tuned models



1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Flamingo - Encoding Images

Reduces the embedding space to 64 →less computation required

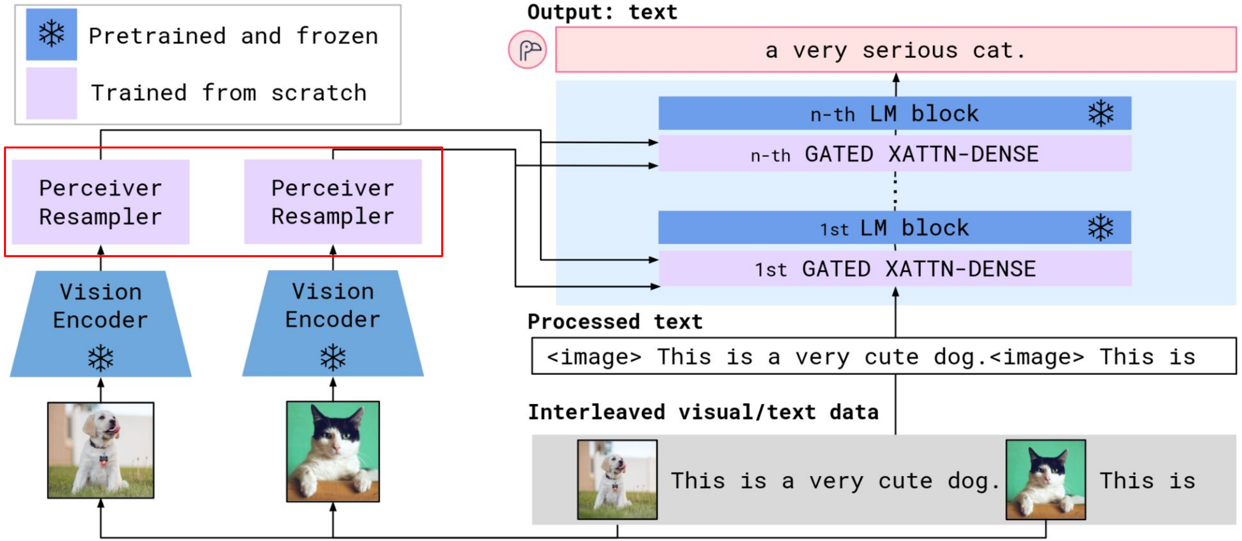
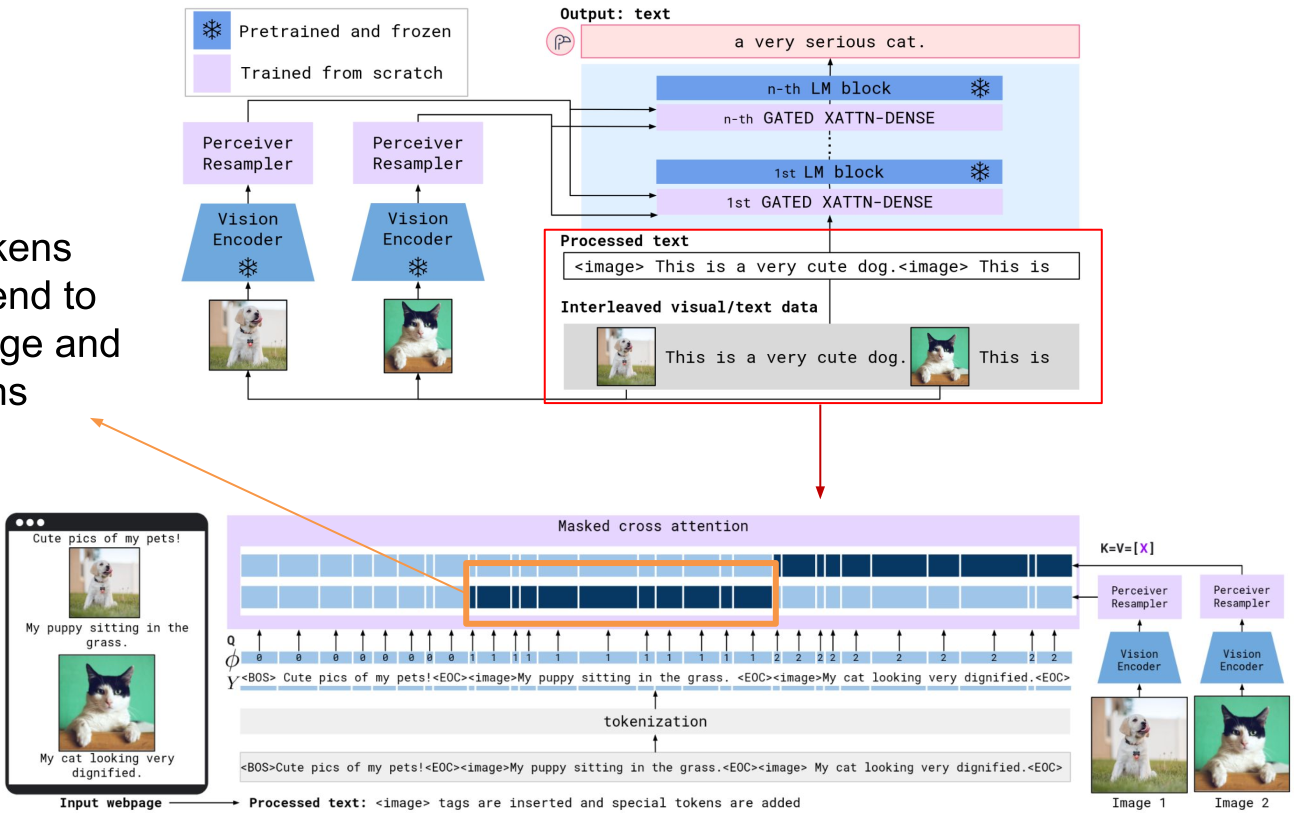


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

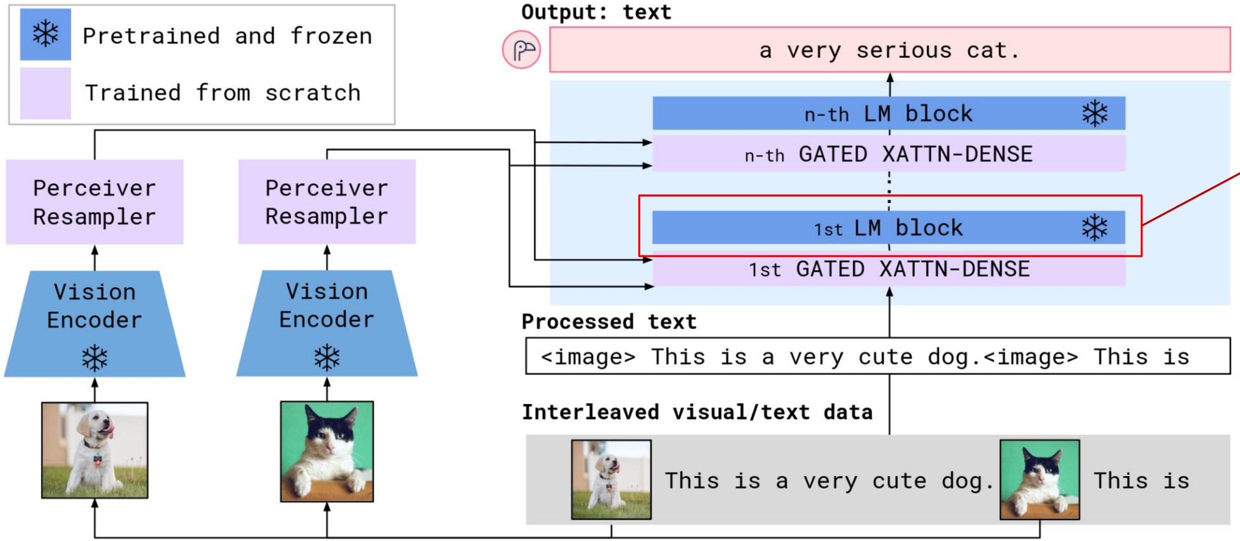
Flamingo - Interleaved Text - Image

Image/Text tokens only cross-attend to preceding image and in-chunk tokens



1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Flamingo - Encoding Interleaved Image-Text



Large Language Model: **Chinchilla**

Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Flamingo - Encoding Interleaved Image-Text

Condition frozen LLM to Visual Input

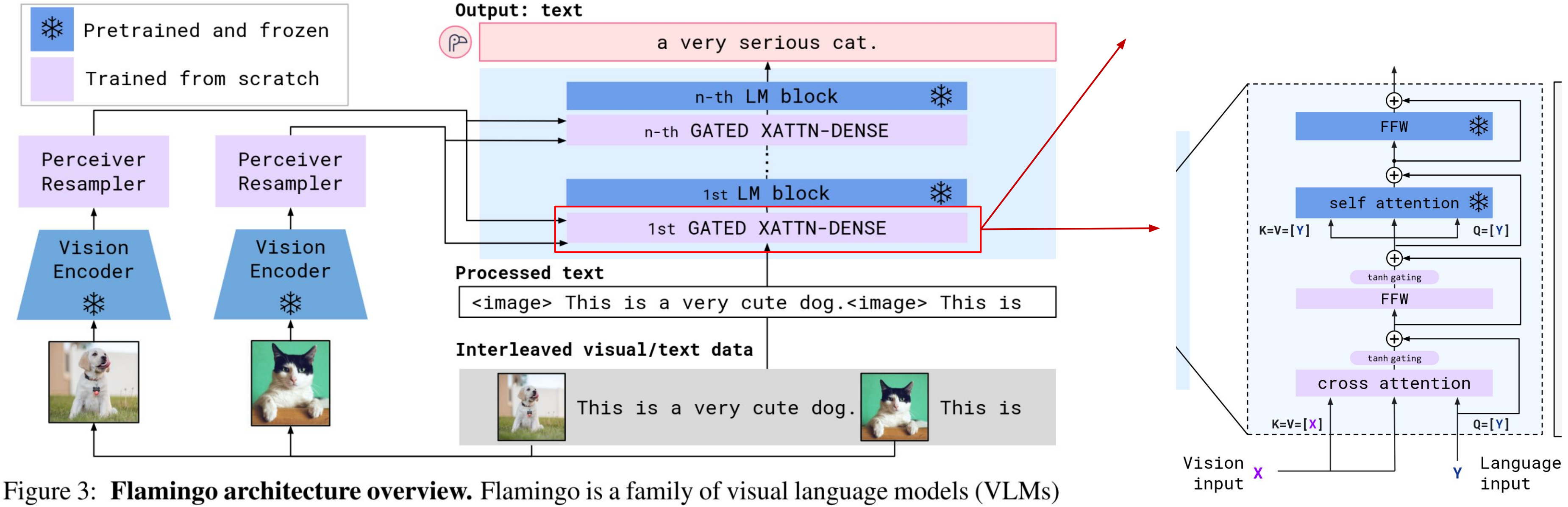


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Flamingo - Training

- Trained layers: **Perceived Resampler** & **Cross-Attention Layers**

- Data:
  - Image - Text: (image, text) pairs & Interleaved Web pages
  - Video - Text: 27 millions of (22 sec video, description) pairs

- Objective:
  - (image, text) or (video,description): trained to generate the caption
  - Interleaved: trained to complete the text conditioned on previous text & images

1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Deep Fusion Models - Outline

- Early multimodal Models
- Fusing modalities at intermediate layers
  - Cross-attention between modalities
- **Fine-grained control** of how modality information flow
- Often **very complex architectures**
- Cross-Attention add large number of parameters
  - **High computational complexity**

1. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

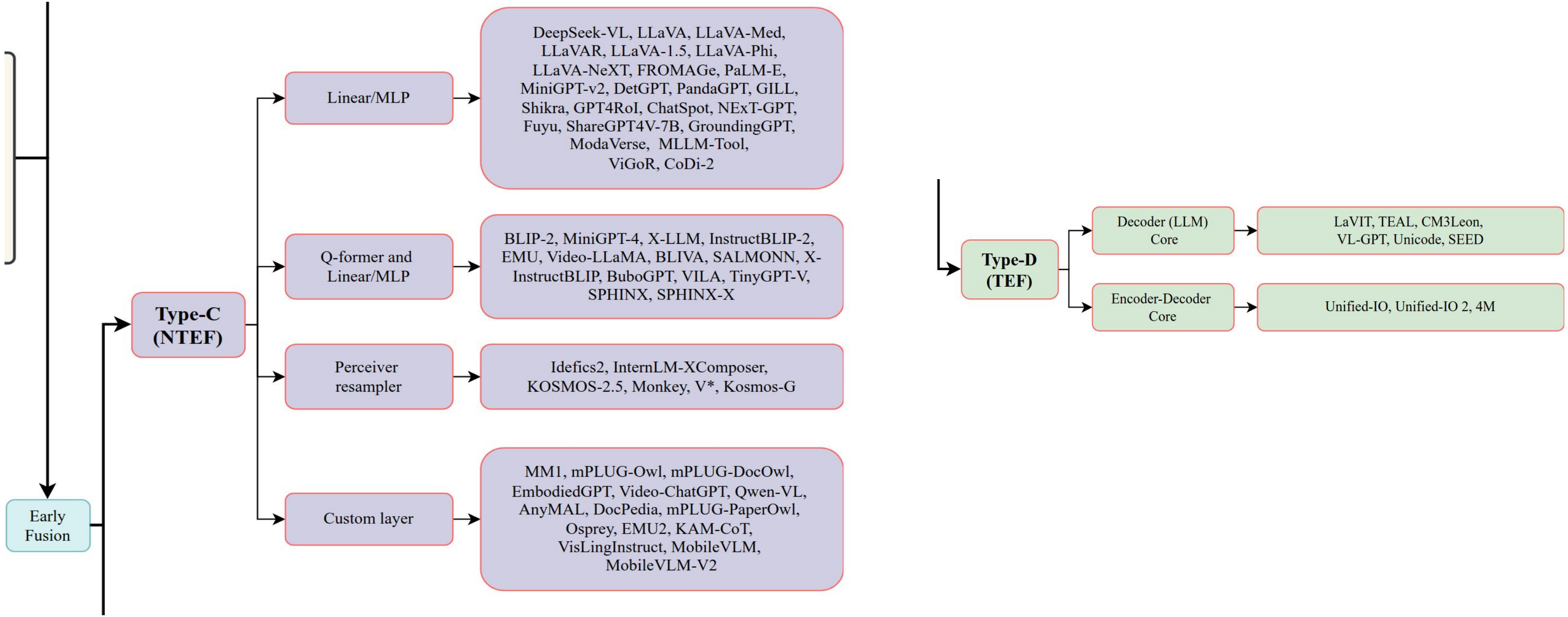
Contents

M **You**
Generate a fun meme about multimodal LLMs like yourself

DALL·E

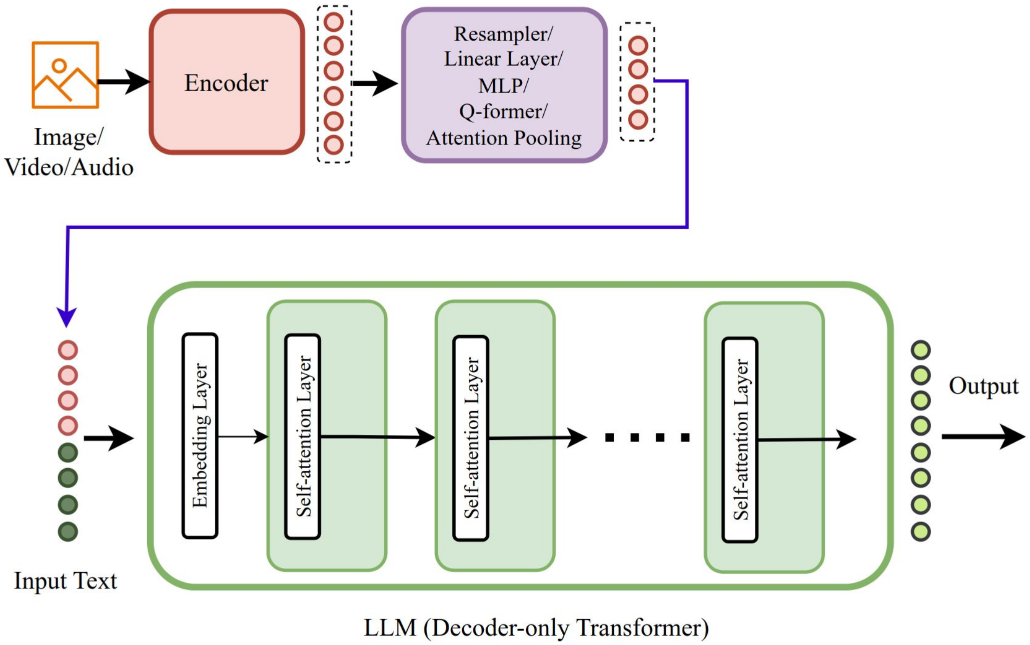
Early Fusion Models

- Models that fuse modalities at early layers at the embedding or tokenizer level



1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

Early Fusion Models
Type C



Non-Tokenized Early Fusion

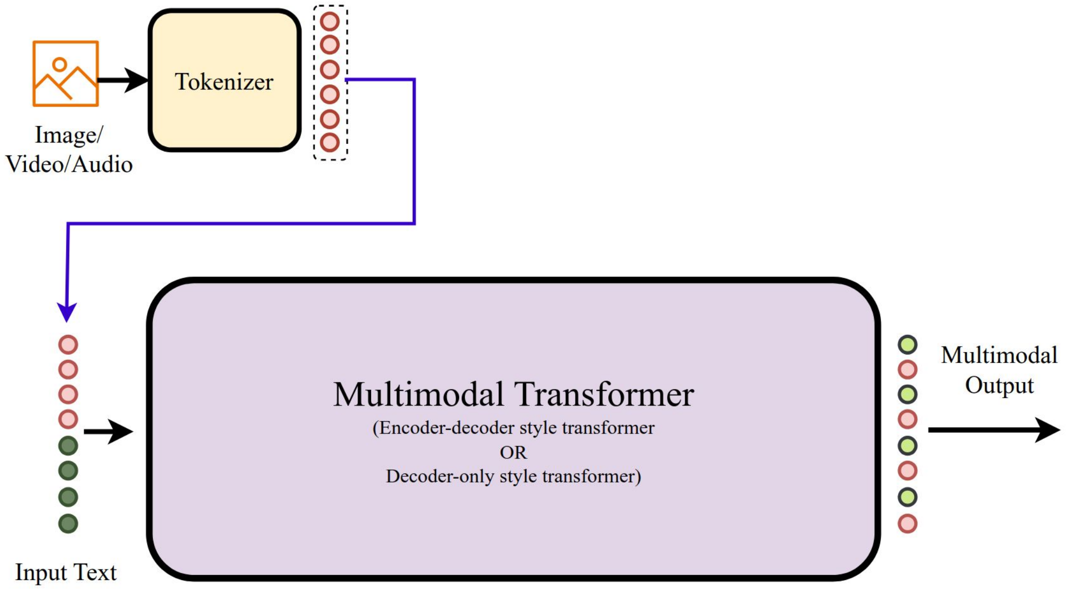
Most prominent multimodal architecture

Modular architecture

Scalable

1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

Early Fusion Models
Type D



Tokenized Early Fusion

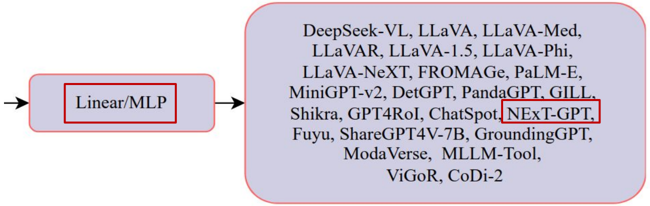
Relies on cross-modal tokenizers or powerful unimodal tokenizers

Scalable

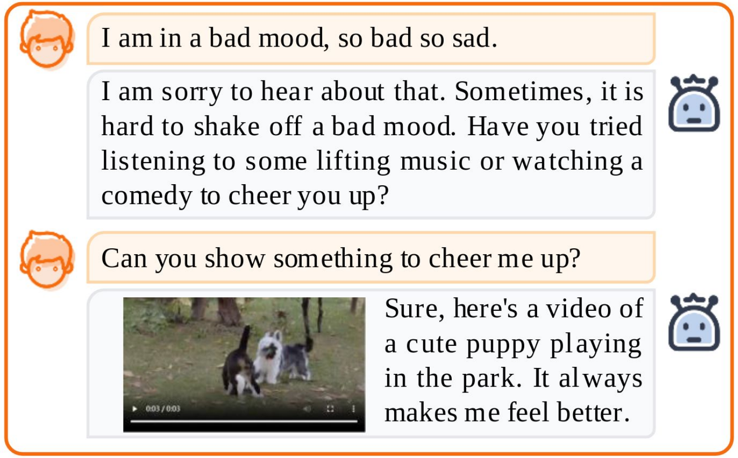
Hard to integrate new modalities (need tokenizer)

1. Wadekar, Shakti N., et al. "The Evolution of Multimodal Model Architectures." *arXiv preprint arXiv:2405.17927* (2024).

Example - NExT-GPT (2024)



- Any Modality → Any Modality



1. Wu, Shengqiong, et al. "NExT-GPT: Any-to-Any Multimodal LLM." Forty-first International Conference on Machine Learning (2024).

Example - NExT-GPT (2024)

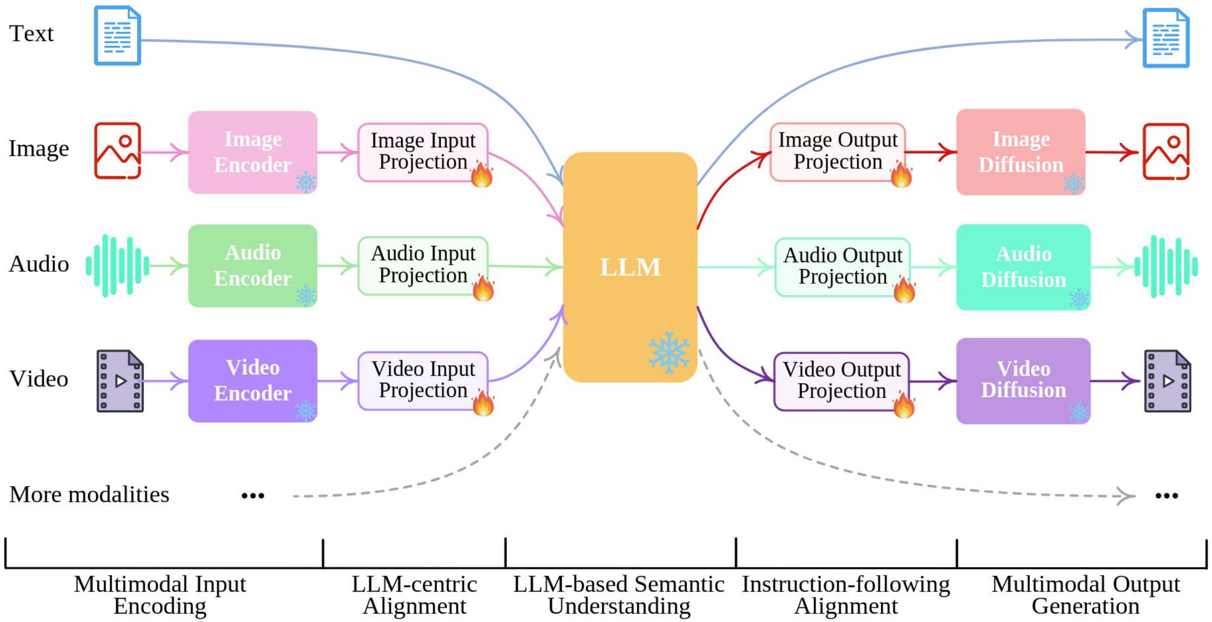


Table 1. Summary of NExT-GPT system configuration. Only 1% of parameters need updating during fine-tuning.

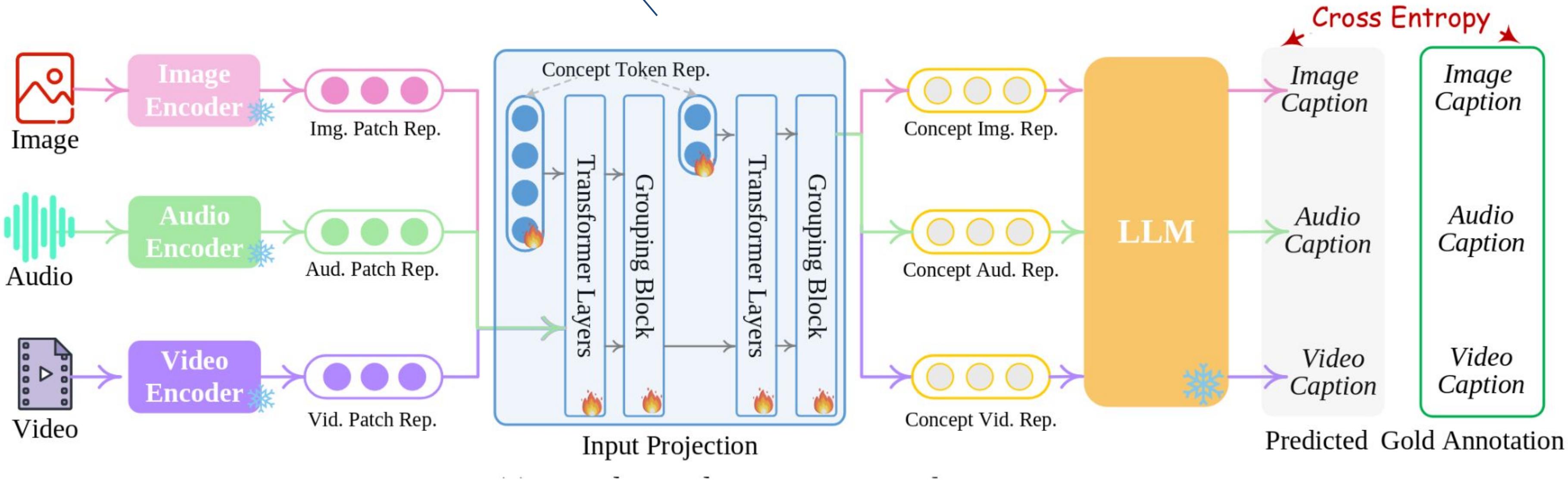
	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—	Vicuna (LoRA	7B❄ 33M🔥)	—	—	—	—
Image	ImageBind	1.2B❄	Grouping	28M🔥			Transformer	31M🔥	SD	1.3B❄
Audio							Transformer	31M🔥	AudioLDM	975M❄
Video							Transformer	32M🔥	Zeroscope	1.8B❄

1. Wu, Shengqiong, et al. "NExT-GPT: Any-to-Any Multimodal LLM." Forty-first International Conference on Machine Learning (2024).

NExT-GPT - Encoding Side

Custom layer to **align** patched representation to **semantic textual tokens**

Loss 1:
Negative log-likelihood of producing signal tokens

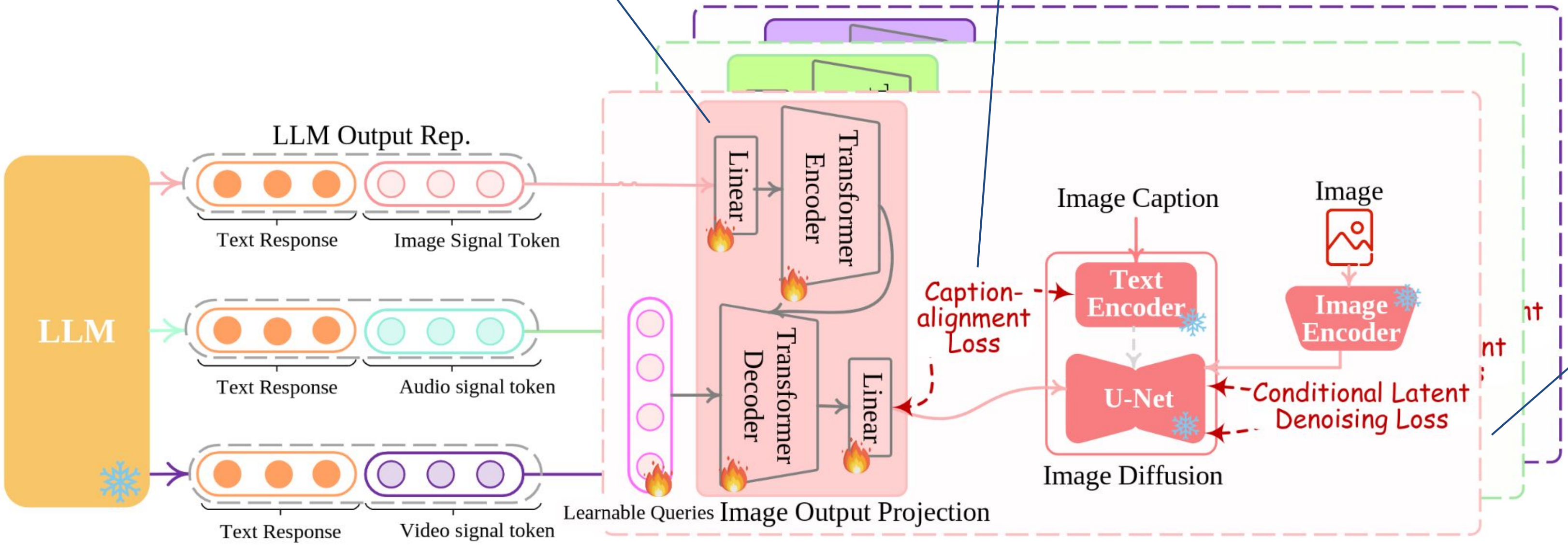


1. Wu, Shengqiong, et al. "NExT-GPT: Any-to-Any Multimodal LLM." Forty-first International Conference on Machine Learning (2024).

NExT-GPT - Decoding Side

Custom layer to **align** LLM output signal to **pretrained diffusion models**

Loss 2:
l2-distance between the hidden states of signal tokens produced by the LLM and the conditional text representation



Loss 3:
Denoising loss of the diffusion model

1. Wu, Shengqiong, et al. "NExT-GPT: Any-to-Any Multimodal LLM." Forty-first International Conference on Machine Learning (2024).

NExT-GPT - Outline

- Type-C model that uses a custom projection layer to align modalities
- Needs **custom losses** to make sure **correct alignment**
- Also needs further Modality-Switching Instruction Tuning to make sure model follows user instructions

1. Wu, Shengqiong, et al. "NExT-GPT: Any-to-Any Multimodal LLM." Forty-first International Conference on Machine Learning (2024).

Contents

You
Generate a fun meme about multimodal LLMs like yourself

DALL·E

Conclusion

- Research in MM-LLMs is recent but wide
- Some researchers see it as an extra-step towards AGI
  - Agents capable of understanding multiple modalities
- Various architectures, but **early fusion** currently dominates research works
- Topics not covered:
  - Multimodal Instruction Tuning (with instructions from gpt-4 for example)

A Word about Limitations

We are still far from AGI!!

- Lack of understanding on generalization to unseen tasks/modality
- Models are prone to **hallucinations** (objects, audio, speech)
- Models perpetuate **social/cultural biases** present in training data

TP
Multimodal Training

M **You**

Generate a fun meme about multimodal LLMs like yourself

DALL-E

Institut des sciences du Digital Management & Cognition

UNIVERSITÉ DE LORRAINE

Loria Laboratoire lorrain de recherche en informatique et ses applications

DEEZER

Multimodal Training

- **Goal:** classify pairs of (image, description) into labels
- **Data:** subset of WebVision dataset

```
Index: 394
Label: bookshop
Text: The highest forms of literature were to be found on the English bookstall. Perhaps these should be in the Self-Help section .
```

Multimodal Training

- **Step 1:** Classification with description only using BERT

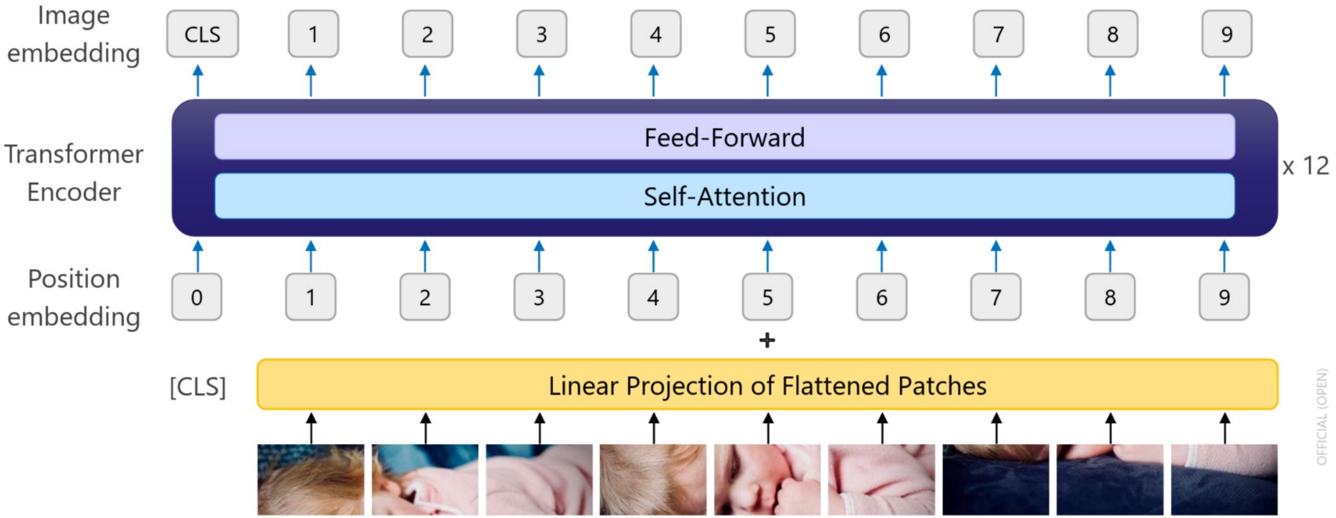
- **Step 2:** Classification with Dual-Encoder (independent) with BERT and ResNet50

- **Step 3:** Classification with Joint-Encoder using Align Before Fuse (ALBEF)

Li, Junnan, et al. "Align before fuse: Vision and language representation learning with momentum distillation." *Advances in neural information processing systems* 34 (2021): 9694-9705.

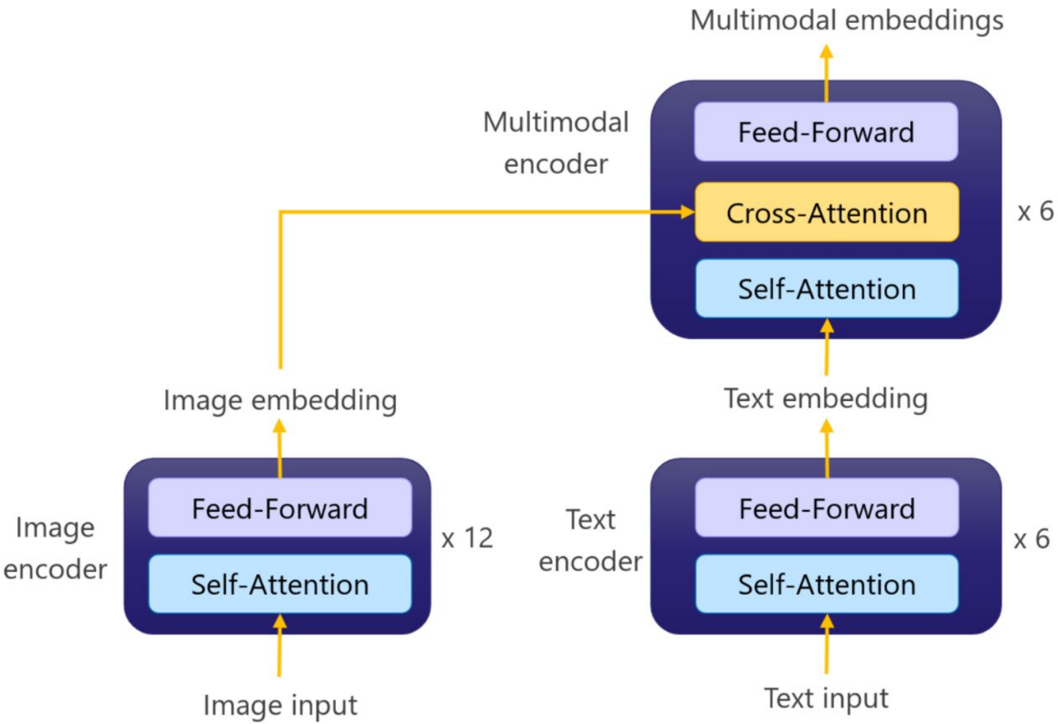
Align Before Fuse (ALBEF)

Image Encoder: Vision Transformer



"An image is worth 16x16 words: Transformers for image recognition at scale." Dosovitskiy, et al. *arXiv:2010.11929*, 2020.

Align Before Fuse (ALBEF)



(Recap) limitations of dual-encoder:

- Architecture only supports linear interactions between image and text
- Pretrained text and image encoders don't have knowledge about text-image pairs because they were pretrained separately on unimodal datasets

Multimodal Training - Organization

- We will use **Google Colab** to build & train models

- Material link:
  https://github.com/gasmichel/Multimodality_M2TAL