Multimodality

Andrey Kuznetsov
PhD, Head of FusionBrain Lab, AIRI

Term 5. Modern Methods and Algorithms of Generative Artificial Intelligence

Outline

- *How the journey started*
- *Fusion approaches*
- *Architecture design approaches*
- *New challenges*

01

How the journey started

Let's turn to the person

The information that a person perceives and which is necessary for making rational decisions is presented in various modalities

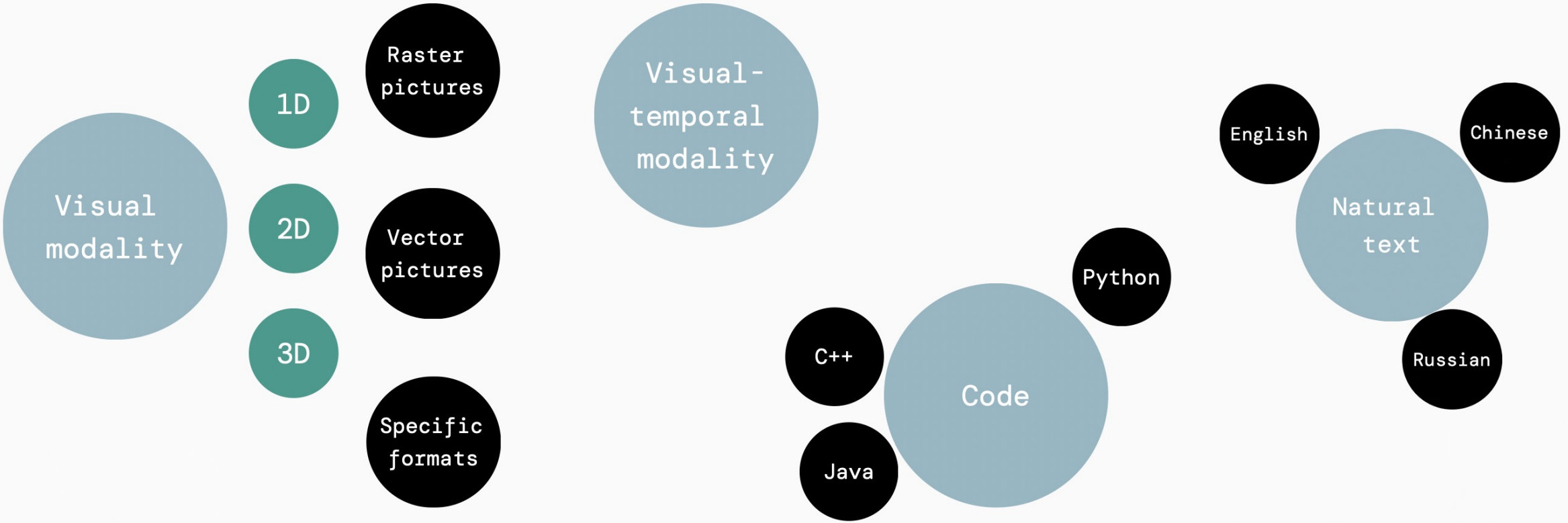
The ability to jointly solve various problems in various domains and modalities is a step towards Artificial General Intelligence

AIRI

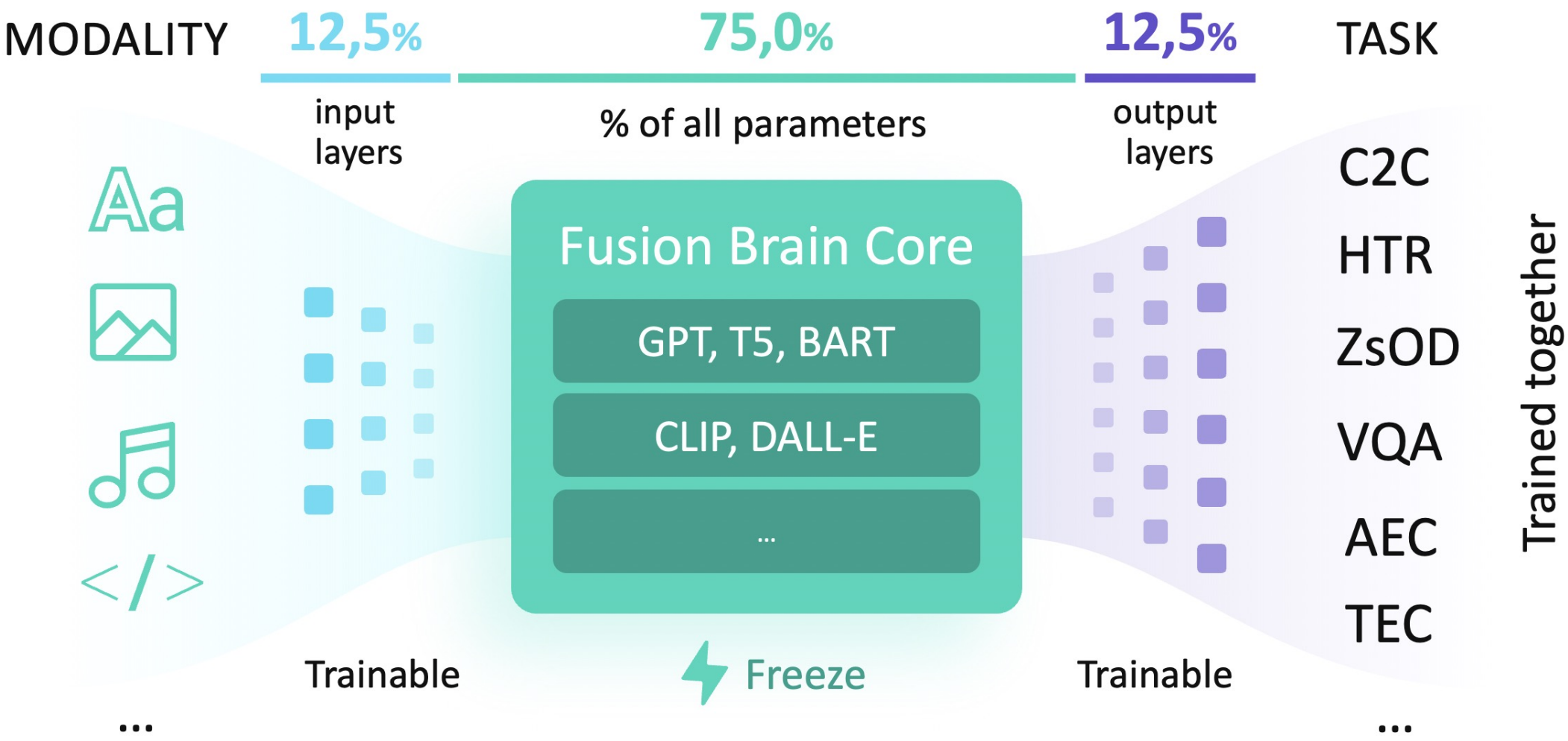
How can MODALITY be defined?

There is NO clear and generally accepted definition

One possible way to formalize the concept of modality is to go **from human perception** of data.
Here we come to hierarchical structure:



Motivation


MODALITY
12,5%
75,0%
12,5%
TASK
input layers
% of all parameters
output layers
Aa
Fusion Brain Core
GPT, T5, BART
CLIP, DALL-E
...
C2C
HTR
ZsOD
VQA
AEC
TEC
Trained together
Trainable
Freeze
Trainable
...
...

Motivation: WHY it is reasonable

Efficient multi-modal multi-task models

WHY we need multi-*

decoder setup	COCO det. mAP	VG det. mAP	VQAv2 accuracy
single-task training	40.6 / –	3.87	66.38 / –
shared (COCO init.)	**40.8** / 41.1	**4.53**	67.30 / 67.47

WHY we need efficiency

Model	#Params
GPT-3	**175 B**
Retrieval-based models	**1 B** (3*BERT-Large)

AIRI

Motivation: WHY it is still a challenge

Model	#params	GLUE	SuperGLUE
RoBERTa-Large ST	8,5B	88.2	76.5
RoBERTa-Large MTL	355M	86.0	78.6
CA-MTL (RoBERTa-Large)	397,6M	**89.4**	**80.0**

Encoder (BERT)-based
Multi-task: **better**

T5 (3B) STL	48B	88.5	86.4
HyperGrid (3B) MTL	3B	88.2	84.7
T5 (11B) STL	176B	**89.7**	**88.9**
HyperGrid (11B) MTL	11B	89.4	87.7

Decoder (T5, GPT)-based
Single-task: **better**

AIRI

A step to multimodality. CLIP

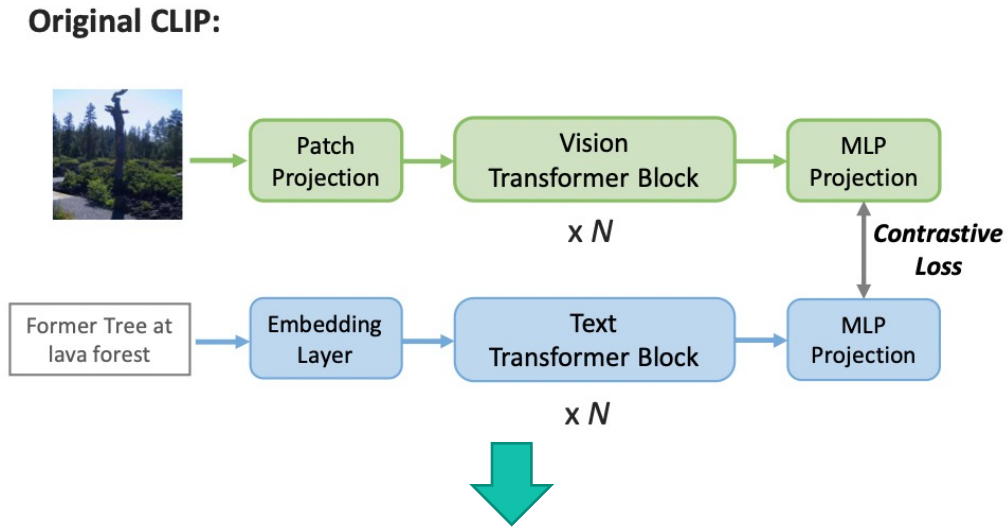
2021 trends:

- → Large pre-trained models (BERT, GPT-3)
- → Multi-modality and multi-tasking (**CLIP**, DALL-E, DALL-E 2, UniT)

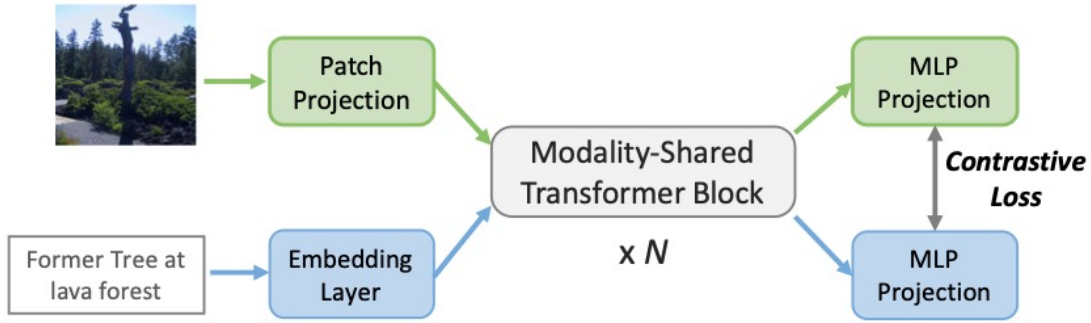
Target modalities: texts, images, sounds and other modalities like videos, programming languages, graphs and time series

Target tasks: NLP, CV and combined tasks like VQA

CLIP[1], 2021

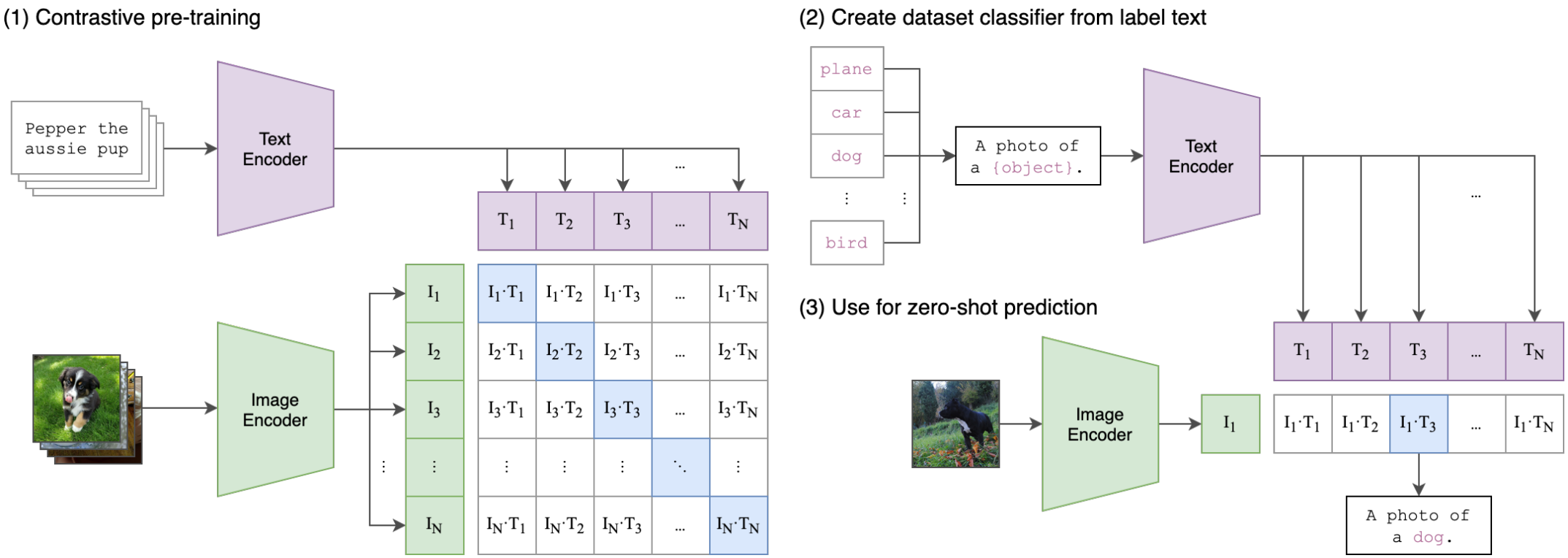


MA-CLIP[2], 2022, ICLR



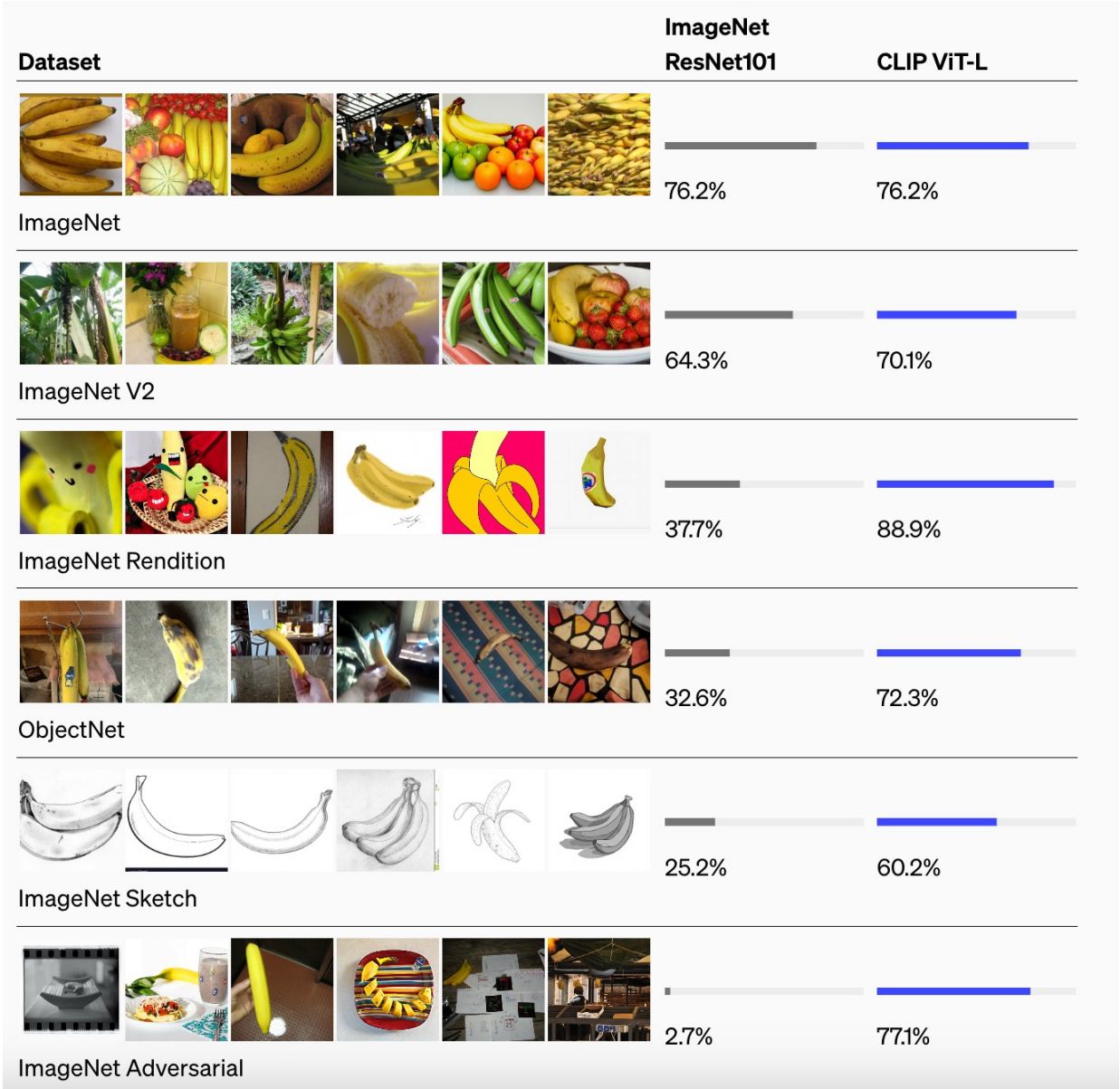
[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." 2021 *(OpenAI)*.
[2] You, Haoxuan, et al. "MA-CLIP: Towards Modality-Agnostic Contrastive Language-Image Pre-training." *2021 (Withdrawn submission to ICLR-2022)*

CLIP. Architecture



CLIP. Results

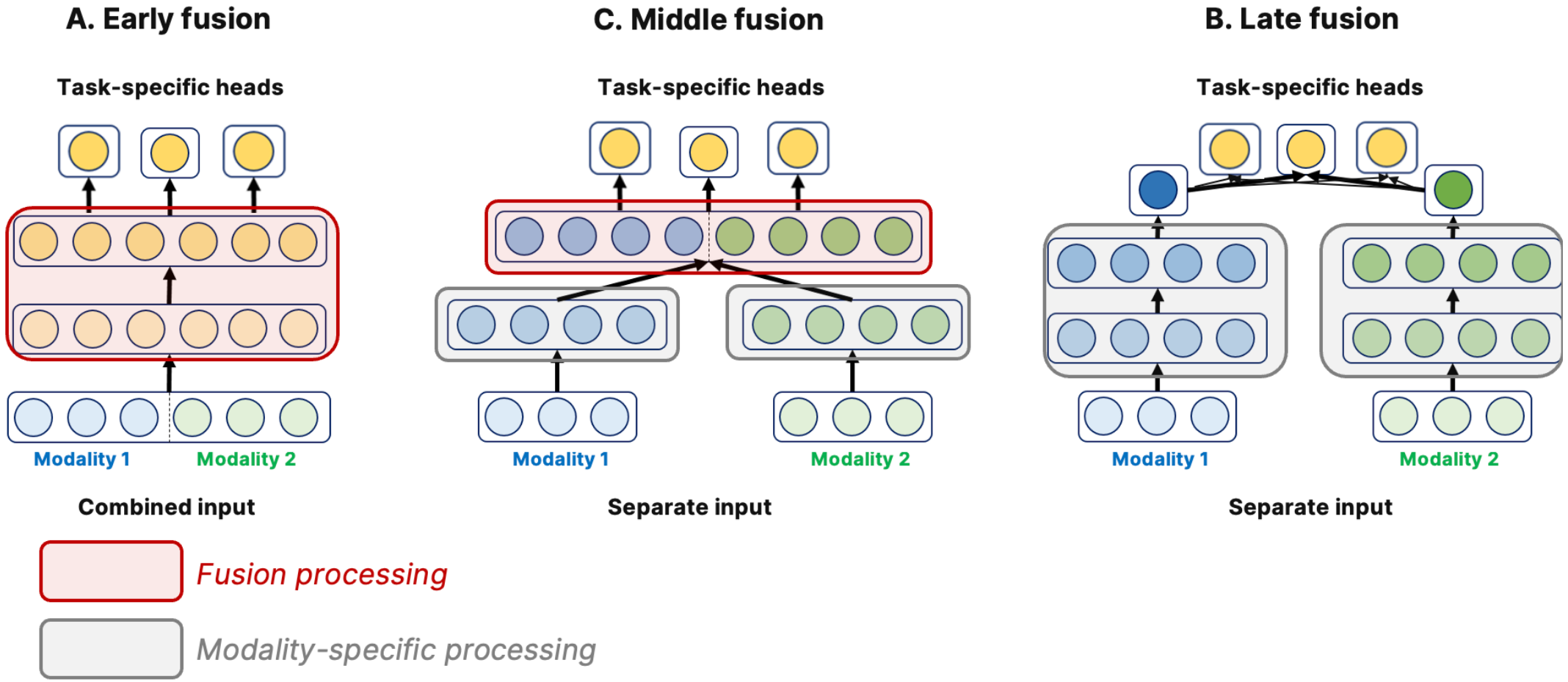
- Standard classification approaches show lower accuracy in zero-shot classification tasks on different benchmarks
- ImageNet pretrain makes the features not robust to new data with new poses, backgrounds, etc.
- Generalization ability is lower for classic approaches



02

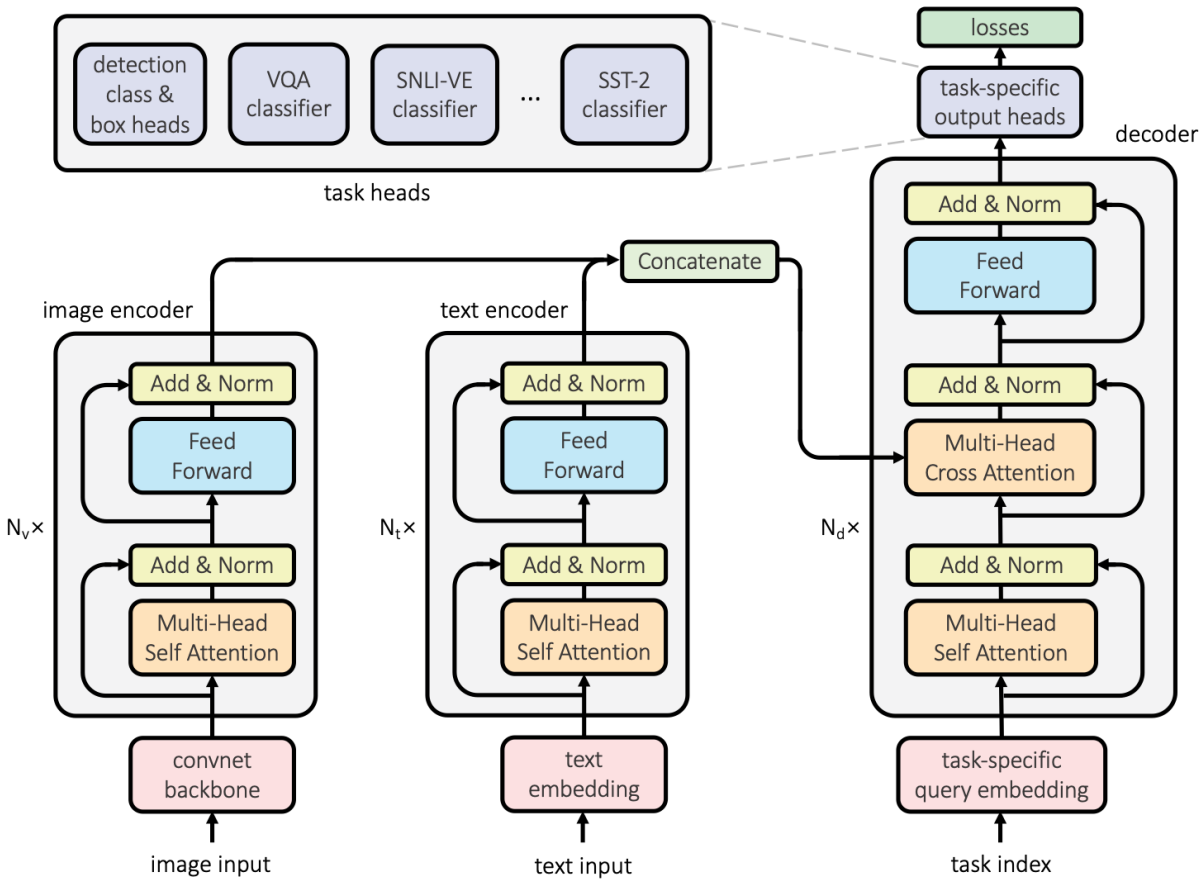
Fusion approaches

How can we fuse modalities?

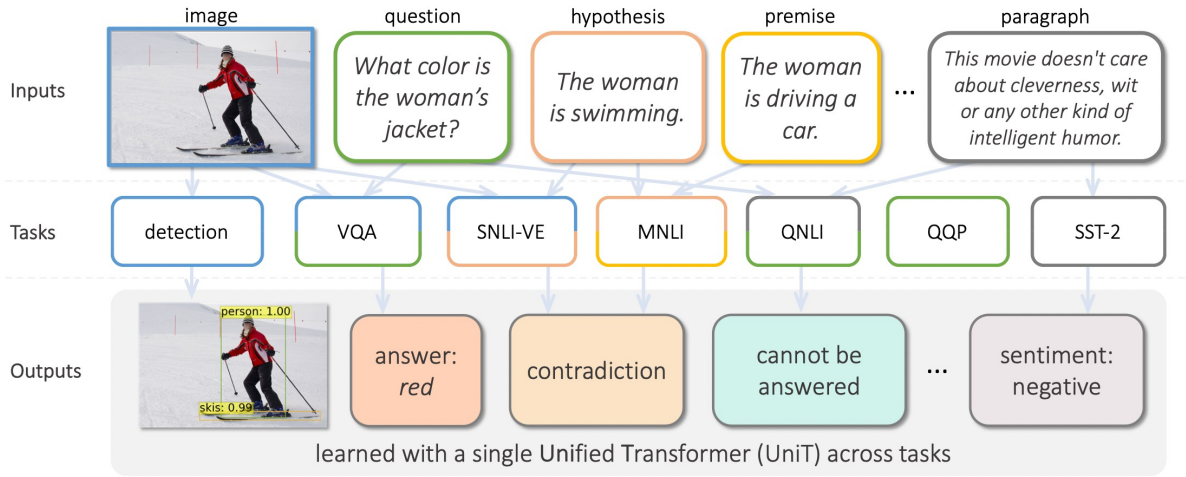


Multimodality. 2021

UniT[1]: via cross-attention



- Its own encoders for each model
- a common decoder, at the end
- task-specific heads (for all tasks, except for object detection, they are a two-layer perceptron)

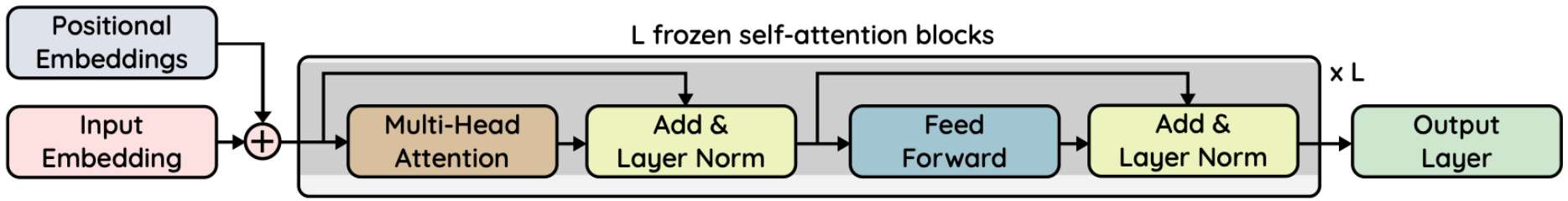


#	decoder setup	COCO det. mAP	VG det. mAP	VQAv2 accuracy	SNLI-VE accuracy	QNLI accuracy	MNLI-mm accuracy	QQP accuracy	SST-2 accuracy
1	UniT – single-task training	40.6	3.87	66.38 / –	70.52 / –	91.62 / –	84.23 / –	91.18 / –	91.63 / –
2	UniT – separate	32.2	2.54	67.38 / –	74.31 / –	87.68 / –	81.76 / –	90.44 / –	89.40 / –
3	UniT – shared	33.8	2.69	67.36 / –	74.14 / –	87.99 / –	81.40 / –	90.62 / –	89.40 / –
4	UniT – separate (COCO init.)	38.9	3.22	67.58 / –	74.20 / –	87.99 / –	81.33 / –	90.61 / –	89.17 / –
5	UniT – shared (COCO init.)	39.0	3.29	66.97 / 67.03	73.16 / 73.16	87.95 / 88.0	80.91 / 79.8	90.64 / 88.4	89.29 / 91.5
6	UniT – per-task finetuning	42.3	4.68	67.60 / –	72.56 / –	86.92 / –	81.53 / –	90.57 / –	88.06 / –
7	DETR [5]	43.3	4.02	–	–	–	–	–	–
8	VisualBERT [31]	–	–	67.36 / 67.37	75.69 / 75.09	–	–	–	–
9	BERT [14] (bert-base-uncased)	–	–	–	–	91.25 / 90.4	83.90 / 83.4	90.54 / 88.9	92.43 / 93.7

[1] Hu, Ronghang, and Amanpreet Singh. "UniT: Multimodal Multitask Learning with a Unified Transformer." 2021 (Facebook).

Multimodality. 2021

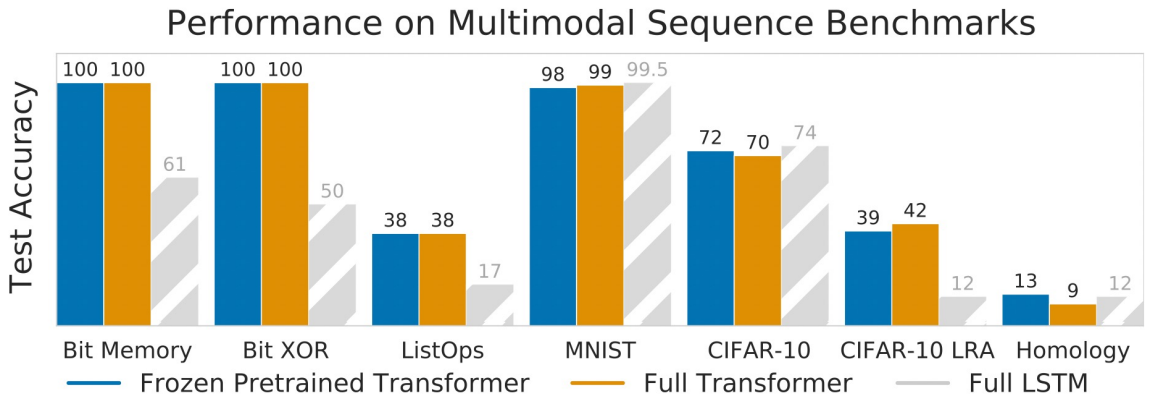
FPT[1]: via frozen MHA/FFN, tunable LN



Feedforward and self-attention layers are frozen

Task	output only	+ layernorm	+ input	+ positions
Bit Memory	76%	94%	100%	100%
Bit XOR	56%	98%	98%	100%
ListOps	15%	36%	36%	38%
MNIST	23%	96%	98%	98%
CIFAR-10	25%	54%	60%	68%
CIFAR-10 LRA	17%	39%	39%	39%
Homology	2%	9%	10%	13%

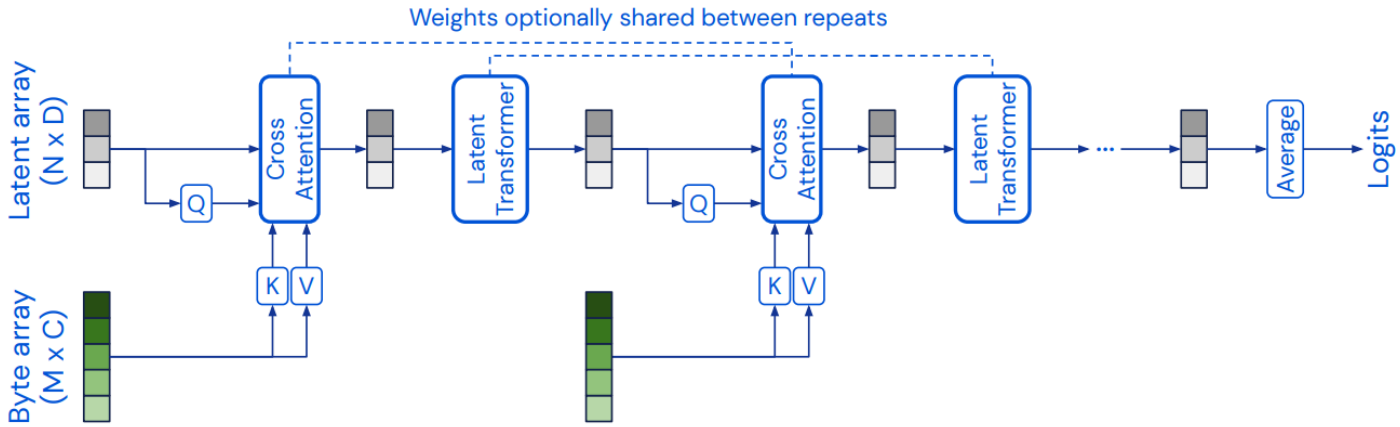
Ablation study



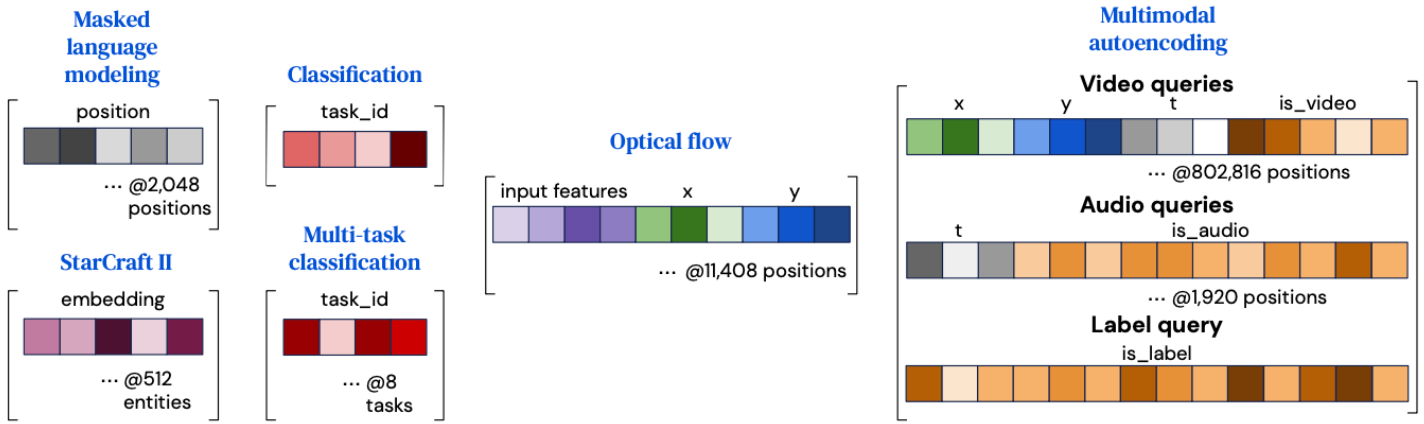
[1] Lu, Kevin, et al. "Pretrained transformers as universal computation engines." 2021 (Google)

Multimodality. 2021

Perceiver[1]: iterative CA



Perceiver IO: output queries



Main idea:

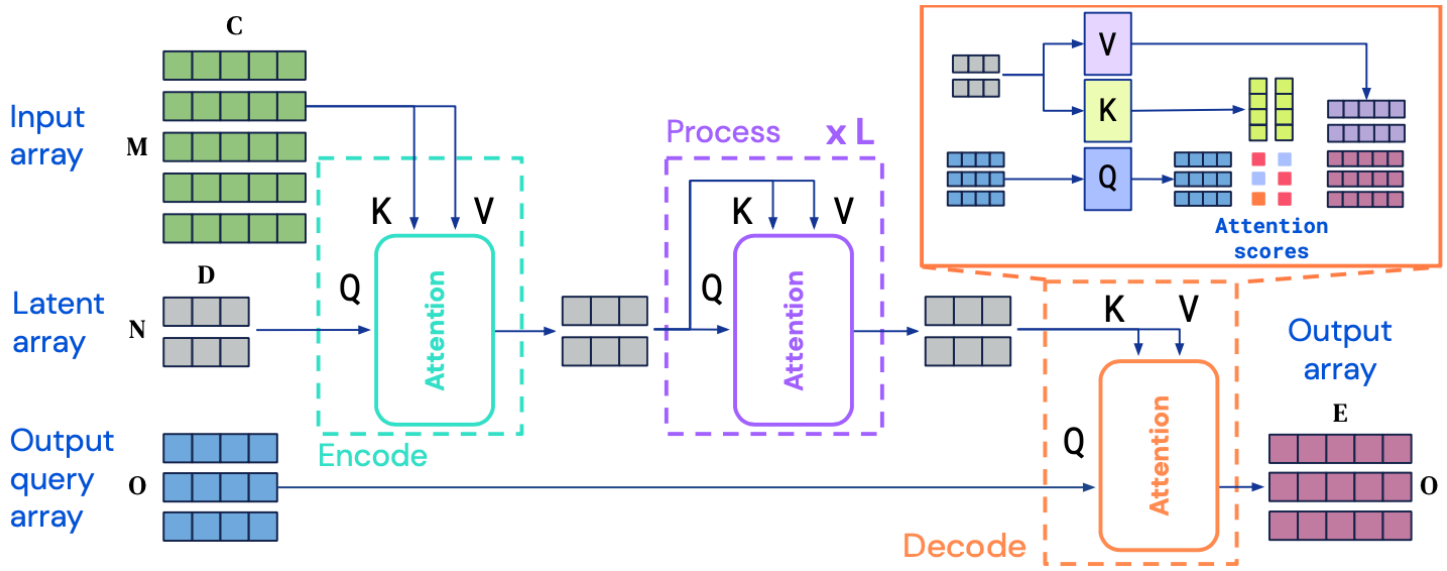
- **Iterative fusion** through **cross-attention** (query – latents, KV - input) allowing **linear** scaling on **input** size (not quadratic)
- Latent transformer is **GPT-2** like
- Weights of CA/SA are **shared**

[1] Jaegle, Andrew, et al. "Perceiver: General perception with iterative attention." 2021 (*DeepMind*)
[2] Jaegle, Andrew, et al. "Perceiver io: A general architecture for structured inputs & outputs." 2021 (*DeepMind*)

AIRI

Multimodality. 2021

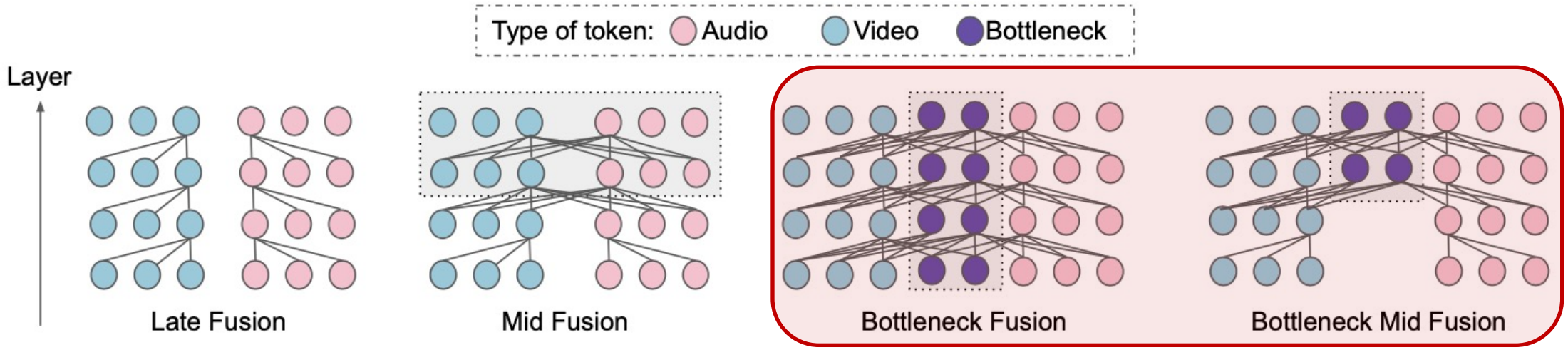
Perceiver IO²: CA on input/output



- **Perceiver IO²** added ability to work with **multitask and different output sizes** via CA where query is output structure, KV – latents (**complexity** – still the **linear** depending on the **output** size)

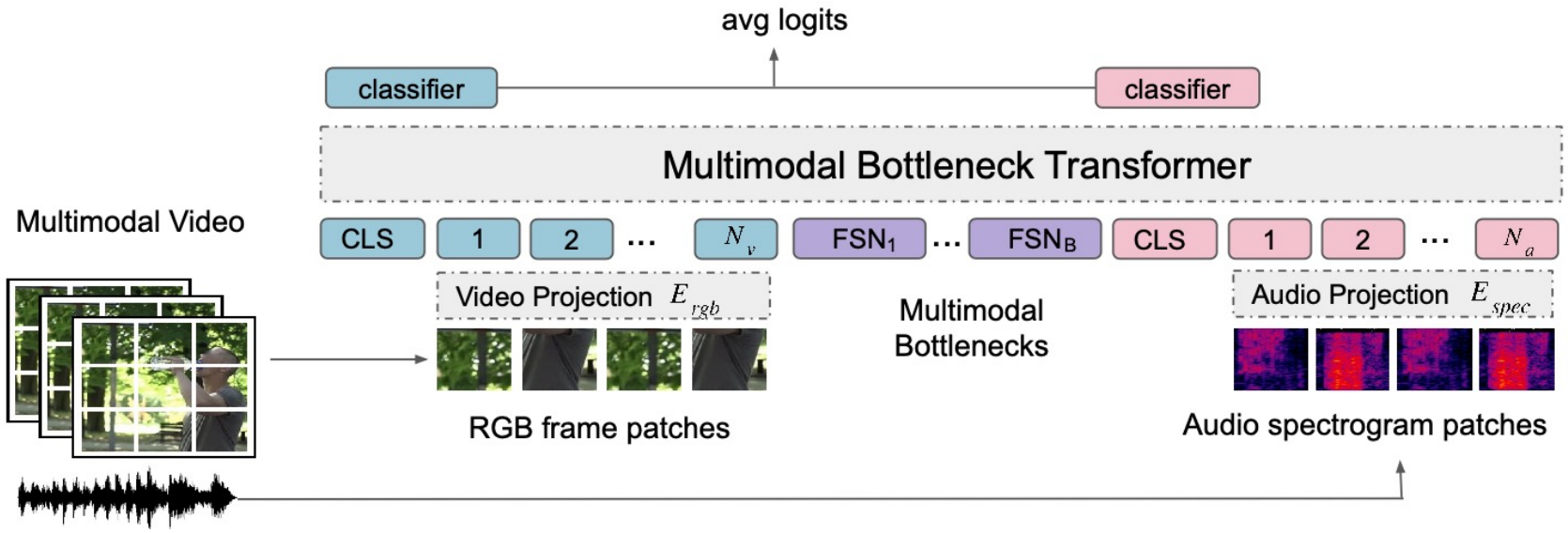
[1] Jaegle, Andrew, et al. "Perceiver: General perception with iterative attention." 2021 (*DeepMind*)
[2] Jaegle, Andrew, et al. "Perceiver io: A general architecture for structured inputs & outputs." 2021 (*DeepMind*)

Multimodality. Through information bottleneck



[1] Nagrani, Arsha, et al. "Attention Bottlenecks for Multimodal Fusion." 2021 (*Google*)

Multimodality. Through information bottleneck



MBT[1] scheme

Main idea:

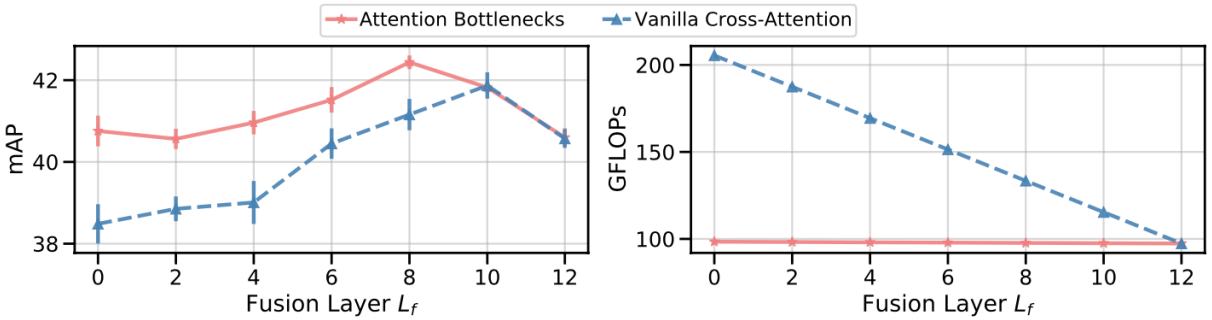
- **Middle-fusion through a small bottleneck** (B = 4 is used)
- **Fusion** is needed **closely to the top**

[1] Nagrani, Arsha, et al. "Attention Bottlenecks for Multimodal Fusion." 2021 (*Google*)
[2] https://www.robots.ox.ac.uk/~vgg/data/vggsound/

Multimodality. Through information bottleneck

VGGSound[2]

Model	Modalities	Top-1 Acc	Top-5 Acc
Chen et al‡ [11]	A	48.8	76.5
AudioSlowFast‡ [34]	A	50.1	77.9
MBT	A	52.3	78.1
MBT	V	51.2	72.6
MBT	A,V	**64.1**	**85.6**



[1] Nagrani, Arsha, et al. "Attention Bottlenecks for Multimodal Fusion." 2021 (*Google*)
[2] https://www.robots.ox.ac.uk/~vgg/data/vggsound/

AIRI

03

Architecture design approaches

Tool-augmented LLM

- → LLM — as **orchestrator**
- → LLM can call specialized pre-trained models**:**
  - → for each modality
  - → for each task

⊕ *Can choose* **SOTA-model for each task**

⊖ **A strong lack of interaction between modalities**

ToolFormer, APIBank, OpenAGI, GigaChat

End-to-end multimodal LLM

- → Modality fusion inside **LLM**
- → **End-to-end** cross-modal training

⊕ **Strong interaction between modalities**

⊖ **High cost and complexity (computational as well) of training**

Large World Model, Flamingo, 4M, JEPA, V-JEPA, RUDOLPH

Modality bridging with pretrained models

- → Using of pretrained **LLM and specialized encoders for each modality**
- → **Training a mapping** from each modality feature space → for LLM's embedding space

⊕ *Optimizing small number of parameters* → **low computational complexity**

⊖ **Reduced inter-modality interation**

FROMAGE, LLaVA, MoE-LLAVA, LanguageBind, OmniFusion

AIRI

LLM — orchestrator

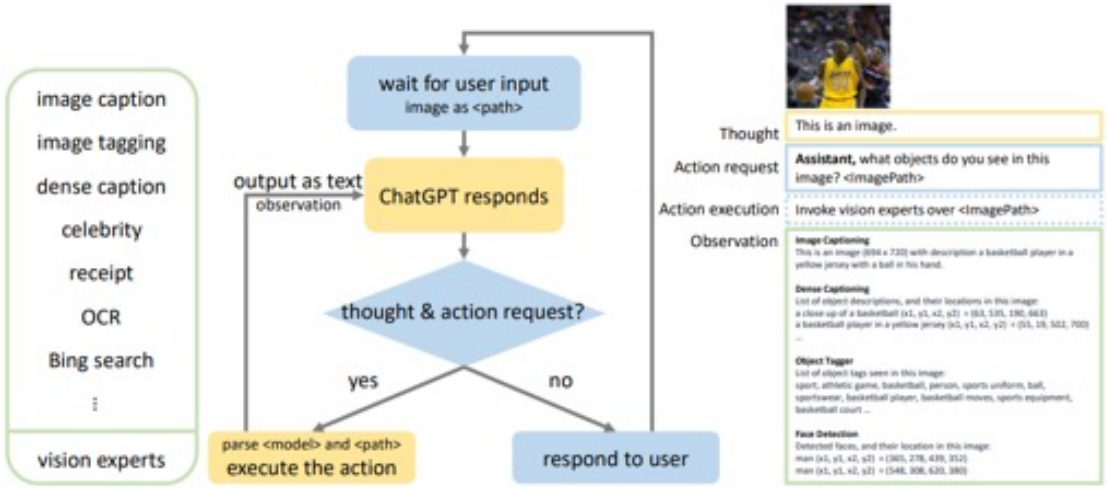
- The most simple way of interaction between LLM and other systems/agents/models/...
- The task is to train the model to properly “prompt” on the input and “parse” the response
- Complex task for an LLM to select experts needed when the number of external models and APIs extends

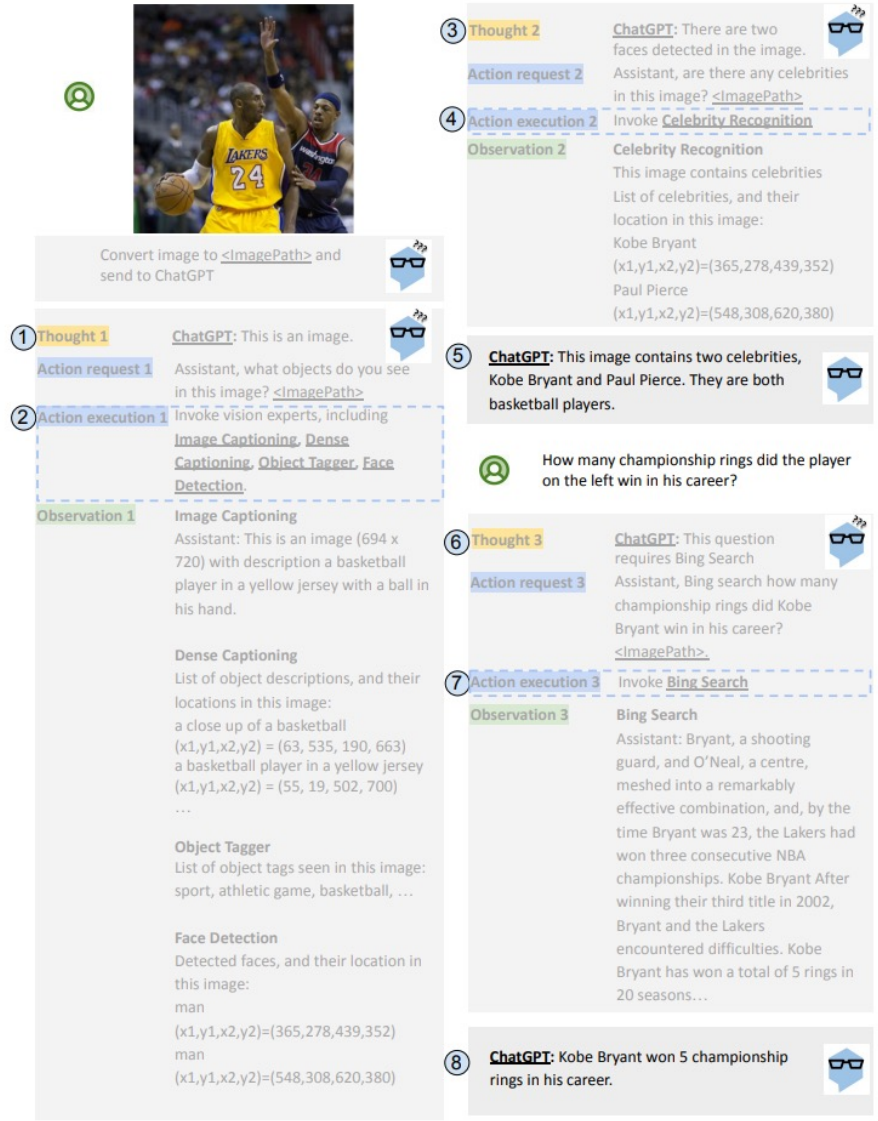
Tasks solved by competitors (MM-REACT)

- *Visual Math & Text reasoning*
- *Spatial/coordinate understanding*
- *Visual planning & prediction*
- *Multi-image reasoning*
- *Multi-Hop Doc Understanding*
- *Video Summarization*
- ...

https://arxiv.org/abs/2303.11381

LLM — orchestrator





Tool-augmented LLM

- LLM — as **orchestrator**
- LLM can call specialized pre-trained models**:**
  → for each modality
  → for each task

⊕ *Can choose* **SOTA-model for each task**

⊖ **A strong lack of interaction between modalities**

ToolFormer, APIBank, OpenAGI, GigaChat

End-to-end multimodal LLM

- Modality fusion inside **LLM**
- **End-to-end** cross-modal training

⊕ **Strong interaction between modalities**

⊖ **High cost and complexity** (computational as well) **of training**

Large World Model, Flamingo, 4M, JEPA, V-JEPA, RUDOLPH

Modality bridging with pretrained models

- Using of pretrained **LLM and specialized encoders for each modality**
- **Training a mapping** from each modality feature space → for LLM's embedding space

⊕ *Optimizing small number of parameters* → **low computational complexity**

⊖ **Reduced inter-modality interation**

FROMAGE, LLaVA, MoE-LLAVA, LanguageBind, OmniFusion

AIRI

Flamingo (2022)

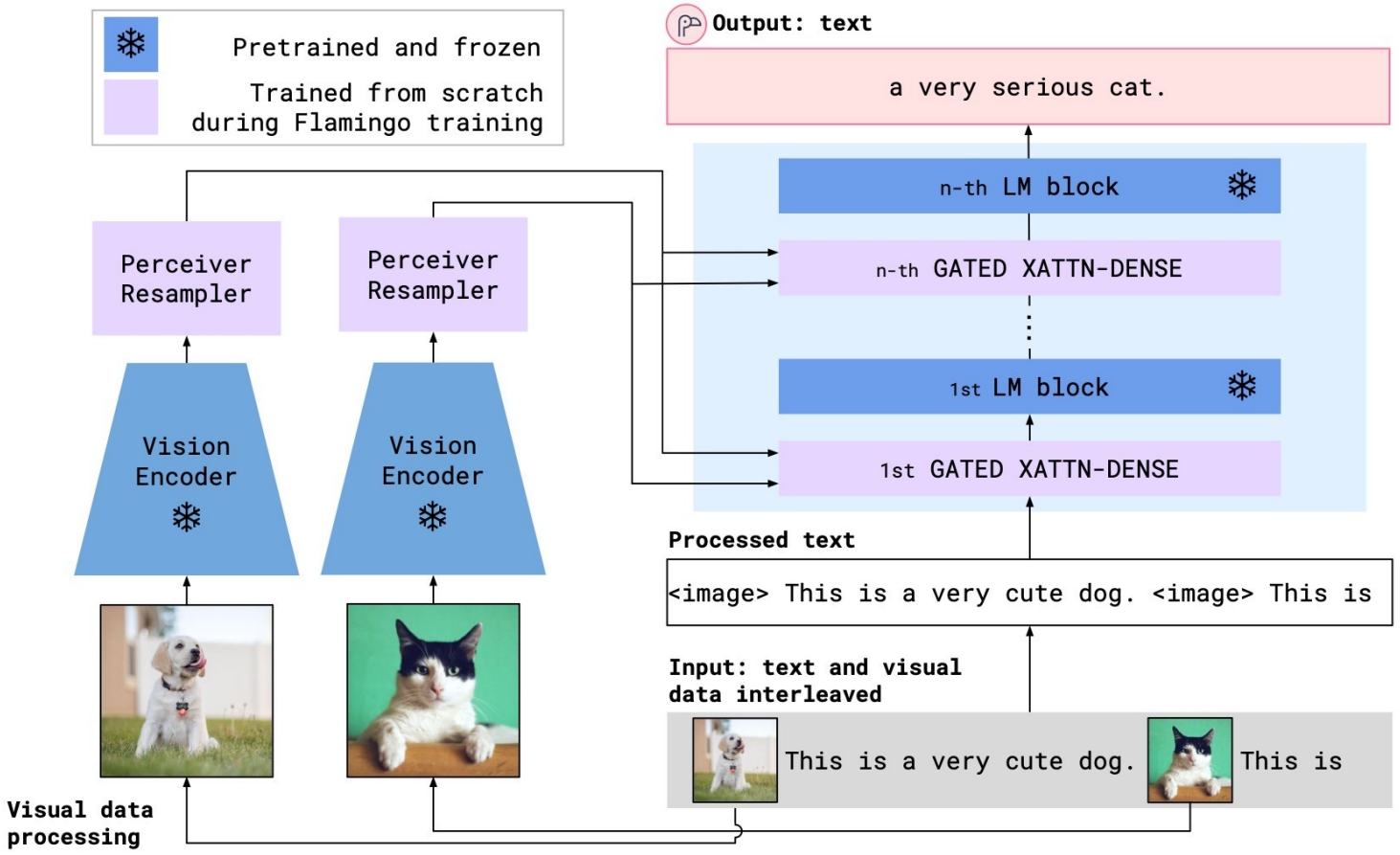
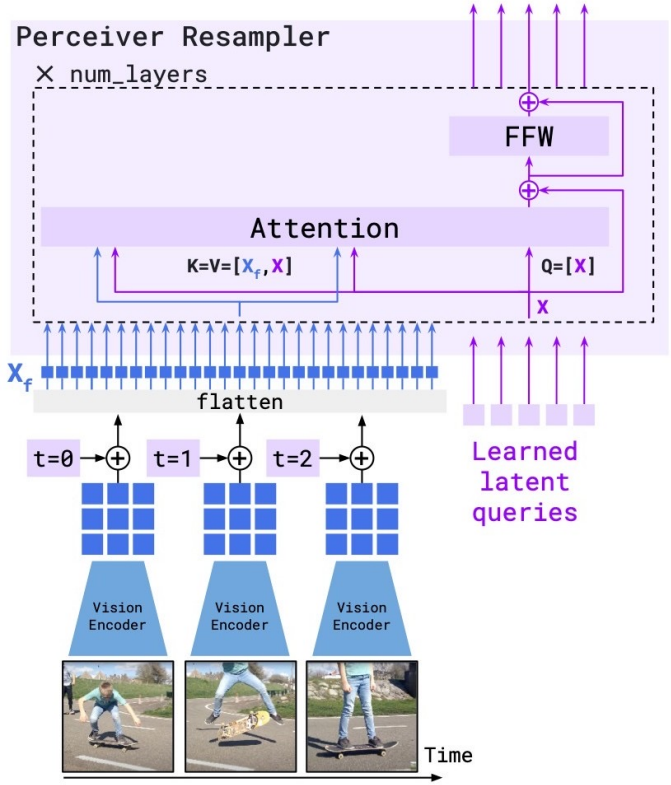


Figure 3 | **Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35, 23716-23736.

Flamingo (2022)




```python
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

Figure 4 | **The Perceiver Resampler** module maps a *variable size* grid of spatio-temporal visual features coming out of the Vision Encoder to a *fixed* number of output tokens (five in the figure), independently of the input image resolution or the number of input video frames. This transformer has a set of learned latent vectors as queries, and the keys and values are a concatenation of the spatio-temporal visual features with the learned latent vectors. More details can be found in Section 3.1.1.

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35, 23716-23736.

RUDOLPH (2022)

RUssian Decoder On Language Picture Hyper-tasking

- image generation from text description (ruDALL-E mode)
- text description generation based on image (image captioning task)
- image ranking according to text description (ruCLIP mode)
- text generation (ruGPT-3 mode)
- image quality and resolution enhancement (super resolution task)
- question answering on images (VQA task)
- ...

3 versions:

350M **1.3B** **2.7B**

Solves tasks based on **two modalities** (test and video)

Кодировщик изображения: Enc **Sber VQGAN**

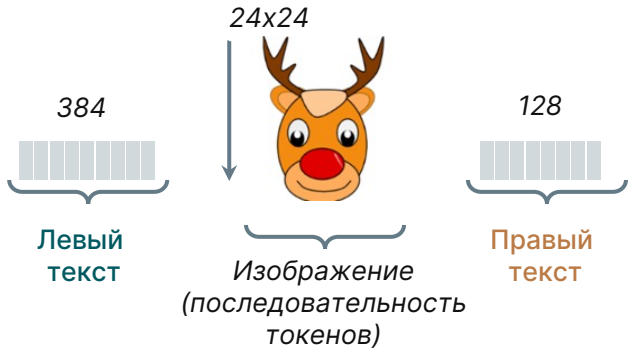
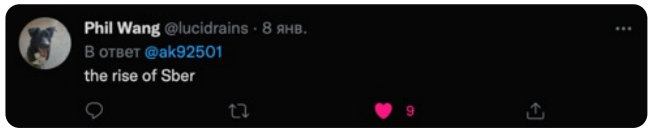


Image decoder: Dec **Sber VQGAN / Image Diffusion**



AIRI

Model training

Autoregressive training.

Predict next tokens based on the previous context (masked):

- **t2i**: image generation based on textual description
- **i2t**: caption generation based on the image
- **t2t**: language modeling (only in the left tokens)

Training data:

- **i2t/t2i**: 119M text-image pairs
- **t2t:** 60M text pieces

AIRI

Special tokens for task understanding

Pre-training

<LT_UNK>

<RT_UNK>

<LT_T2I> - *text-to-image*

<LT_I2T> - *image-to-text*

<LT_T2T> - *text-to-text*

<RT_I2T> - *image-to-text*

Fine-tuning (75/25)

<LT_TQA>

<RT_TQA>

<LT_MQA>

<RT_MQA>

<LT_VQA>

<RT_VQA>

<LT_CAP>

<RT_CAP>

<LT_GEN>

Inference

Standard approach:

<LT_UNK>

<RT_UNK>

——————————————————

2 approach:

input: text + image

<LT_T2I>

<RT_I2T>

input: text

<LT_UNK>

<RT_T2T>

Special tokens for task understanding

During inference the model has no a-priori information about the task.

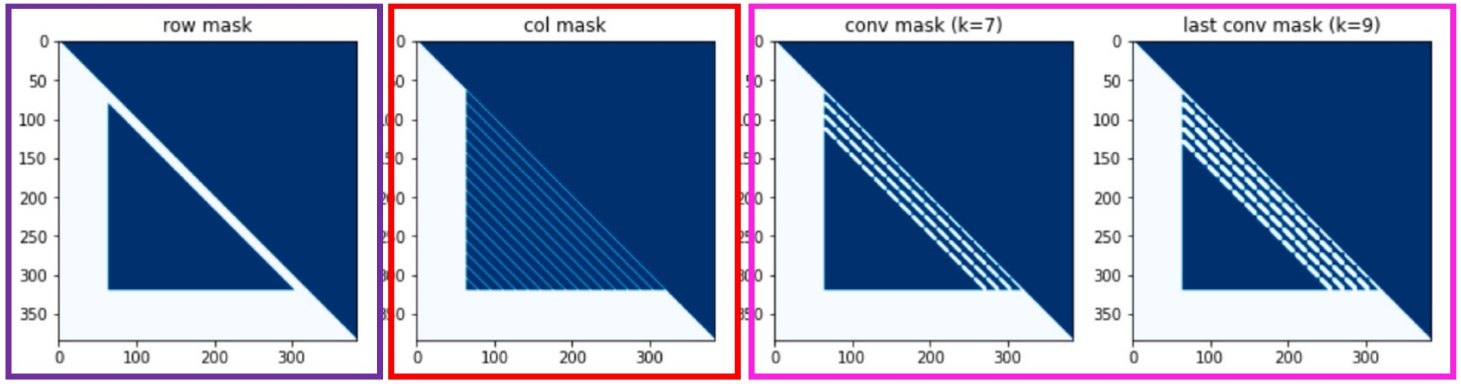
Add *unknown* tokens for each input sample:

- `<LT_UNK>`
- `<RT_UNK>`

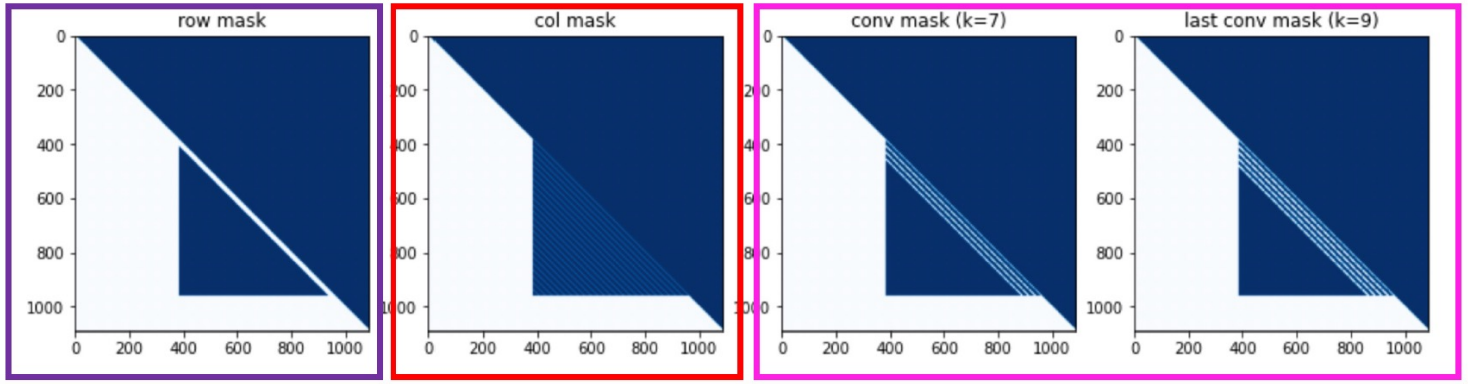
AIRI

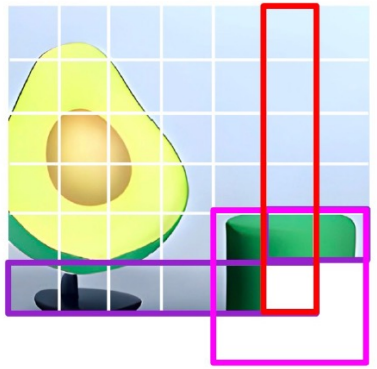
Attention masks

350M:



2.7B:





Model finetuning

Text QA

Answer the question based on text: ‘The oldest higher education institution of the city - Ryazan State University was opened in 1915 as a women's teacher's institute and was named after S A Esenin in 1985. Question: When was Ryazan State University opened?’.

In 1915.

Math QA

Solve linear equation for y: 6 multiply on y - 54 это -72. Options: A. -4; B. -20; C. -5; D. -3

D. -3

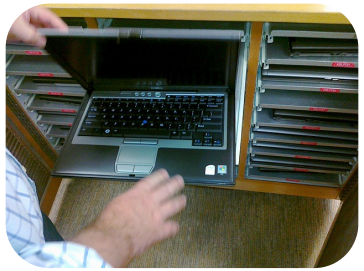
Text Recognition

Detect text on image

Beauty salon (студия красоты)

Visual QA

Answer the question based on image: Is the computer switched on?



No

Image Captioning

Describe the image

A cat sitting in the car.

Image Generation

Generate an image. Silver balls lying on the ground around a smaller red ball

AIRI

Tasks examples

Text QA

Left text

Answer the question. Question: Who is Cheburashka?

Right text

Answer: Cheburashka is one of the most famous characters from Soviet television, a small plush toy you can see in the shop.

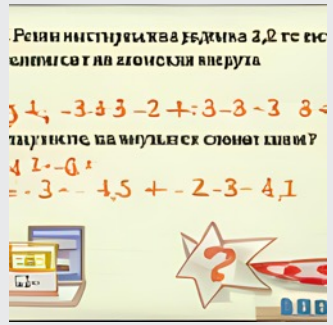
Math QA

Left text

Solve the maths problem. If y is multiplied by 2, you get 16. What is equal to y? Options: a) 1; b) 2; c) 8; d) -8

Right text

Answer: c 8



Text2Text

Logical QA

Left text. Solve the logic problem. Mary's dad has 5 daughters: Chichi, Chacha, Chuchu and Cheche. What is the name of the fifth daughter?

Right text. Answer: Mary

Math-Logical QA

Left text. Choose the correct answer to the question from the given answer choices.

Lucy and David competed against each other in a game of Ring the Bell. David scored 120 points and Lucy and David together scored 235 points. How many points did Lucy score?

Options: a) 100; b) 117; c) 115; d) 105

Right text. Answer: 115

AIRI

Text2Image (+ bonus: text completion)

Image Generation

Left text. Draw the image according to the text. Landscape: coniferous forest in the sunlight

Right text. and snow

Left text. Generate an image based on the description. Field with dandelions in Van style

Right text. oil

Image2Text

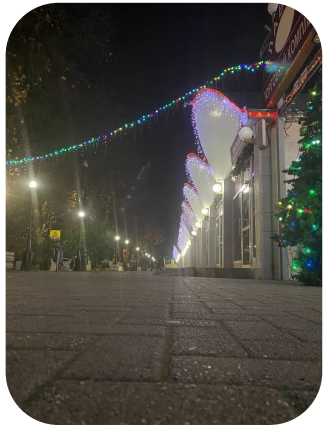
Image Captioning

Left text. Describe the image

Right text. Answer: the picture shows two penguins on a cliff

Visual QA

Left text. Analyse the picture and answer the question. What time of day is depicted?



Right text. Answer: night time

Text Recognition in the Wild

Left text. Detect text on image

Right text. Answer: то самое тесто

AIRI

Tool-augmented LLM

- → LLM — as **orchestrator**
- → LLM can call specialized pre-trained models**:**
  - → for each modality
  - → for each task

⊕ *Can choose* **SOTA-model for each task**

⊖ **A strong lack of interaction between modalities**

ToolFormer, APIBank, OpenAGI, GigaChat

End-to-end multimodal LLM

- → Modality fusion inside **LLM**
- → **End-to-end** cross-modal training

⊕ **Strong interaction between modalities**

⊖ **High cost and complexity** (computational as well) **of training**

Large World Model, Flamingo, 4M, JEPA, V-JEPA, RUDOLPH

Modality bridging with pretrained models

- → Using of pretrained **LLM and specialized encoders for each modality**
- → **Training a mapping** from each modality feature space → for LLM's embedding space

⊕ *Optimizing small number of parameters* → **low computational complexity**

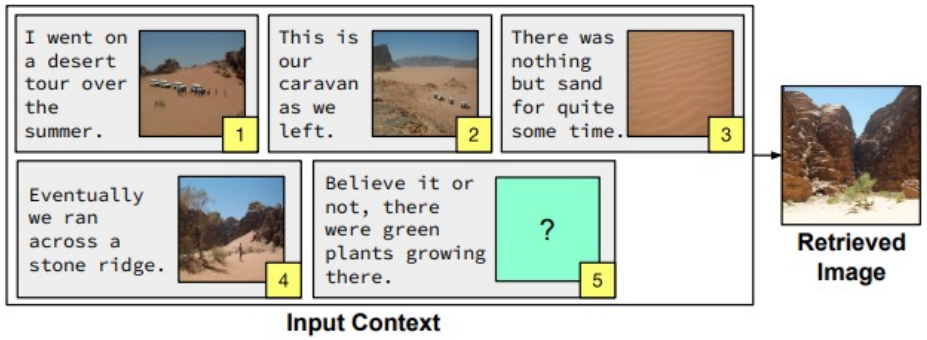
⊖ **Reduced inter-modality interation**

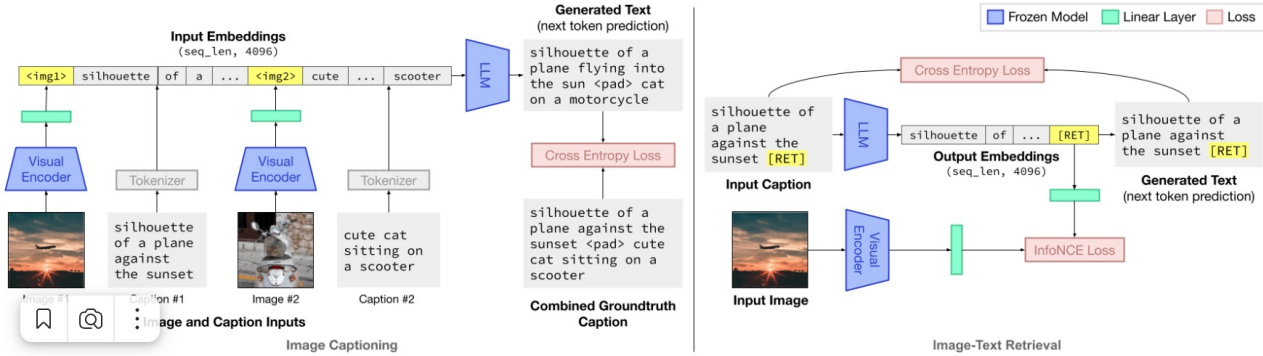
FROMAGE, LLaVA, MoE-LLAVA, LanguageBind, OmniFusion

AIRI

FROMAGE

- Frozen Retrieval Over Multimodal Data for Autoregressive Generation
- Frozen LLM and visual encoder are used
- 2 tasks on pretrain — image captioning and image-text retrieval
- visual embeddings to text space linear mapping
- [RET] embedding linear mapping through the text space to the groundtruth image embedding (contrastive approach)





Grounding Language Models to Images for Multimodal Inputs and Outputs

Model	Trainable Params	Finetuning Data	IT2T					T2I		
			NDCG	MRR	R@1	R@5	R@10	R@1	R@5	R@10
ViLBERT (Lu et al., 2019)	114M	3.1M	11.6	6.9	2.6	7.2	11.3	-	-	-
CLIP ViT-L/14 (Radford et al., 2021)	300M	400M	10.9	8.5	3.1	8.7	15.9	17.7	38.9	50.2
Flamingo (Alayrac et al., 2022)	10.2B	1.8B	**52.0**	-	-	-	-	Incapable		
ESPER (Yu et al., 2022b)	4M	0.5M	22.3	**25.7**	14.6	-	-	Incapable		
FROMAGe (ours)	5.5M	3.1M	16.5	22.0	**17.6**	**20.1**	**25.1**	**20.8**	**44.9**	**56.0**

- Language model — OPT-6.7B
- Visual encoder— CLIP ViT-L/14
- Training — 18k steps with bs = 180, 1 day on 1×A6000

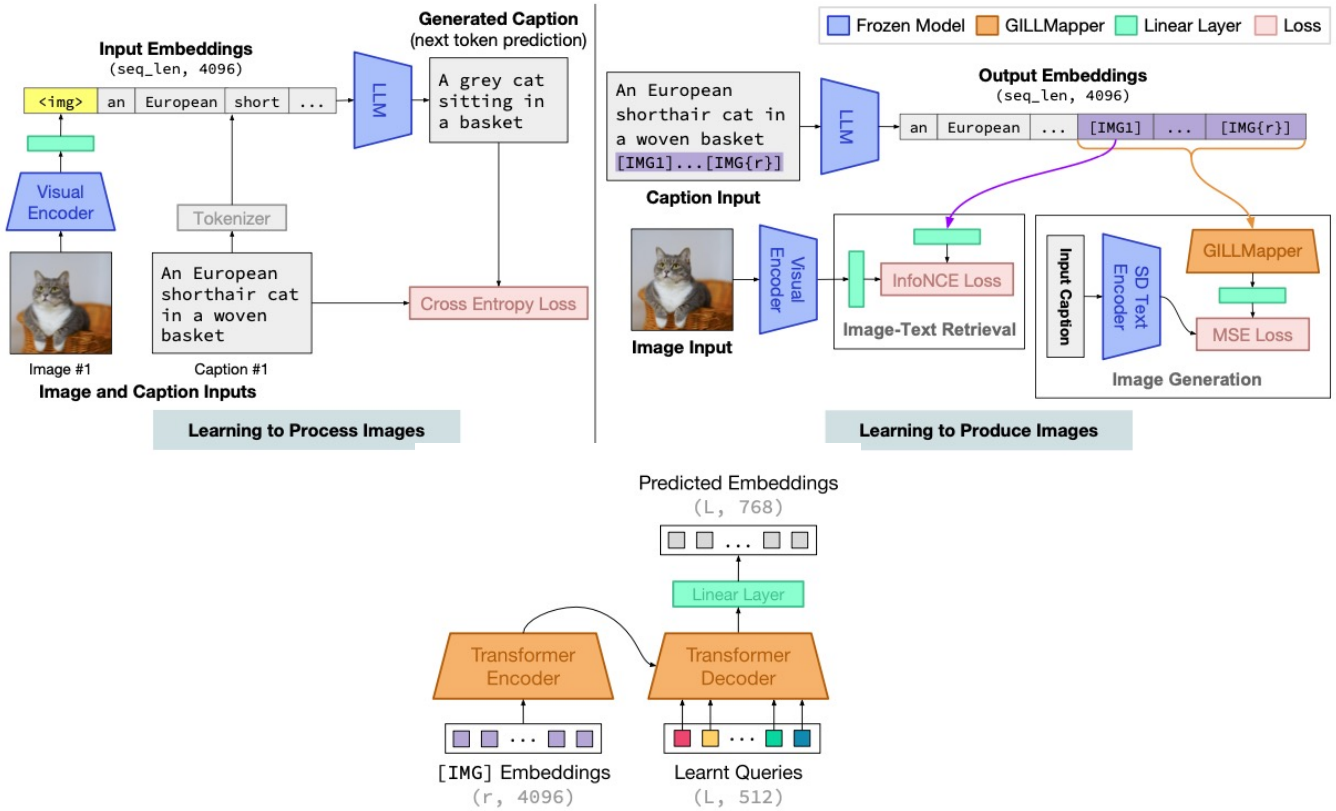
https://arxiv.org/abs/2301.13823

GILL

- Generating Images with Large Language Models
- Generator — Stable Diffusion 1.5

Pretrain on 4 tasks:
1) Image understanding
2) Special token [IMG] generation
3) Image generation — GILLMapper, encoder-decoder transformer with 4 layers
4) Image retrieval

Train 3 linear mappings, [IMG] embedding matrix and GILLMapper



GILLMapper architecture

- Language model — OPT-6.7B
- Visual encoder— CLIP ViT-L/14
- Image generator — SD 1.5
- Trainable parameters— 50M
- Training — 20k steps with bs = 200, 2 days on 2×A6000

https://arxiv.org/abs/2305.17216

OmniFusion

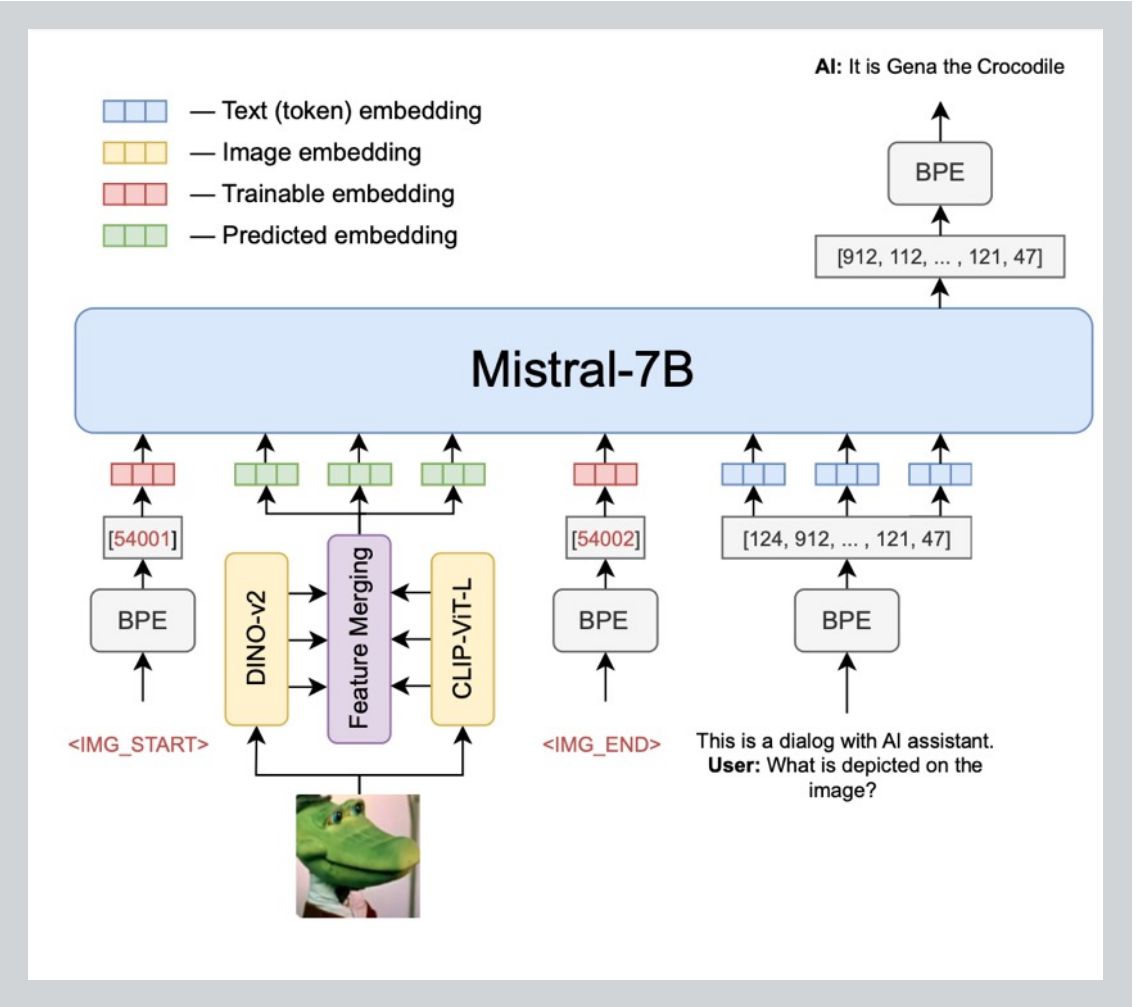
Omnis (lat.) — comprehensive **Fusion** (en.) — amalgamation

Key features

- *The first multimodal conversation model in Russia*
- *Deals with two modalities: text, images*
- *Two-stage learning process*
  - *Pretrain on large dataset of pairs «text + image»*
  - *Finetune on downstream*

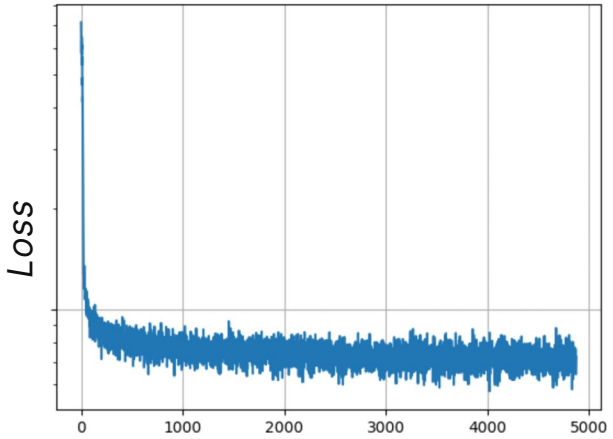
LLM: GigaChat, Mistral

Visual encoder: CLIP-ViT-Large, InternViT, encoder fusion

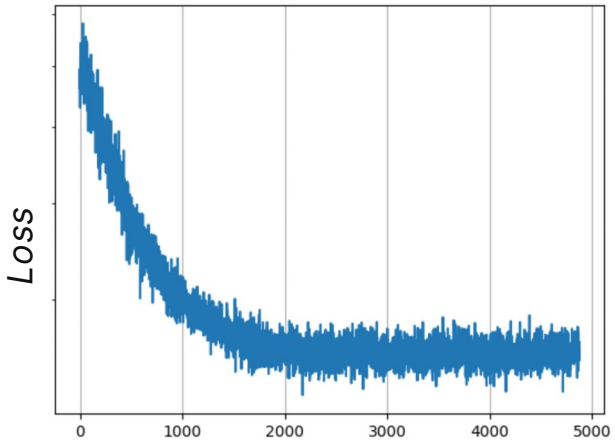


OmniFusion training

1st stage (pretrain, high learning rate value)



Second stage (SFT, low learning rate value)



Datasets

Image captioning (pretrain)

EN

- *CC3M filtered*
- *COCO Captions*
- *LLaVAR-Pretrain*
- ...

~2M

Images in MM dialogue (SFT)

RU

- *capcha_dialogs*
- *qa_val_llava_rlhf_syn*
- *celeb_hq_qa*
- *coco_qa_chatt*
- ...

~200k

EN

- *VQAv2*
- *GQA*
- *OKVQA*
- *LLaVAR-Instruct*
- ...

~700k

Overall training time: ~ 48 hours on 8 A100 GPUs

AIRI

Benchmark datasets

- **Viz-wiz -** real questions composed by people with low vision (Q: photo description)
- **MM-Vet -** complex multimodal dataset (Q: open-ended, GPT-4 used to evaluate responses)
- **POPE -** hallucination detection in multimodal dialogues
- **MMBench** - various dialogue sets using images (GPT-4 is used to evaluate the correctness of responses)

Q: What will the girl on the right write on the board?
GT: 14

Required capabilities:
Recognition
Spatial awareness
OCR
Math

Benchmark datasets

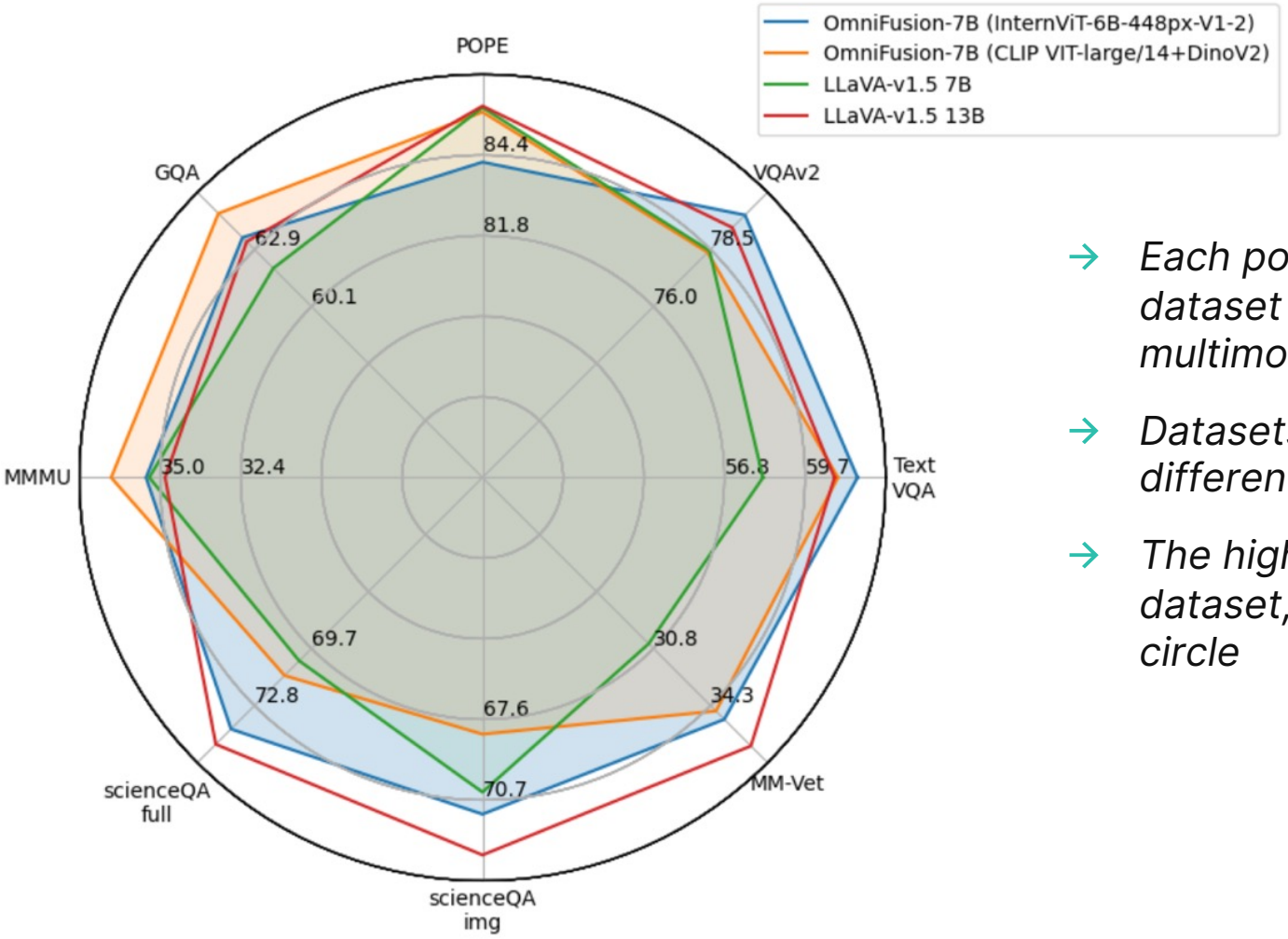
- **Llava-bench** - *small benchmark from the LLaVA paper (synthetic dialogues based on image descriptions, created with GPT4)*
- **Text-VQA** - *OCR benchmark (Q: on text in pictures)*
- **ScienceQA** - *various questions on lecture material in different domains (Q: with answer choice)*
- **VQA v.2** - *open-ended questions about pictures (Q: requires deep understanding of the picture, logical reasoning ability of the model)*

Question: Can you explain this meme in detail?

Answer: This meme features an image of an oven tray with fried chicken nuggets arranged in the shape of the continents on a world map, resembling a bird's-eye view of Earth. Above the image, there is a caption that reads "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is." The humor in this meme comes from the juxtaposition of the earnest caption with the amusing representation of Earth using chicken nuggets.

Llava-bench

Experimental results

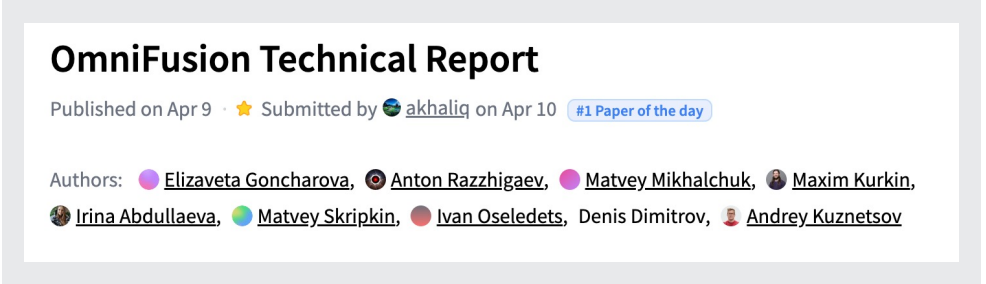


- *Each point on the circle is a benchmark dataset to assess the quality of the multimodal model*
- *Datasets are a set of image questions from different domains*
- *The higher the quality of the model on the dataset, the closer the point is to the outer circle*

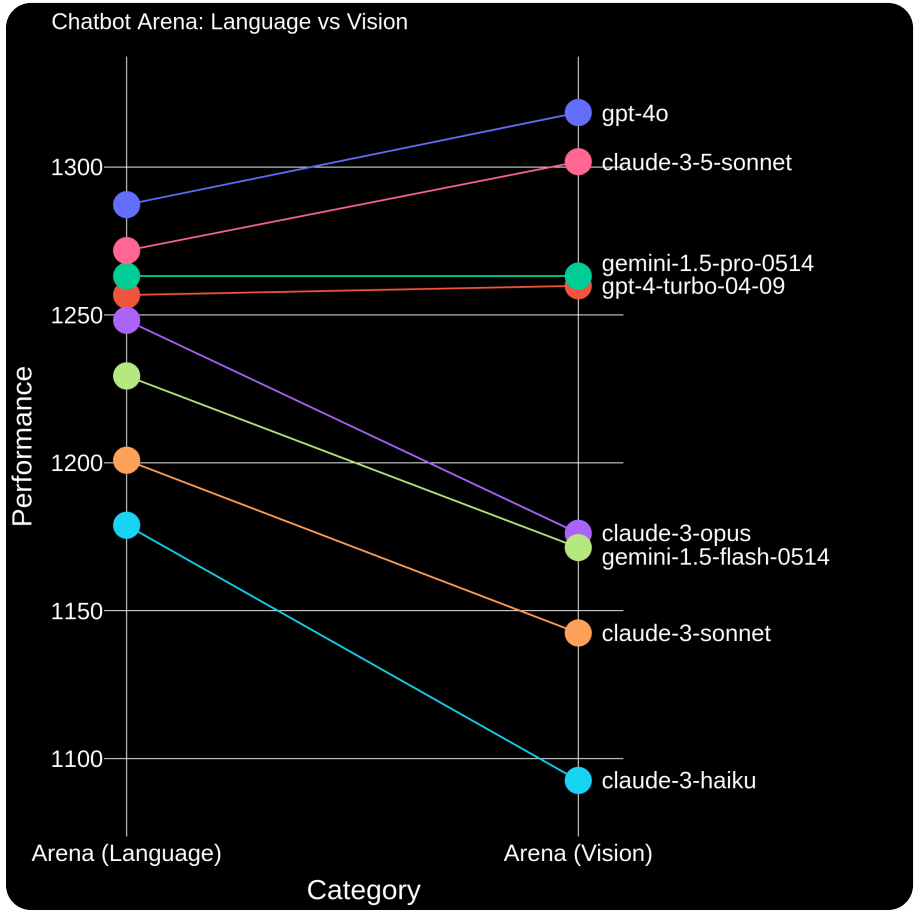
Key points in numbers

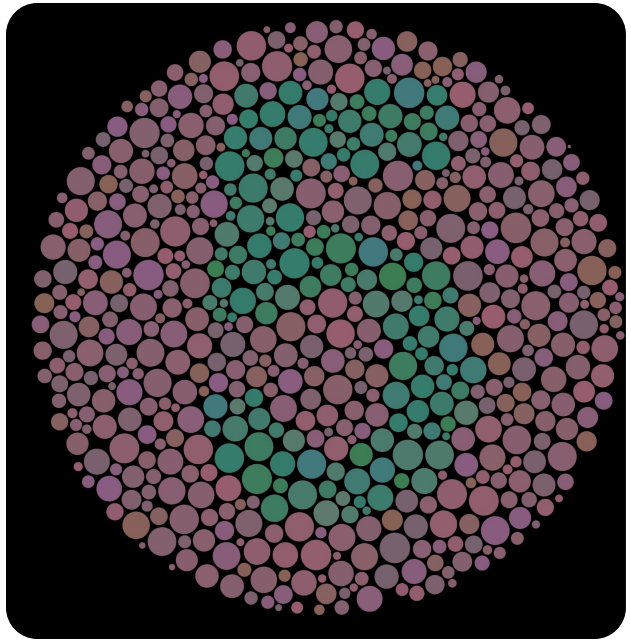
- It is able to describe images, answer questions about them, recognise text, etc.
- Innovative architectural solutions have led to a significant increase in the quality of the model
- The quality exceeds most similar open-source foreign solutions, and on a number of tasks even larger models in terms of the number of parameters
- First place in Daily Papers on Hugging Face on release date 10 April 2024
- Model posted in open-source

OmniFusion 1.1 (upd. 07.24)	VS	LLaVA-v1.5
70,73	TextVQA	58,20
74,13	ScienceQA	70,42
73,37	MMBench	64,30
37,20	MathVista	25,40
42,20	MMMU	35,30



VLM nowadays





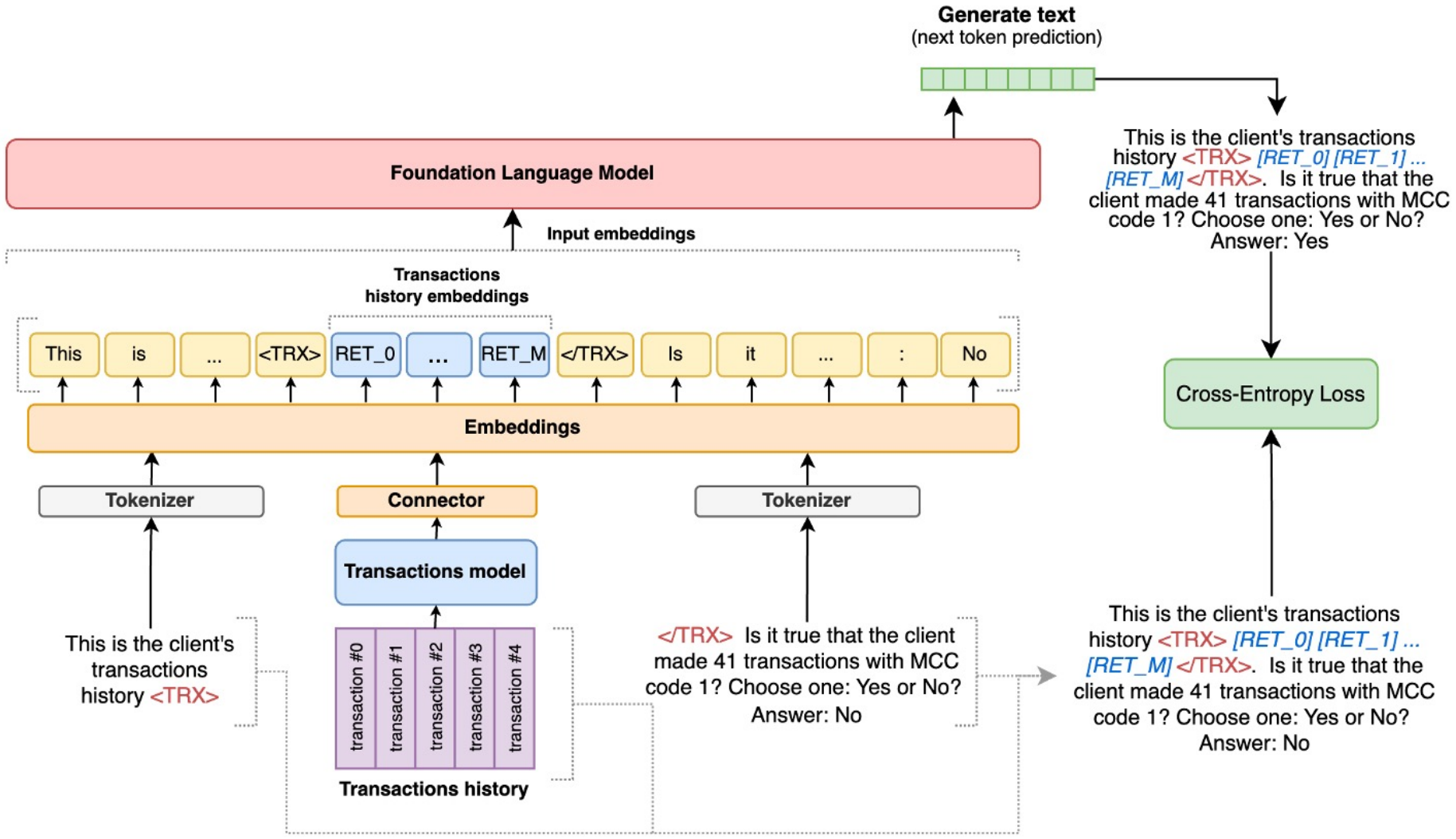
User: what is the number?

Gemini 1.5 Pro: The number hidden in the image is **8**.

Claude 3 Haiku: There is no number visible in this image. The image shows a colorful abstract pattern composed of various sized colored circles arranged in a circular pattern against a dark background.

https://lmsys.org/blog/2024-06-27-multimodal/

Other modalities: LLM + event sequences

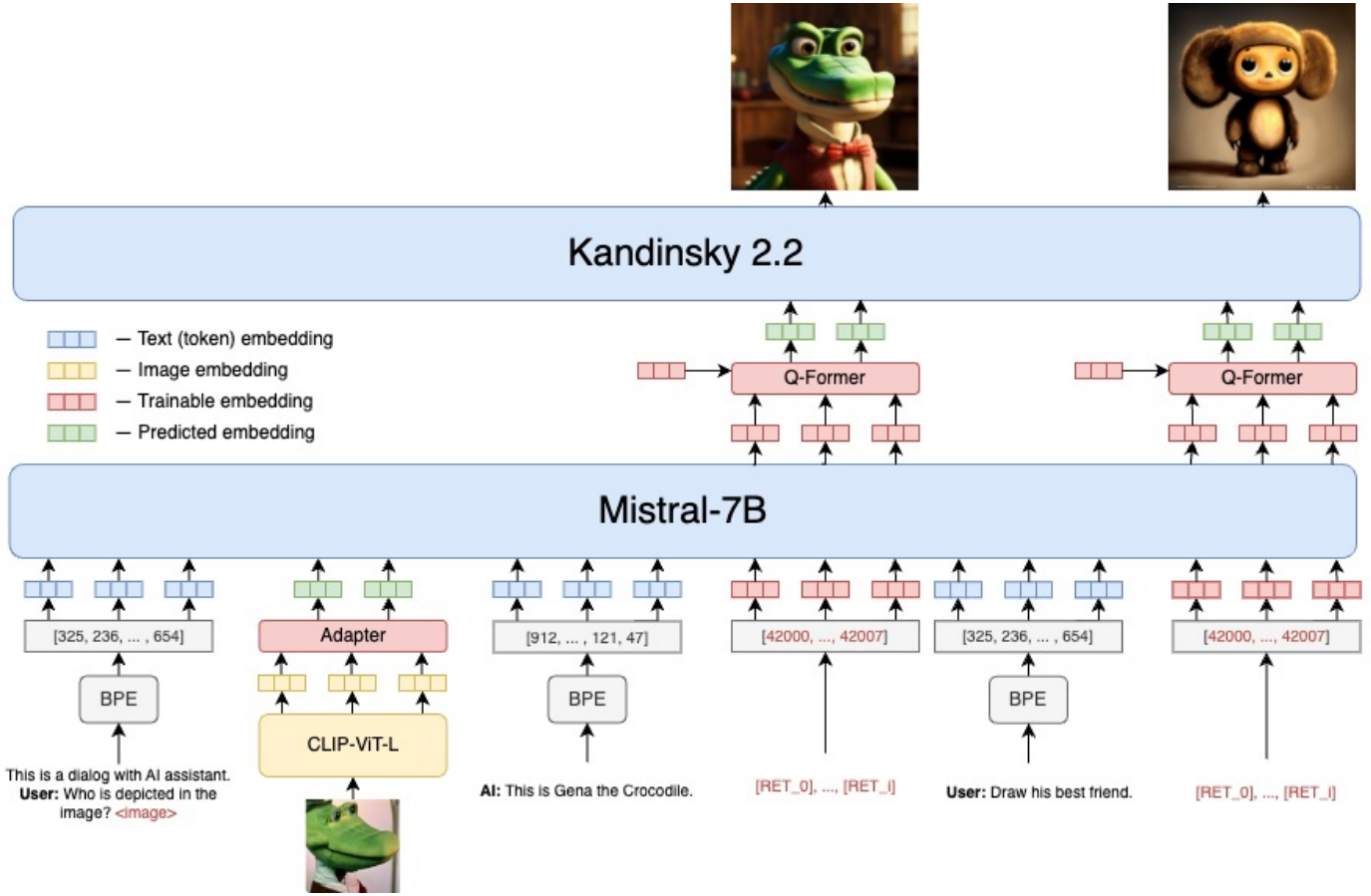


04

New tasks

OmniFusion for image generation and editing

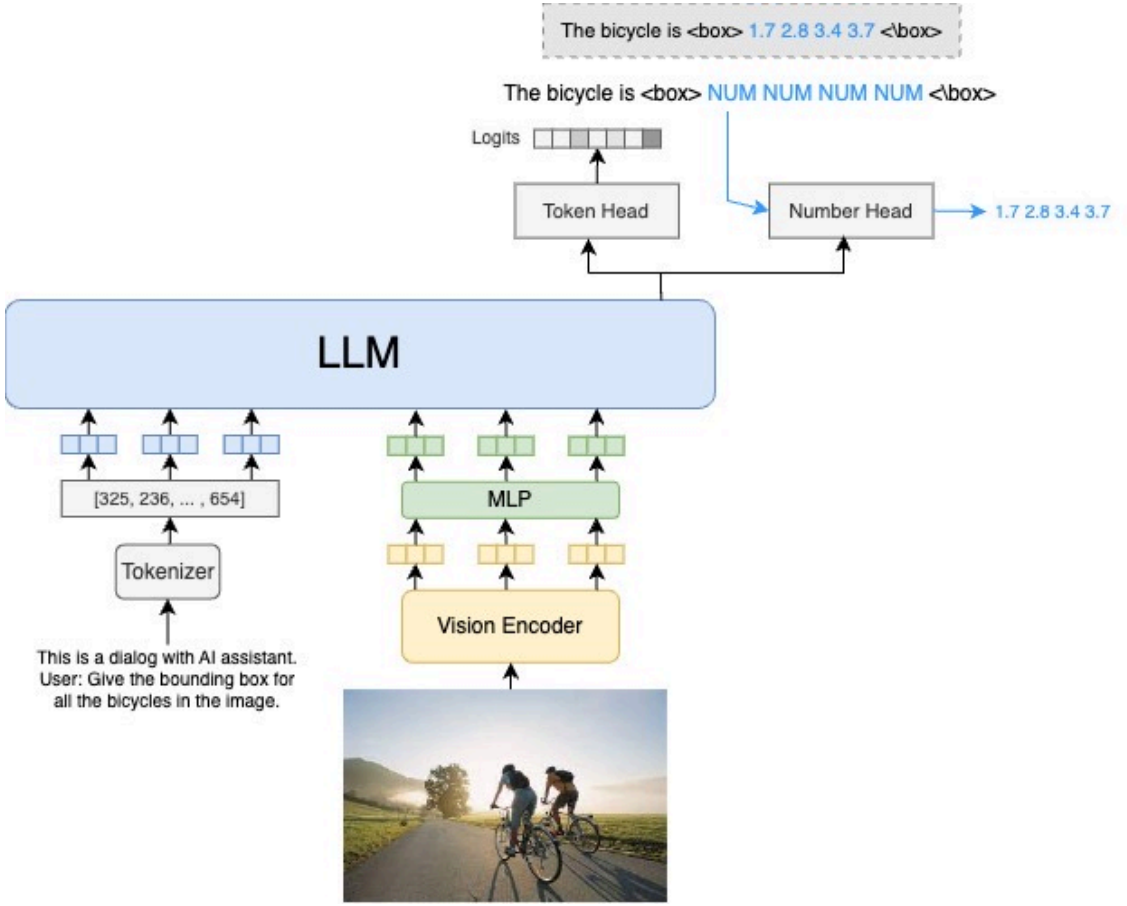
- Specific mapper to convert LLM text embeddings to input text embeddings of text-to-image Kandinsky 2.2
- Architecture — transformer
- [RET] token generation ~10 ms



OmniFusion for object detection

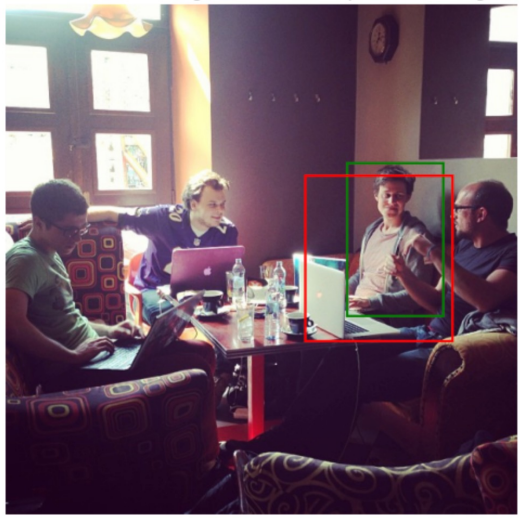
<image/> U: question A: <box> NUM NUM NUM NUM <\box>

- LLM processes text tokens
- **NUM** tokens are processed with a regression head, predicting a float coordinate

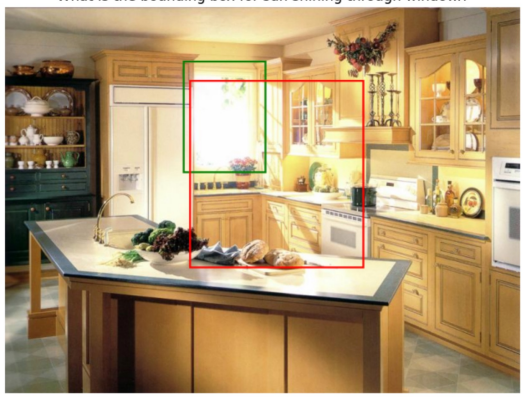


Visual Grounding quality estimation

Provide the bounding box for second person from right.



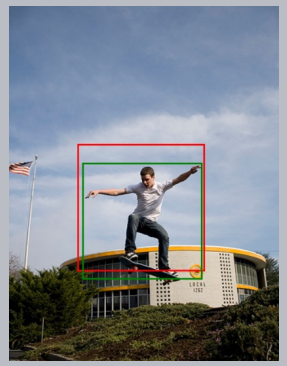
What is the bounding box for sun shining through window.?



Type	Model	RefCOCO val	RefCOCO test-A	RefCOCO test-B
Multitask models	OFA-L* (Wang et al., 2022b)	79.96	83.67	76.39
	Shikra-7B (Chen et al., 2023a)	87.01	90.61	80.24
	Qwen-VL (Bai et al., 2023)	89.36	92.26	85.34
	Ferret-13B (You et al., 2023)	89.48	92.41	84.36
	CogVLM-Grounding	**92.76**	**94.75**	**88.99**
	OmniFusion-Grounding	91.2	92.59	85.05
Grounding-oriented models	G-DINO-L (Liu et al., 2023e)	90.56	93.19	89.05
	UNINEXT-H (Lin et al., 2023a)	92.64	93.43	91.46
	ONE-PEACE (Wang et al., 2023a)	92.58	94.18	89.26

Redraw in blue

OmniFusion + Kandinsky.
Image generation



Where is a jumping skater?

His position is: [0.258, 0.344, 0.723, 0.652], accuracy: 76%

Object detection

AIRI

Contacts

Contacts

PhD,
Head of FusionBrain Lab, AIRI
Executive director on data science, Sber AI
kuznetsov@airi.net

@KUZNETSOFF87

@COMPLETE_AI