Deep Learning

Deeper, Better, _____ , Stronger than Machine Learning

Valentin Barriere

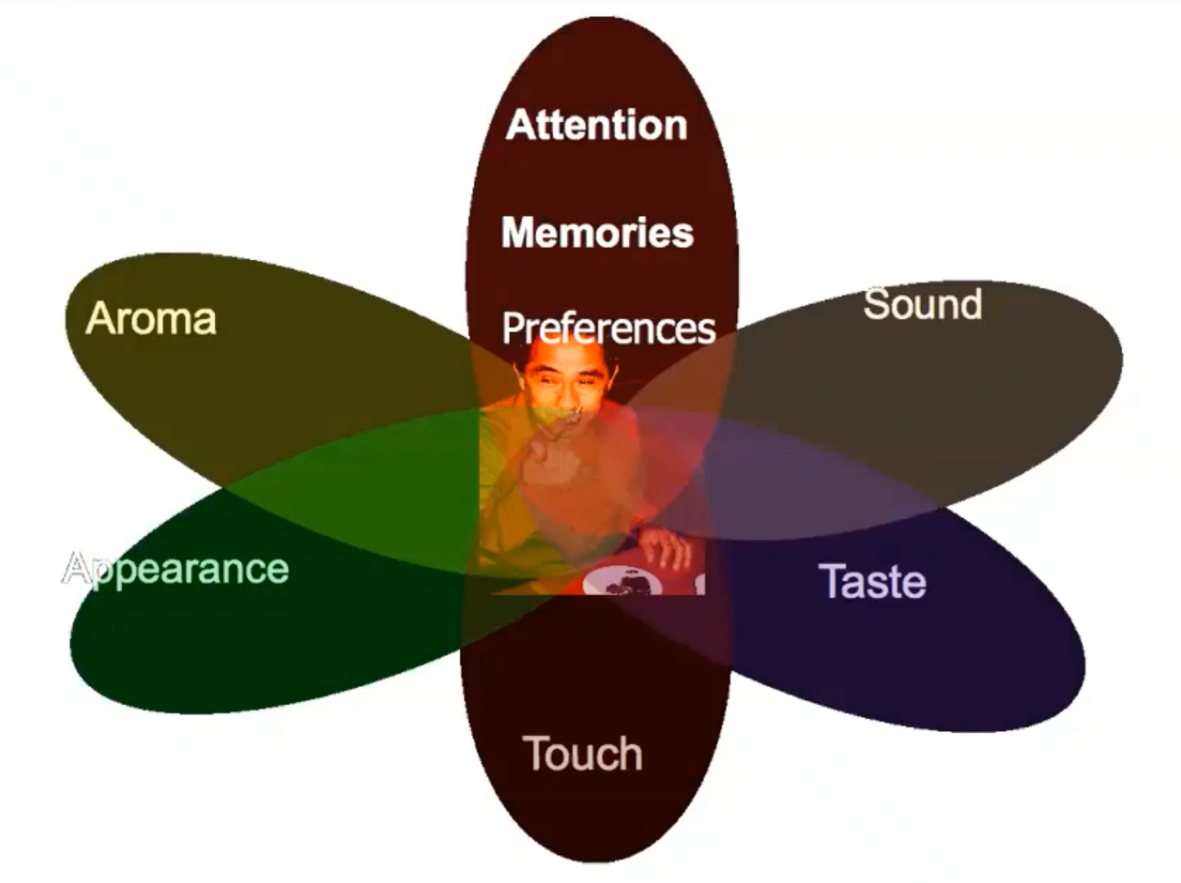
Universidad de Chile – DCC

CC6204, Primavera 2024

Multimodal Models

What is multimodal?

Modalities can refers to the human senses



Background: Multimodality

Human Intelligence and Artificial Intelligence, collect and process the information before taking a decision

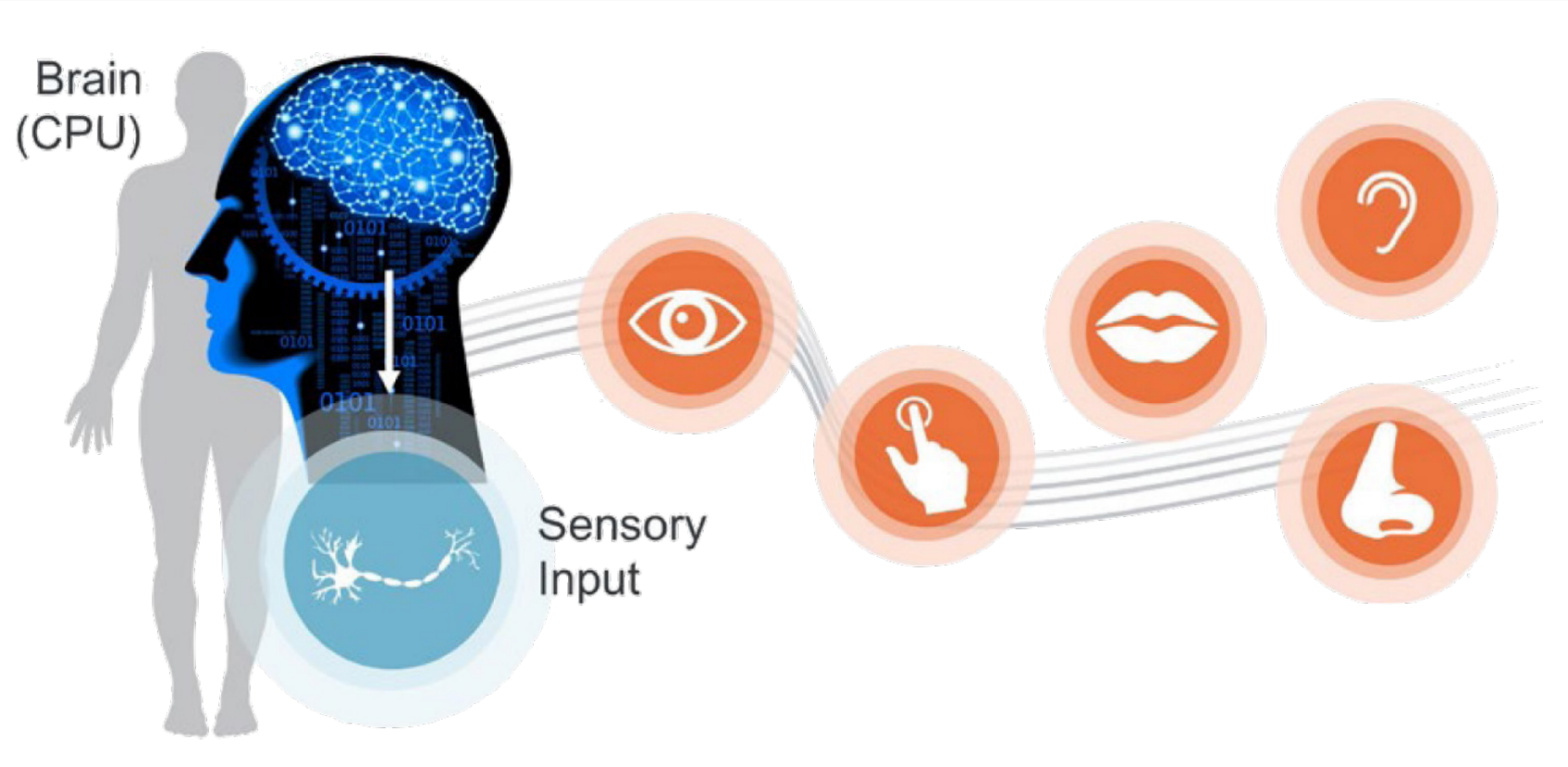


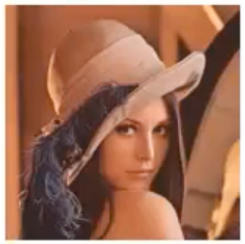
Figure 1: We use MModality in everyday life, this is how humans are communicating

Background: Multimodality

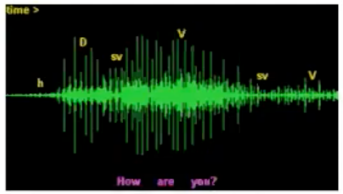
- **Language**: verbal content contains words, but also pragramatics and syntax

Language

- **Vision**: gesture, facial expressions, gaze, body language, ...


Vision

- **Speech**: prosody, vocal expressions, ...



Speech

But also: touch, physiologocal (ECG), mobile (location), social (friends in a network)...

Background: Multimodality

Basic definitions

Modality refers to a certain type of information, and it's representation format in which information is stored. It is transmitted via a medium, which is a system of communication/transmission.

Human Communication

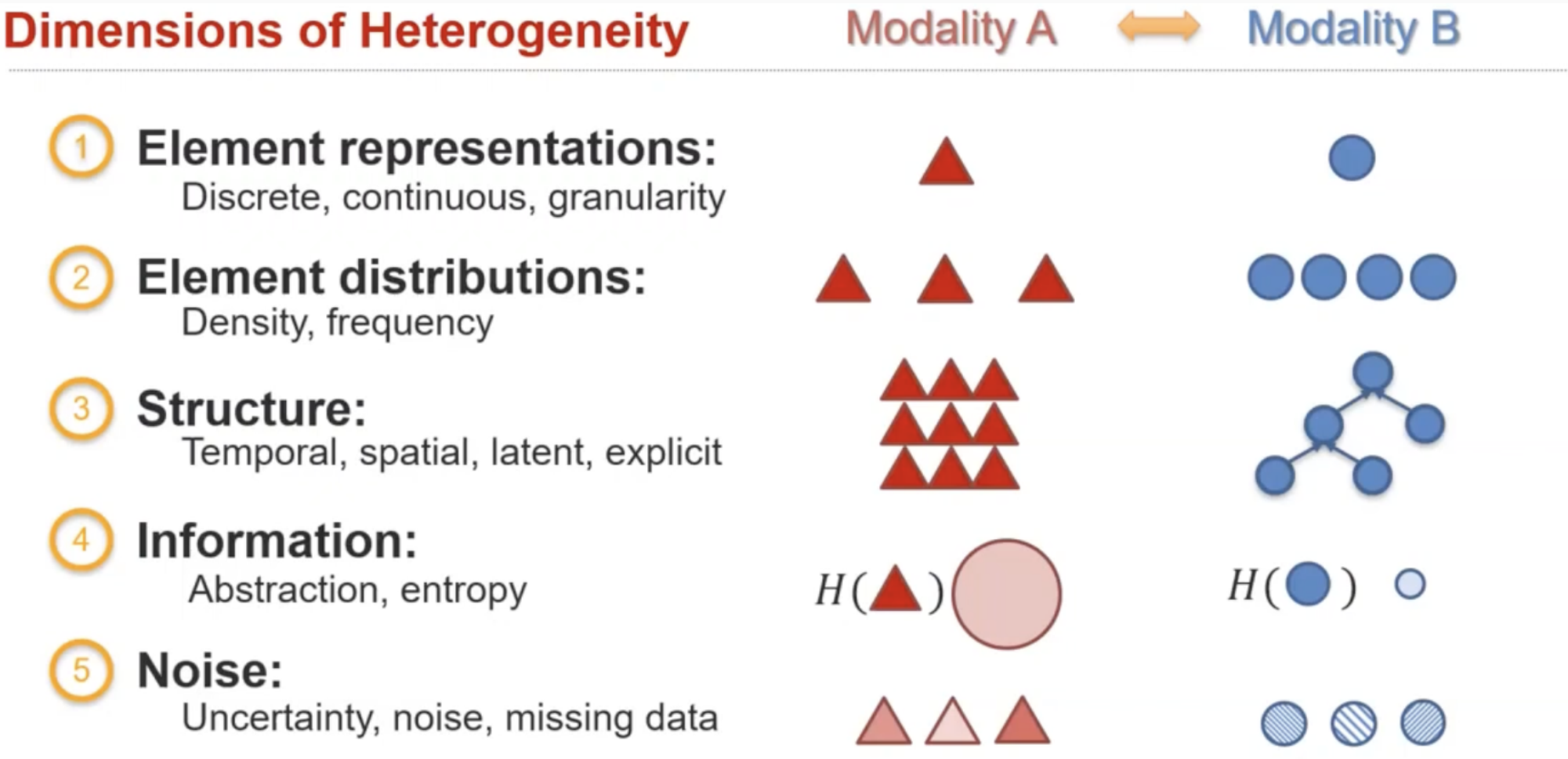
Transmission of a multimodal signal subject to the interpretation of individuals.

Depending on the different modalities that carry the information, we speak of vocal, verbal, facial, gestural signals, ...

Use of these signals to detect different aspects of the speaker: Emotion [24], Traits [14], or Employability [5].

They are heterogeneous and interconnected!

Modalities

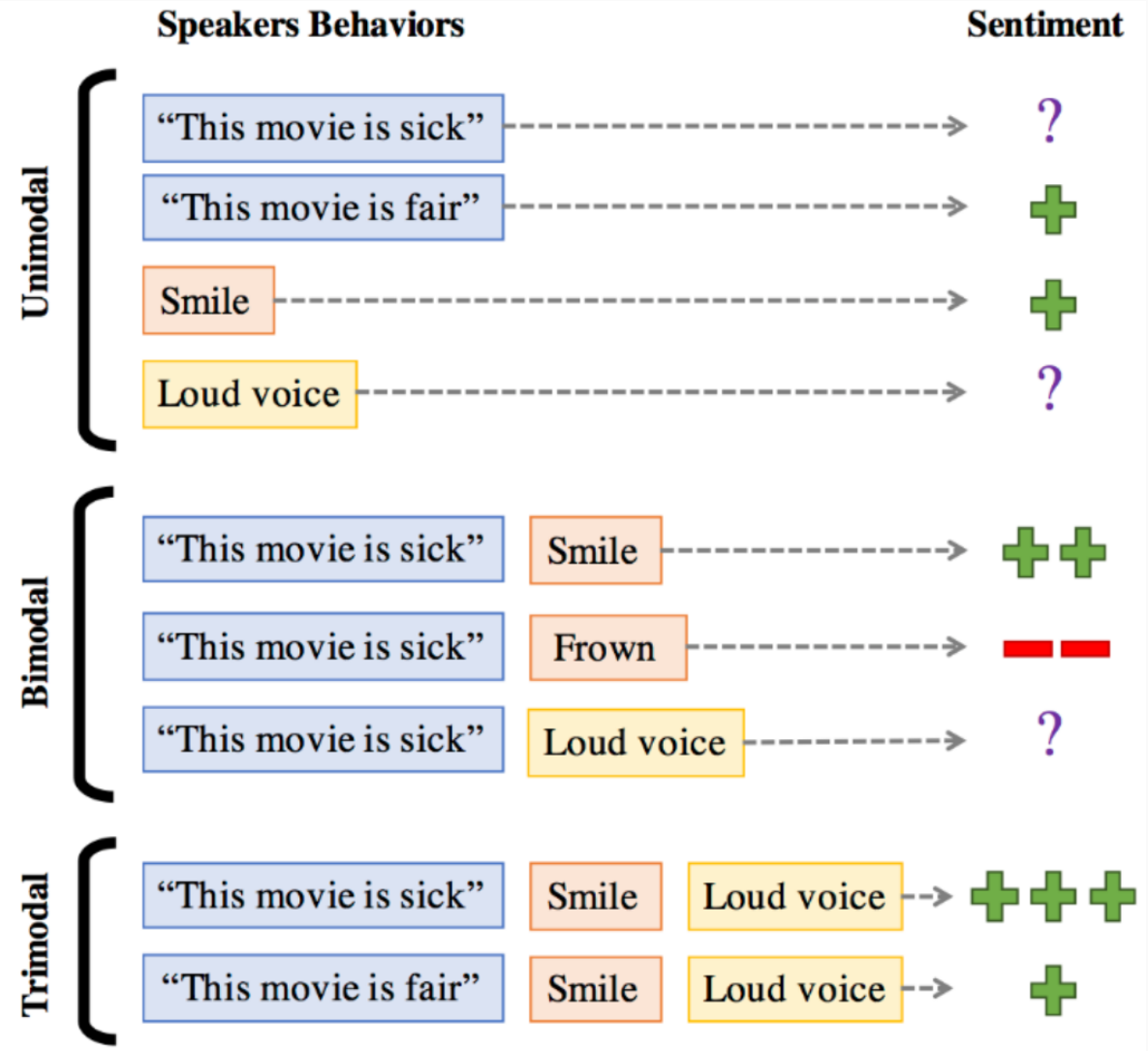


But they are also interconnected (McGurk effect):

`https://www.youtube.com/watch?v=2k8fHR9jKVM`

Complementarity

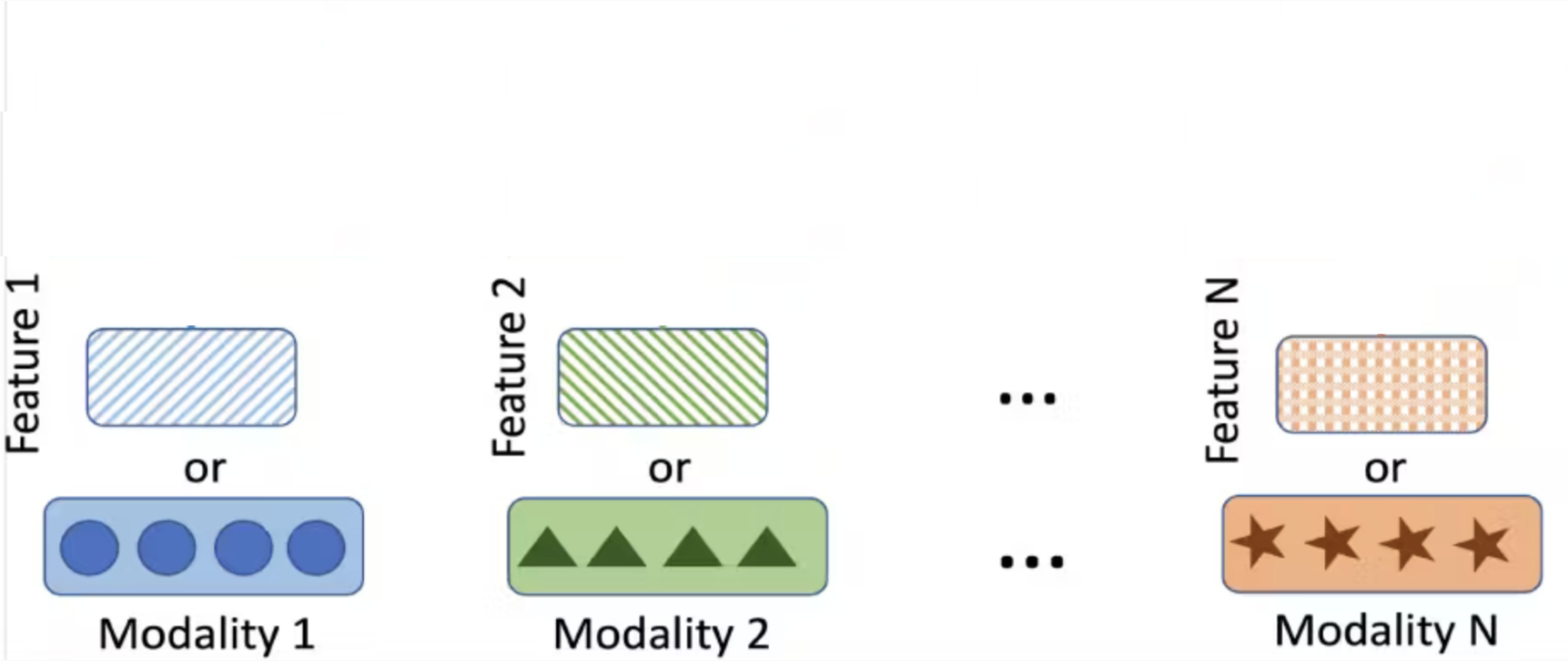
Different modalities can be consensual, complementary, one can help to reinforce the other, etc...



Representation and Encoding

Encoding

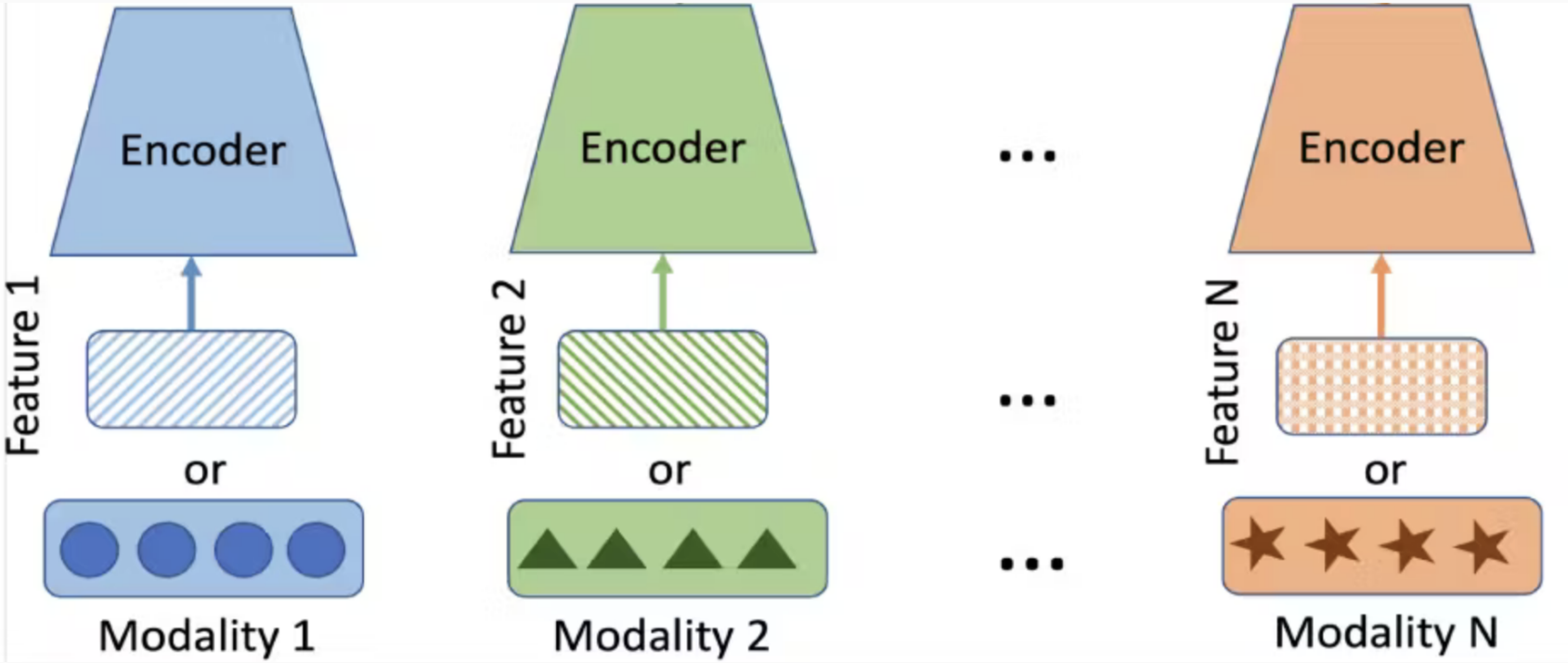
You can encode multimodal data using different methods: extract audio descriptors, extract image descriptors, text descriptors, etc...



Representation and Encoding

Encoding

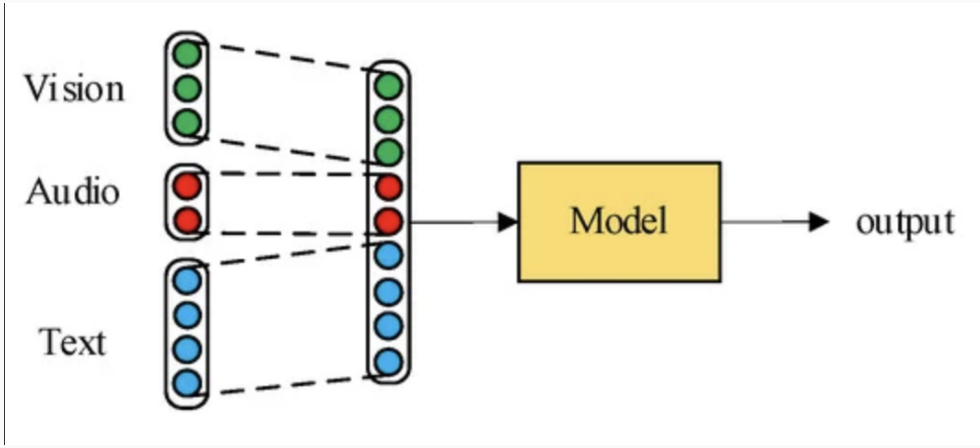
You can encode multimodal data using different methods: extract audio descriptors, extract image descriptors, text descriptors, etc...



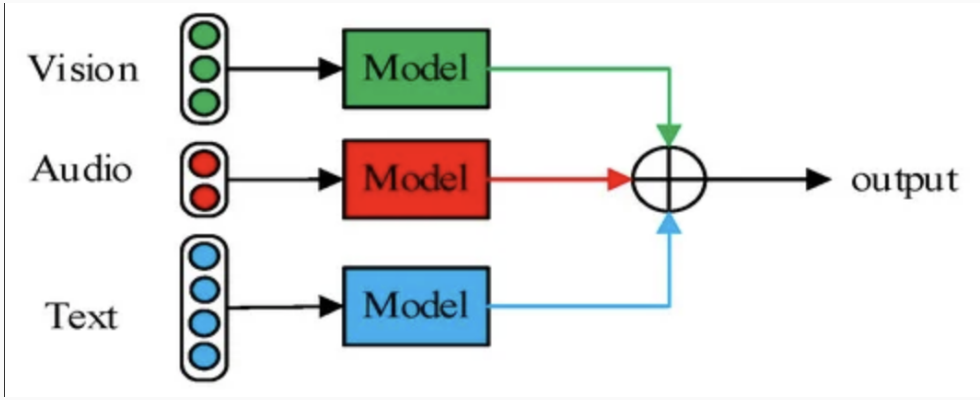
Also possible using networks: a pre-trained CNN to encode the image, a pre-trained transformers or word embedding to encode the text, etc...

Basic Fusions

- Early Fusion: before processing



- Late Fusion: after the prediction

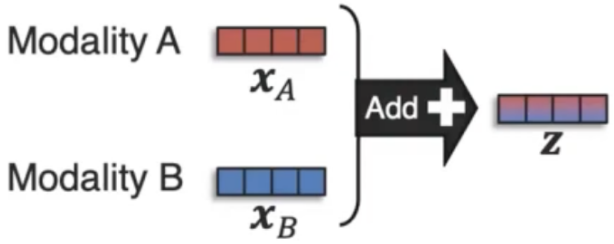


Historically

When the descriptors were extracted manually, the fusion of the unimodal representation before or after processing by a model. Nowadays, it is just big networks extracting the features and fusing them inside

Model-based Fusions I: Different type of fusions

- Additive Fusion:

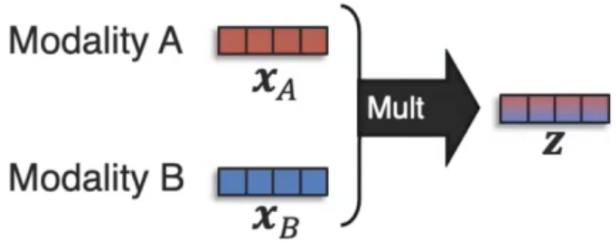


Additive fusion:

$$z = w_1 x_A + w_2 x_B$$

⇒ 1-layer neural network can be seen as additive

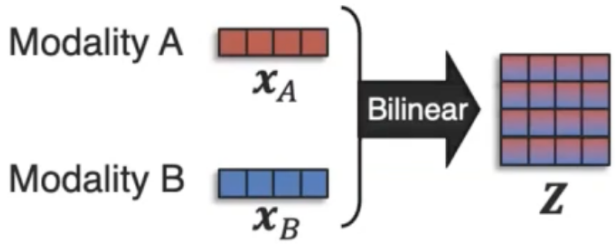
- Multiplicative Fusion:



Simple multiplicative fusion:

$$z = w(x_A \times x_B)$$

- Bilinear Fusion:



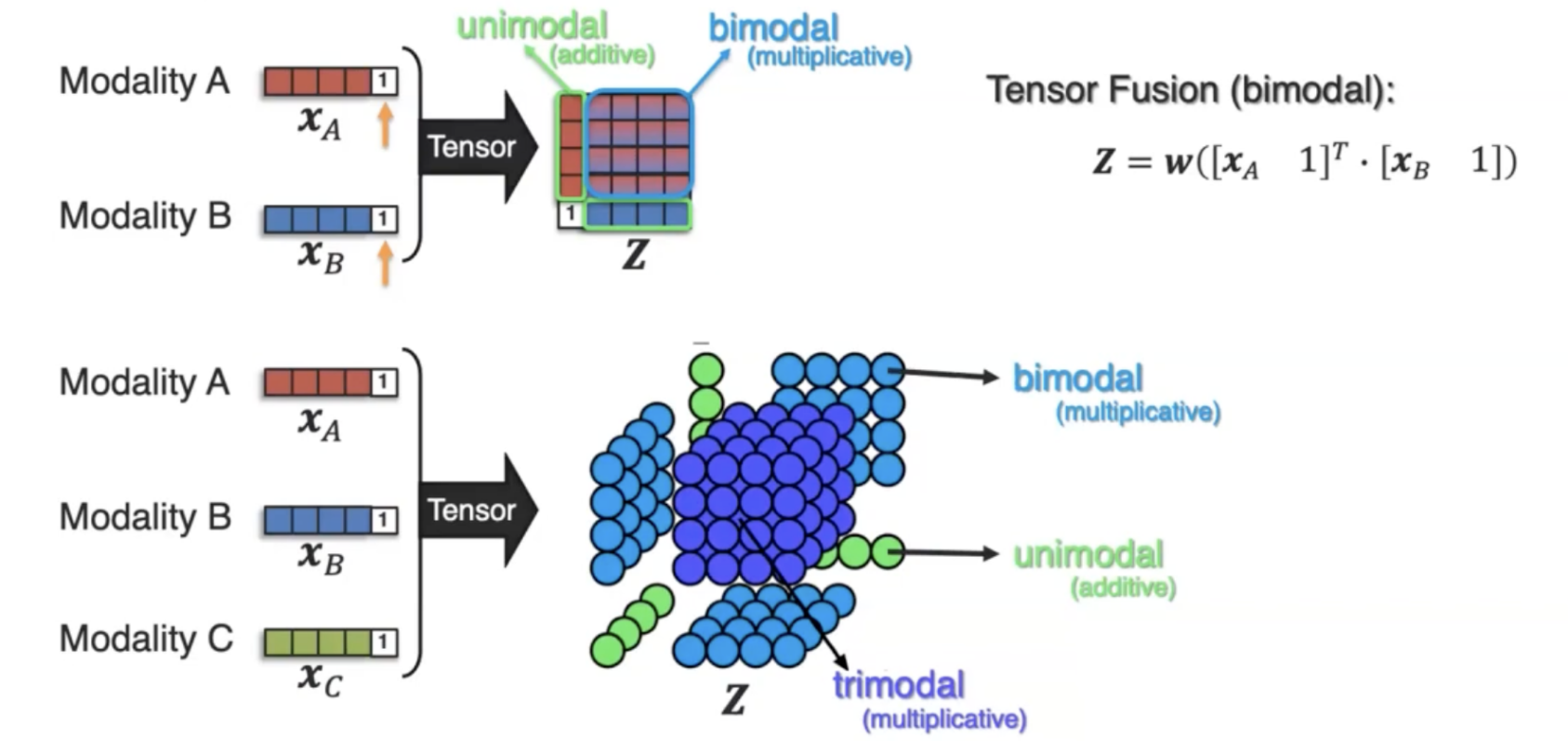
Bilinear Fusion:

$$Z = W(x_A^T \cdot x_B)$$

Remember: The data already has the interactions in it, the model is just trying to learn this

Model-based Fusions II

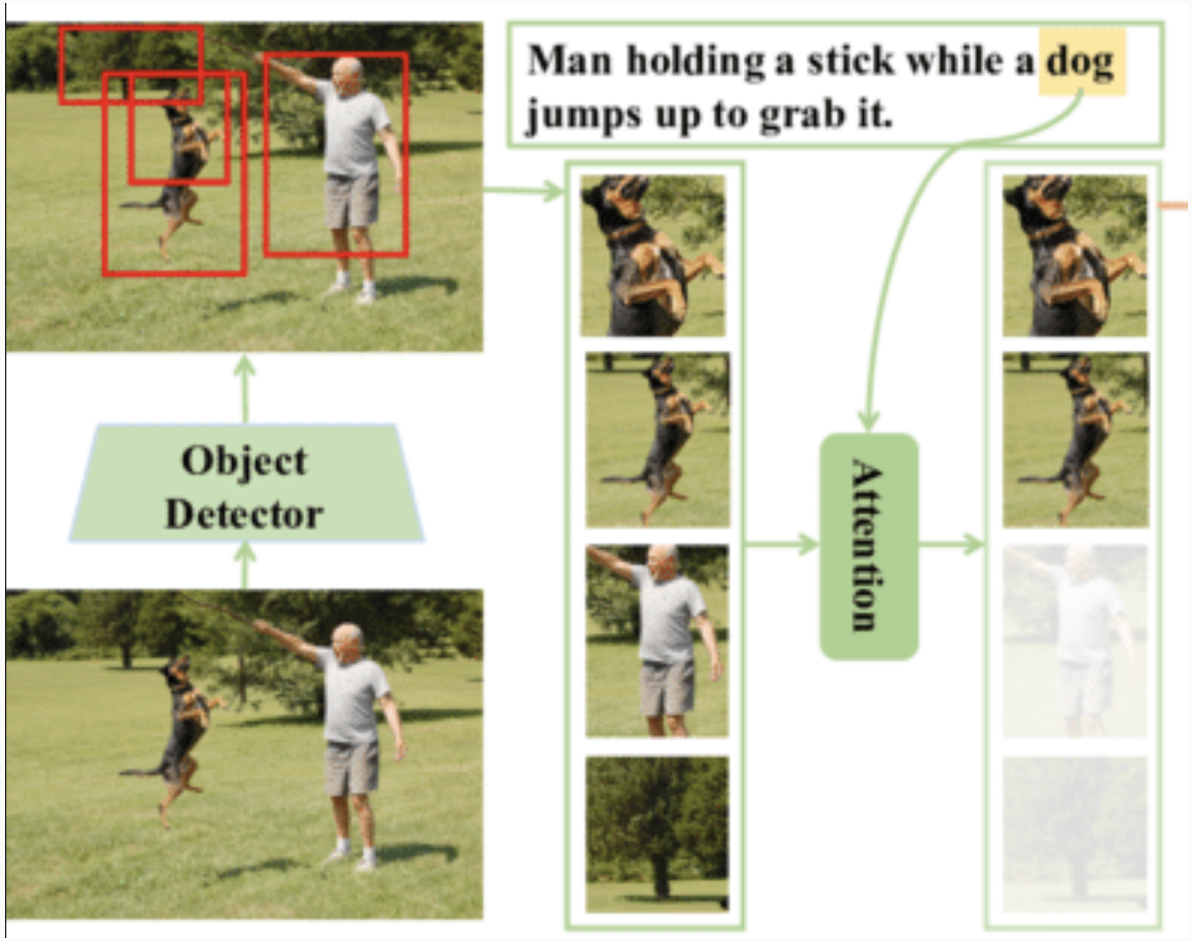
You can also create more complex fusion modules, but be careful about the complexity and number of parameters!



More in papers [16, 30, 7, 15, 18]

Model-based Fusions III

It is possible to use attention-like gating mechanisms, and in particular cros-modal attention. This helps to focus on the right part of the modalities when looking for something specific!



Model-based Fusions IV

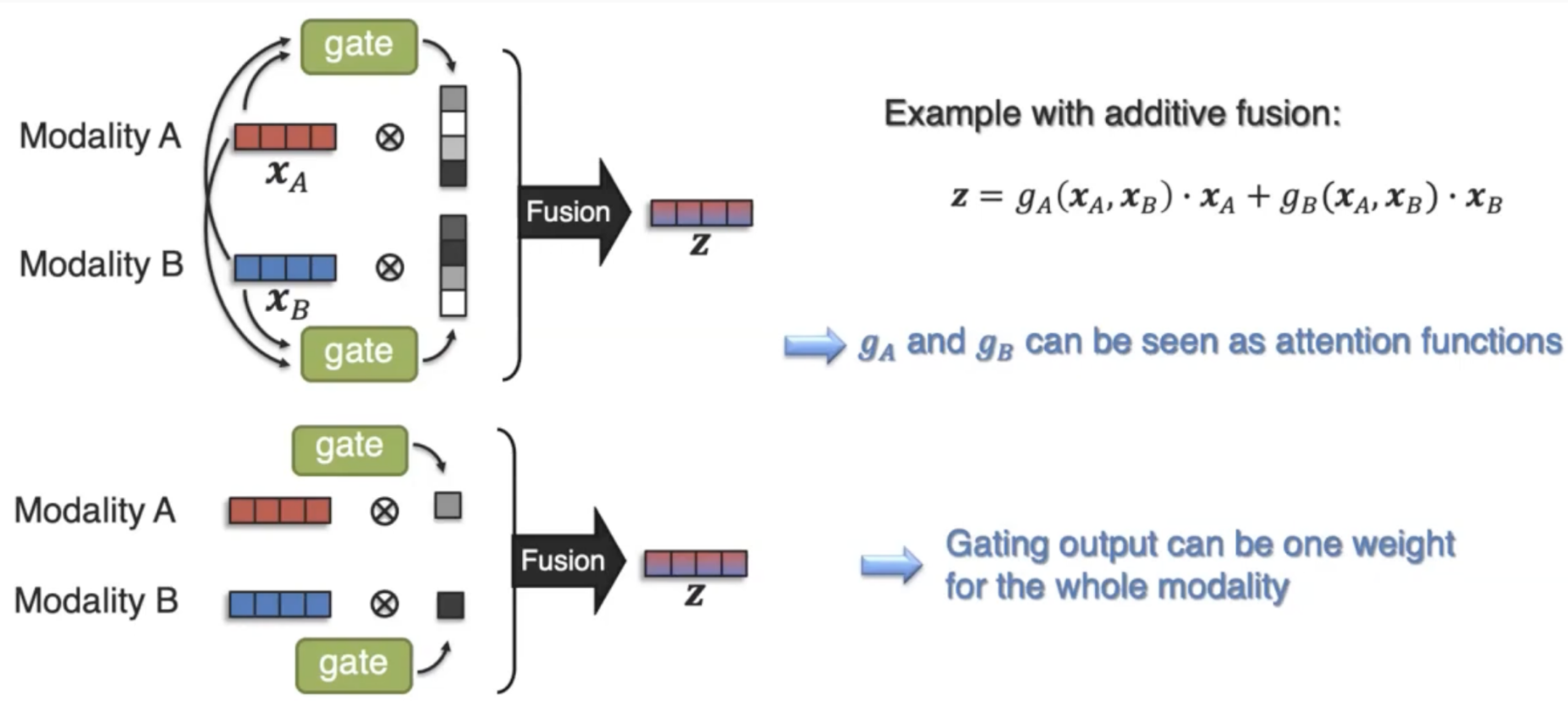


Figure 2: Cross-modal gating attention [2]

It can be seen as a way to:

- prevent unwanted signal from propagating forward (negative; *gating*)
- select the preferable signal to move forward (positive; *attention*)

Tasks

Output \ Input	Image	Text	Image & Text
Image	Vision tasks 	Text-based image generation, image retrieval Baby pandas walking on the grass 	Text-guided image editing, referring segmentation Panda on the far left 
Text	Image captioning  A group of baby pandas on the grass	NLP tasks ... world's rarest mammals, the giant [MASK] . Only about 1,500 of these black-and-white bears...	Visual question answering How many pandas are there?  There are 8 pandas
Image & Text	 How many pandas are there?	Visual Dialogue There are 8 pandas	Show me the image of a movie with a panda 

Courtesy of Javiera Castillo-Navaro

"Old School" Datasets

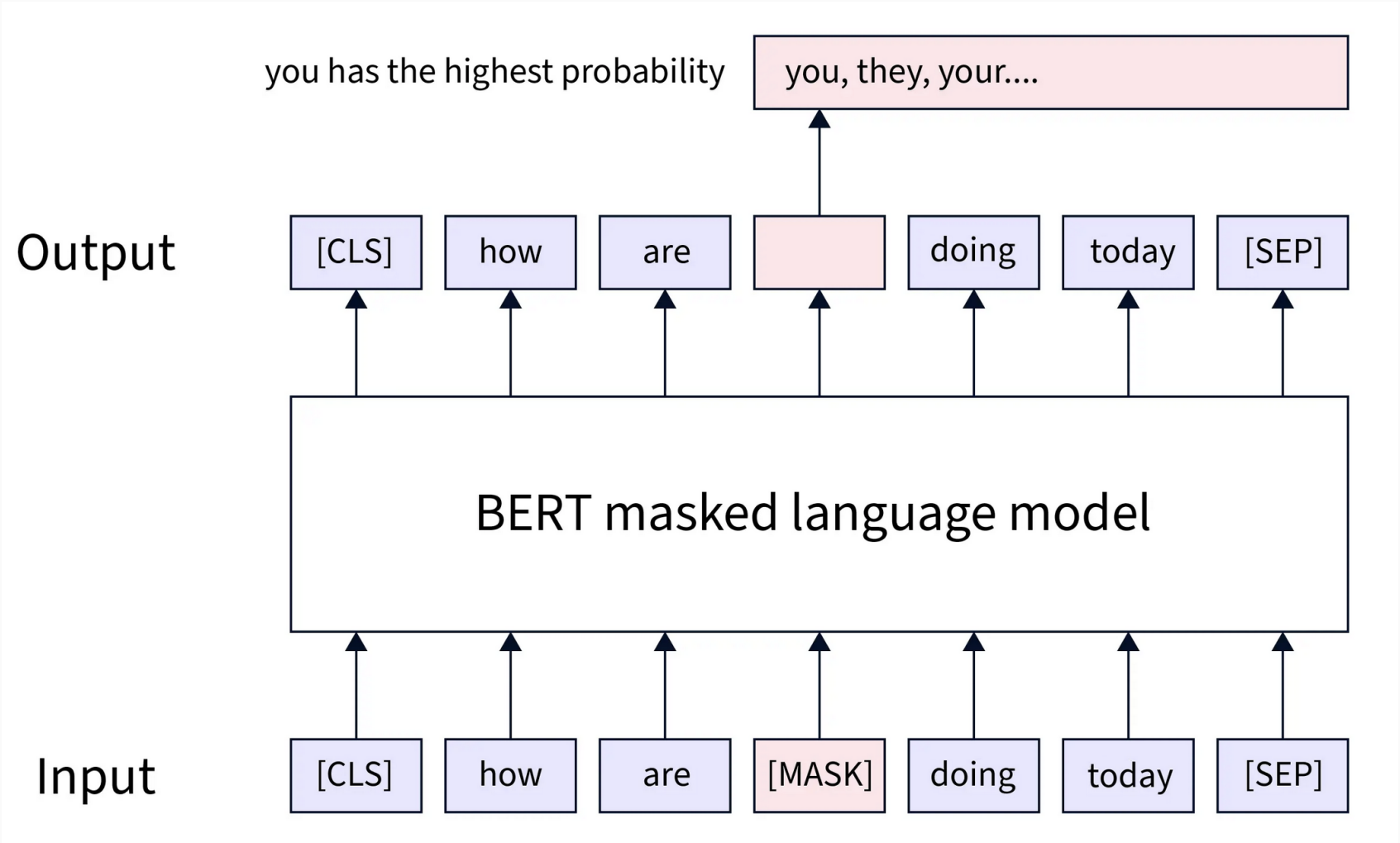
Figure 3: These datasets were used to train the multimodal models back in the days

For generative models, we will see that new data is available nowadays.

Courtesy of Javiera Castillo-Navaro

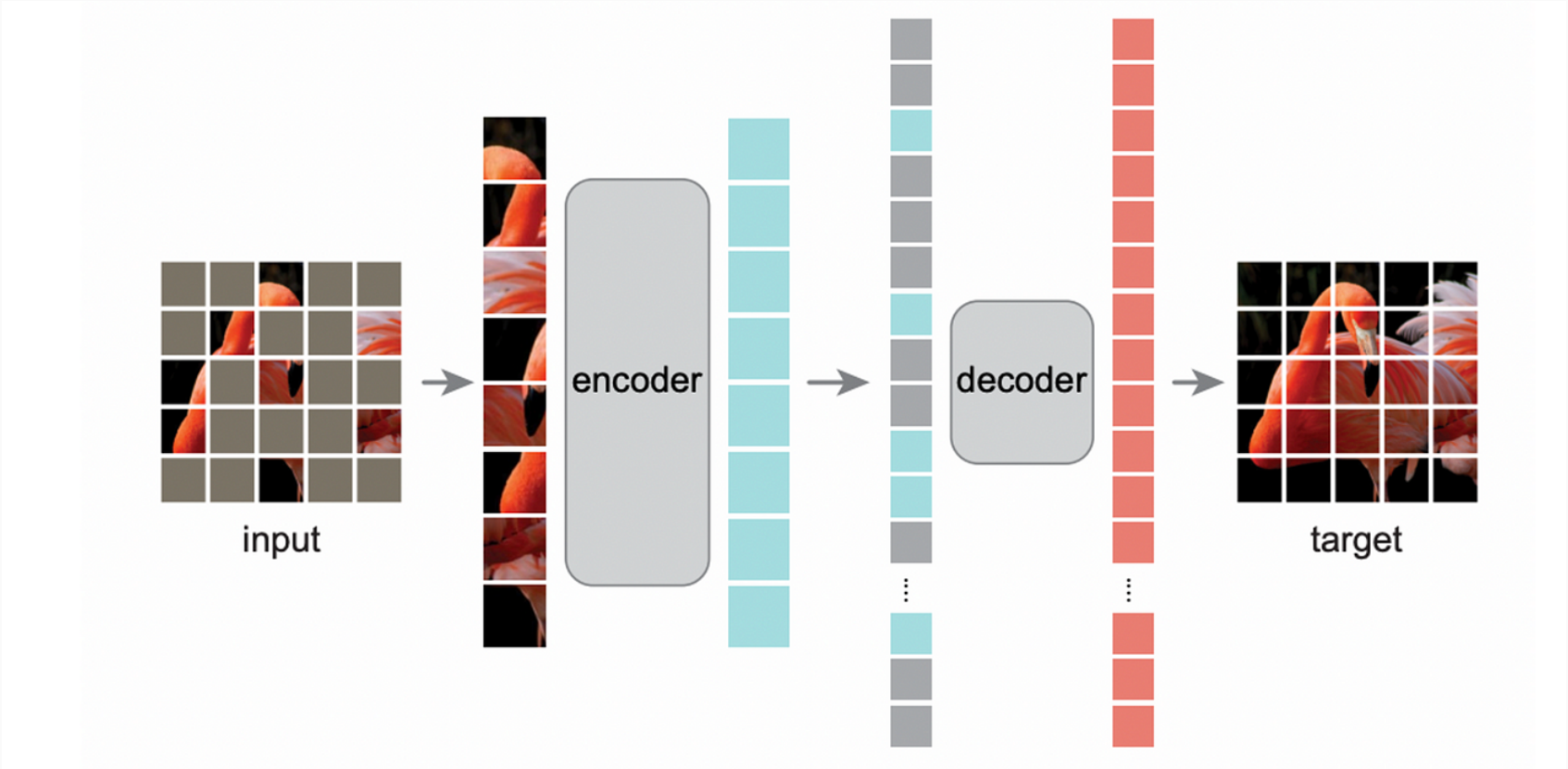
Fusion using Transformers

In the end, the transformers architecture functions well to fuse the different modalities, even their raw data is very different.



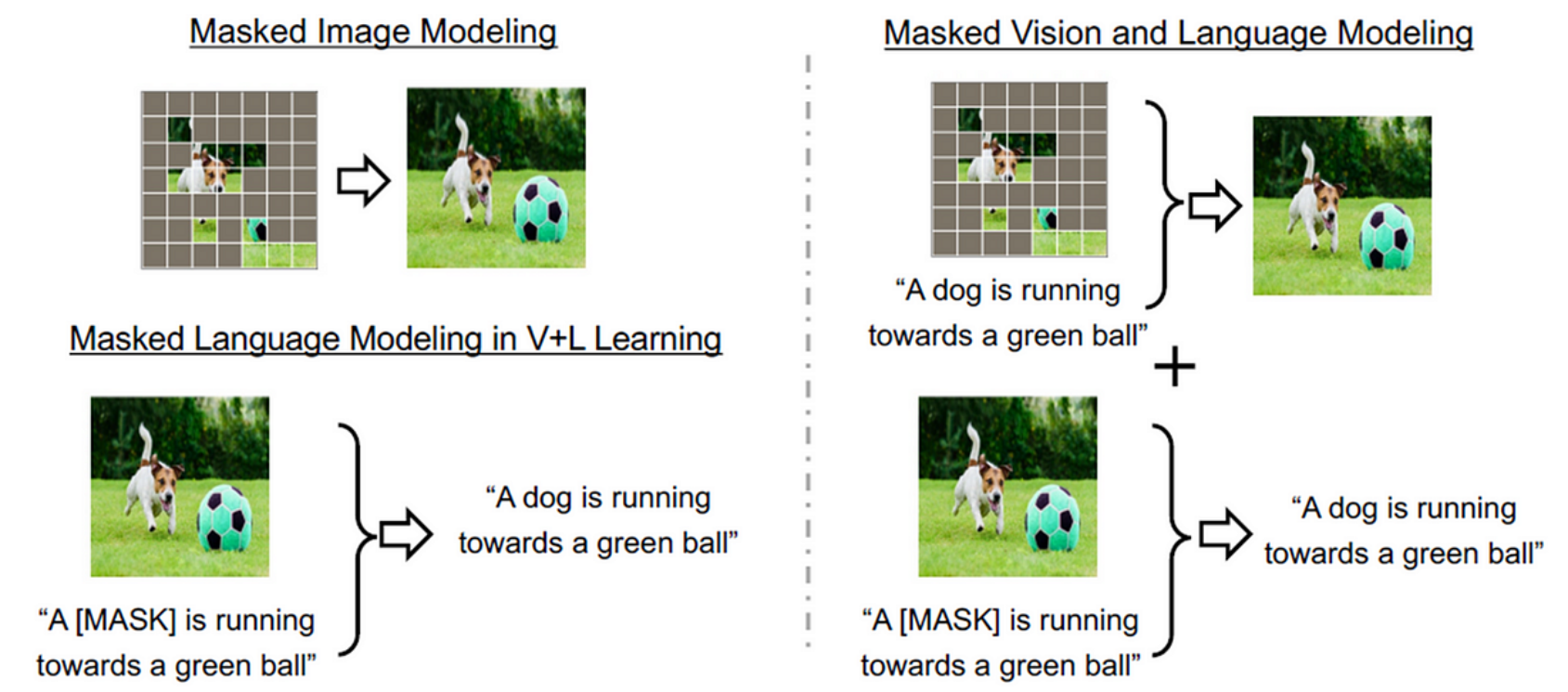
Fusion using Transformers

In the end, the transformers architecture functions well to fuse the different modalities, even their raw data is very different.



Fusion using Transformers

In the end, the transformers architecture functions well to fuse the different modalities, even their raw data is very different.



First LMMs: LXMERT

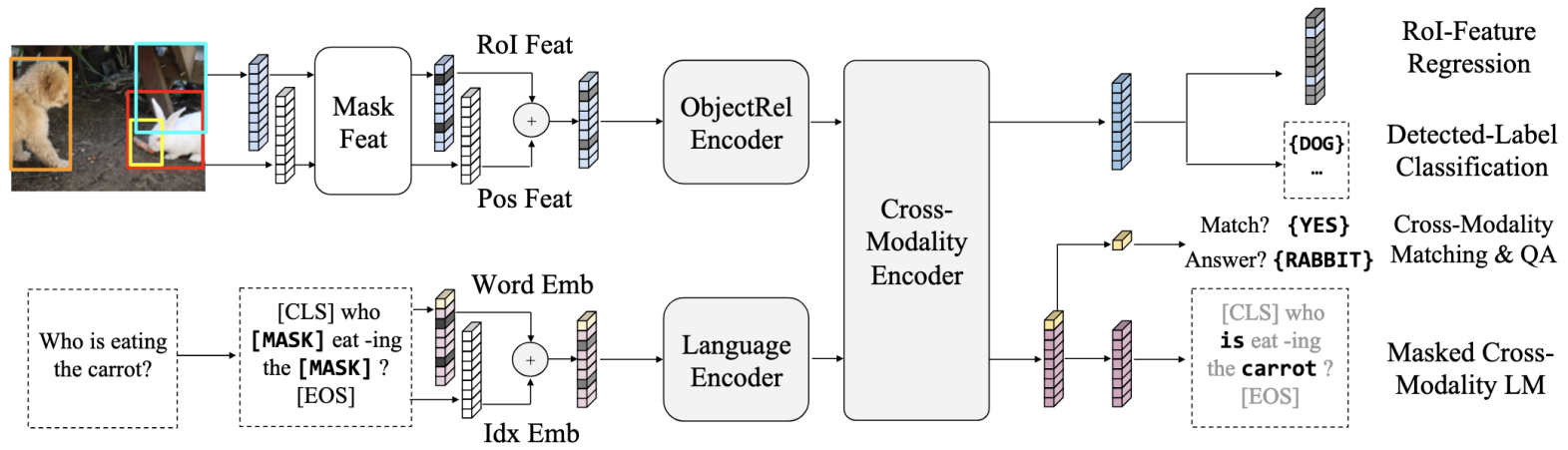
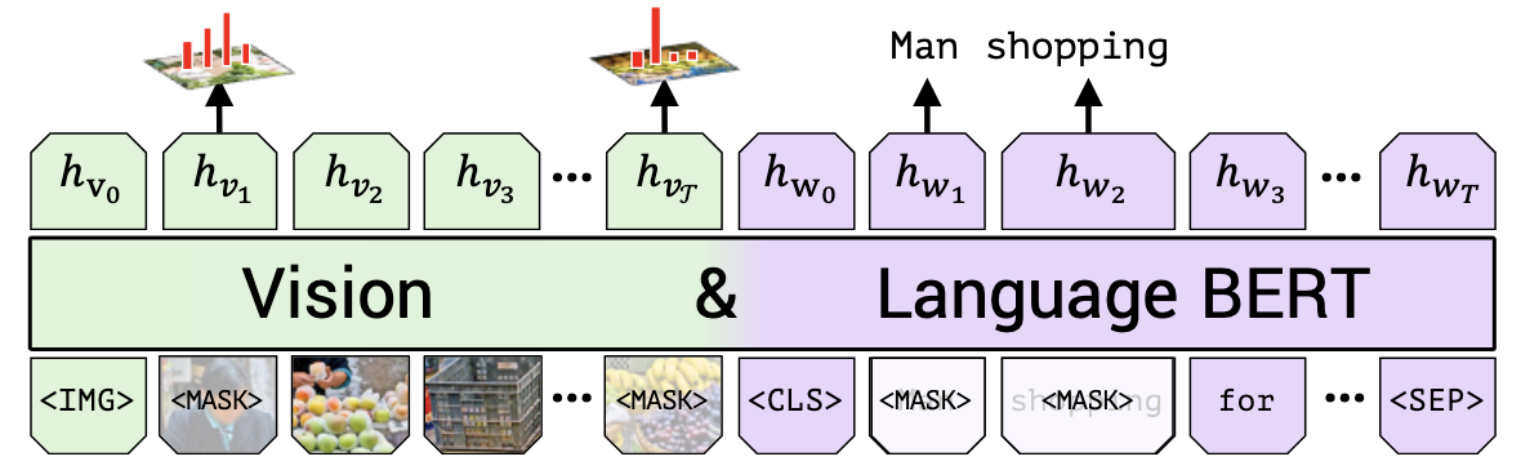


Figure 2: Pre-training in LXMERT. The object RoI features and word tokens are masked. Our five pre-training tasks learn the feature representations based on these masked inputs. Special tokens are in brackets and classification labels are in braces.

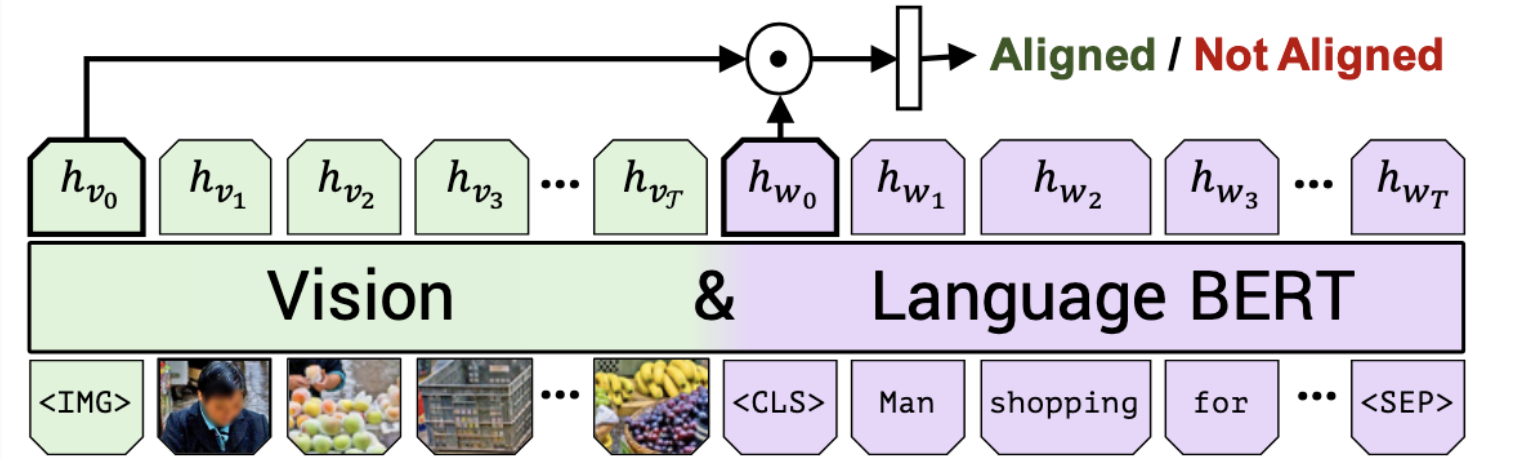
The first MModal Transformers [26]

- Use a trained faster-RCNN to extract Region of Interests (RoIs) embeddings, and learn to classify it
- Mask words for language, or dimensions for vision
- Cross-modal matching and VQA tasks

First LMMs: ViLBERT



(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

First LMMs: ViLBERT

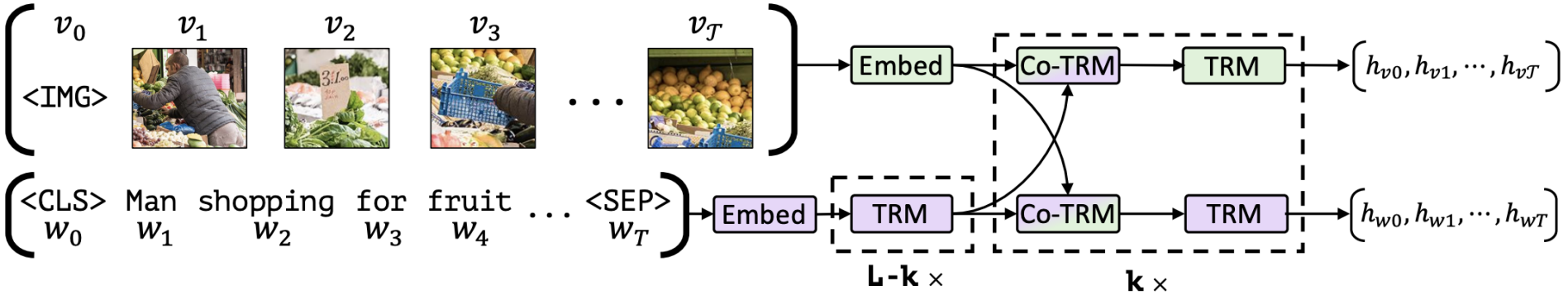
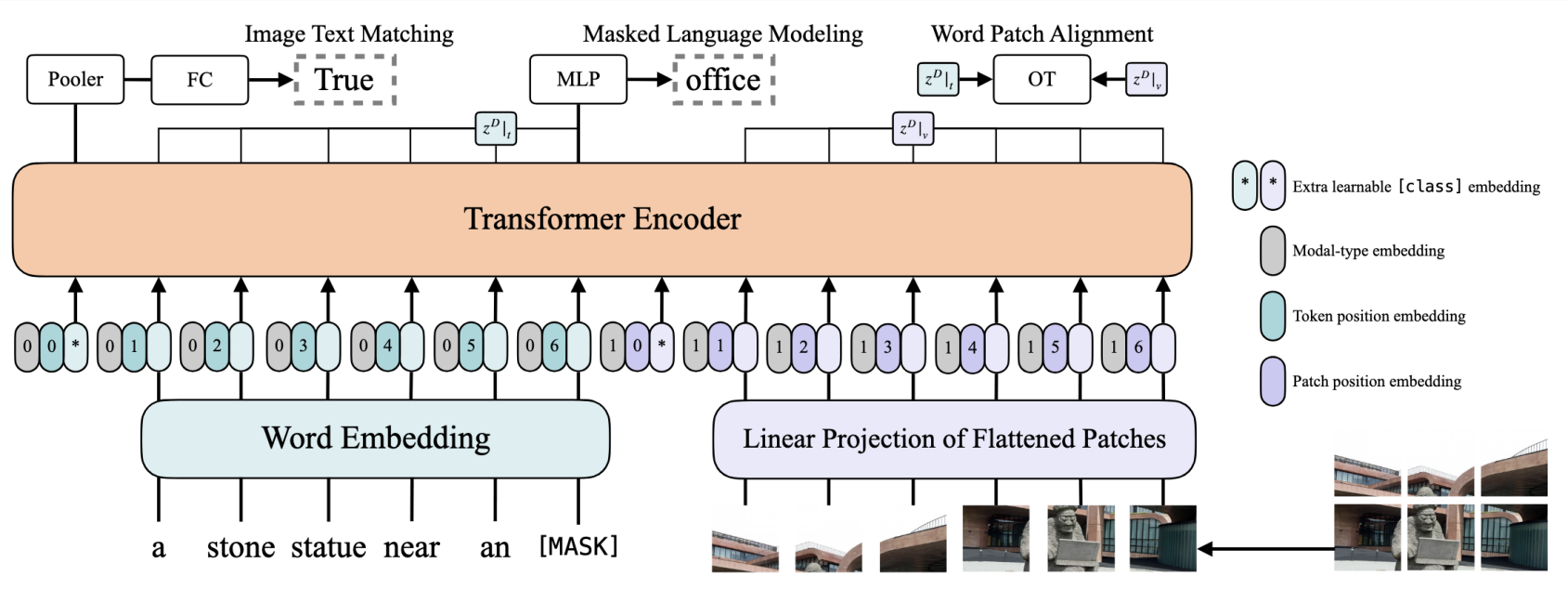


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

The ViLBERT model has a cross-attention transformers layers, using the queries and keys/values different modalities, allowing for interaction between the modalities in the inside layers [19]

First LMMs: ViLT

The Vision Language Transformers does not use any CNN to encode the image, showing that flattening the image before transformer layers is good enough! [8]



It uses three losses, similar to what we've seen beforehand.

Coordination

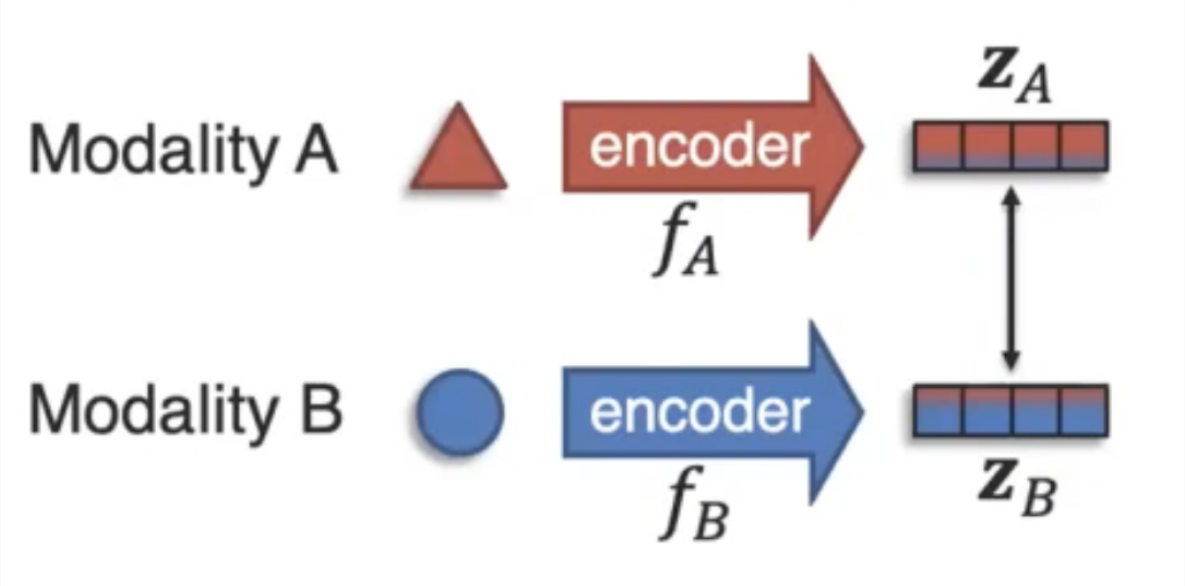


Figure 4: Coordination: align the modalities in a latent common space is a clever way to train a multimodal model in a self-supervised way.

Coordination

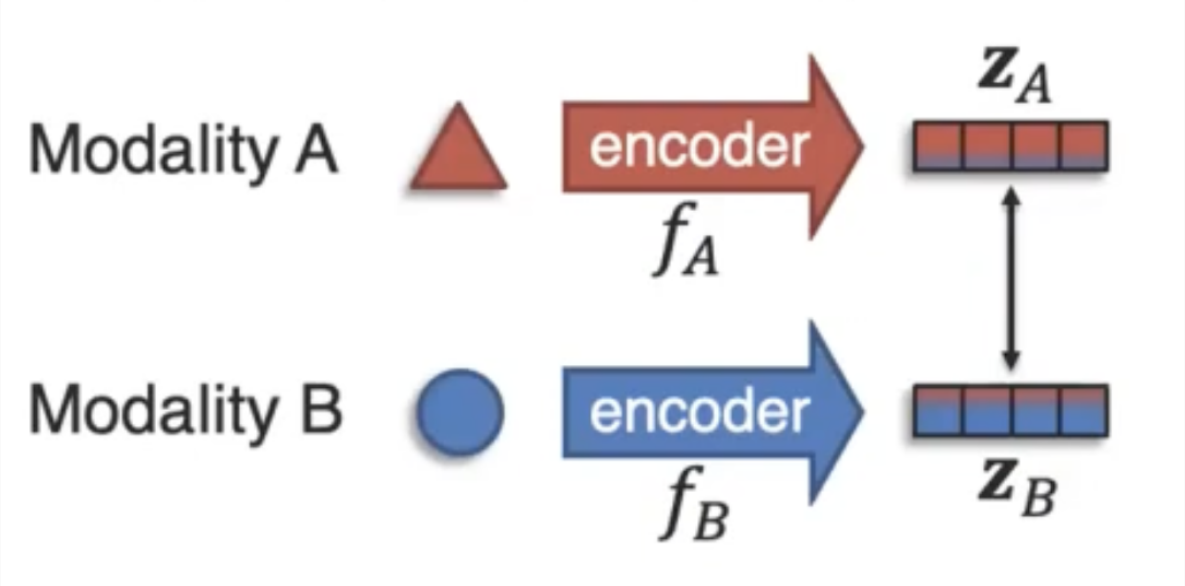
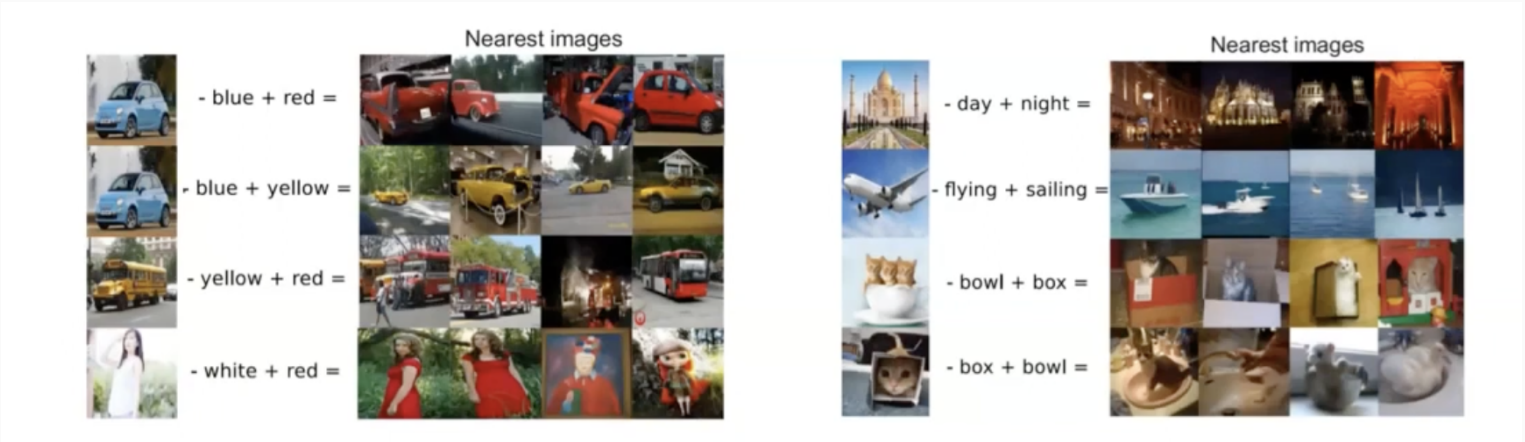


Figure 4: Coordination: align the modalities in a latent common space is a clever way to train a multimodal model in a self-supervised way.



Contrastive Learning: Example on Images

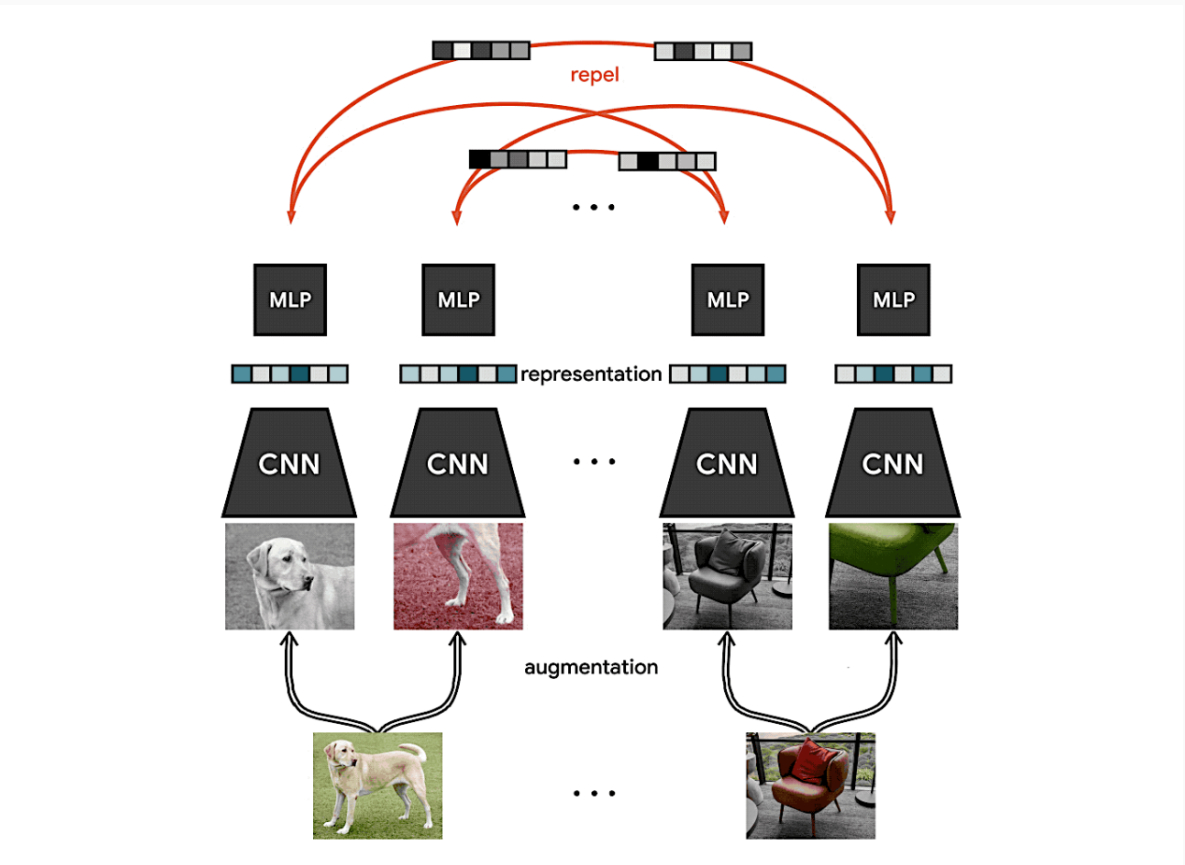


Figure 5: SimCLR is an example of pre-training on ImageNet

Principle

Representations from parts of the same images are getting closer while representations of the patches from different images are getting pushed apart. Can also work with classes!

CLIP: Contrastive Language-Image Pre-training

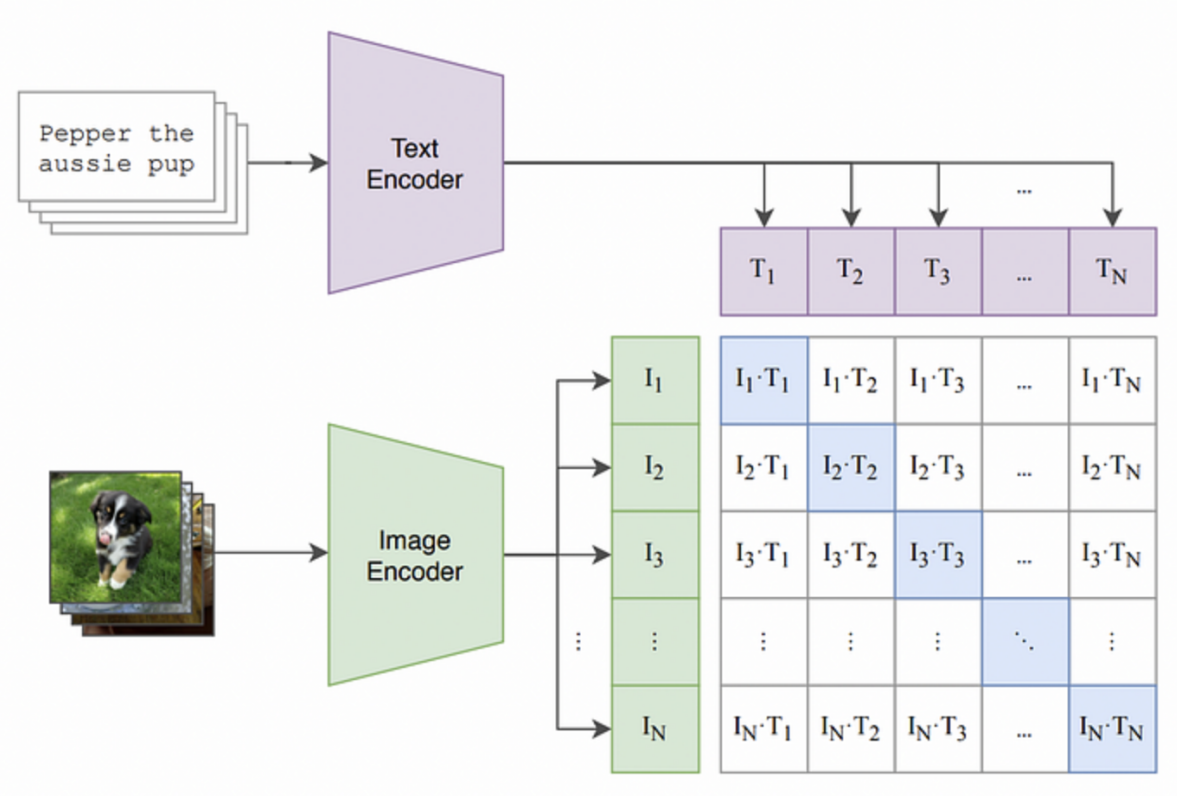


Figure 6: CLIP has been trained over 400M image-text pairs

Principle

Learn a shared multimodal embedding space, and maximize/minimize the distance between the embeddings belonging together. Loss equivalent to learn a function that maximizes the mutual information between the modalities [22].

CLIP: zero-shot

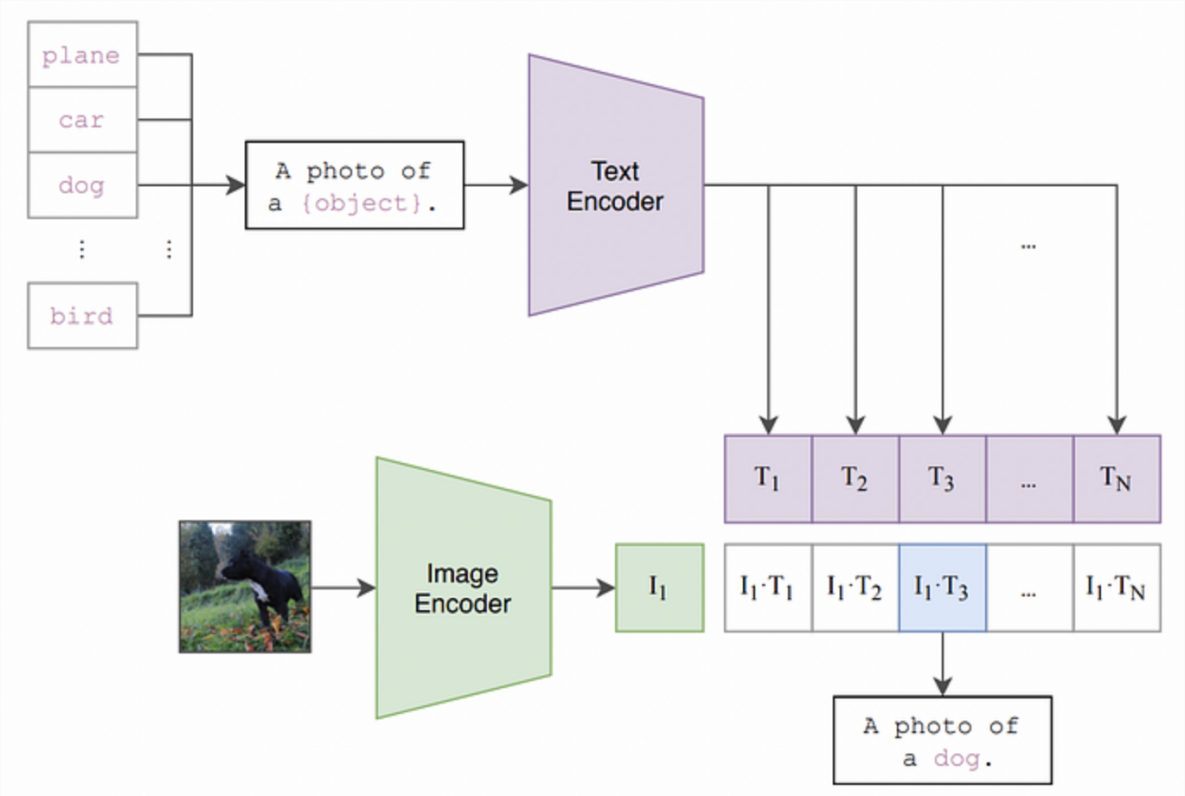


Figure 7: CLIP is very good at zero-shot using text templates

However, this is really dependant of the data and classes (it sucks for OOD such as Satellite data).

More on CLIP here

BLIP: Bootstrapping Language-Image Pre-training

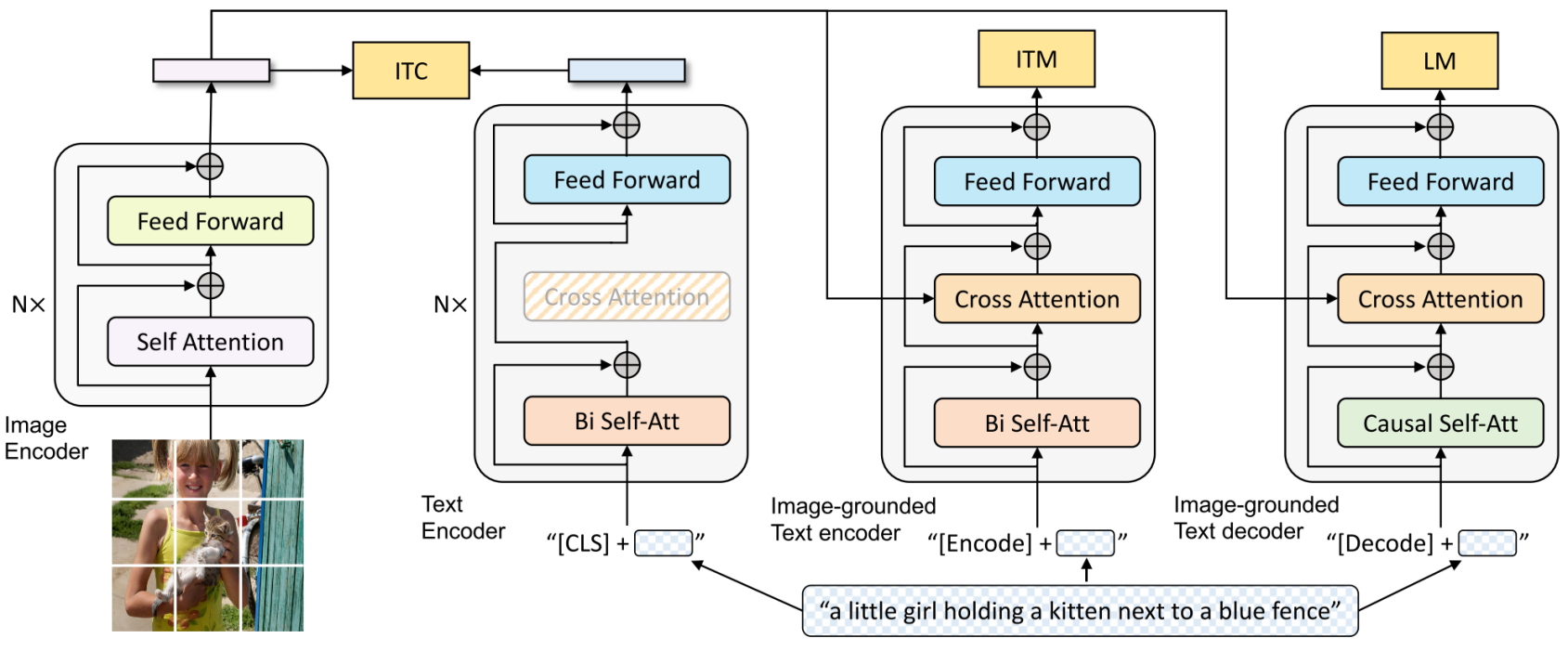


Figure 8: Unimodal encoders + multimodal image-grounded text-encoders [12]

Three losses

Image-text contrastive (ITC) loss for the unimodal encoders. For the multimodal encoders, an image-text matching (ITM) to distinguish between positive and negative image-text pairs, and a Language Model (LM) loss to generate text. **Also bootstraps the dataset to clean noisy examples** (filter bad captions).

Text to image generator: CLIP + VQ-GAN

Figure 9: Artificial images generation using CLIP + VQ-GAN

Text to image generator: CLIP + VQ-GAN

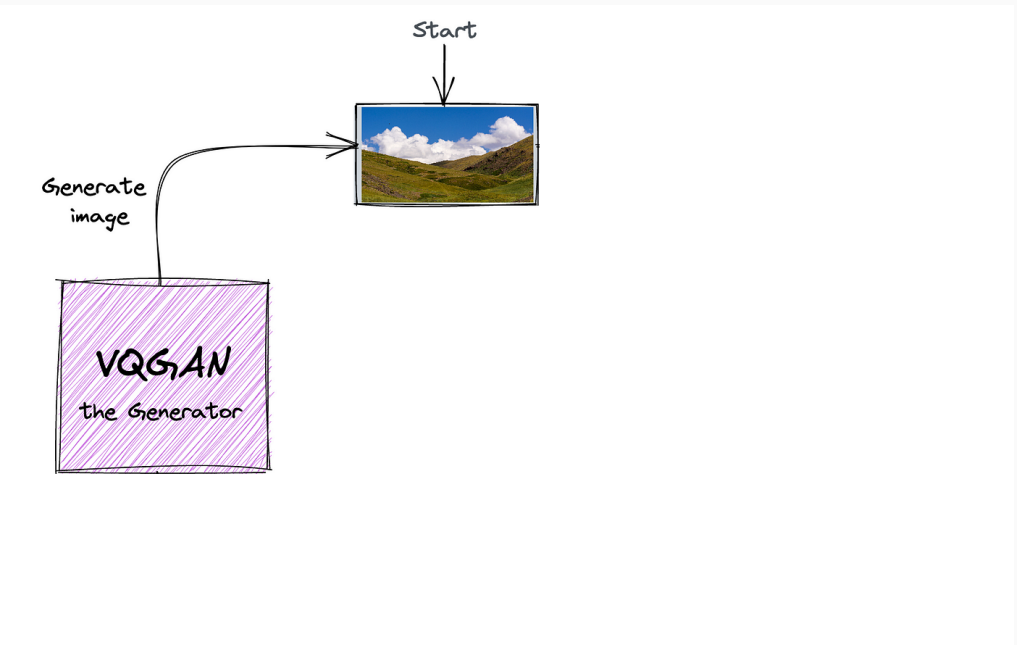


Figure 10: Artificial images generation conditioned on text

Text to image generator: CLIP + VQ-GAN

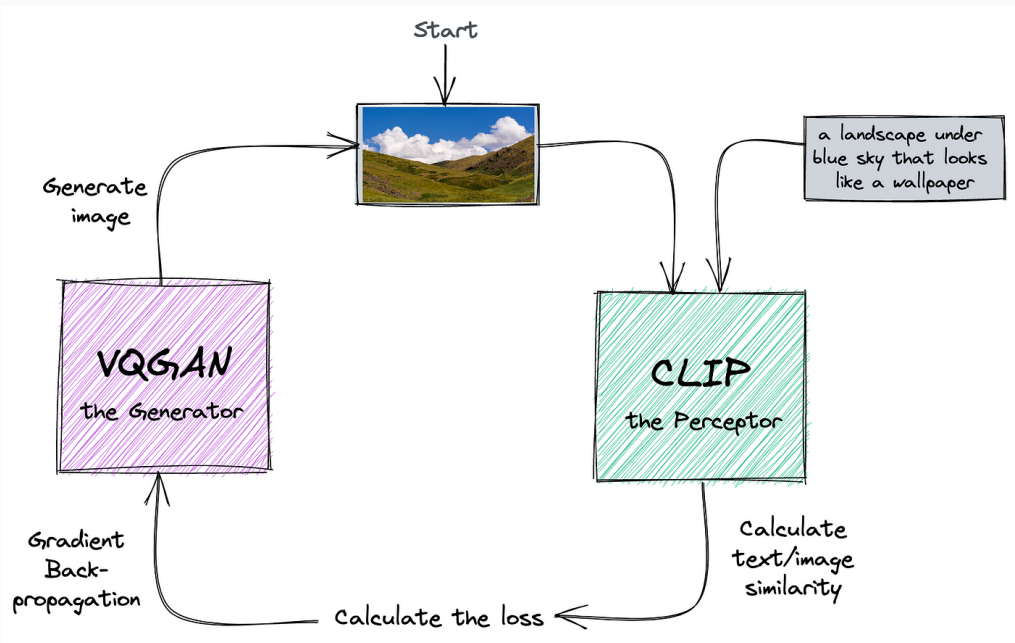


Figure 10: Artificial images generation conditioned on text

Text to image generator: CLIP + VQ-GAN

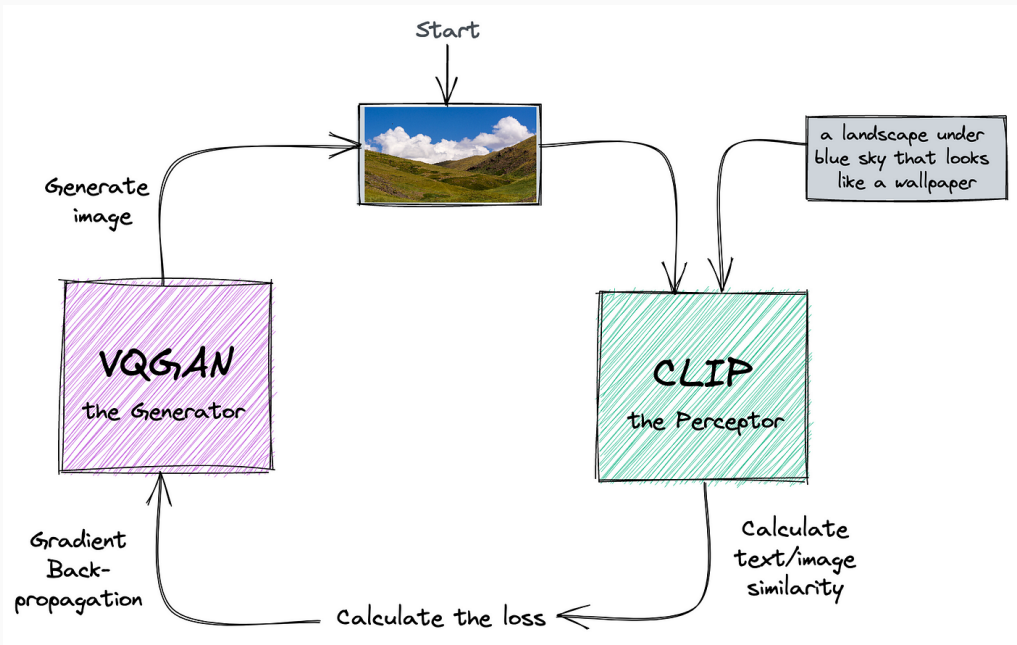
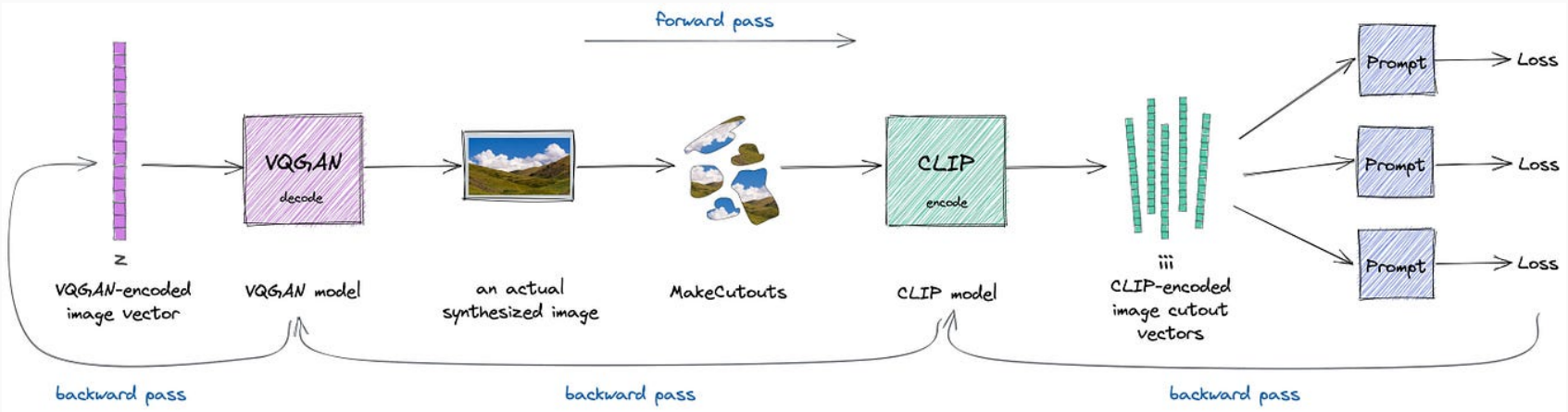


Figure 10: Artificial images generation conditioned on text



Interesting explanations and a notebook with code explanations

Diffusion: New advances in text-to-image I

Diffusion models allow to generate images conditioned on text

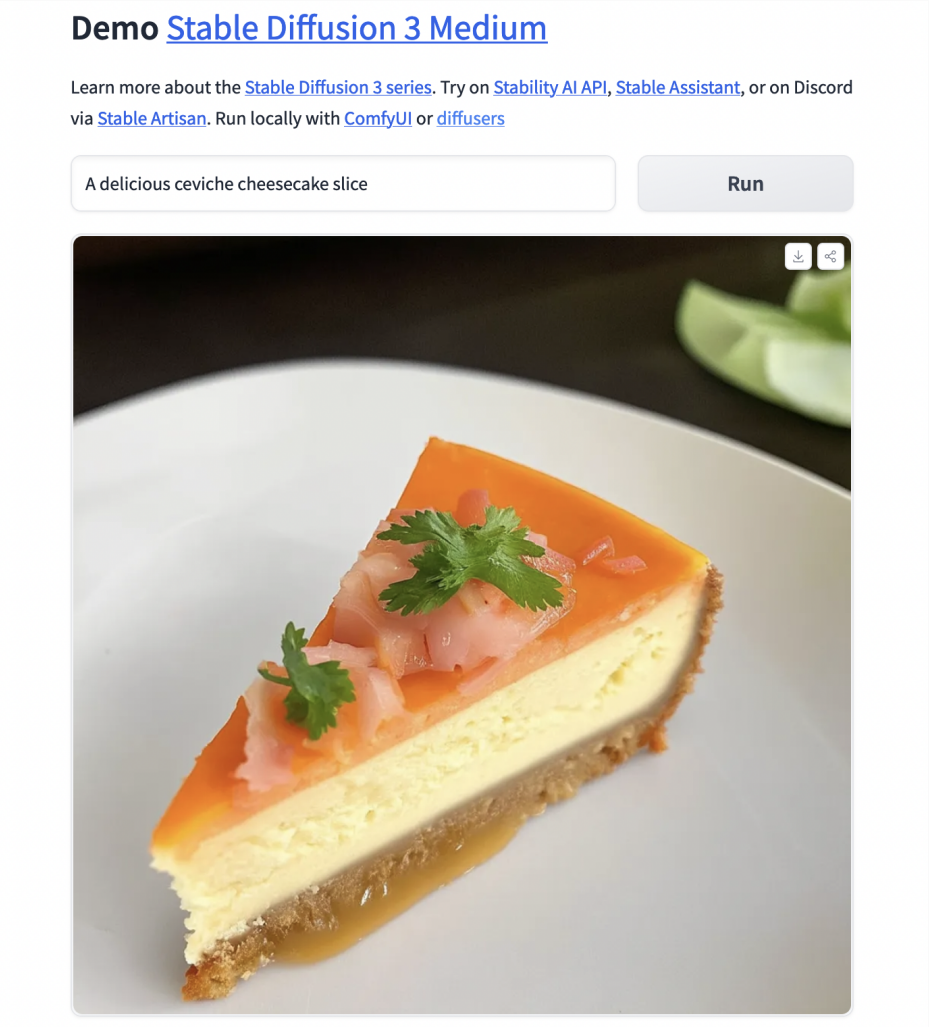


Figure 11: Examples of funny images obtained with Diffusion models

Diffusion: New advances in text-to-image II

Definition

Probabilistic models that learn the data distribution by modeling the reverse of a diffusion process (adding noise step-by-step) to generate new data points.

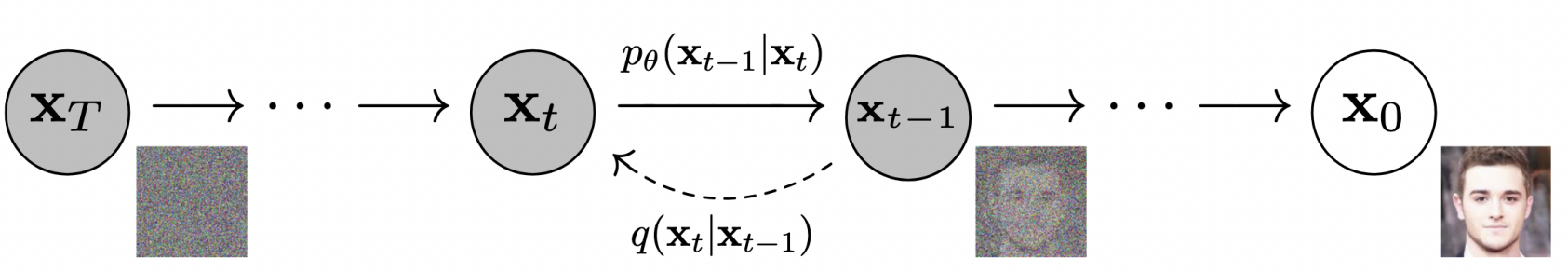


Figure 12: Denoising an image [6]

Process

- **Forward**: Add small amounts of noise to the data, which gradually becomes noisier over time steps, until resembling random noise
- **Reverse**: Learn to gradually remove noise to recover the original data, starting from pure noise and generating realistic data.

More here

Diffusion: Stable Diffusion

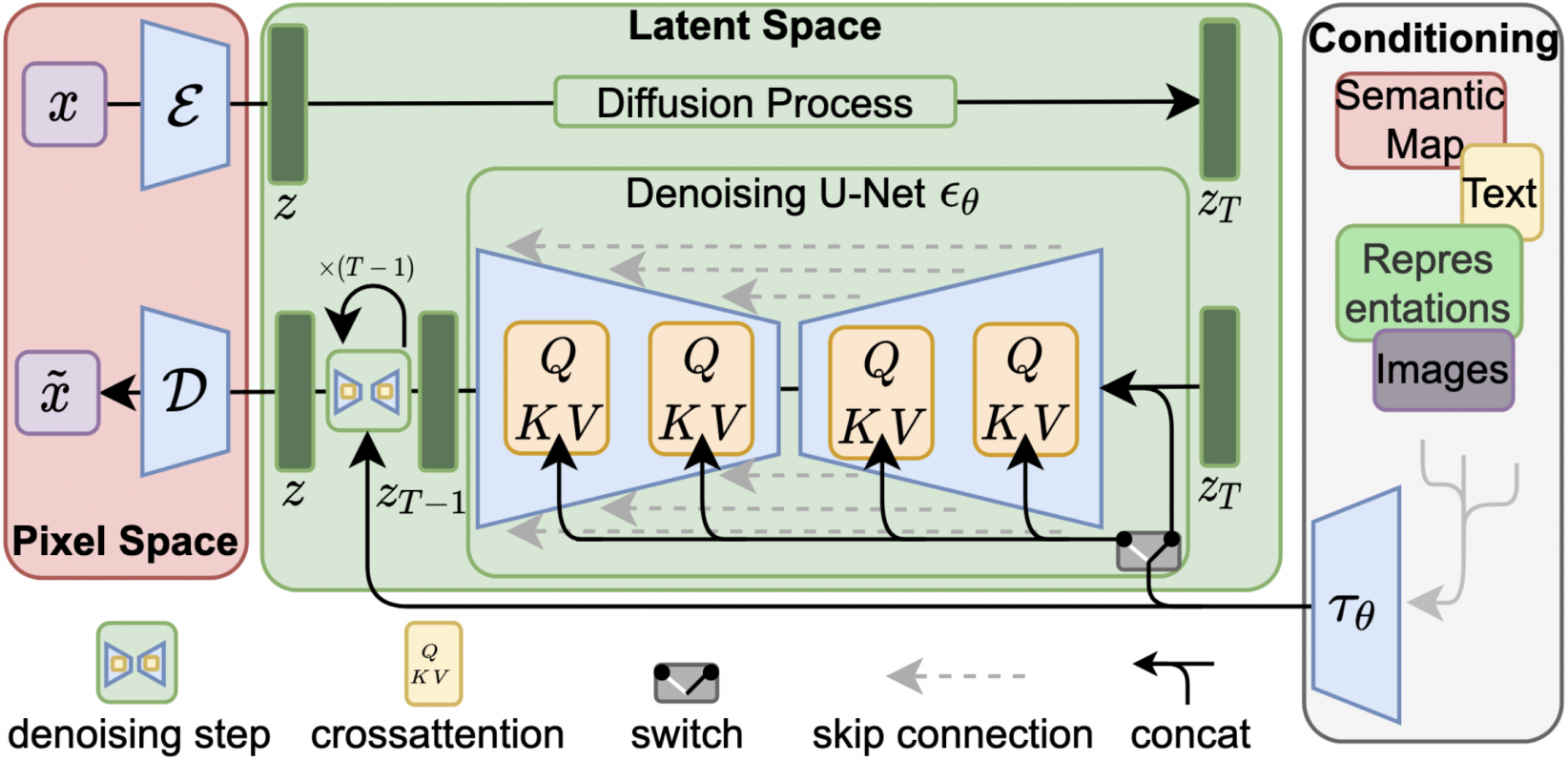


Figure 13: Stable Diffusion architecture [23]

- Use a Denoising Network (e.g., U-Net) to reduce the noise at each step.
- Conditioning (e.g., text embedding) helps guide the model to generate a specific output using x-attention mechanism.

LLM-based: Frozen encoders

- Learn image embeddings to be aligned with a **frozen language model**
- Goal: encode images **into the word embedding space of a LLM**
- The LLM should generate captions for those images
- Very good a few-shot

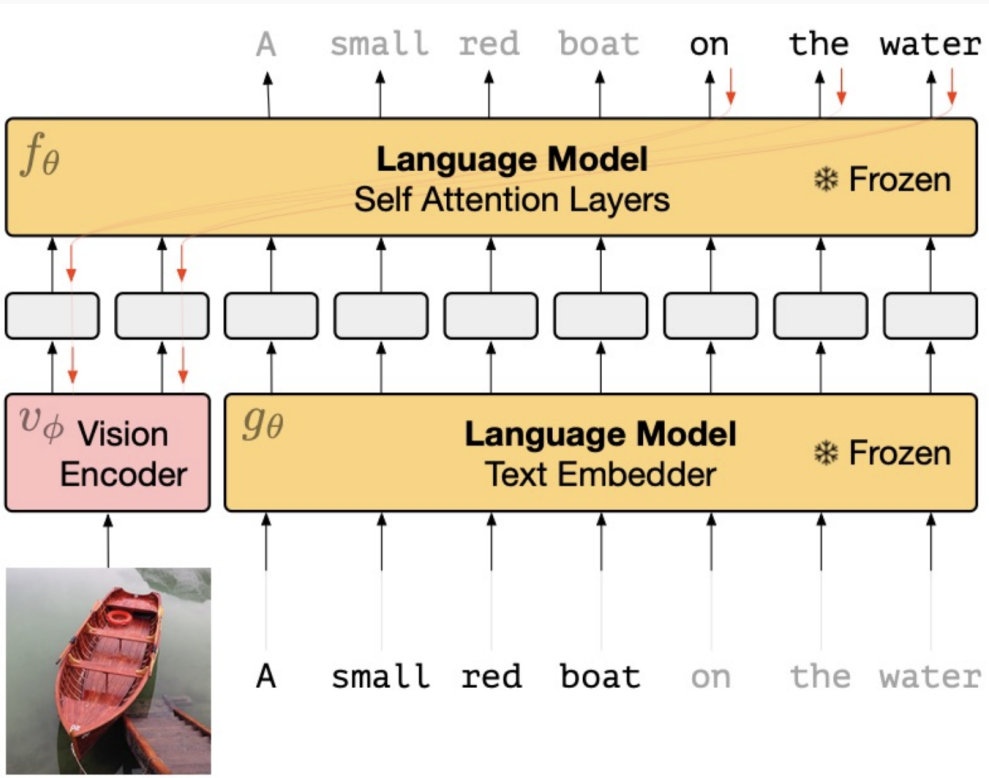


Figure 14: Few shot examples from the model called Frozen [27]

LLM-based: BLIP2 and BLIP3

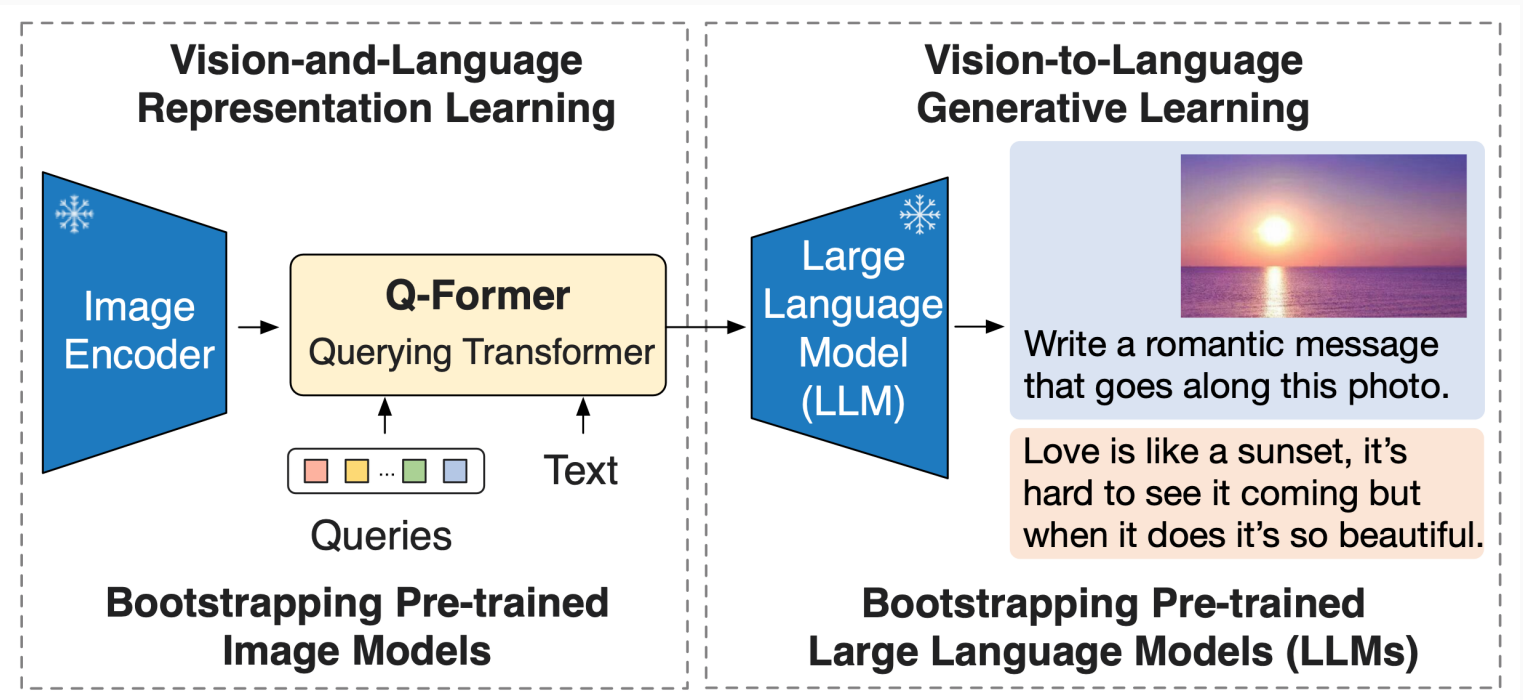


Figure 15: BLIP 2/3 [11, 29] use frozen encoders, just learning the Q-Former

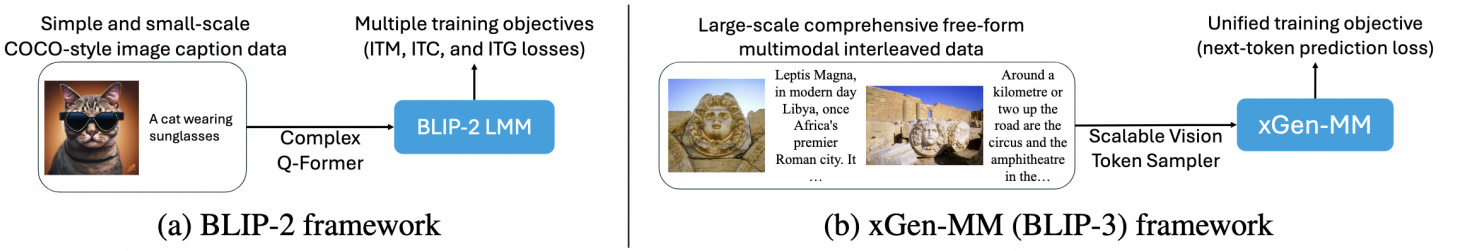


Figure 1: **We introduce xGen-MM (BLIP-3)**, a framework (b) for developing Large Multimodal Models (LMMs). Our framework improves upon BLIP-2 (a) [1] by (1) increasing the richness, scale, and diversity of training data, (2) replacing the Q-Former layers with a more scalable vision token sampler, and (3) simplifying the training process via the unification of the training objectives to a single loss at every training stage. The resulting suite of LMMs can perform various visual language tasks and achieve competitive performance across benchmarks.

LLM-based: xGen-MM (BLIP3) I

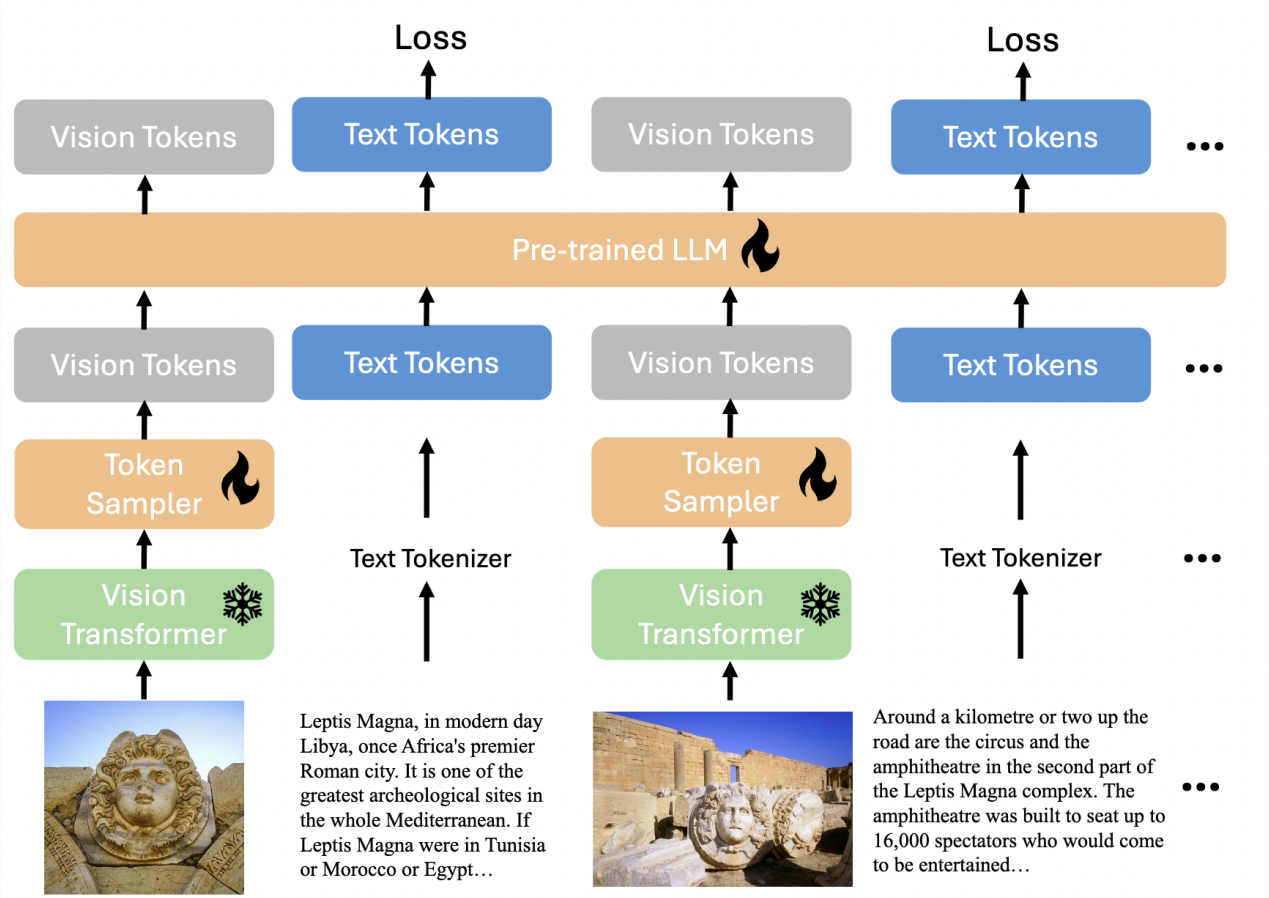


Figure 2: **Overview of the xGen-MM (BLIP-3) framework.** Free-form interleaved images and texts from the ensembled interleaved and caption datasets are input into the framework, with each modality undergoing a separate tokenization process to be fed into the pre-trained LLM in natural order. A standard auto-regressive loss is then applied to the text tokens. The Vision Transformer is kept frozen during training, while all other parameters, including the token sampler and the pre-trained LLM, are trained.

Figure 16: BLIP 3 use frozen encoders, but now relies only on one task of Language Modeling. Models available here

LLM-based: xGen-MM (BLIP3) II

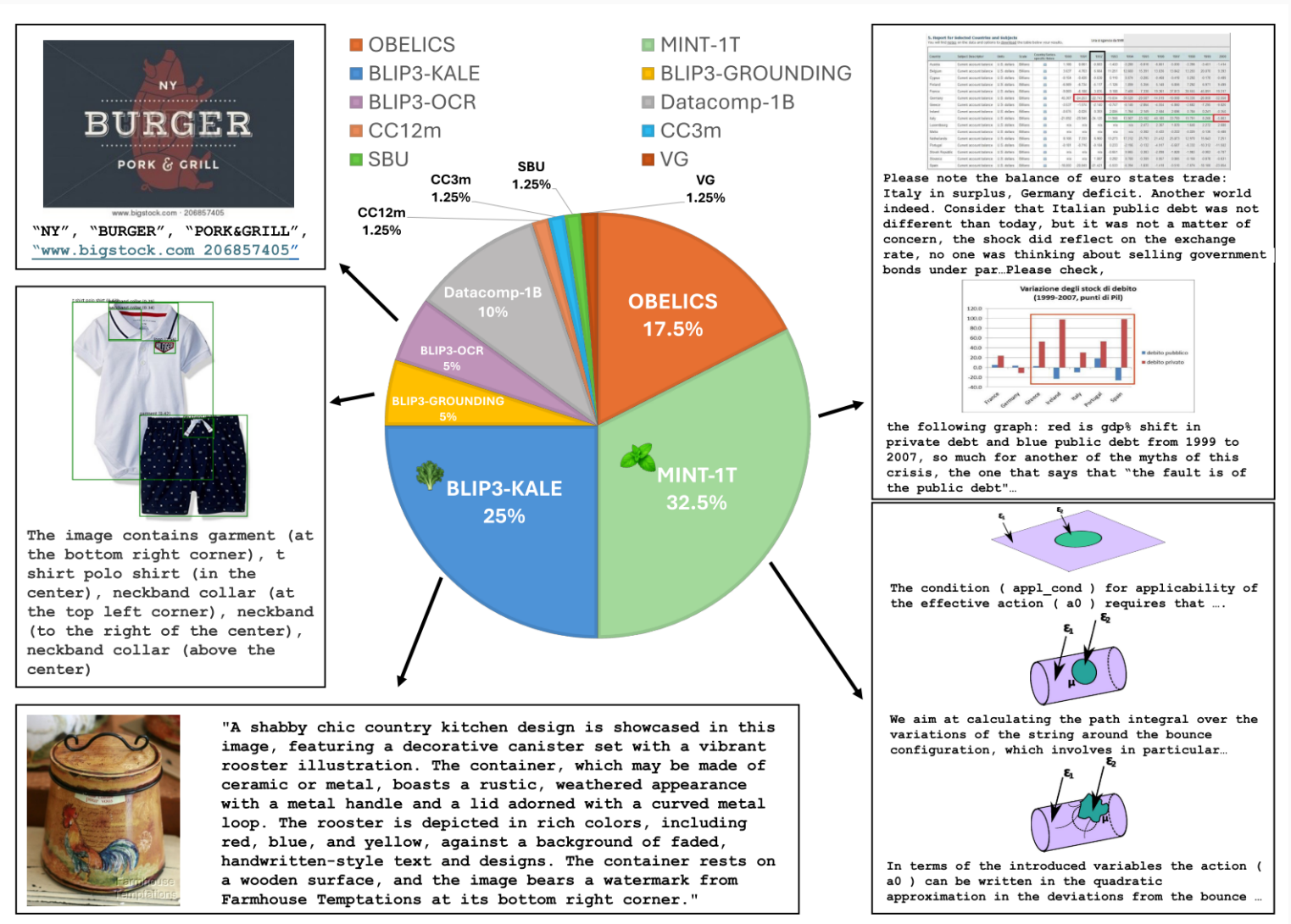


Figure 17: BLIP3 training set contains interleaved multimodal data (mix of sequences of images and text) of very diverse content

LLM-based: Flamingo

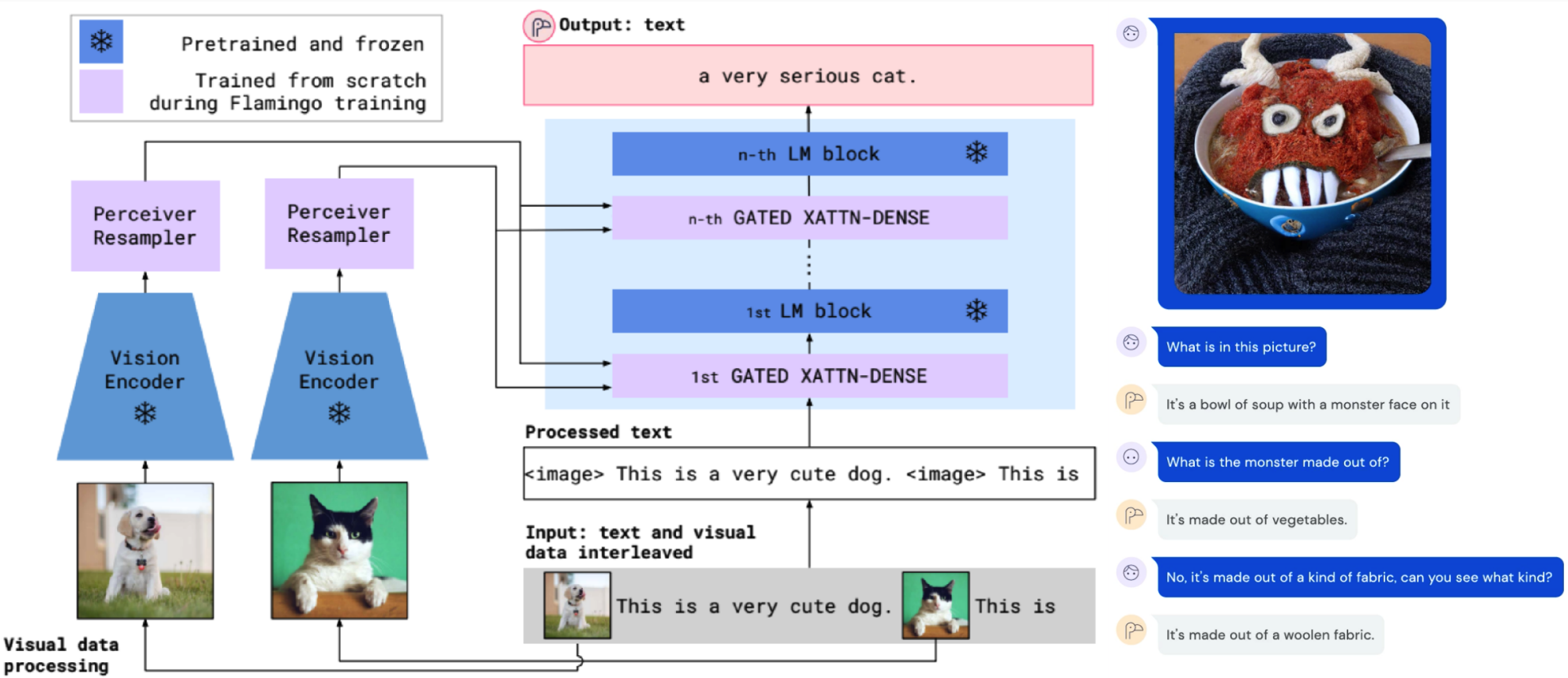
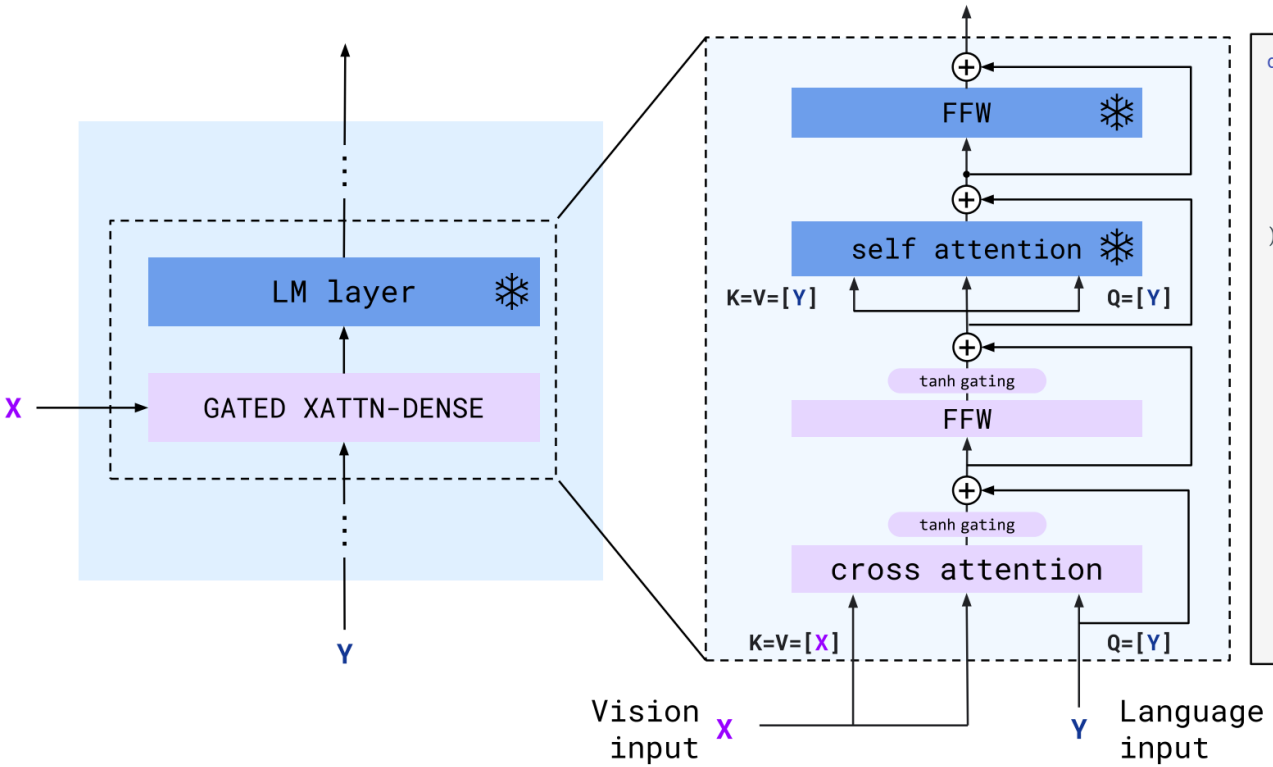


Figure 18: Flamingo was the first LMM that can converse with a human [1]

- Parts trained from scratch, part pre-trained and frozen
- Cross-modal gating mechanism
- Interleaved text with images

LLM-based: Flamingo II



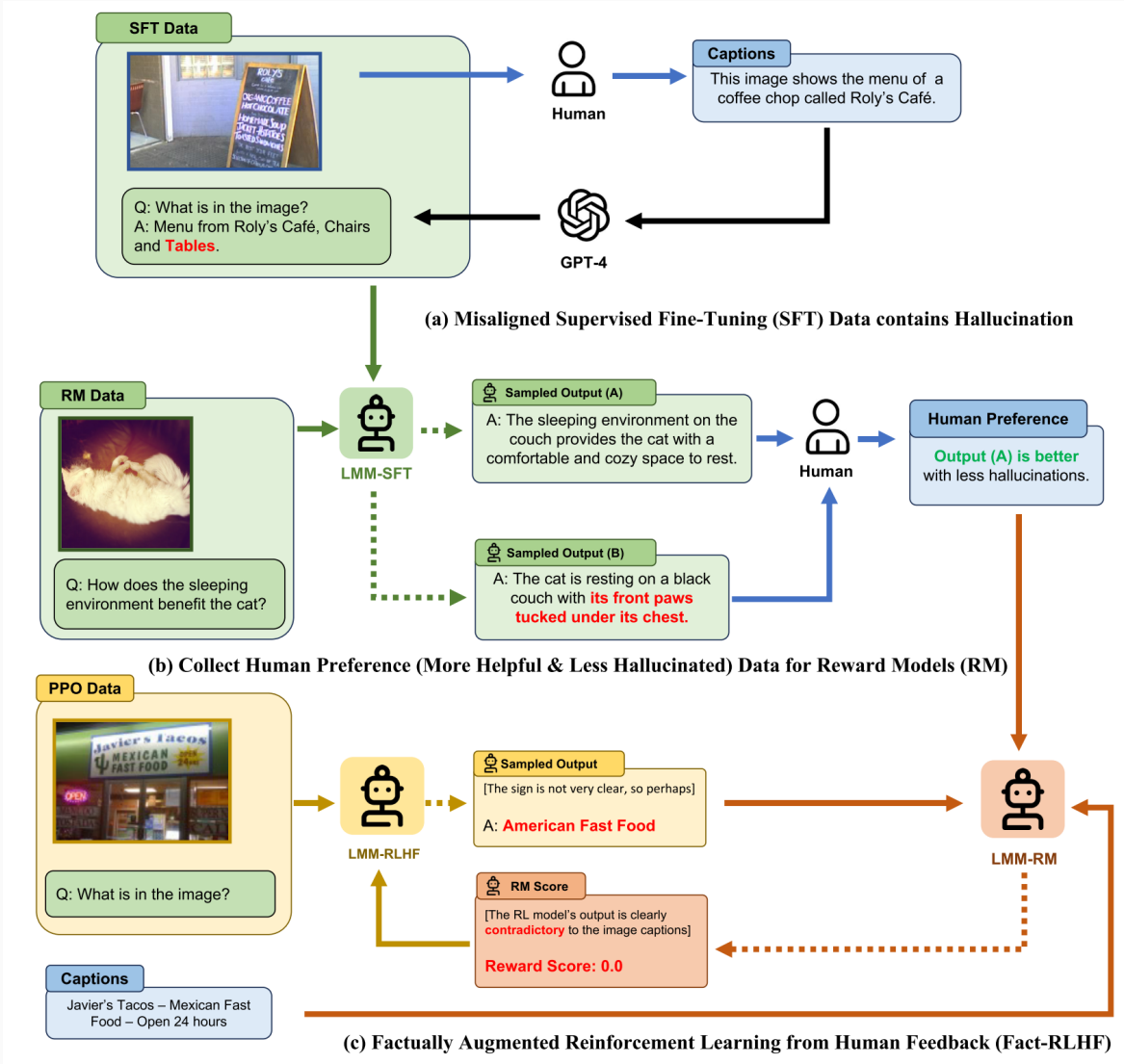

```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

Figure 19: Gated attention mechanism of Flamingo

The LM layers are frozen to preserve its quality obtained from a very large pre-training.

LLM-based: Large Language and Vision Assistant (LLaVA)

- Open-source model based on LLama, adapting it to multimodal data [17]
- Projection matrix from the supplied image using CLIP and feeds it to Llama
- Pre-training of the projection matrix, then fine-tuning adding the LLM weights
- Possible to RLHF [25]



Open-source datasets: OBELICS

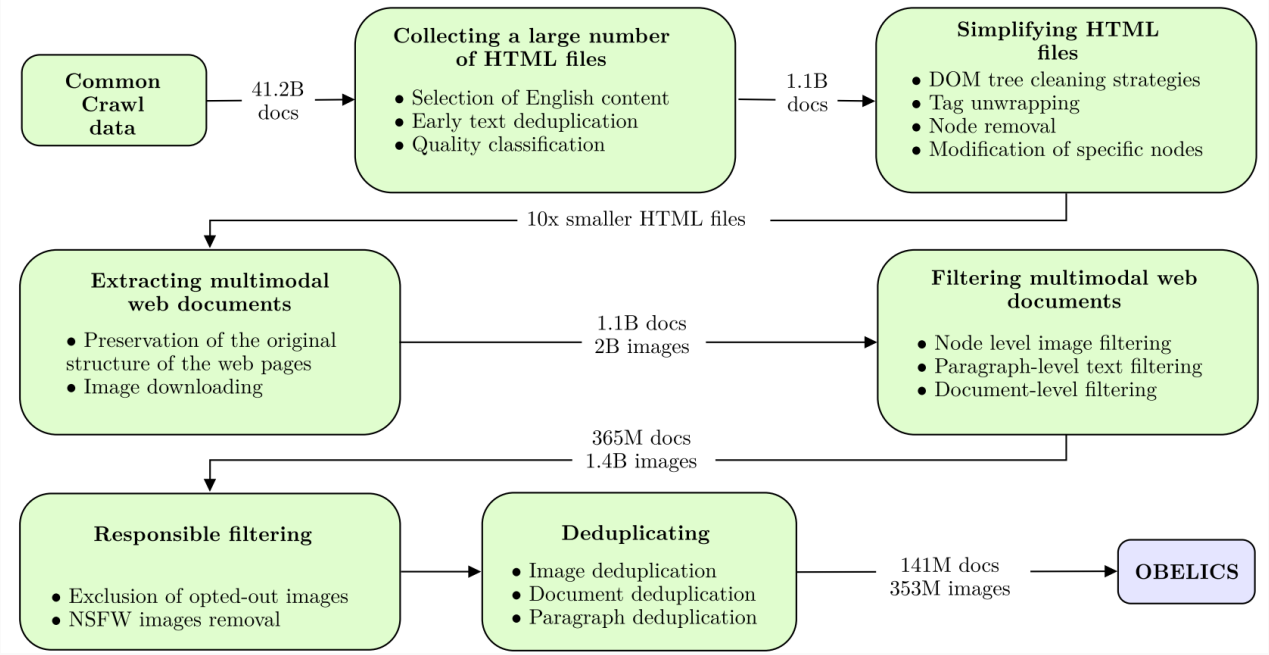


Figure 20: OBELICS [9] is an open, massive and curated collection of interleaved image-text web documents, containing 141M documents, 115B text tokens and 353M images

A serie of models called IDEFICS that were trained on this dataset are available open-source here.

Towards video

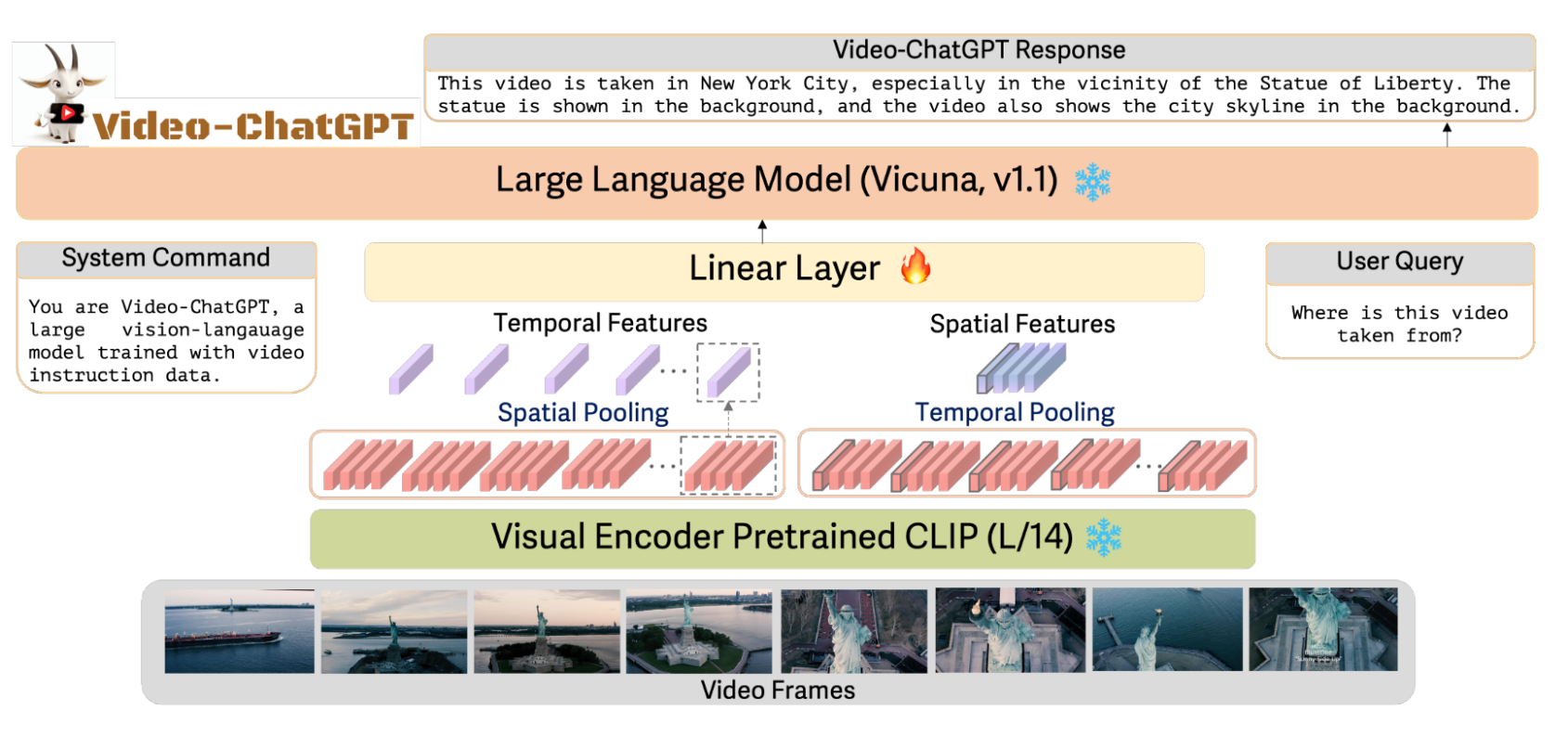
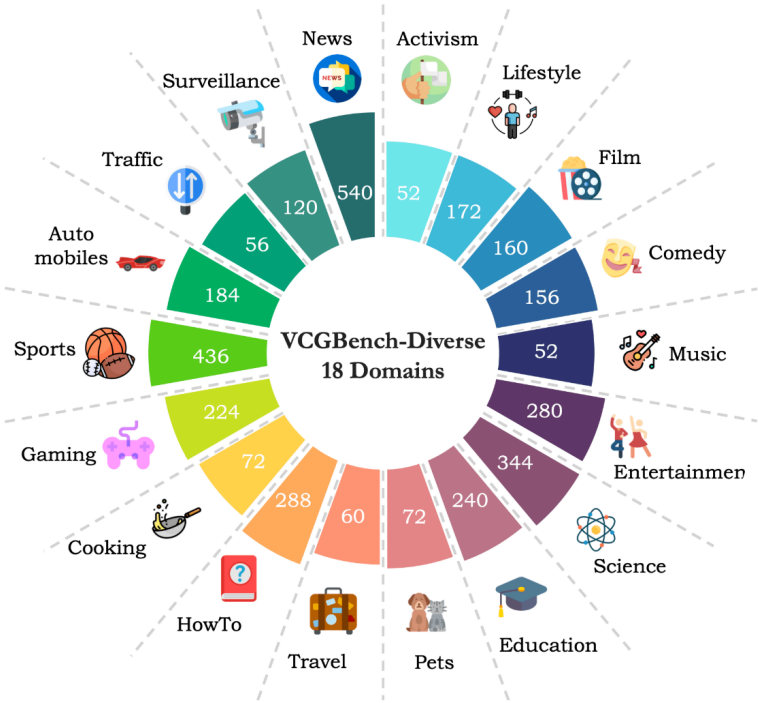


Figure 21: Architecture of Video-ChatGPT, visual encoder extract features that are pooled spatially and temporally [21]

Others architectures available including audio [4, 10, 31, 28, 3, 20]

How to evaluate the video generative LLM?



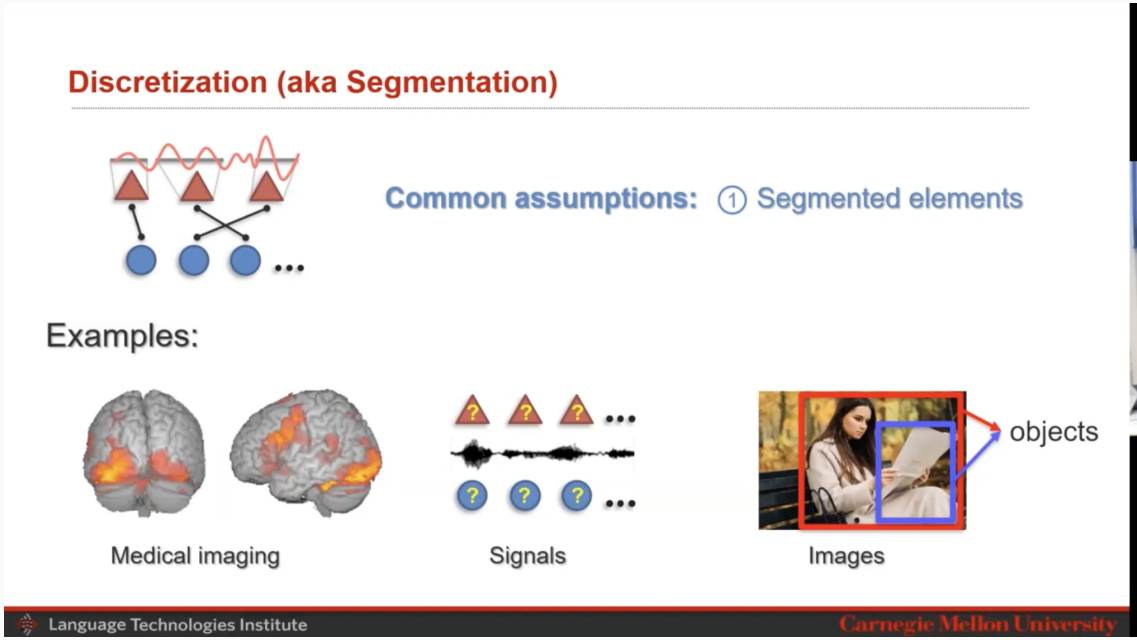
Category	Description	#	Domains
	5 Video Capturing Methods		
Stable Settings	Videos shot in stable, predictable environments with minimal camera movement.	1200	Cooking, How-to, Education
Dynamic Settings	Videos with significant camera movement requiring adaptation to rapid context shifts.	448	Sports, Traffic, Travel
Fixed Cameras	Videos recorded from stationary cameras, providing consistent viewpoints for monitoring purposes.	124	Surveillance, Automobile
Professional Quality	Professionally produced videos with high visual and audio quality, and controlled lighting.	1608	News, Film
Variable Quality	Informal videos with varying quality, often using handheld devices, captured in spontaneous settings.	124	Lifestyle, Pets
	6 Reasoning Complexities		
Sequential Understanding	Requires comprehension and following of a series of steps or actions in order.	828	Cooking, How-to, Education
Predictive Reasoning	Involves understanding and predicting outcomes of dynamic, intricate action sequences.	180	Sports, Gaming
World Knowledge	Demands integration of broader contextual information and world knowledge to interpret video content.	848	Science, News
Causal Reasoning	Focuses on understanding cause-and-effect relationships within the video.	340	Surveillance, Activism
Emotional Reasoning	Involves interpreting stories, character motivations, and emotional subtexts.	1080	Entertainment, Film, Comedy
Analytical Reasoning	Requires critical analysis and interpretation of complex information or situations.	228	Traffic, Automobile

Figure 22: VCGBench-Diverse video conversational benchmark [20]

- 18 broad video categories, with 4,354 QA pairs
- tasks: dense video captioning, spatial and temporal understanding, and complex reasoning
- five video-capturing methods, ensuring diversity and robust generalization and six reasoning complexities

Want to learn more about MModal ML?

The awesome Multimodal Machine Learning course from CMU made by LP Morency and Paul Piu Liang will get you as far as possible!



`https://www.youtube.com/@LPMorency/videos`

You can also check these incredible papers [16, 13]

Questions?

References i

J. B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. C. T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan.
Flamingo: a Visual Language Model for Few-Shot Learning.
In *Advances in Neural Information Processing Systems*, volume 35, 2022.

J. Arevalo, T. Solorio, M. Montes-Y-Gómez, and F. A. González.
Gated multimodal units for information fusion.
In *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 2017.

References ii


S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu.
VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset.
(NeurIPS), 2023.


B. He, H. Li, Y. K. Jang, M. Jia, X. Cao, A. Shah, A. Shrivastava, and S.-N. Lim.
MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding.
In *CVPR*, pages 13504–13514, 2024.


L. Hemamou, G. Felhi, V. Vandenbussche, J.-c. Martin, and C. Clavel.
HireNet : a Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews.
In *AAAI*, 2019.

References iii


J. Ho, A. Jain, and P. Abbeel.
Denoising diffusion probabilistic models.
In *Advances in Neural Information Processing Systems*, volume 2020-Decem, pages 1–25, 2020.


Y.-H. Hubert Tsai, P. Pu Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov.
Learning Factorized Multimodal Representations.
In *ICLR*, number Pas publie, 2019.


W. Kim, B. Son, and I. Kim.
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision.
In *Proceedings of Machine Learning Research*, volume 139, pages 5583–5594, 2021.

References iv

H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh.
OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents.
In *Advances in Neural Information Processing Systems*, volume 36, pages 1–20, 2023.

F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li.
LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models.
2024.

J. Li, D. Li, S. Savarese, and S. Hoi.
BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.
2023.

References

J. Li, D. Li, C. Xiong, and S. Hoi.
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.
(2), 2022.

P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. J. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood, R. Salakhutdinov, and L. P. Morency.
Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework.
Advances in Neural Information Processing Systems, 36(NeurIPS):1–43, 2023.

References vi


P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency.
Multimodal Language Analysis with Recurrent Multistage Fusion.
In *EMNLP*, 2018.


P. P. Liang, A. Zadeh, and L.-P. Morency.
Multimodal Local-Global Ranking Fusion for Emotion Recognition.
2018.


P. P. Liang, A. Zadeh, and L.-P. Morency.
Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions.
(1):1–65, 2022.


H. Liu, C. Li, Y. Li, and Y. Jae.
Improved Baselines with Visual Instruction Tuning.
In *Neurips*, 2023.


Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency.
Efficient Low-rank Multimodal Fusion with Modality-Specific Factors.
In *ACL*, 2018.


J. Lu, D. Batra, D. Parikh, and S. Lee.
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.
In *NIPS'19*, number NeurIPS, pages 1–11, 2019.

References viii


M. Maaz, H. Rasheed, S. Khan, and F. Khan.
VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding.
pages 1–18, 2024.


M. Maaz, H. Rasheed, S. Khan, and F. S. Khan.
Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models.
In *ACL*, volume 1, pages 12585–12602, 2024.


A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever.
Learning Transferable Visual Models From Natural Language Supervision.
2021.


R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer.
High-Resolution Image Synthesis with Latent Diffusion Models.
CVPR, 2022.


S. Sahay, S. H. Kumar, R. Xia, J. Huang, and L. Nachman.
Multimodal Relational Tensor Network for Sentiment and Emotion Classification.
2018.


Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell.
Aligning Large Multimodal Models with Factually Augmented RLHF.
pages 1–26, 2023.

References x


H. Tan and M. Bansal.
LXMERT: Learning Cross-Modality Encoder Representations from Transformers.
In *EMNLP*, pages 5099–5110, 2019.


M. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill.
Multimodal Few-Shot Learning with Frozen Language Models.
2021.


A. J. Wang, L. Li, K. Q. Lin, J. Wang, K. Lin, Z. Yang, L. Wang, and M. Z. Shou.
COSMO: COntrastive Streamlined MultimOdal Model with Interleaved Pre-Training.
2024.

L. Xue, M. Shu, A. Awadalla, J. Wang, and A. Yan.
xGen-MM (BLIP-3): A Family of Open Large Multimodal Models.
pages 7–11, 2024.

A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency.
Tensor Fusion Network for Multimodal Sentiment Analysis.
In *EMNLP*, 2017.

Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar.
Learning Video Representations from Large Language Models.
In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2023-June, pages 6586–6597, 2023.