Montreal, June 15–19

DLMI2026

Advanced Concepts in Deep Learning – Part I: attention, transformers, foundation models

Nicolas Thome
Prof. at Sorbonne University ISIR Lab, MLIA TEAM
On leave at ILLS, CNRS, Montreal

Fonds de recherche Nature et technologies Québec

cnrs

itech santé

CIMS Centre d'imagerie moléculaire de Sherbrooke

CHUS Centre hospitalier universitaire de Sherbrooke

SORBONNE UNIVERSITÉ

MICCAI

ÉTS Le génie pour l'industrie

CREATIS

UDS Université de Sherbrooke

ILLS International Laboratory on Learning Systems

ISIR Institut des Systèmes Intelligents et de Robotique

MLIA Machine Learning & Deep Learning for Information Access

Transformers everywhere since 2017

NLP: BERT, GPT-3/4, Chat-GPT, *etc*

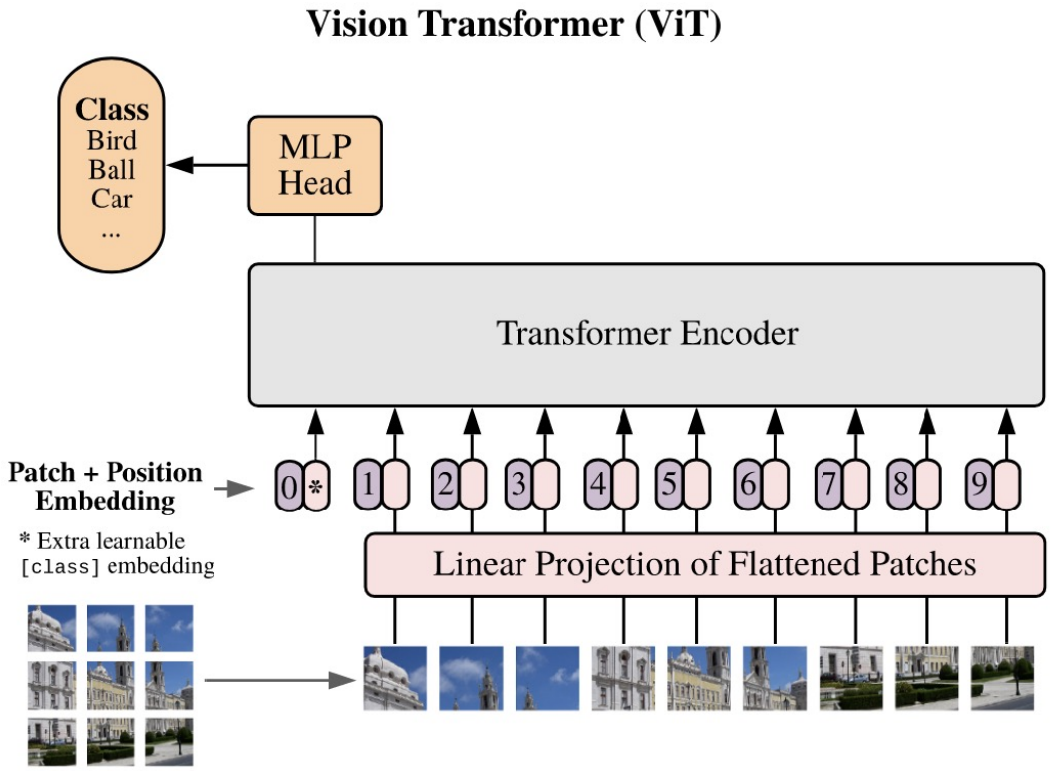
explain quantum theory to a child in the style of snoop dogg

Yo, little homie, let me break it down for you
Quantum theory, it's a crazy thing, it's true
See, the world is made up of tiny, tiny particles
That are always moving, and changing, and actin' wild

These particles, they don't always behave like we think
Sometimes they're waves, and sometimes they're points, it's a trip
They can be in two places at once, and they can be in two states
It's a mind-blowing concept, but that's the quantum fate

So don't be scared, little homie, it's all good
Quantum theory, it's just the way the world works, it's understood
It's a wild ride, but it's a cool one too
Quantum theory, it's the way the world does what it do.

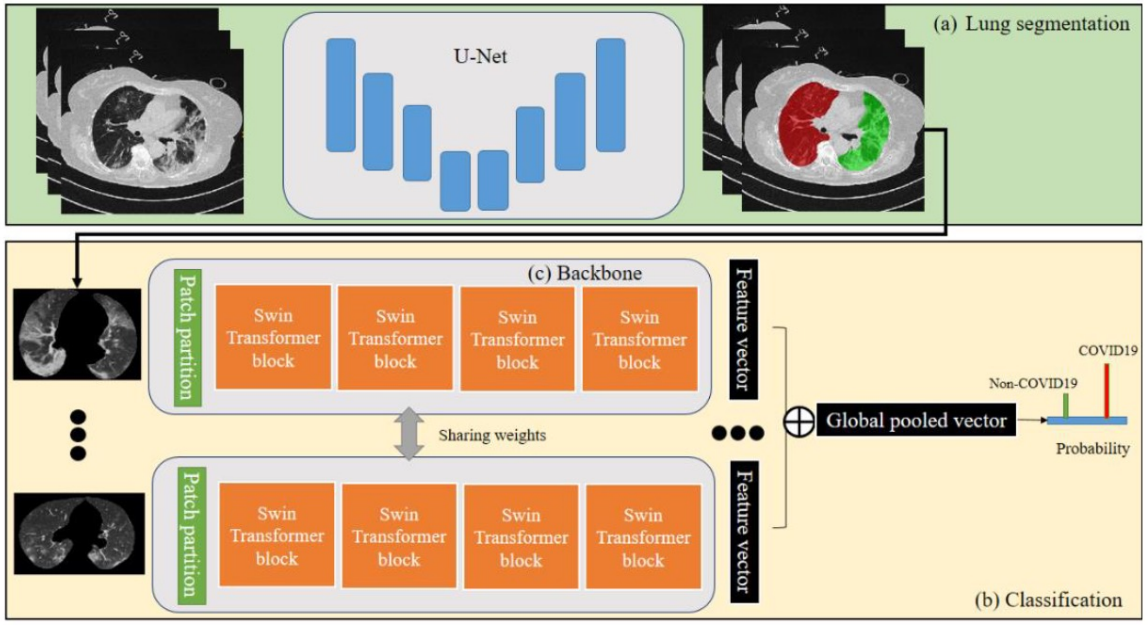
Vision since ’21: Vision Image Transformer (ViT)



Transformer in medical image analysis

Used in various contexts and tasks

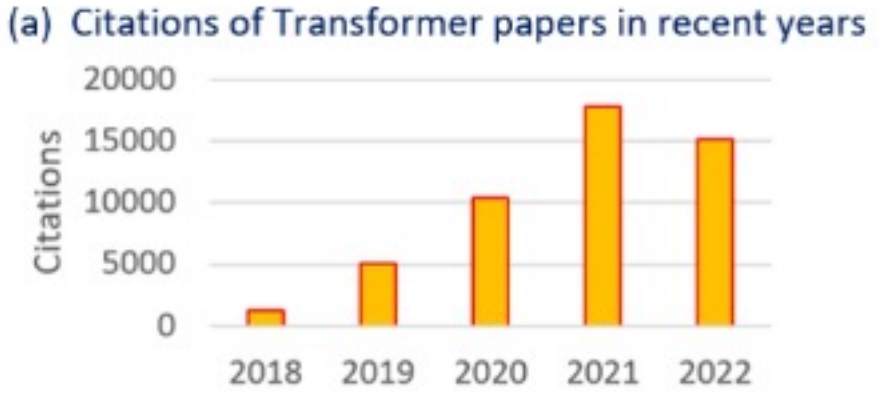
- Image classification, detection, *e.g.* COVID, Semantic segmentation
- Image Registration
- Image Generation
- Im-2-im translation

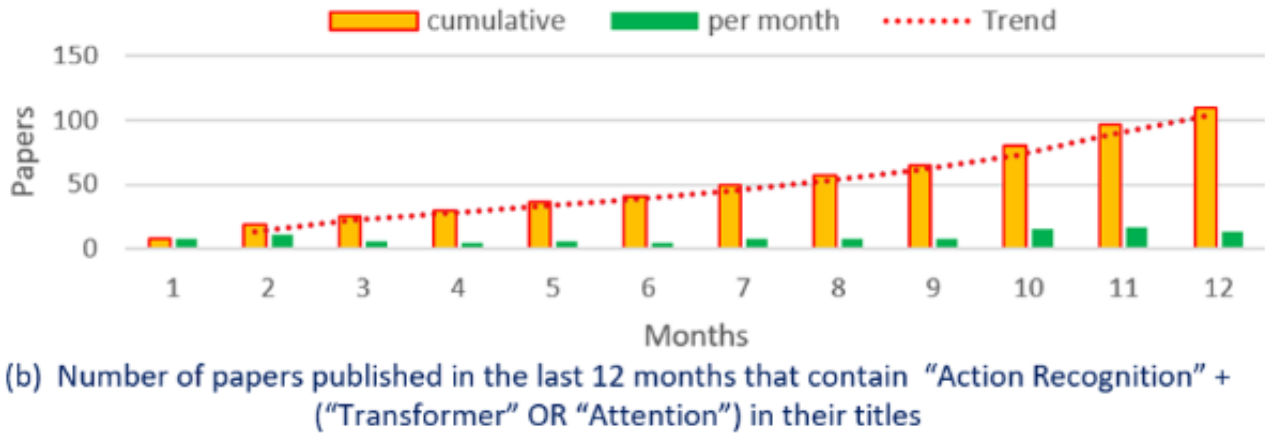


Zhang L, Wen Y. Mia-cov19d: A transformer-based framework for covid19 classification in chest cts. arXiv, 2021.

Focus on this talk

- Paper on transformer every day…





(b) Number of papers published in the last 12 months that contain "Action Recognition" + ("Transformer" OR "Attention") in their titles

- By no means exhaustive literature review

Intelligent Medicine 3 (2023) 59–78

Contents lists available at ScienceDirect

Intelligent Medicine

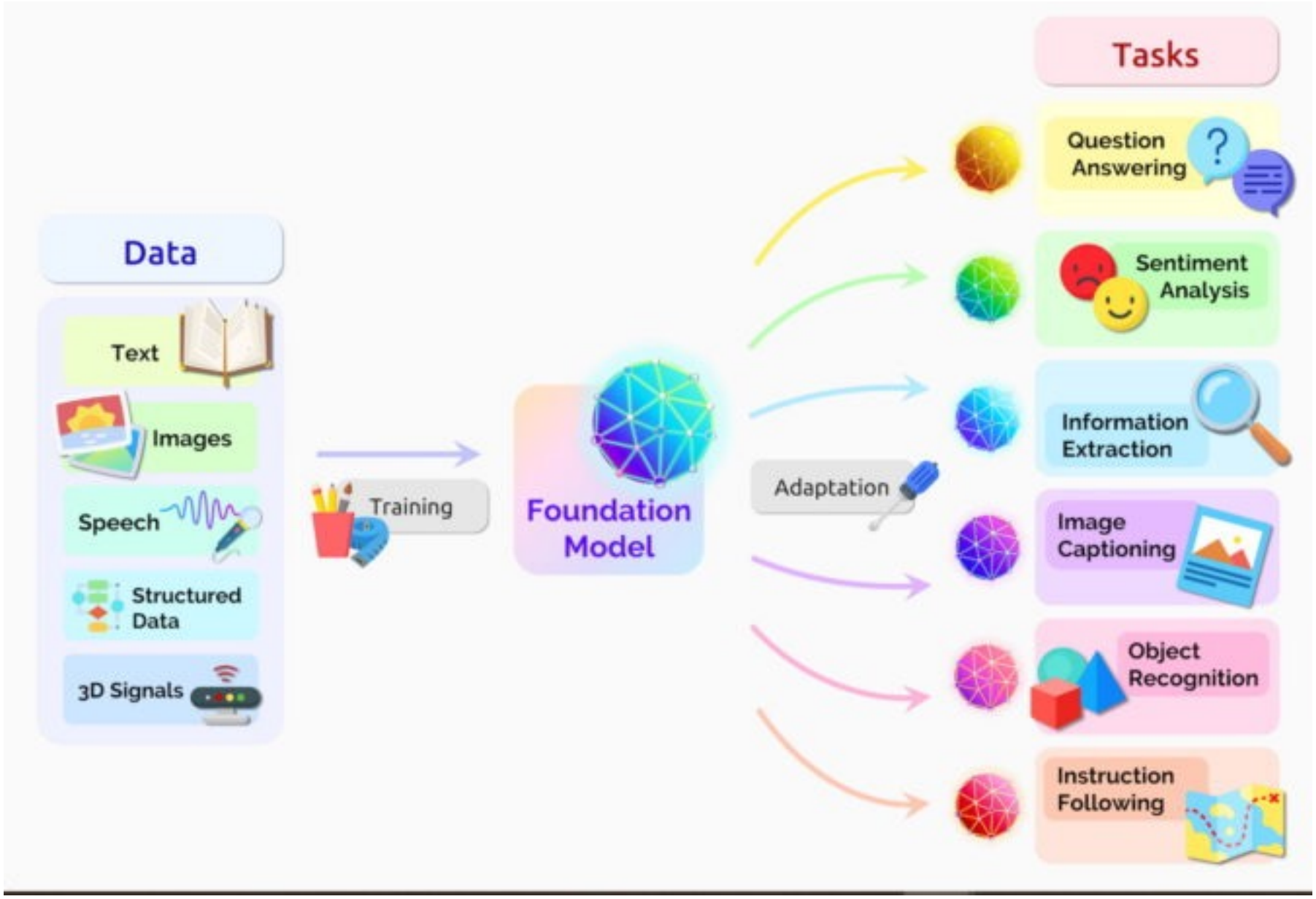
journal homepage: www.elsevier.com/locate/imed

Review

Transformers in medical image analysis

Kelei He[1,2,#], Chen Gan[2,#], Zhuoyuan Li[1,2,#], Islem Rekik[3,4,#], Zihao Yin[2], Wen Ji[2], Yang Gao[2,5], Qian Wang[6,*], Junfeng Zhang[1,2,*], Dinggang Shen[6,7,8,*]

Foundation models



- Internet-scale pre-training
- Based on transformers
- Applicable to a wide range of downstream tasks

Focus on this talk

1. Transformers: building blocks
2. Transformers in vision & medical image segmentation
3. Foundation Models

From sequence to set

- Break down input into *tokens*, i.e. vectors
- Structured sequence of elements → a **set** of tokens, no order
  - Token: primitives, elementary elements of data
    - Text: token are e.g. words
    - Image: token are e.g. patches

Text

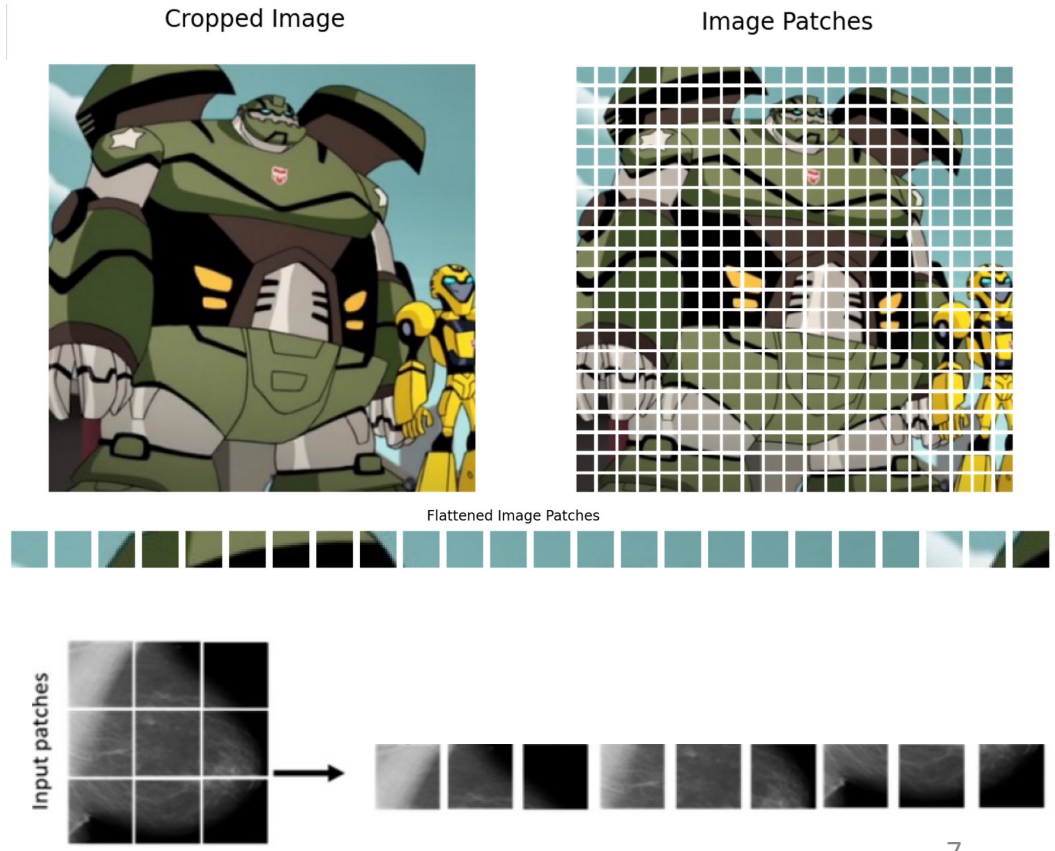
"Hello I love you"

Tokenization

"Hello", "I", "love", "you"

"love", "Hello", "I", "you"

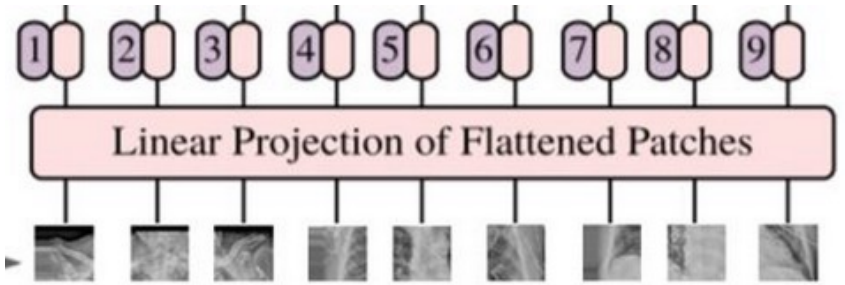
"I", "love", "Hello", "you"



Input embedding

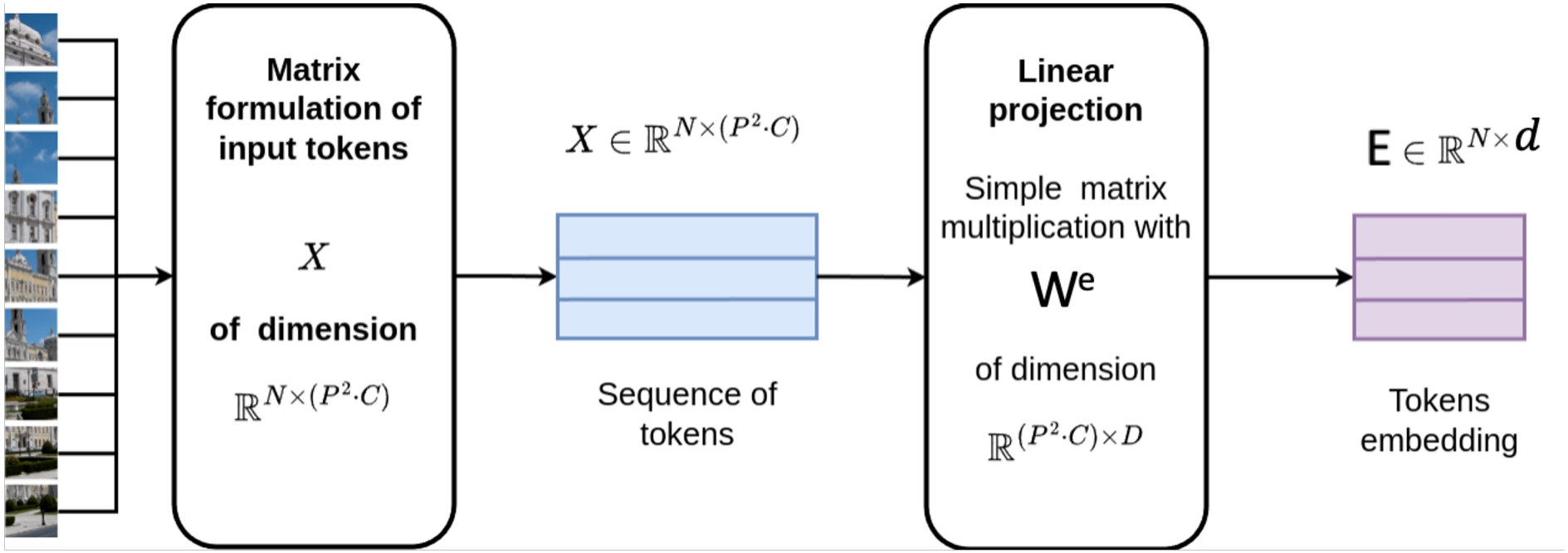
- Token: input vector in $\mathbb{R}^t$
  - Word: t =|V|, V vocabulary
  - Image patch: t = $\mathbb{R}^{PxPxC}$, P patch size, C # channels

- Input embedding: linear projection $\mathbb{R}^t \rightarrow \mathbb{R}^d : e_i = x_i W^e$



Input embedding

$$E = X W^e$$



Positional encoding

- Structed input (text, image) → set of token:
  - Permutation invariant
  - Loosing structural information from data
- Recovering structure: **positional encoding (PE)**
  - Mapping token position t to a vector $\mathbf{p}_t \in \mathbb{R}^d$
  - Seminal PE: sinusoidal

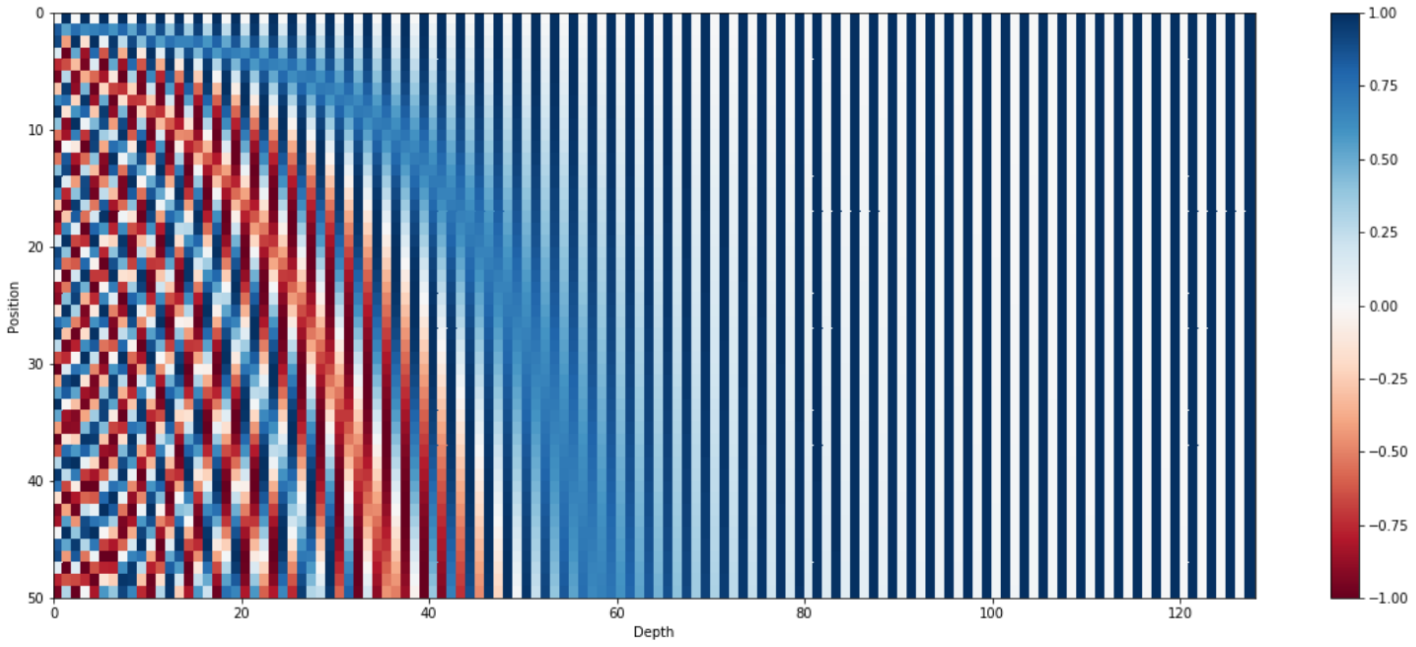
$$\overrightarrow{p}^{(i)} := \begin{cases} \sin(\omega_k . t), & \text{if } i = 2k \\ \cos(\omega_k . t), & \text{if } i = 2k+1 \end{cases}$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

$$\overrightarrow{p} = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}_{d \times 1}$$

Sinusoidal positional encoding

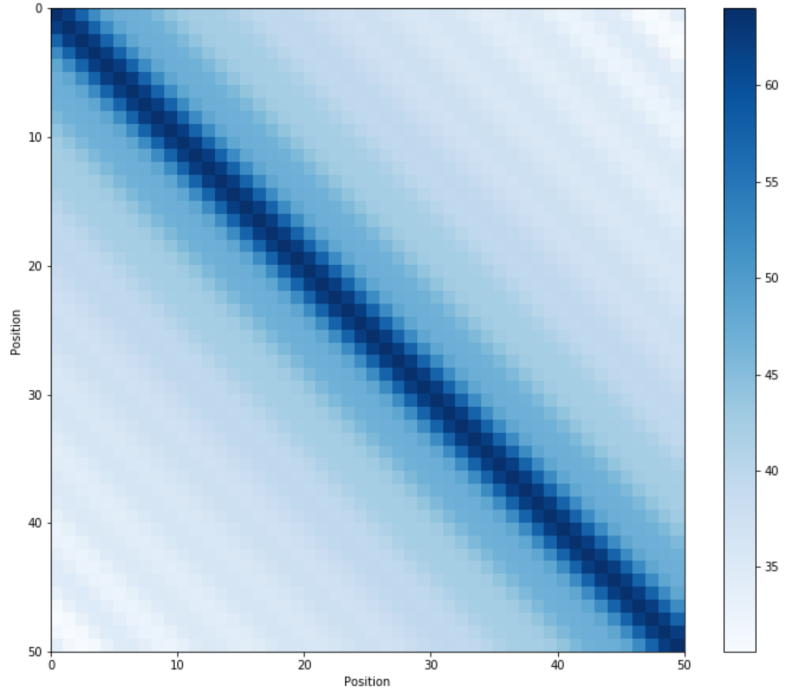
- Unique vector $\mathbf{p}_t$ for each position t
- $p_t(i) \in [-1;1]$: natural normalization



d=128, max length of token set = 50

- Model relative position
- Positional similarity:

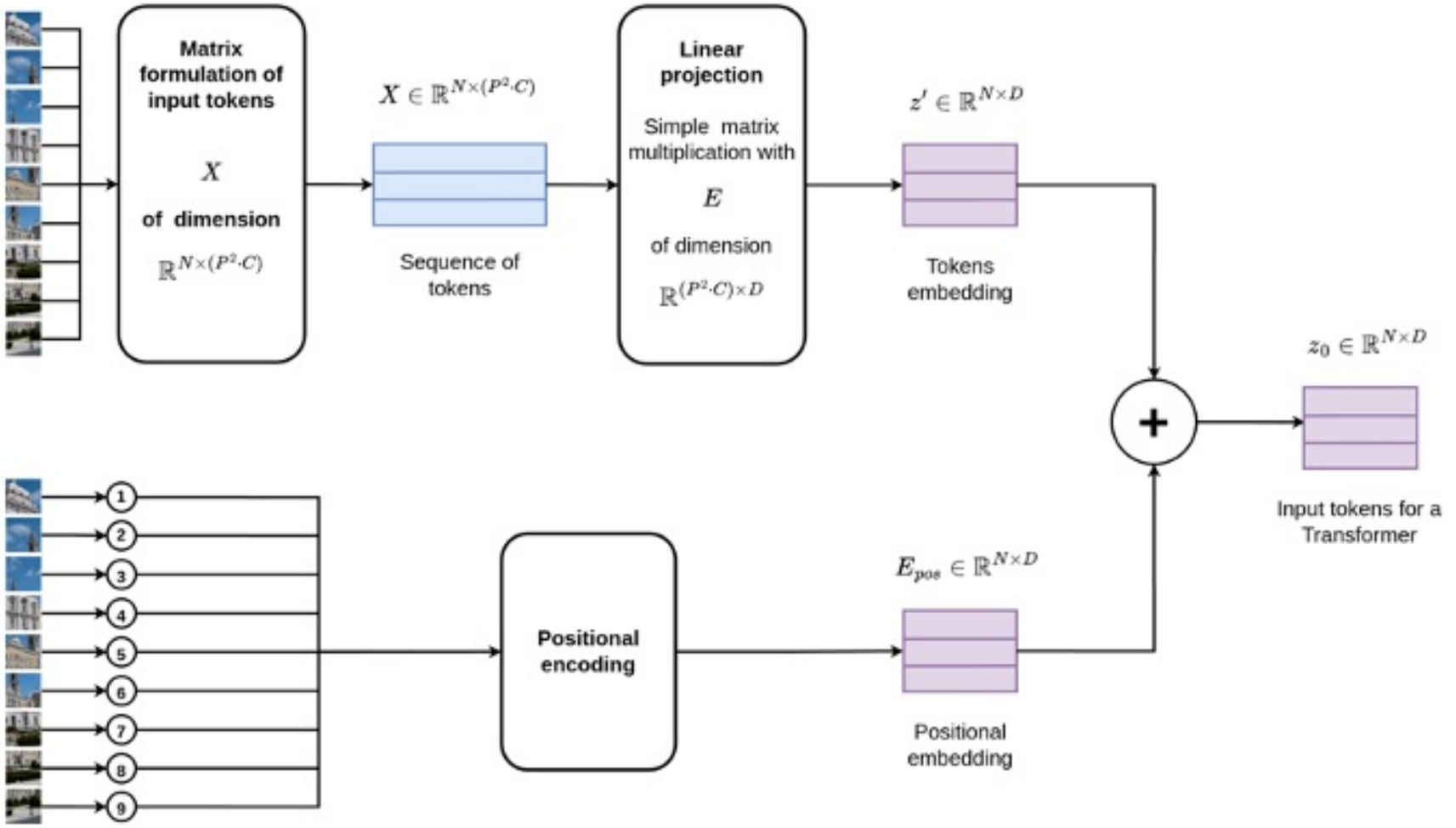
$$K = PP^t$$



Positional encoding

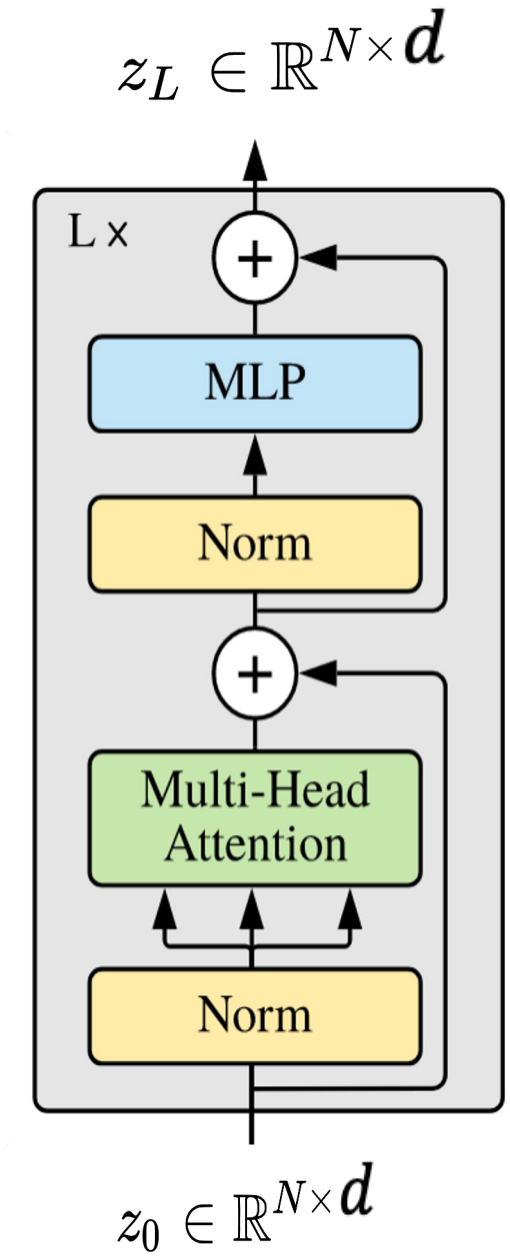
- Other possible encoding, can be learned
- Final embedding :

$$E_{\text{pos}} \in \mathbb{R}^{N \times D} \sim \mathcal{N}(0, 0.02)$$



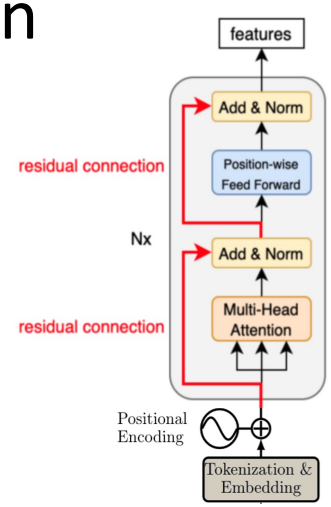
=> Input of transformer!

Transformer [1]: the encoder



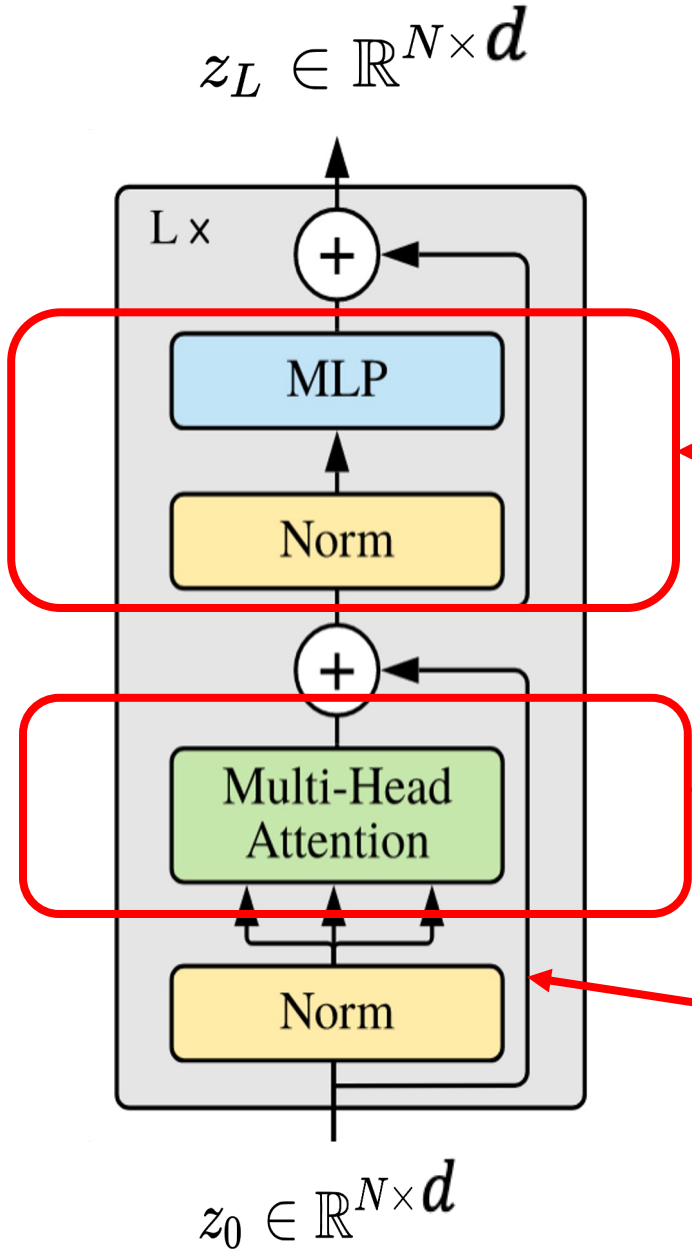
- A stack a N transformer blocks
  - Input a set of embedded tokens
  - Output: a set of re-embedded tokens

- Note: Pre (vs Post) normalization in most modern archis (≠ [1])



[1] Attention Is All You Need. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. NeurIPS 2017.

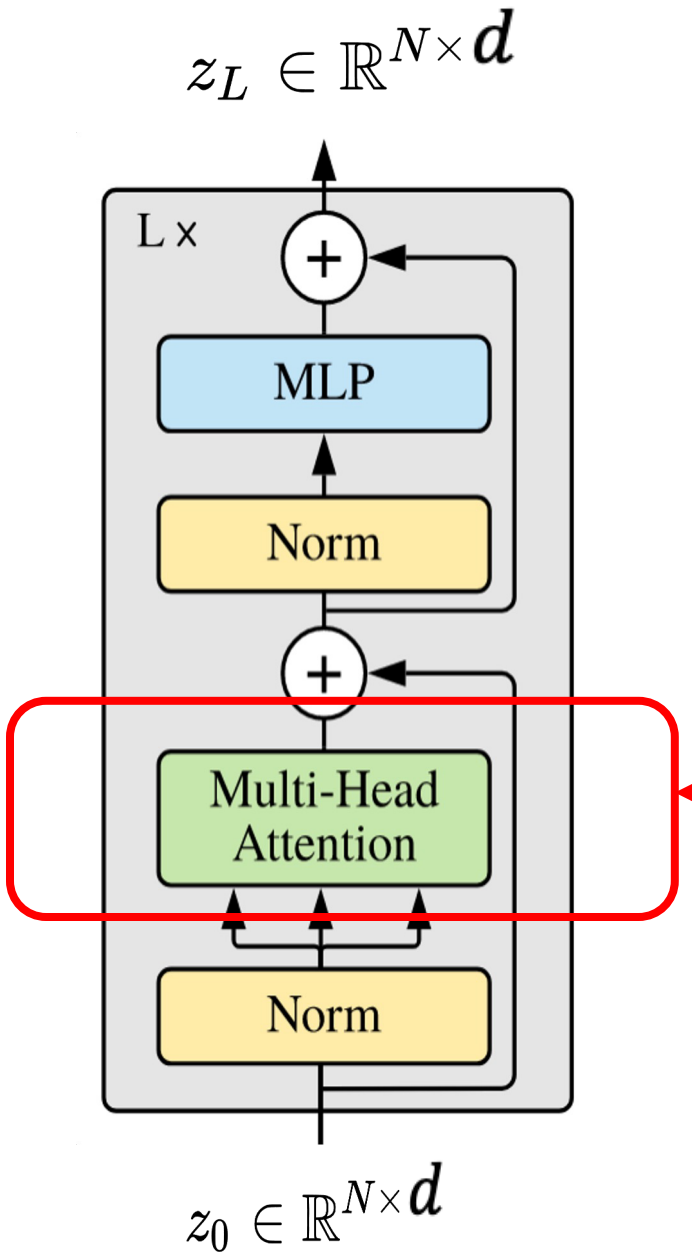
Transformer: the encoder



- **Transformer block**
- **Intra-token computation**
  - MLP, normalization
- **Inter-token computation**
  - Attention
- **Residual connections**
  - Propagate gradients + PE

[1] Attention Is All You Need. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. NeurIPS 2017.

Transformer: the encoder



- **Transformer block**

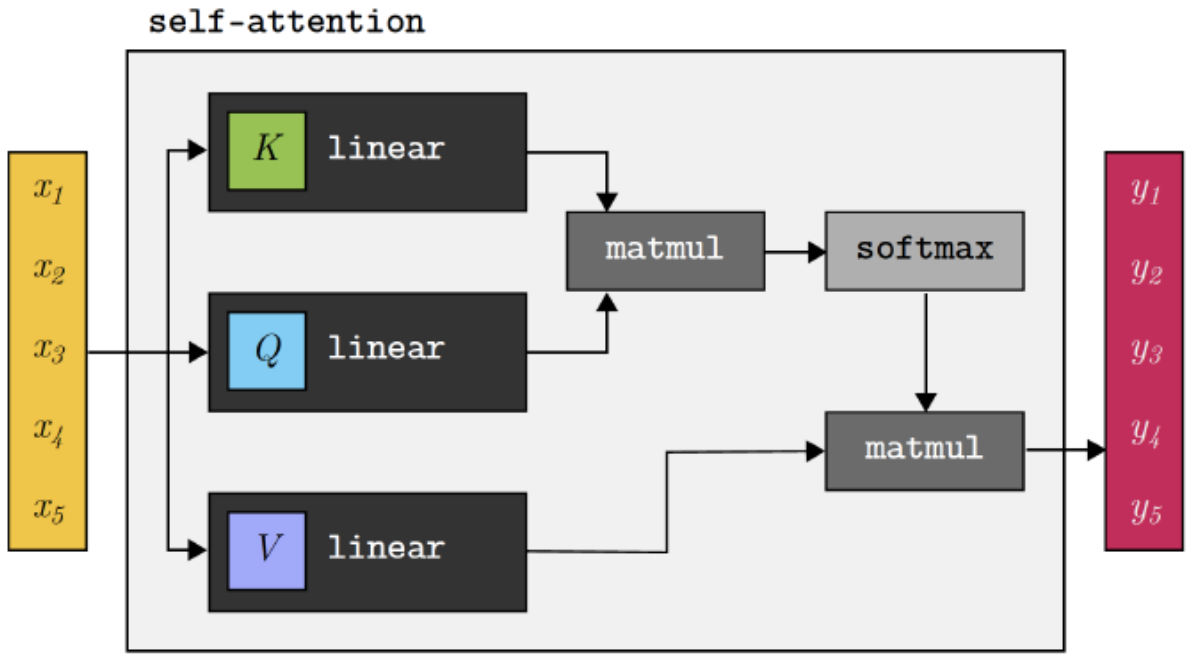
- **Intra-token computation**
  - MLP, normalization

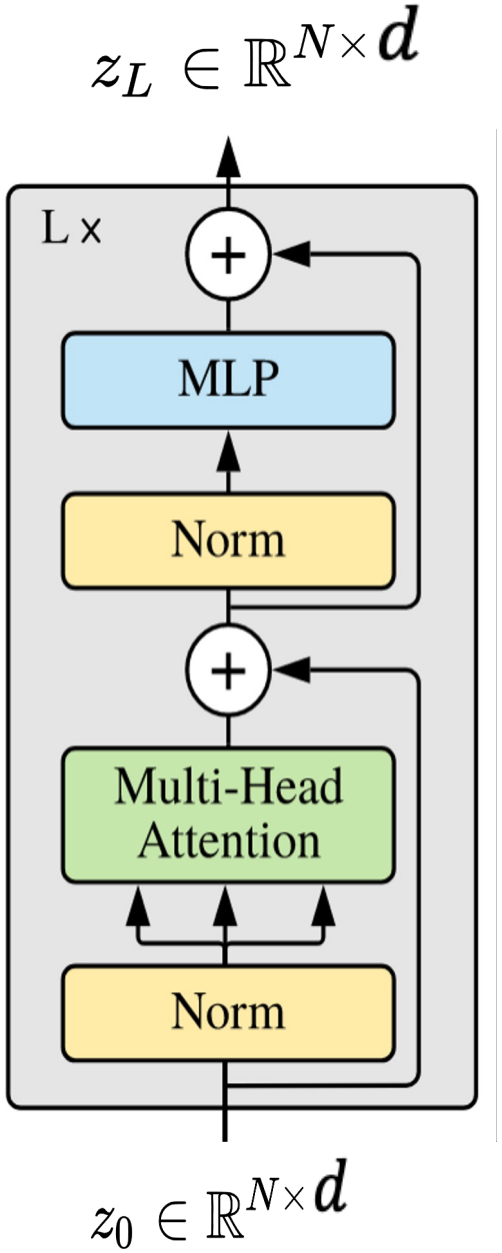
- **Inter-token computation**
  - Attention

- **Residual connections**
  - Propagate gradients + PE

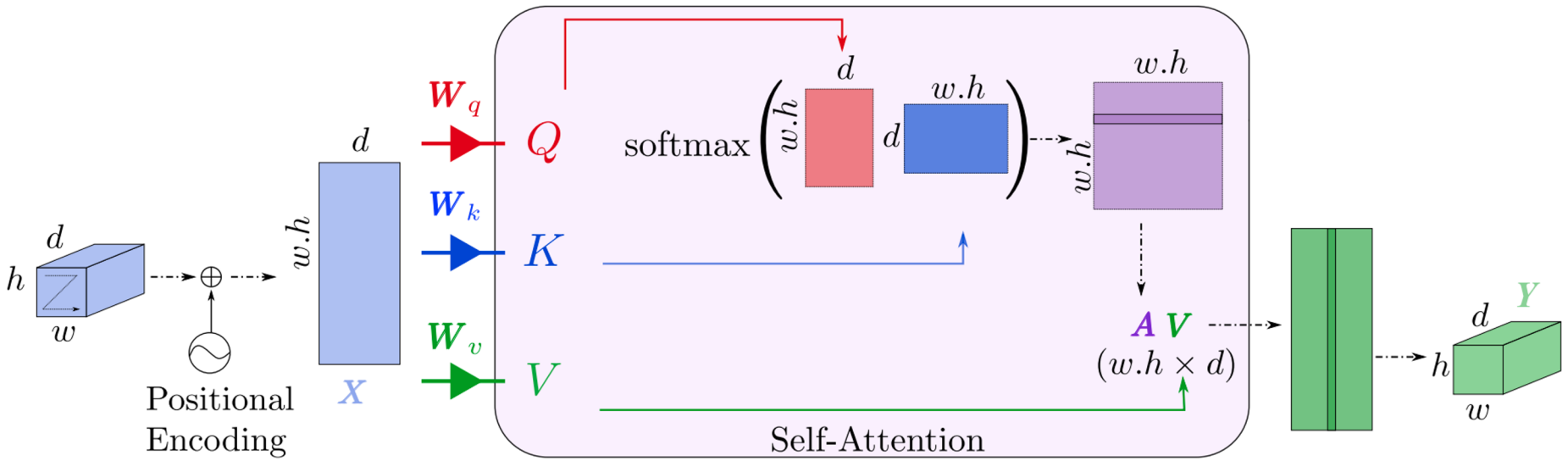
Transformer: self attention

- **The most important and specific module in transformers**
- Project the input set into 3 sets
  - Query: sought info
  - Key: context elements
  - Value: retrieved





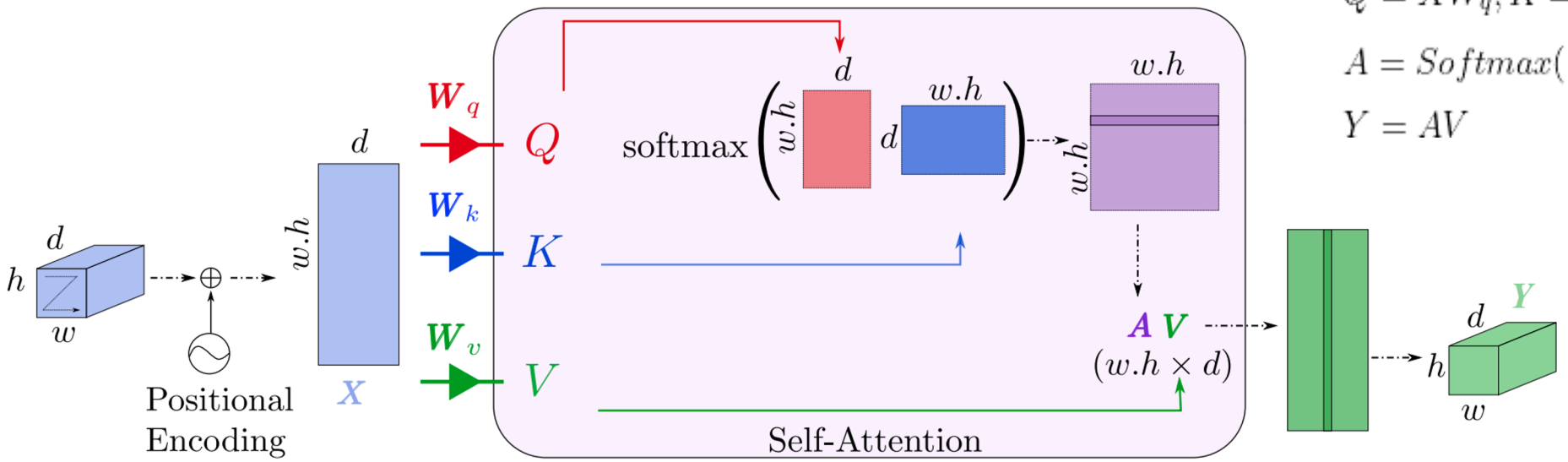
Self-attention



$$X \in \mathbb{R}^{wh \times d}, W_q \in \mathbb{R}^{d \times d}, W_k \in \mathbb{R}^{d \times d}, W_v \in \mathbb{R}^{d \times d}$$
$$Q = XW_q, K = XW_k, V = XW_v$$
$$A = Softmax(\frac{QK^T}{\sqrt{d}})$$
$$Y = AV$$

Self-attention: conclusion

$$X \in \mathbb{R}^{wh \times d}, W_q \in \mathbb{R}^{d \times d}, W_k \in \mathbb{R}^{d \times d}, W_v \in \mathbb{R}^{d \times d}$$
$$Q = XW_q, K = XW_k, V = XW_v$$
$$A = Softmax(\frac{QK^T}{\sqrt{d}})$$
$$Y = AV$$



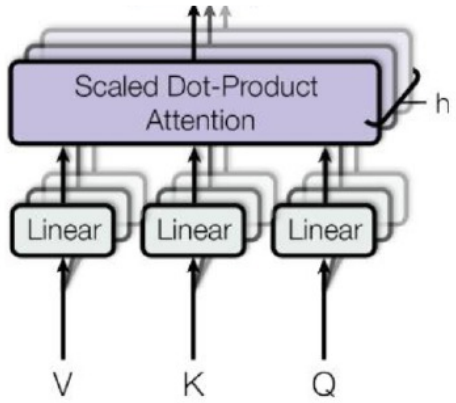
- Each token y_i in Y: computed a linear combination of v_i
  - Enables to model **<u>global interactions</u>** between v_i tokens: full contextual information
  - ≠ ConvNets in vision, interactions limited by the size of the receptive field
  - ≠ RNNs for sequence processing, interactions limited by vanishing gradients
- **Self attention: $O(N^2)$ complexity**
  - Expensive (or impossible) for large N

Multi-headed attention

- High-level idea: multiple self-attention in parallel

- Each head: attend to different parts

- Combine the heads' outputs
  - Concatenation
  - Use a linear layer: desired output size

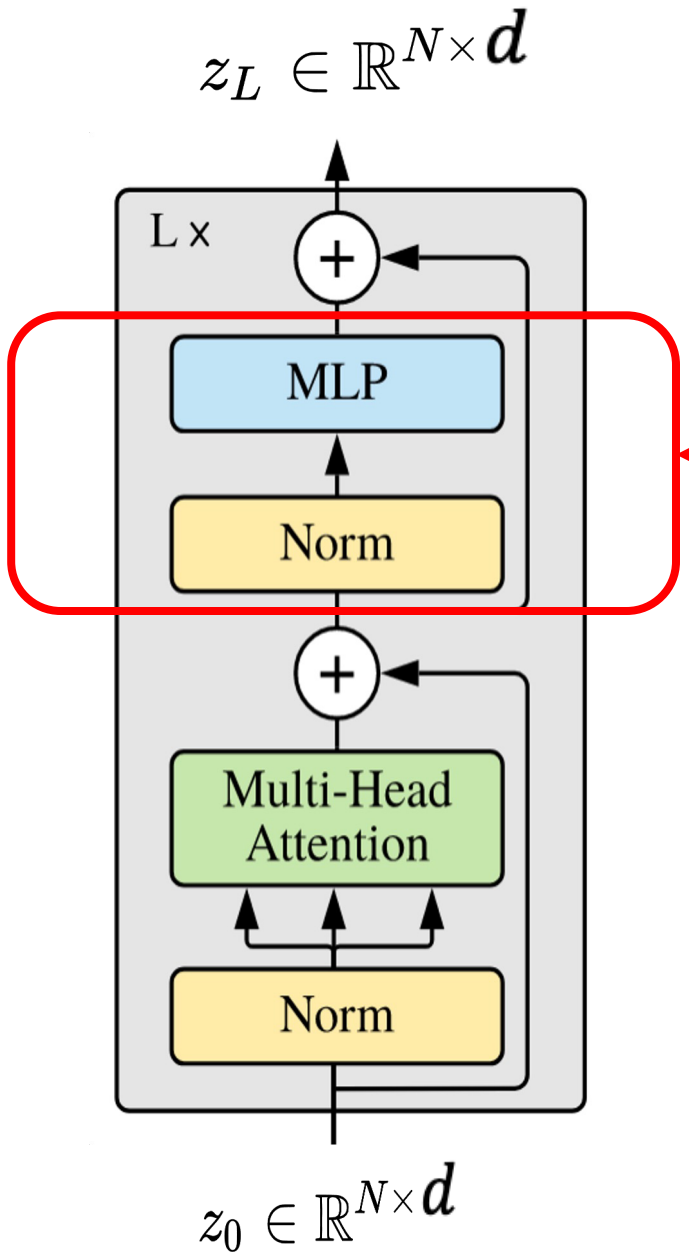


[Vaswani et al. 2017]

Wizards of the Coast, Artist: Todd Lockwood

Credit: Anna Goldie

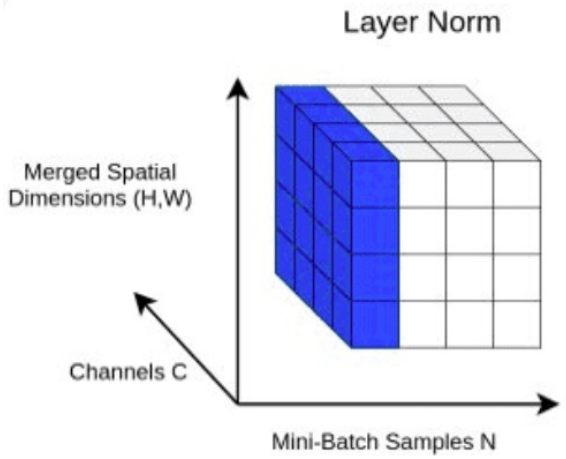
Transformer: the encoder



- **Transformer block**
- **Intra-token computation**
  - MLP, normalization
- **Inter-token computation**
  - Attention
- **Residual connections**
  - Propagate gradients + PE

Layer normalization

- **Intra-token computation layers operate on each token separately**
- Normalization on joint channel and spatial dimensions



$$\mu_n = \frac{1}{K}\sum_{k=1}^{K} x_{nk}$$

$$\sigma_n^2 = \frac{1}{K}\sum_{k=1}^{K} (x_{nk} - \mu_n)^2$$

$$\hat{x}_{nk} = \frac{x_{nk} - \mu_n}{\sqrt{\sigma_n^2 + \epsilon}}, \hat{x}_{nk} \in R$$

$$\mathrm{LN}_{\gamma,\beta}(x_n) = \gamma \hat{x}_n + \beta, x_n \in R^K$$

β, γ
learnable parameters

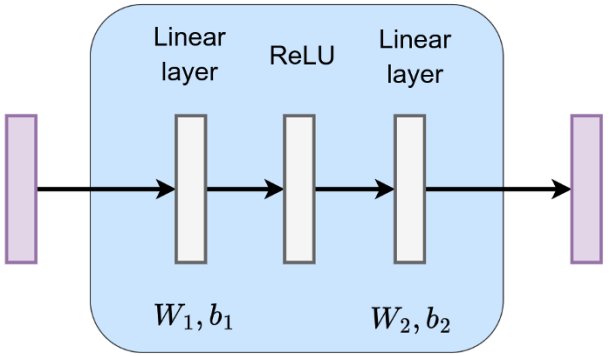
- Stabilize training, faster convergence

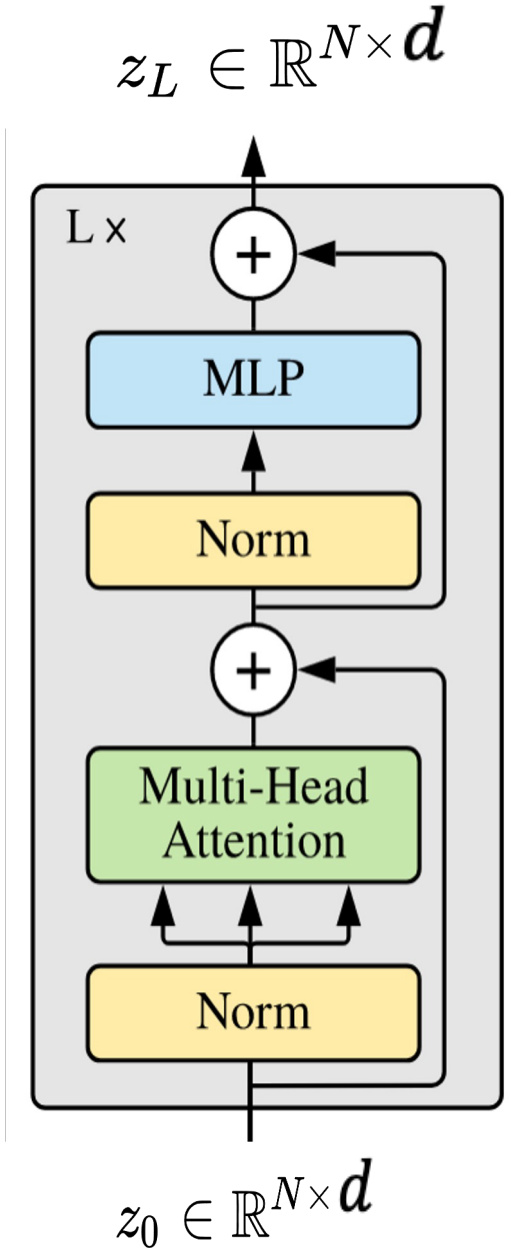
Feed-Forward Network (MLP)

- Applied to each token separately and identically

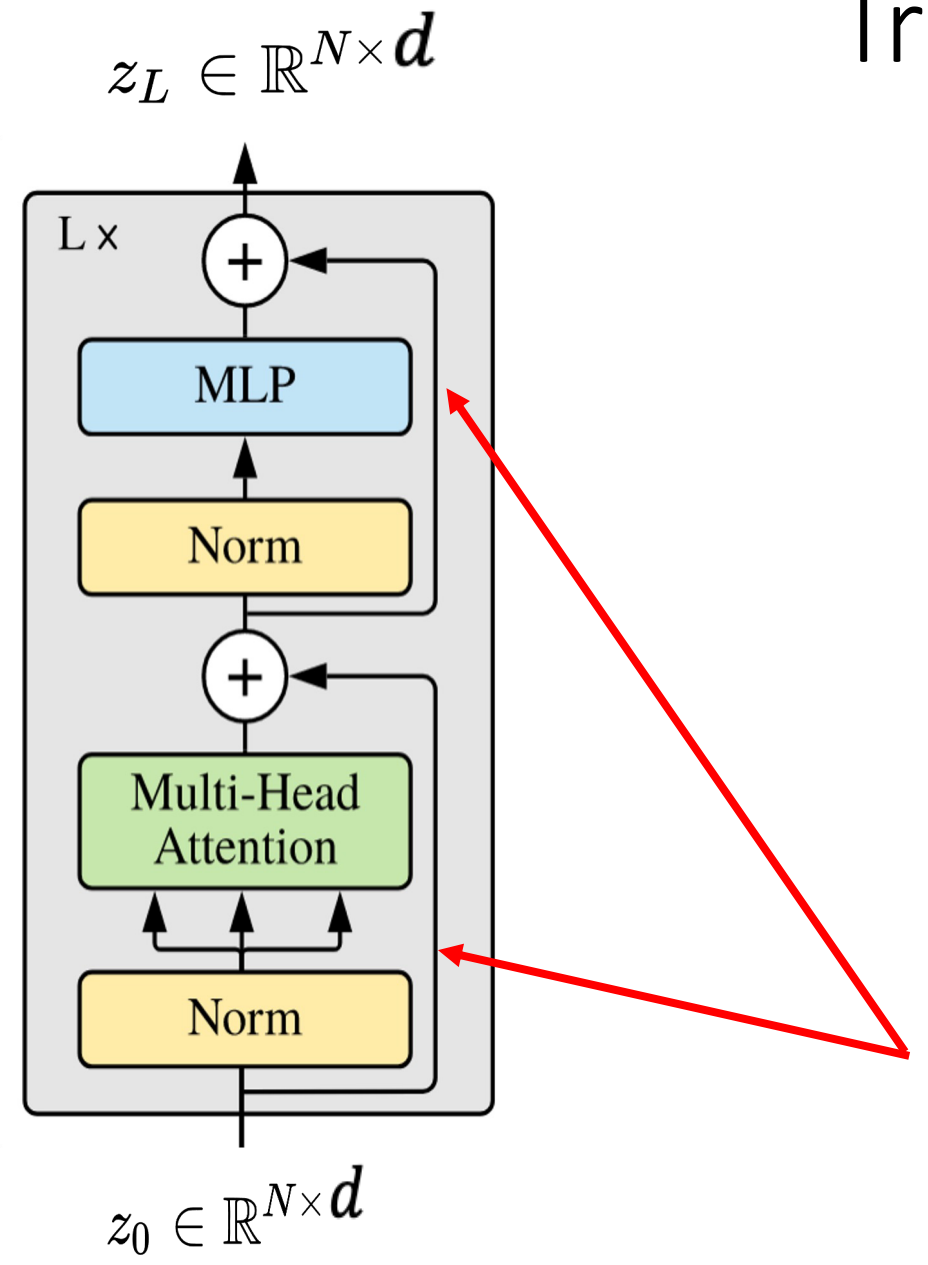
- Feed-Forward Network (MLP)
  - Add non-linearity
  - Refine each token's representation

$$\tilde{z}_{l,i} = \mathrm{LN}(z_{l,i})$$
$$\mathrm{MLP}(\tilde{z}_{l,i}) = \max(0, \tilde{z}_{l,i}W_1 + b_1)W_2 + b_2$$





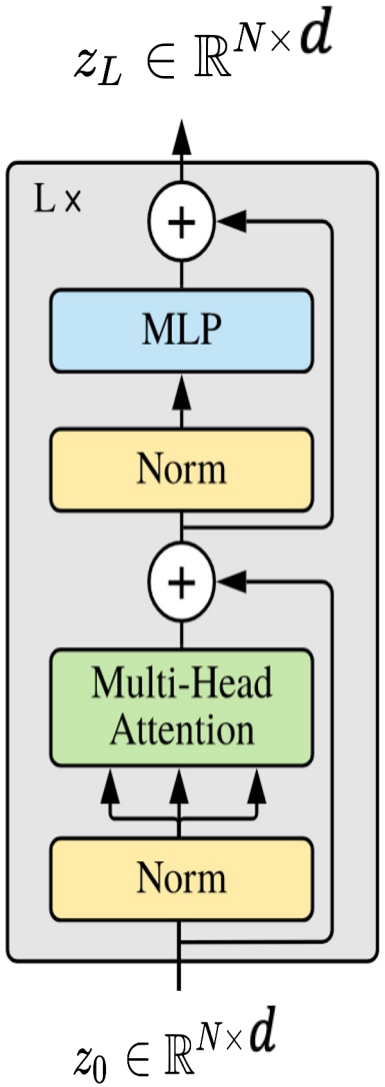
Transformer: the encoder



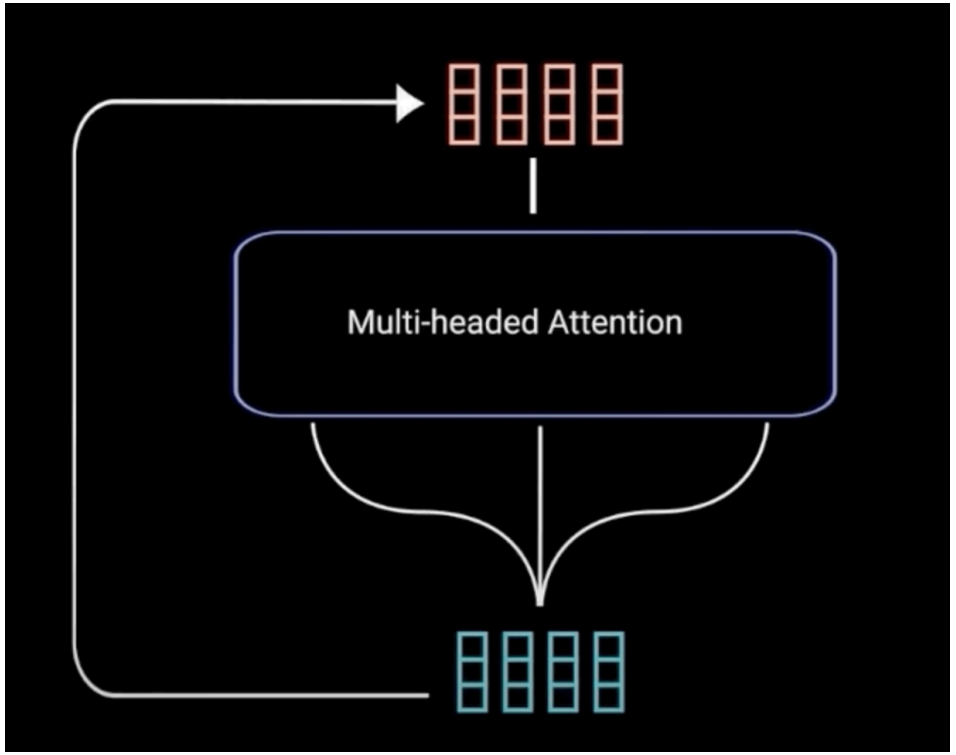
Transformer block

- **Intra-token computation**
  - MLP, normalization
- **Inter-token computation**
  - Attention
- **Residual connections**
  - Propagate gradients + PE

Residual connections

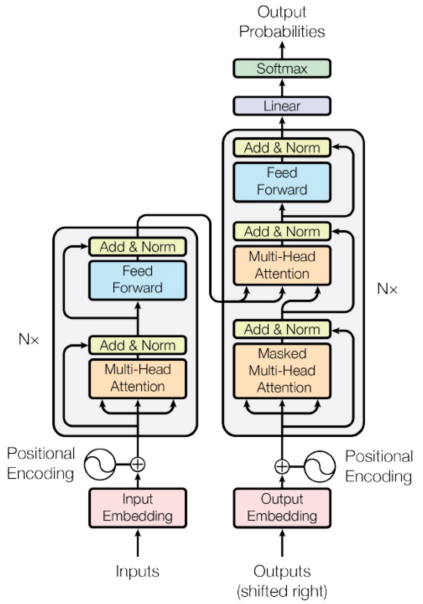


- Better gradient flow (vanishing gradients)
- Leverage input encoding, *e.g.* PE



Beyond transformer encoder: Cross-attention (CA)

- When using CA? Attention between different set of tokens
  - Encoder vs decoder (translation in NLP)
  - Modality A (image) vs modality B (text)



Example:

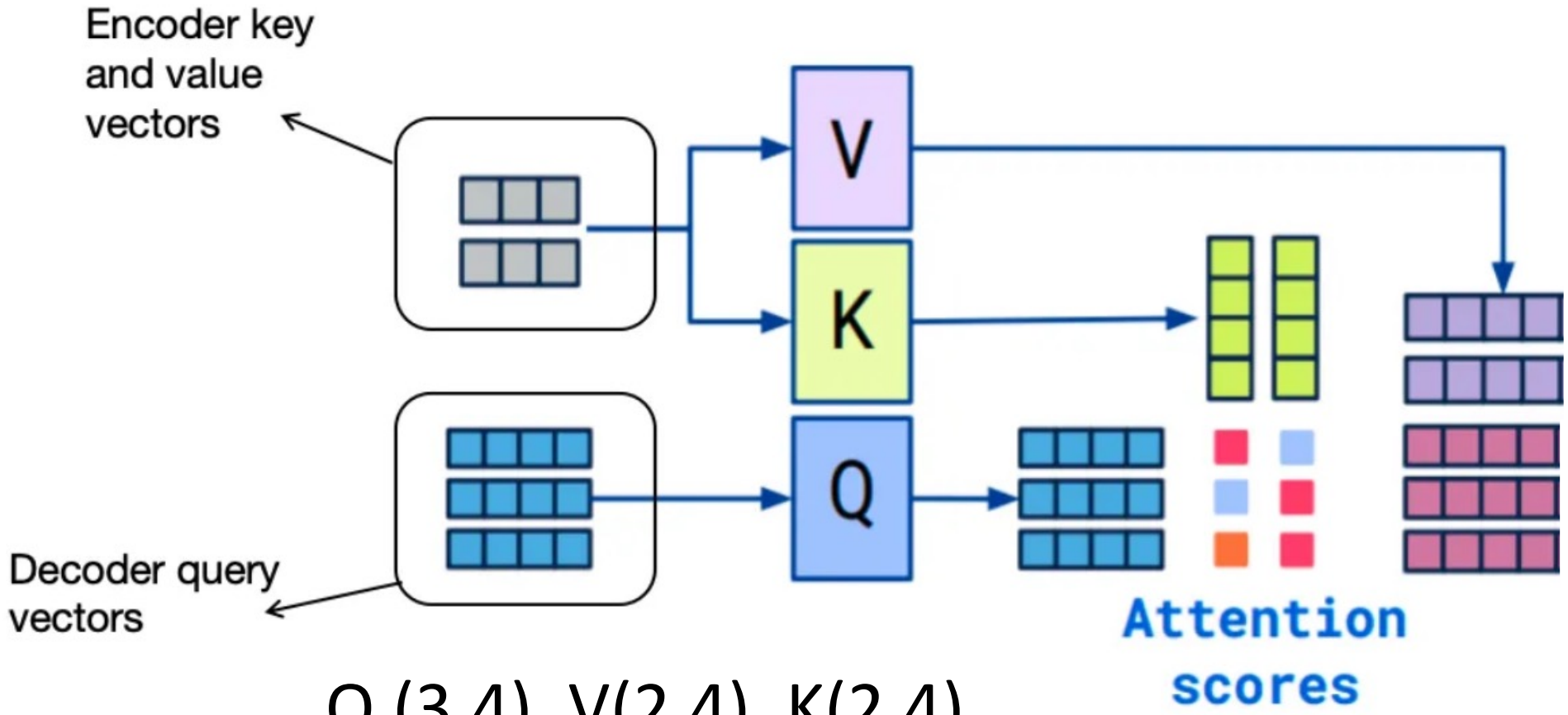
- Modality A => 2 tokens of dim 3
- Modality B => 3 tokens of dim 4
- Modality B: query
- Modality A: key, value

=> CA: re-embedding of B wrt A

A (3,2)

B (3,4)

Cross-attention



Q (3,4), V(2,4), K(2,4)

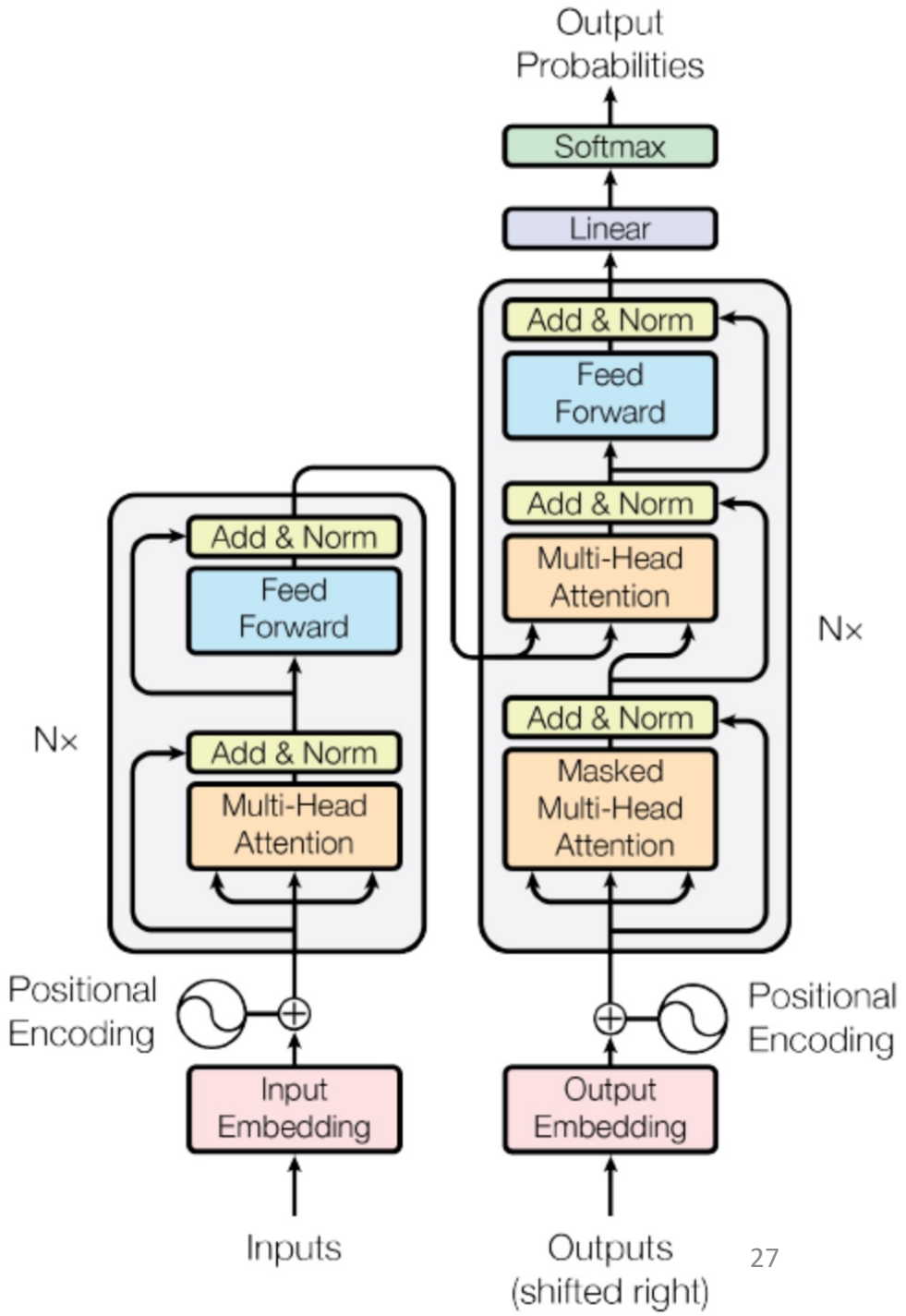
=>A = QK^T => A(3,2)

=> Y = AV => Y(3,4)

Transformer: conclusion

- **Importance of attention: global interactions between tokens**

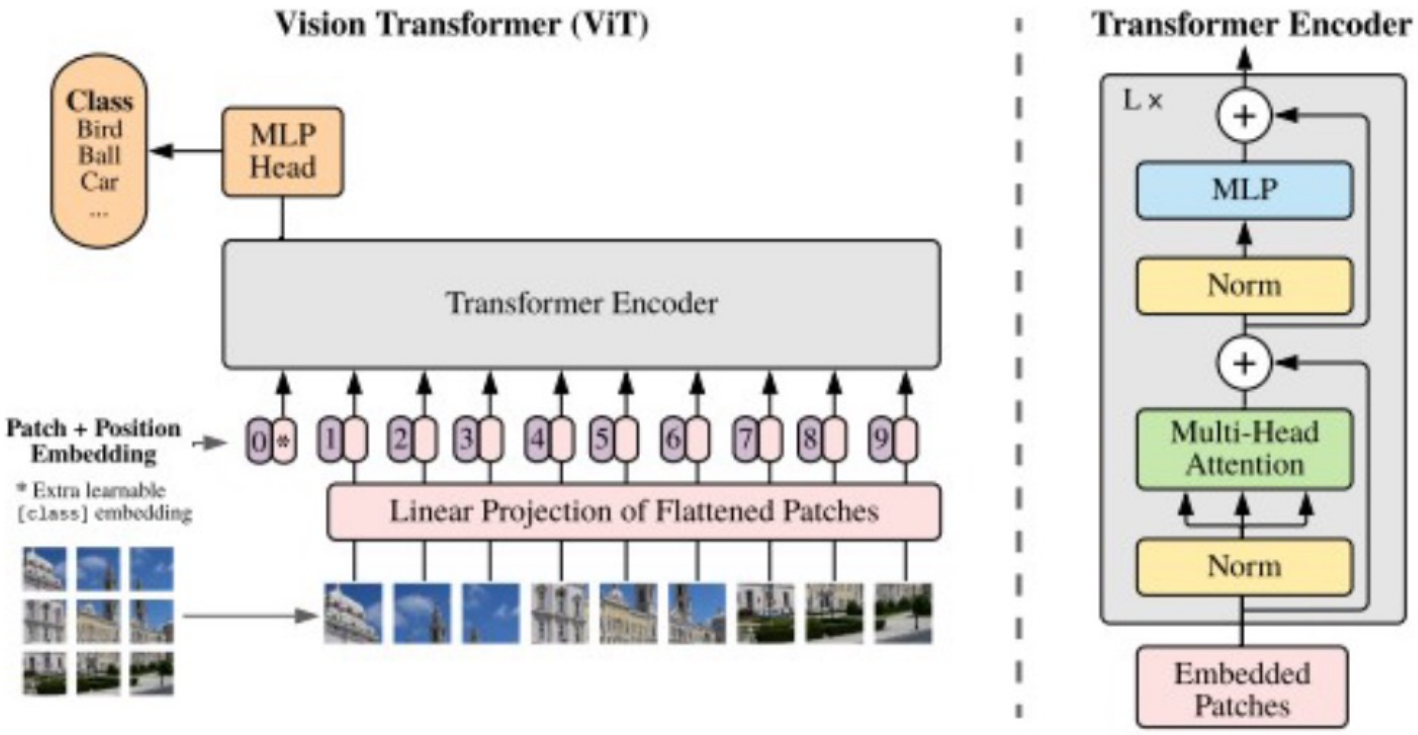
- On the other hand, relaxes inductive biases
  - e.g. ConvNets translation equivariant
    - vs transformers permutation equivariant
  - More flexibility to learn adequate mapping
  - Needs more data



Focus on this talk

1. Transformers: building blocks
2. Transformers in vision & medical image segmentation
3. Foundation Models

Vision Image Transformer (ViT) [2]

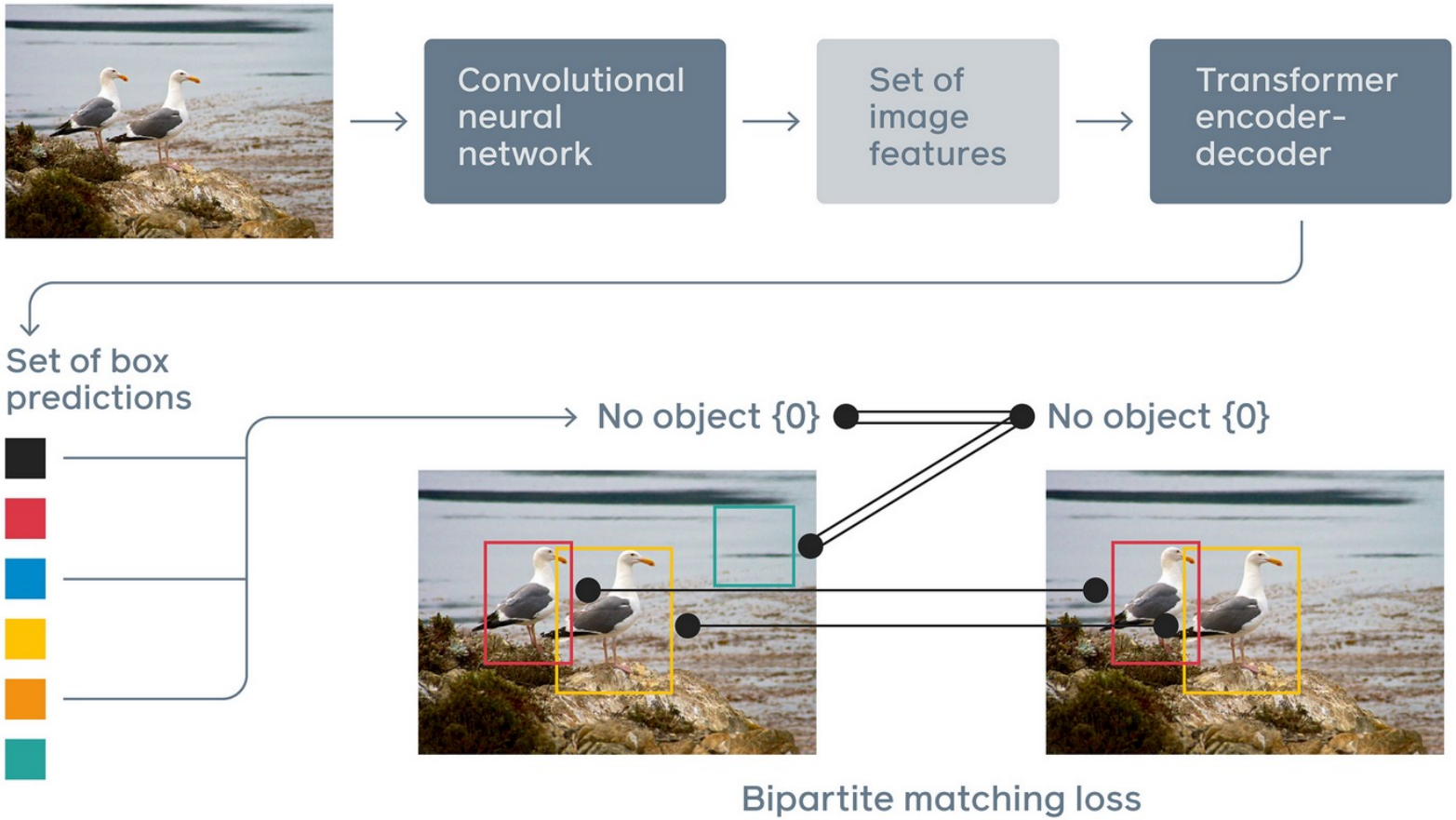


- Direct application of transformer's encoder for images
- Learned on JFT (300.10^6 images)
- Extra learnable token: used for class prediction
  - "Learned" pooling wrt visual tokens

[2] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. ICLR 2020.

ViT demo

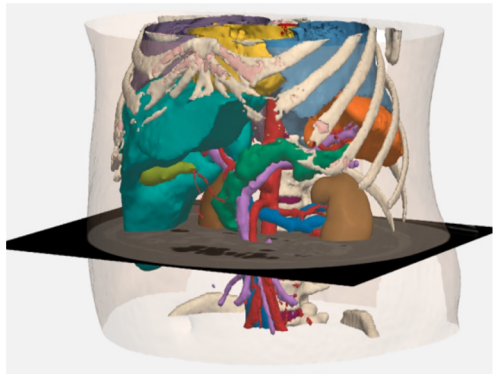
Detection Transformer (DETR) [3]

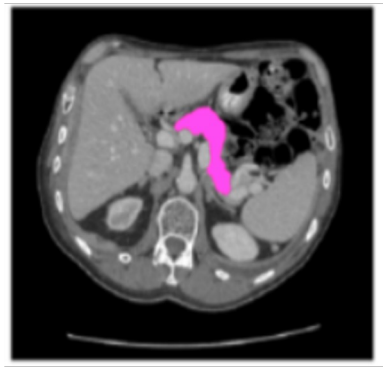


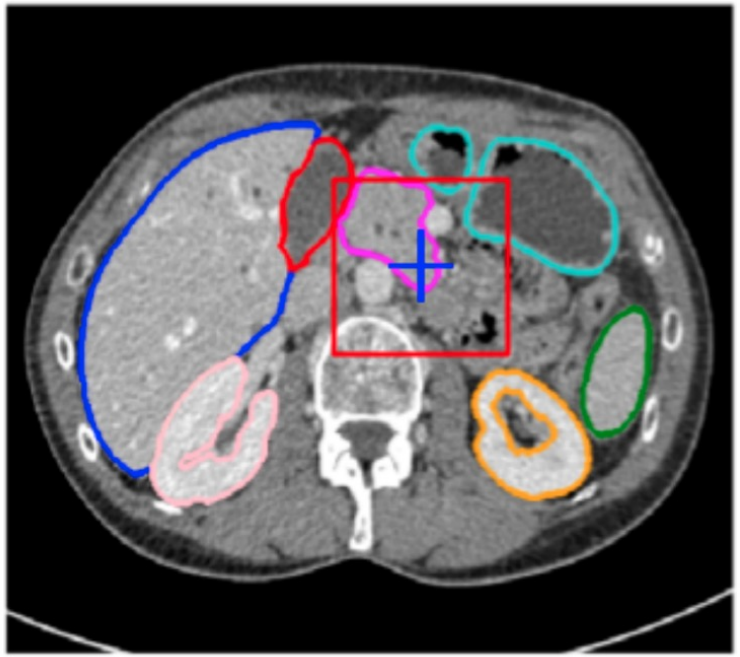
[3] End-to-End Object Detection with Transformers. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. ECCV 2020.

Transformer in medical image segmentation

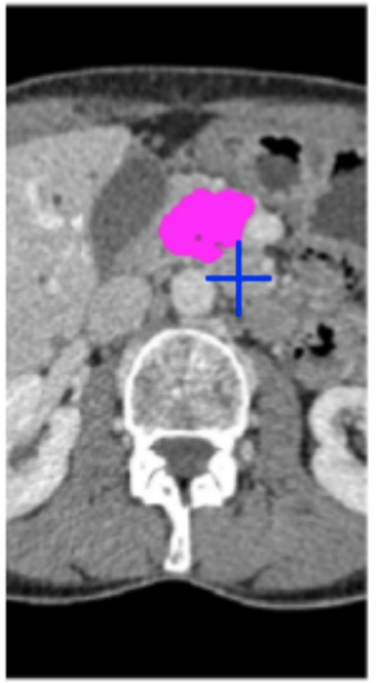
Motivation: U-Net [4] unable to represent full context







a) Ground Truth



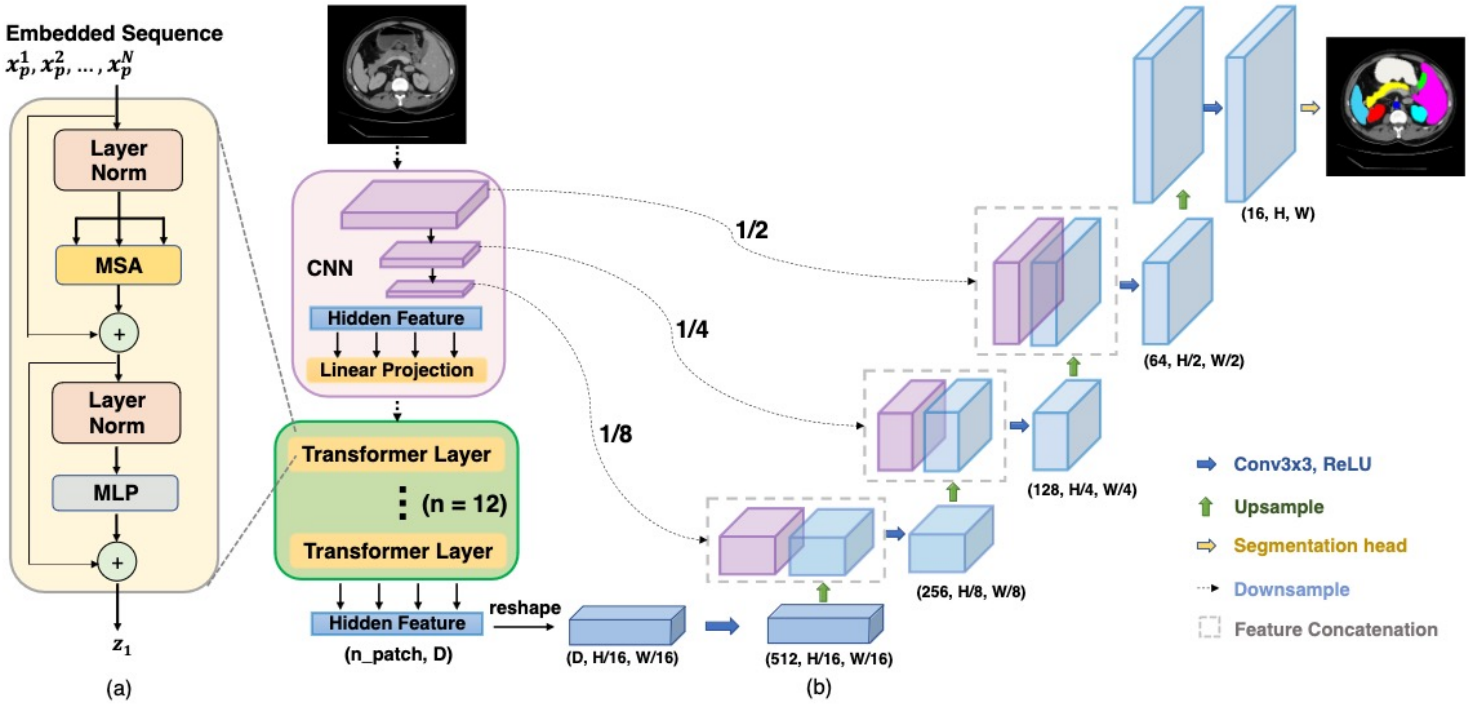
c) U-Net

Segmentation example with U-Net's receptive field (red square)

[4] O. Ronneberger, P. Fischer, and T. Brox. U-net : Convolutional networks for biomedical image segmentation, 2015.

Trans U-Net [5], U-Transformer [6]

- Seminal works for using transformers in medical image segmentation
- Adding self-attention on the bottleneck of a U-Net
  - Inspired from non-local networks [7]

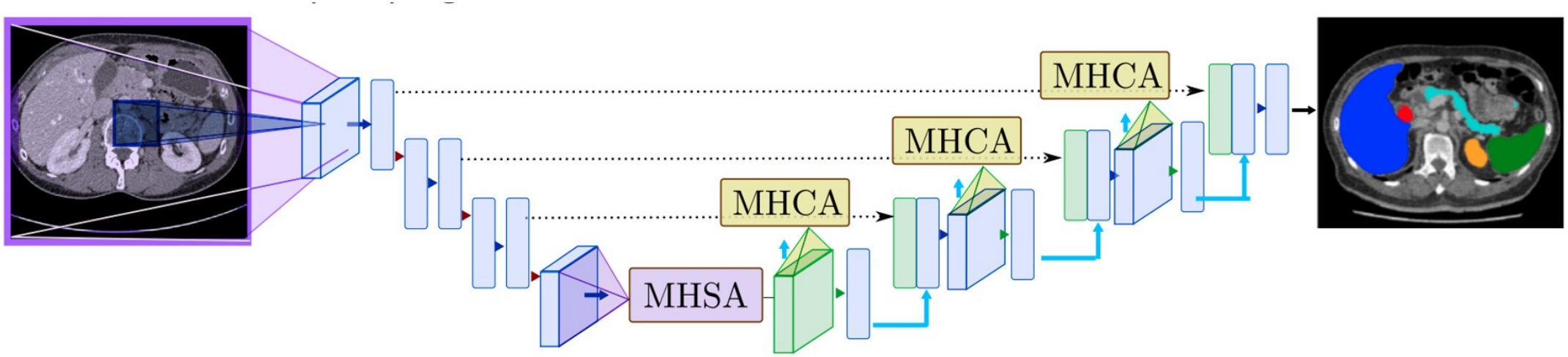


Trans U-Net architecture

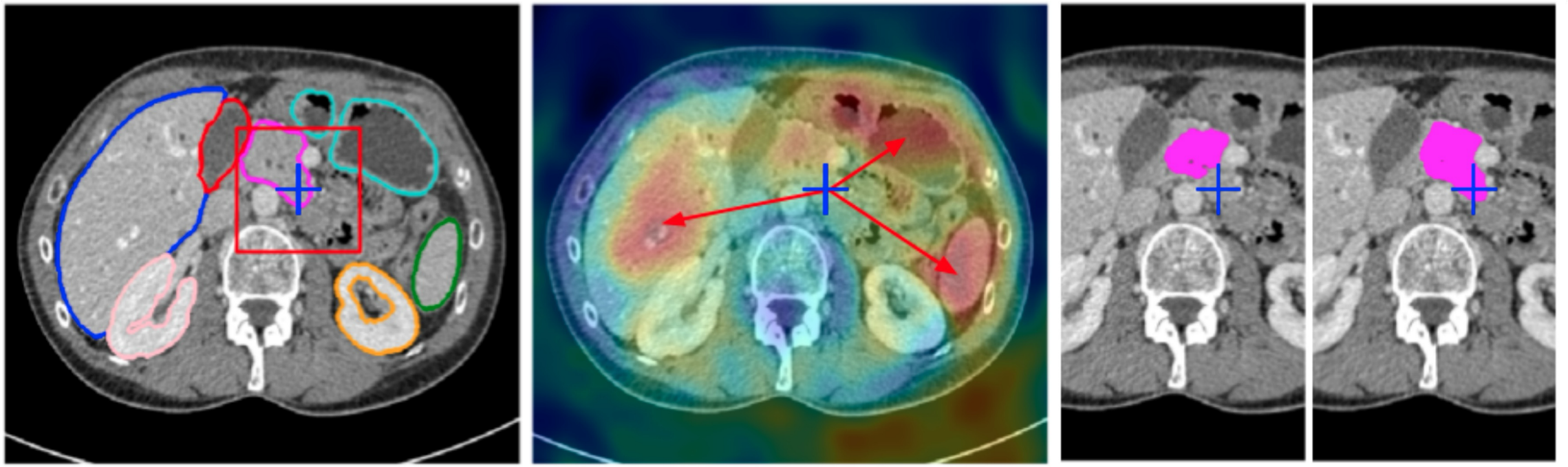
[5] TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. J. Chen et.al. arXiv, Feb 2021.
[6] U-Net Transformer: Self and Cross Attention for Medical Image Segmentation. O. Petit, N. Thome, C. Rambour, L. Soler. arXiv, March 2021.
[7] Non-local Neural Networks. X. Wang, R. Girshick, A. Gupta, K. He. CVPR 2018.

U-Transformer [6]

- **U-Transformer**: self and cross attention in medical image segmentation
  - Self-attention in bottleneck (MHSA)
  - Cross attention to improve super-resolution in skip connections (MHCA)



Results



a) Ground Truth b) Attention map c) U-Net d) U-Transformer

Segmentation example with U-Net's receptive field (red square) and U-Transformer's attention map.

Transformer in segmentation: efficiency

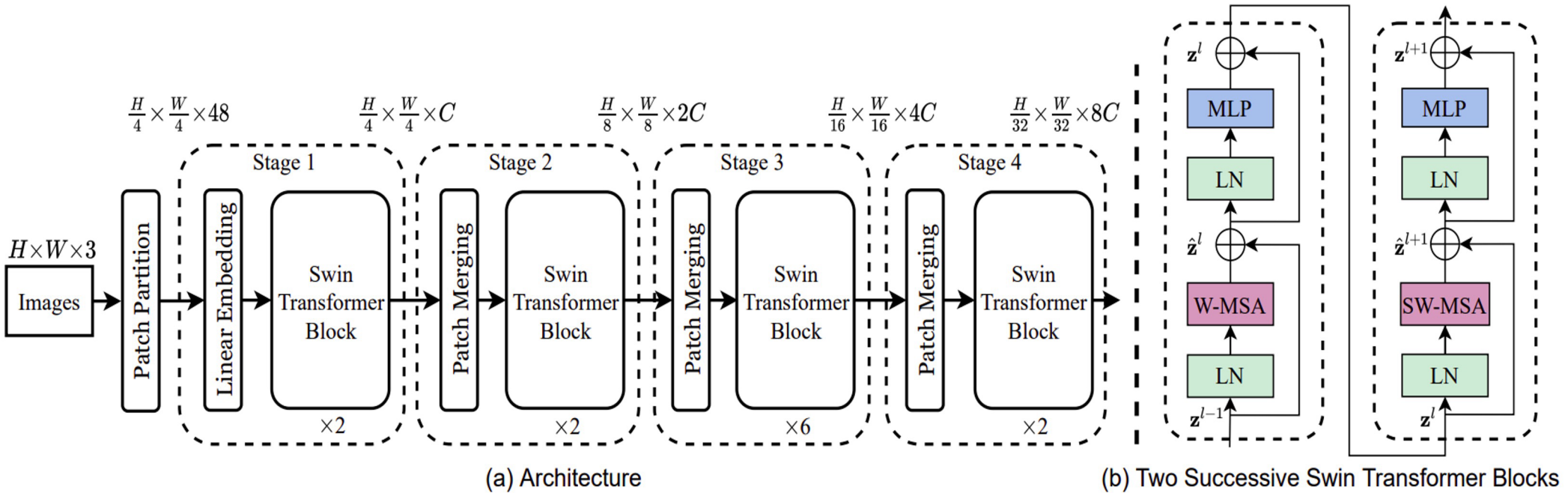
- **Advantages of transformers come with drawbacks**
  - Global context through self-attention vs **$O(N^2)$ complexity wrt # tokens**
- Dense prediction tasks, e.g., segmentation: N => problem exacerbated

- Solution for making self-attention more efficient: very hot topic
  - Hardware-aware optim: *FlashAttention* (2022), …
  - Compressed internal representation: *Perceiver* (2021), …
  - Linear attention approx: *Linformer* (2020), *Performer* (2021), state-space models (SSM, 2022) …
  - **Architectural constraints, e.g., spatial priors in segmentation => local global attention**

Swin-Transformer [8]: Multi-resolution transformer

- Patch merging (~pooling) => larger receptive field
- Enables full attention in larger regions

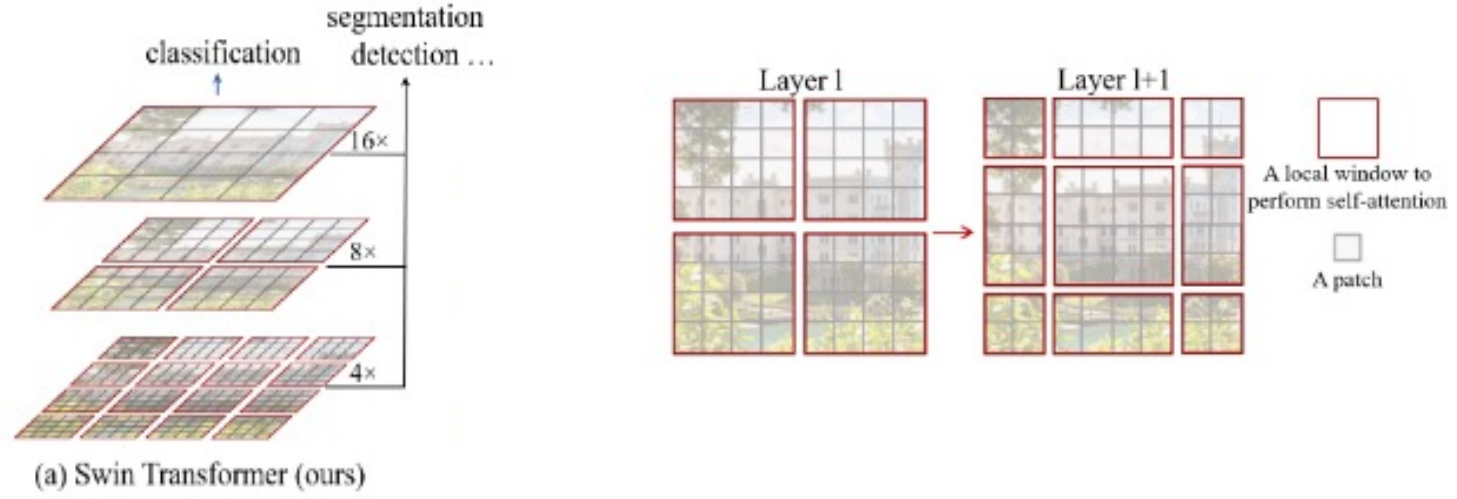
W-MSA: Windows Multi-head Self-Attention
SW-MSA: Shifted W-MSA



[8] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo. ICCV 2021

Swin-Transformer [8]: Multi-resolution transformer

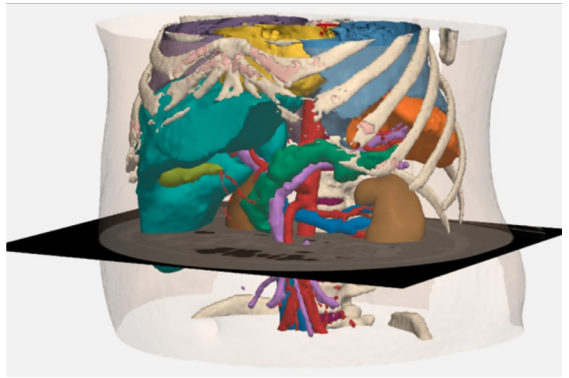
- Local attention in lower-layers, high-resolution
  - Local attention => efficient, but non global!
- Shifted windows at layers l/l+1 ⇒ communication between windows



But no full attention in high resolution feature maps!

[8] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo. ICCV 2021

3D medical image segmentation



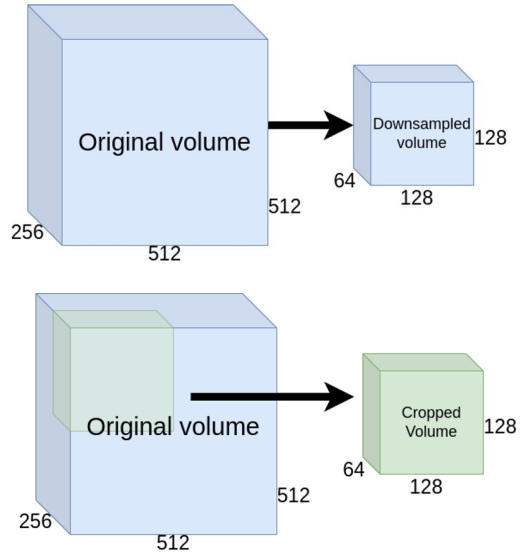
Challenges

- **Size of the input=> Large memory requirements**
  - Vanilla U-Net: 180Gb image size 512x512x256
  - Transformers on input 3D images completly impossible
- **Common strategies to reduce the memory footprint:**

- Downsampling ⇒ **Drop in quality**

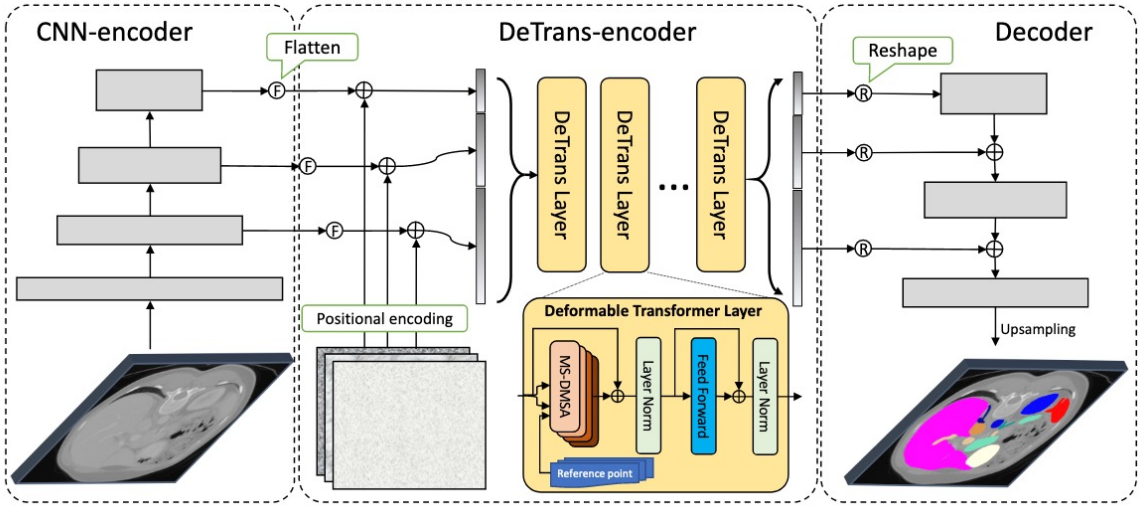
- Limited model size
- Train on 2D slices
- Train on patches

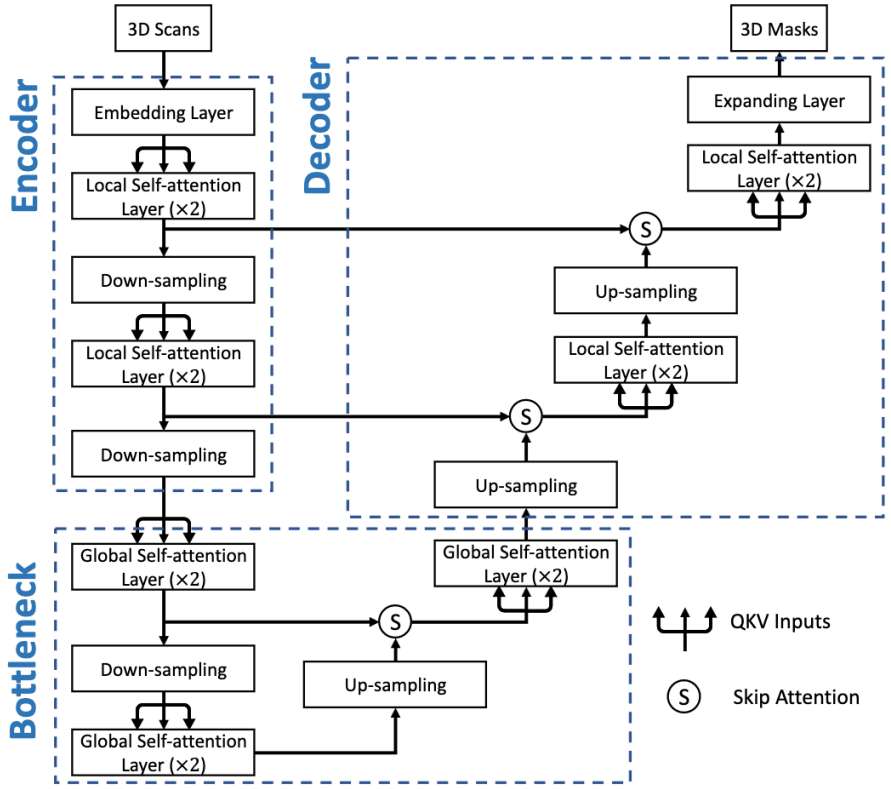
⇒ **No full contextual information**



3D medical image segmentation: patch approaches

- Swin-UNet [9], nn-Former [10]: 3D version of Swin-transformer
- CoTR: **Co**nvolutional NN and **Tr**ansformer [11] => hybrid conv transformer archi



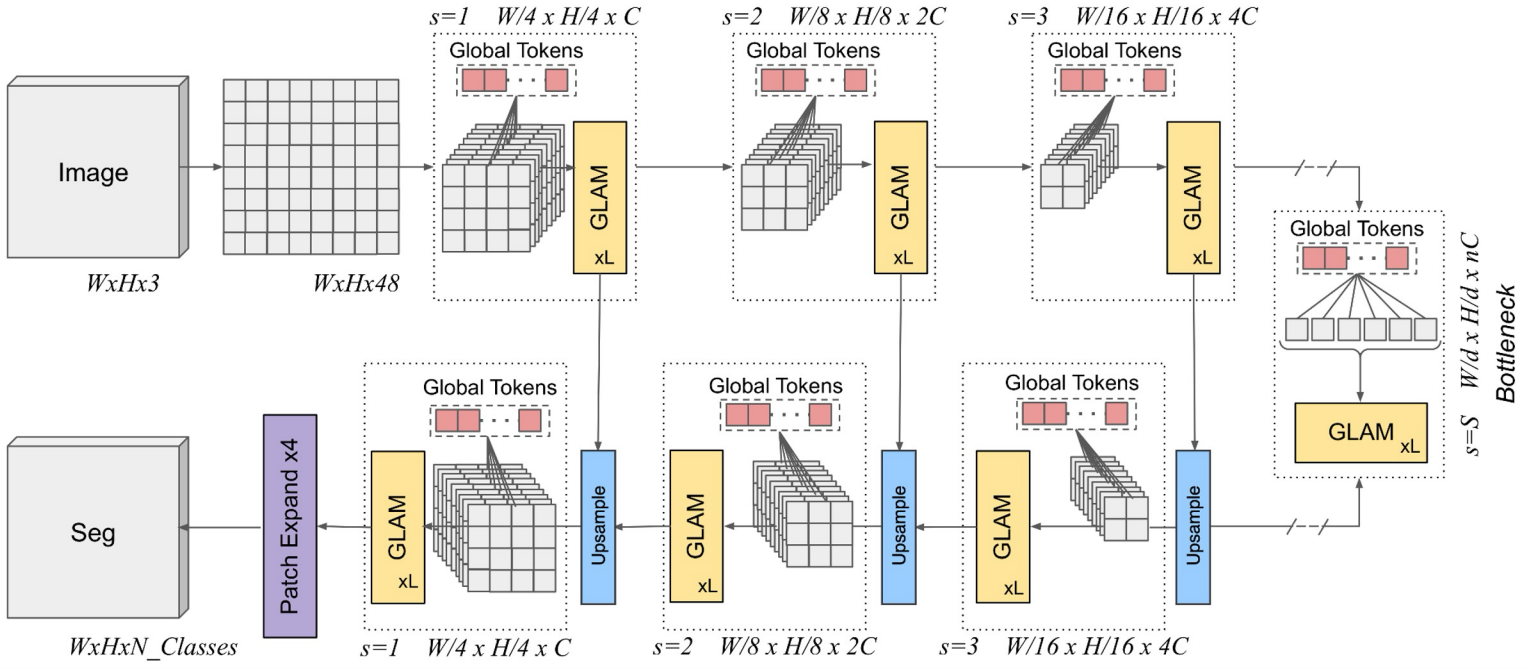


But no full attention in high resolution feature maps!

[9] Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang. Arxiv, May 2021.
[10] nnFormer: Interleaved Transformer for Volumetric Segmentation. H.Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu. Arxiv, September 2021.
[11] CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. Y. Xie, J. Zhang, C. Shen, Y. Xia. MICCAI 2021

Global attention in multi-resolution transformers (GLAM) [12]

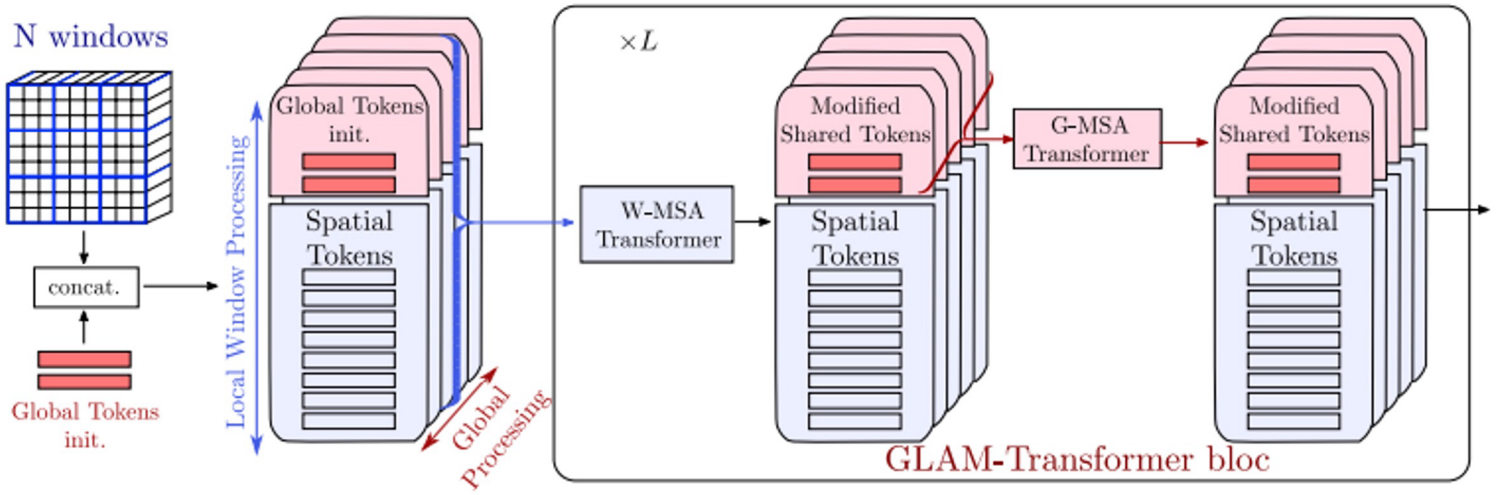
- Architecture based on hierarchical transformer (e.g. Swin, nn-Former)
  - Can also be included in any multi-resolution model (e.g. Conv)
  - GLAM Motivation: Full attention even in high-resolution features



[13] Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation. L. Themyr, C. Rambour, N. Thome, T. Collins, A. Hostettler. WACV 2023.

GLAM block

- Define learnable global tokens in each window, cf CLS in VIT
  - Window self-attention (W-MSA): attention between visual and global tokens
  - Global attention (G-MSA) between global token
- G-MSA: indirection between all visual tokens
  - Break computational complexity of full attention between visual token
  - But enables full indirect interaction between them

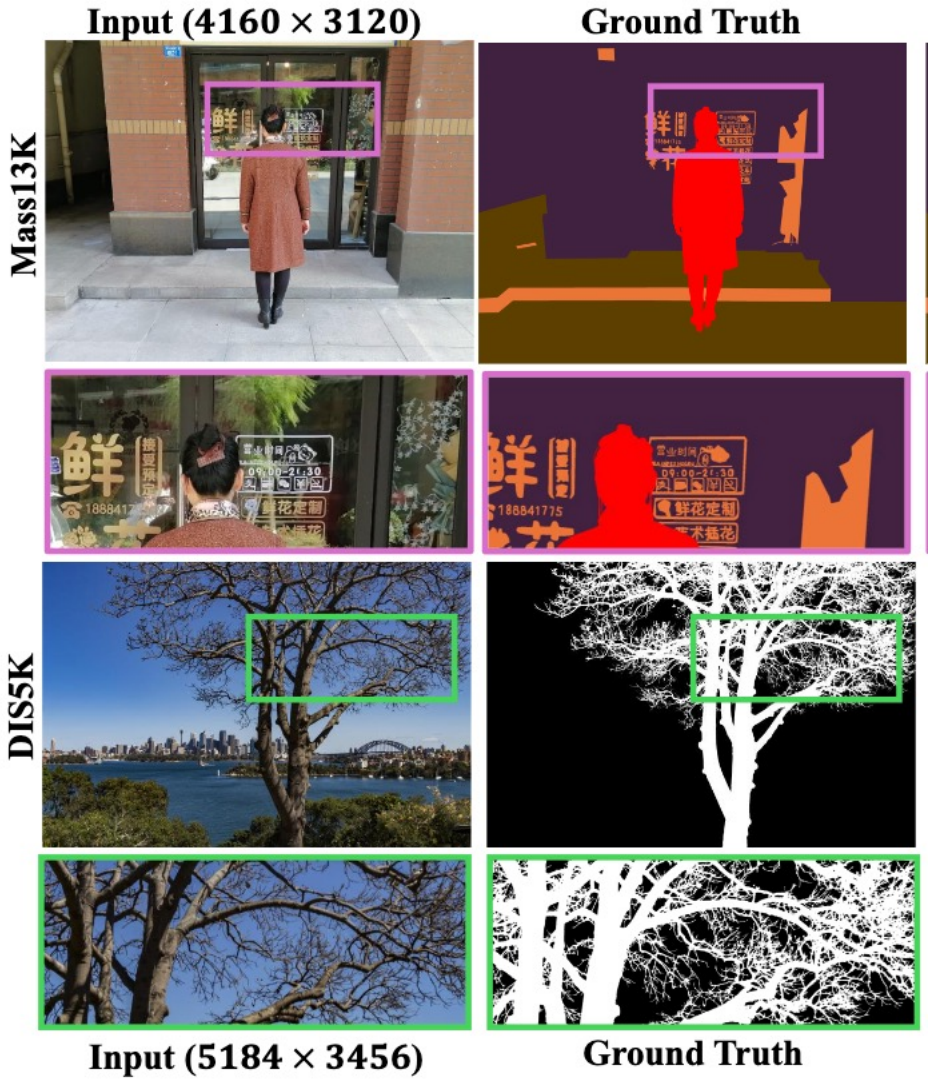


Transformers for segmentation: conclusion

Ongoing work & perspectives

- Several archis extension to improve 3D segmentation
  - Unet-TR, U-Net TR++
  - SegFormer 3D

- Full attention + fine-grained pixel/voxel segmentation still an open question
  - Recent datasets: Mass13K (C. Xie et.al., CVPR'25), DISK5K (X. Qin et.al., ECCV'22)

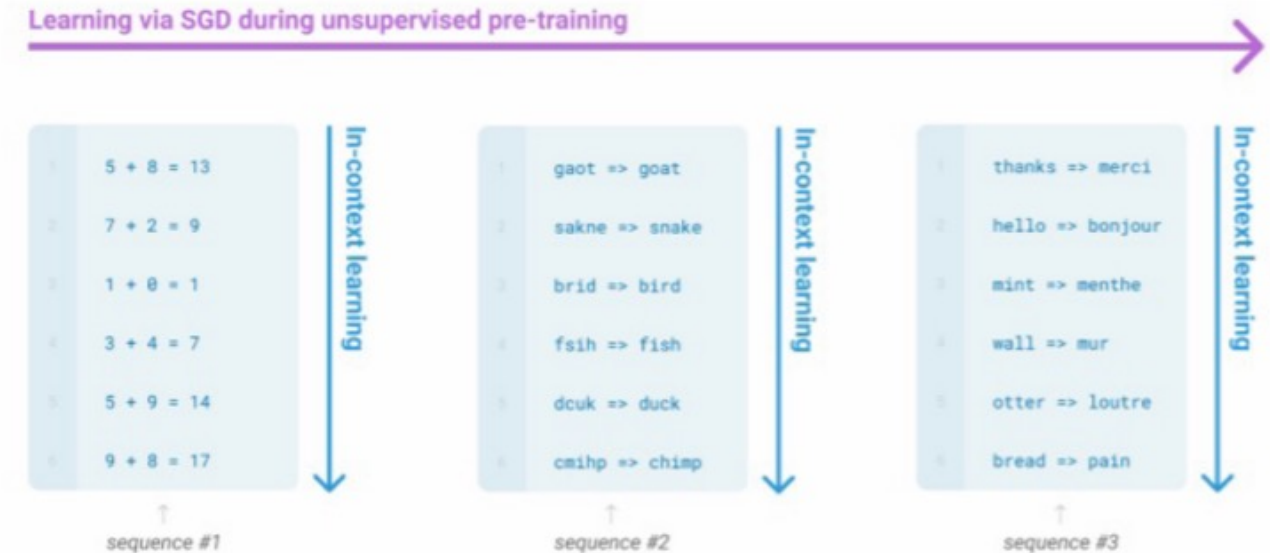


Focus on this talk

1. Transformers: building blocks
2. Transformers in vision & medical image segmentation
3. Foundation Models

Foundation models

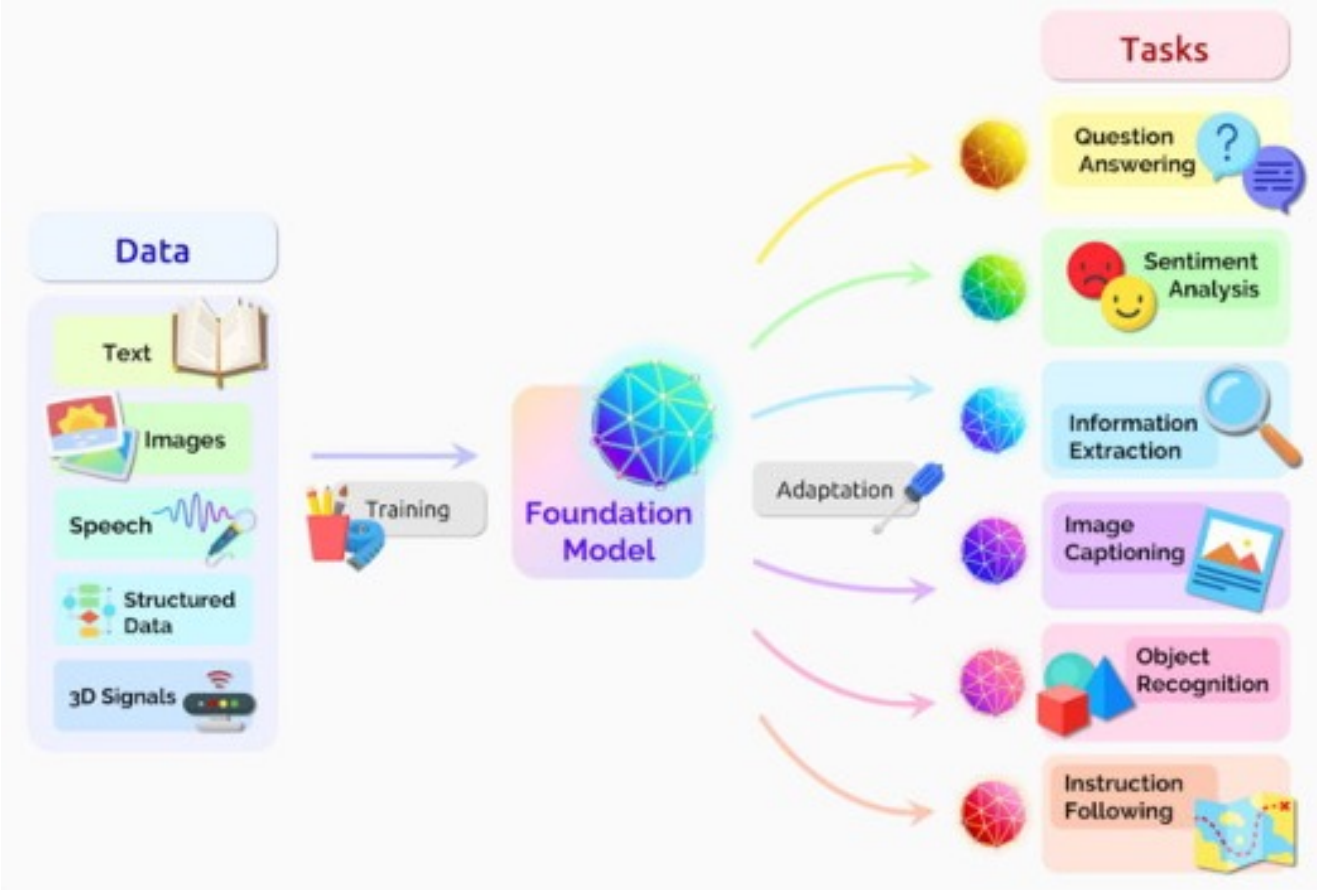
- Models pre-trained on large scale datasets
- Able to solve various, unspecified downstream tasks
  - In-context learning => “zero-shot generalization“
- Training recipe in NLP: Train huge transformers, e.g. GPT-3/GPT-4
  - Pre-training: next word prediction on huge scale datasets
  - Instruction fine-tuning, RLHF
  - Inference: prompted (“in-context learning”) with emerging properties



Foundation models: from NLP multi-modal data

Transformers, tokens => homogeneous way to represent multi-modal data

- Reasoning over perceptual data: image, audio, etc
- Multi-modal inputs AND outputs
- (Cross)-Attention used for representing multi-modal info

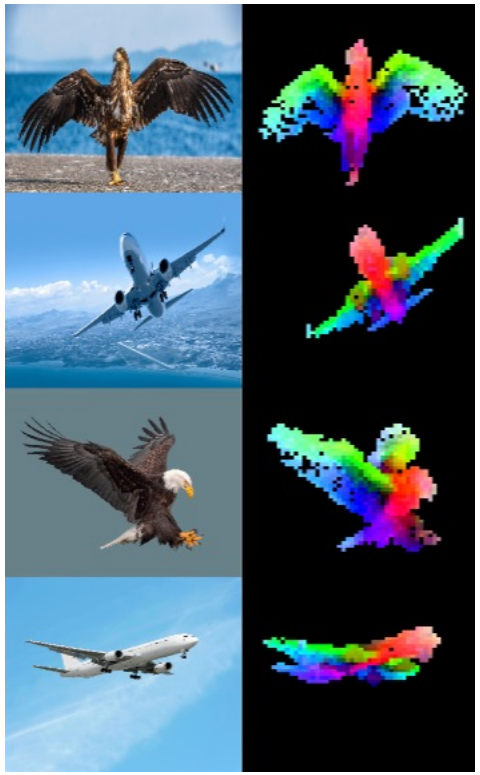


Some foundation models

- NLP: text encoder (BERT), LLMs: Chat-GPT and others (Gemini, Deep-seek, Claude, etc)

deepseek & Claude

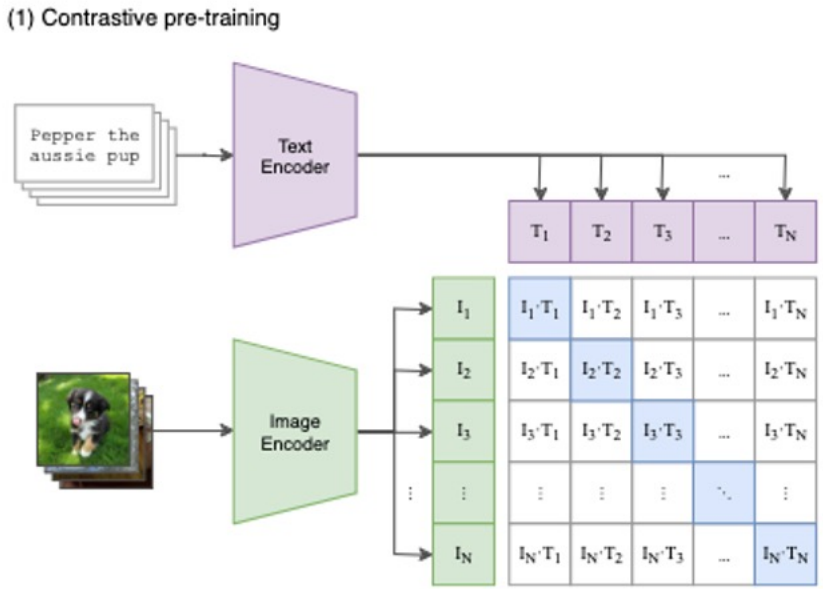
- Vision : image encoders (DINO)



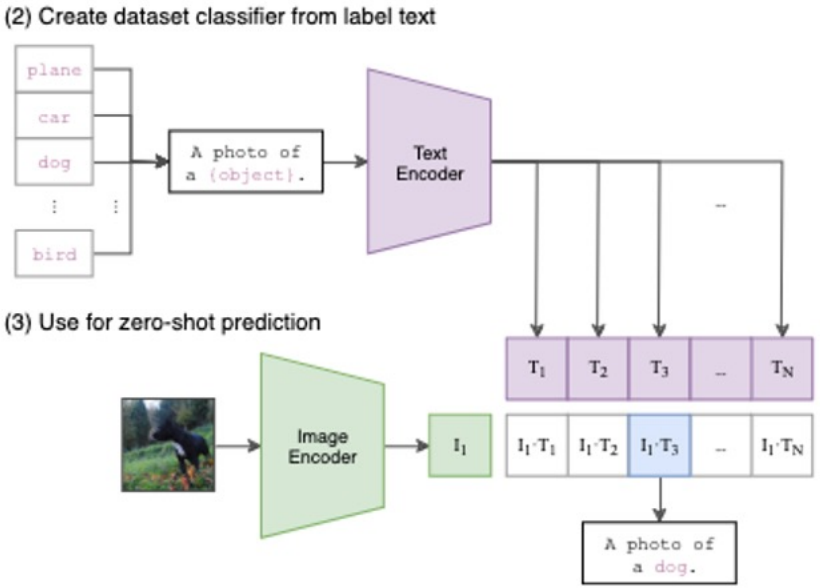
- Vision and language: CLIP (details soon)
- Times series: Chronos, Time-PFN

Vision-Language Models (VLMs): CLIP [13]

- Contrastive Language-Image Pre-training (CLIP): image/ text encoder, alignment



- Inference, e.g. zero-shot image generalization

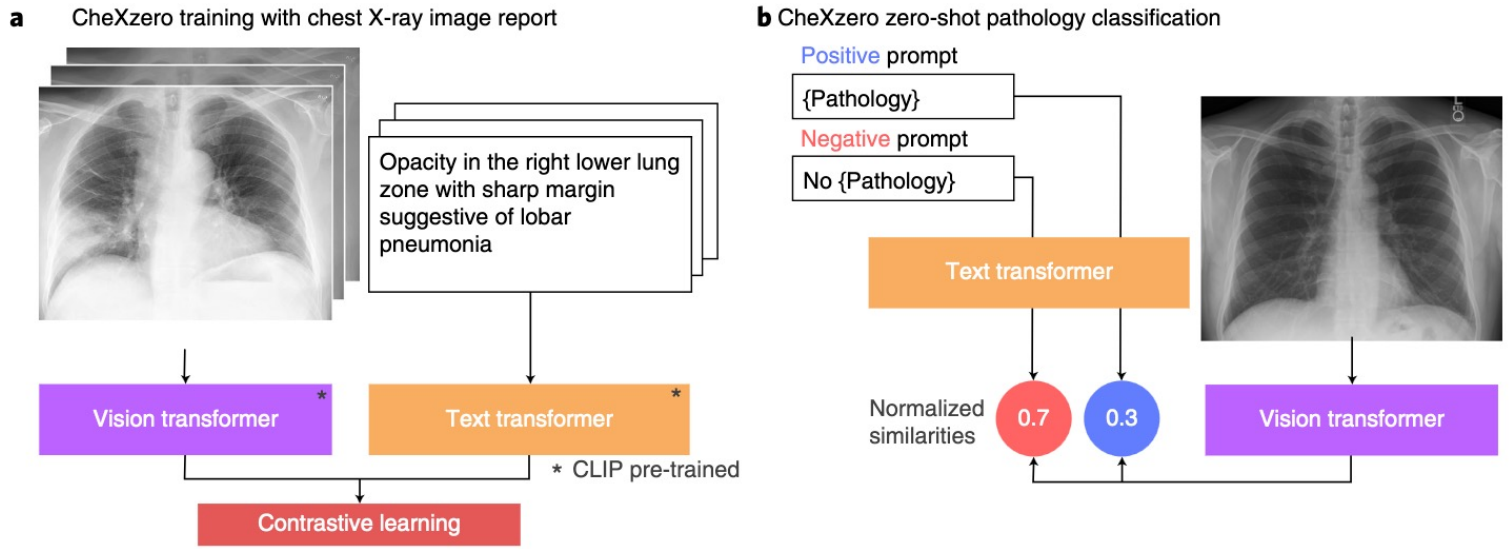


- Several extension for medical data, e.g., MEDCLIP [14]

[13] Learning Transferable Visual Models From Natural Language Supervision. 1. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. ICML 2021
[14] Contrastive Learning from Unpaired Medical Images and Text. Z. Wang, Z. Wu, D. Agarwal, J. Sun. MedCLIP: EMNLP 2022

Zero-shot generalization with VLMs

- CheXzero [15]: POC of CLIP-based model



- Foundation models in healthcare [16]

For example, a clinician might say, "Check these chest X-rays for Omicron pneumonia. Compared to the Delta variant, consider infiltrates surrounding the bronchi and blood vessels as indicative signs"[40].

[15]). E. Tiu, E. Talius, P. Patel, C.P. Langlotz, A.Y. Ng, P. Rajpurkar. Nature Biomedical Engineering volume, 2022
[16] Foundation models for generalist medical artificial intelligence. M. Moor, O. Banerjee, Z.S.H. Abad, H. M. Krumholz, J. Leskovec, E.J. Topol, P. Rajpurkar. Nature volume 616, pp 259–265, 2023.

Multi-modal LLMs (MLLMs) & foundation models

MLLMs for dealing with images and text:

DALL-E [17]: image decoder

[17] . Zero-Shot Text-to-Image Generation A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever. ICML 2020.

Flamingo [18]: text decoder

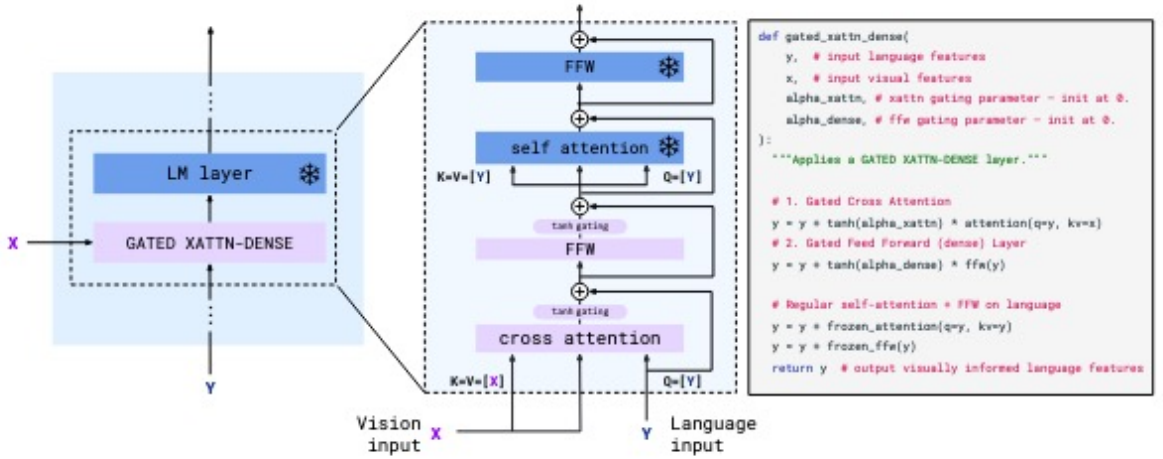
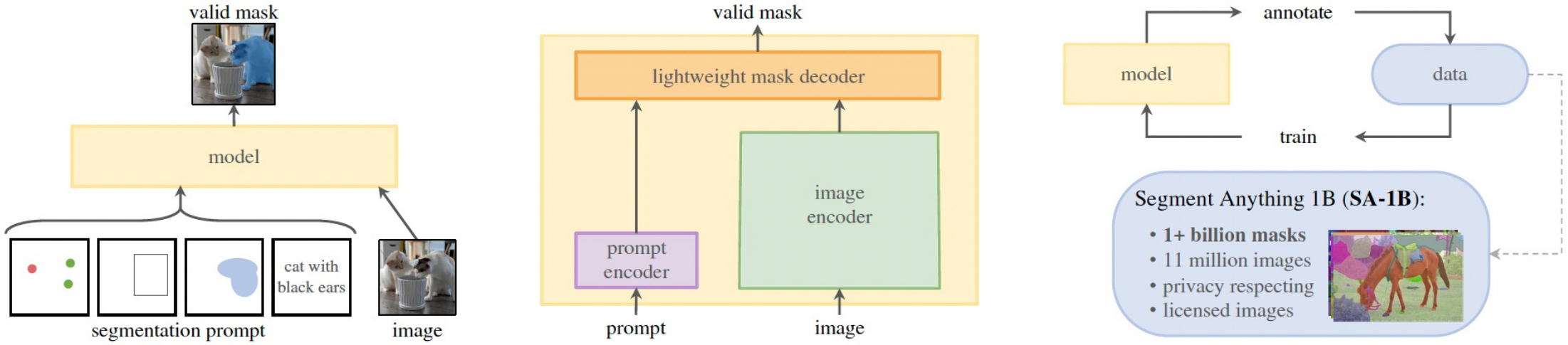


Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

[18] Flamingo: a Visual Language Model for Few-Shot Learning. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan. . NeurIPS 2022

Foundation models in segmentation: SAM

- Segment Anything Model (SAM) [19]
- How to design an effective foundation model for segmentation?
  a) Promotable segmentation with points, BB, or text
  b) Segmentation model: prompt/image encoders + mask decoder
  c) Data engine for interactive data collection & annotation

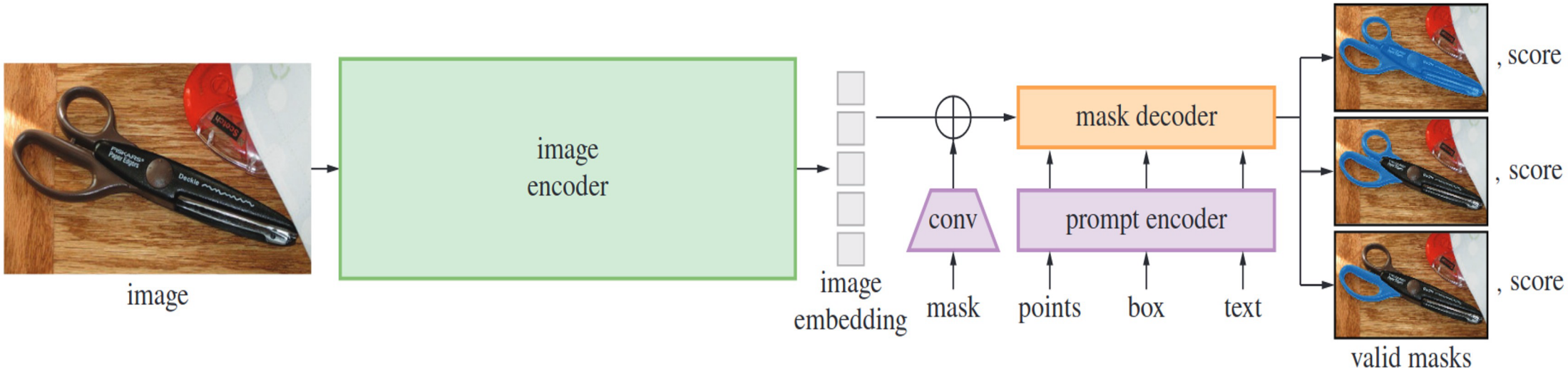


(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

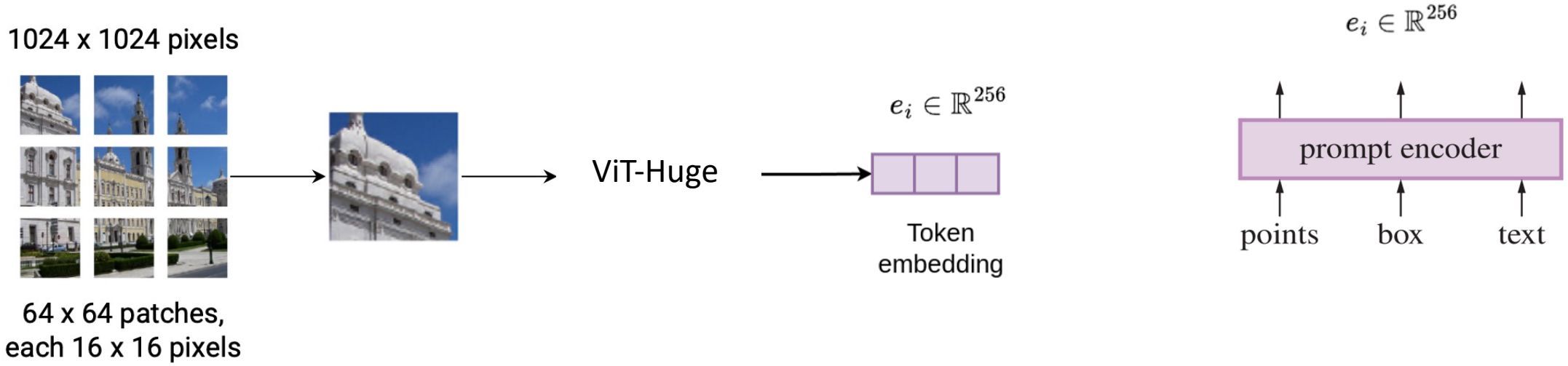
[19] Segment Anything. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.Y. Lo, P. Dollár, R. Girshick. ICCV 2023.

SAM Model

- Relatively simple architecture
- Interactive segmentation using prompts
- Accounts for ambiguous masks based on high-level prompt



SAM encoders transformers



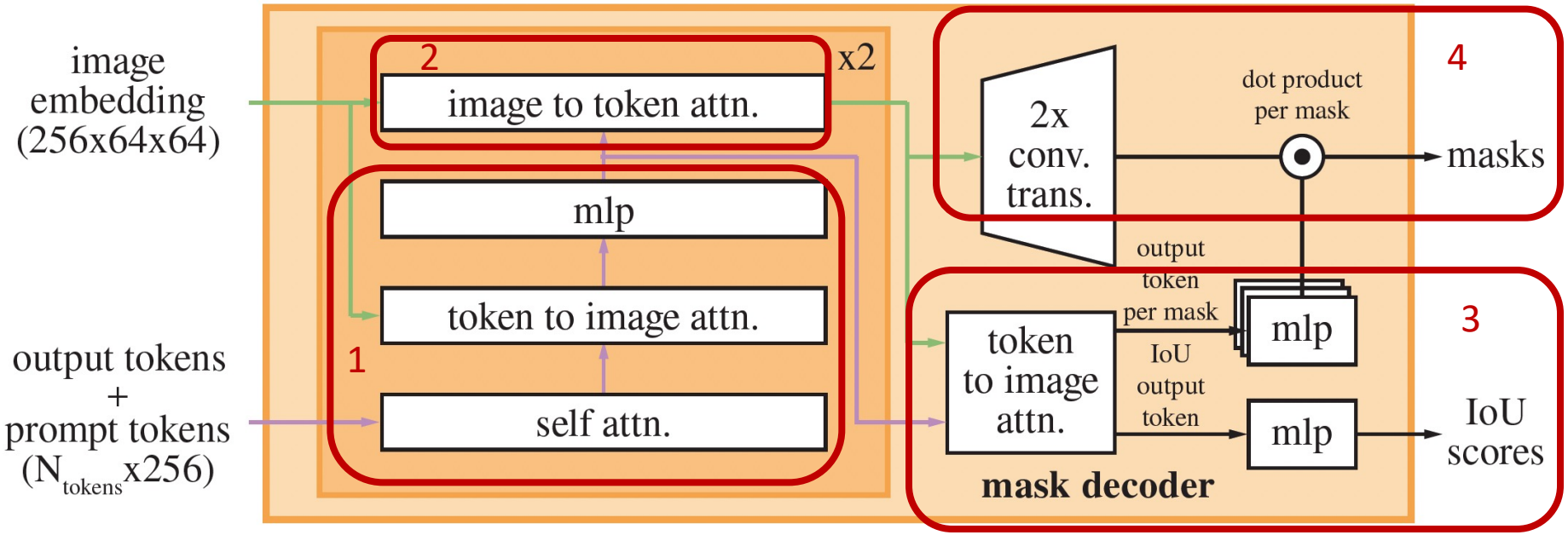
- Image encoder => 16x16 tokens of dim 256, frozen
- Prompt encoder
  - Different geometric prompt types, tokens: 256 dim => learned
    - Points: positional encoding (PE)+ learn embedding of foreground/background
    - BB: for top/left & right/bottom points: PE + learned embedding for “top/left corner”
  - Text, pretrained text Transformer (CLIP), frozen

SAM decoder

Flow: bidirectional attention

1. Prompt (+output tokens, OT): SA, CA (2img) + MLP
2. Img embed: CA (2token)
3. Right : token CA -> 3 MLPs (from OT) for ambiguous pred + MLP for IoU
4. Right: upsampler, gating with MLP output => predicted masks

Learning: focal loss + Dice + l2 for IoU



SAM Example

Point prompt

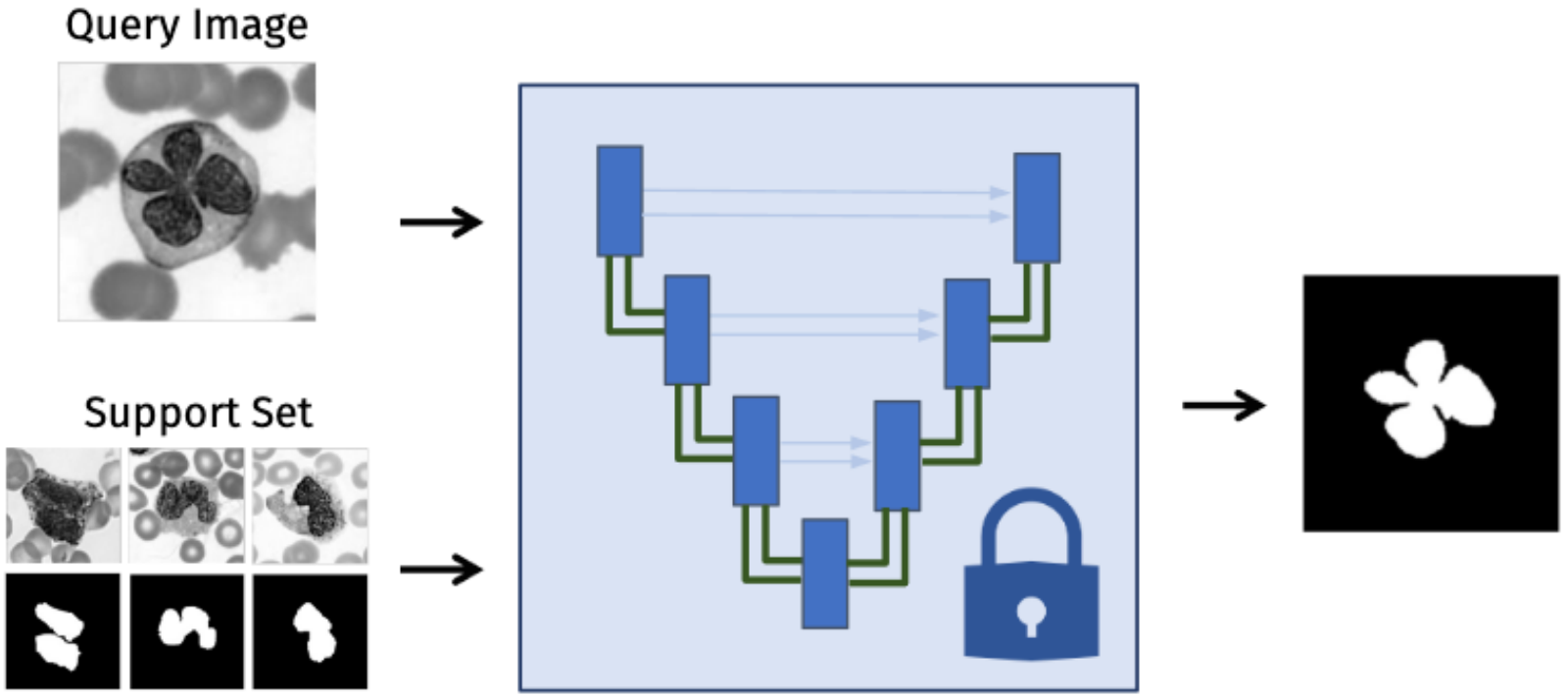
No prompt

Several extensions for medical images, e.g., MedSAM2 [20]

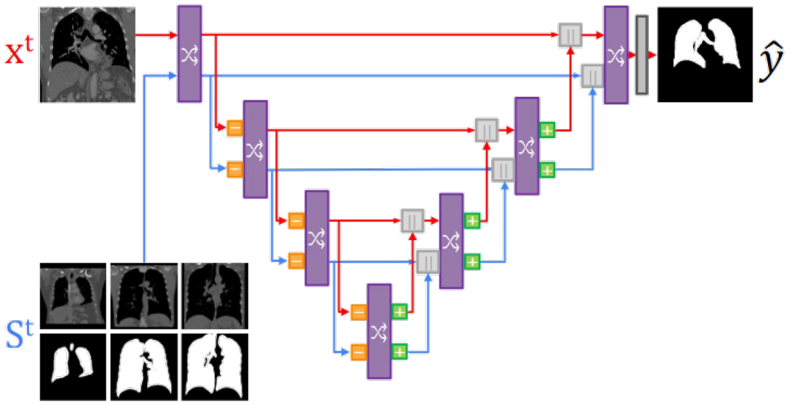
[20] MedSAM2: Segment Anything in 3D Medical Images and Videos. J. Ma et al. Arxiv, 2025.

Foundation models for few-shot segmentation

- Large-scale pre-training for few-shot learning
  - Optimize a model to be effective when only few annotated samples available
- **The UniverSeg family:** input image + small context of segmented masks

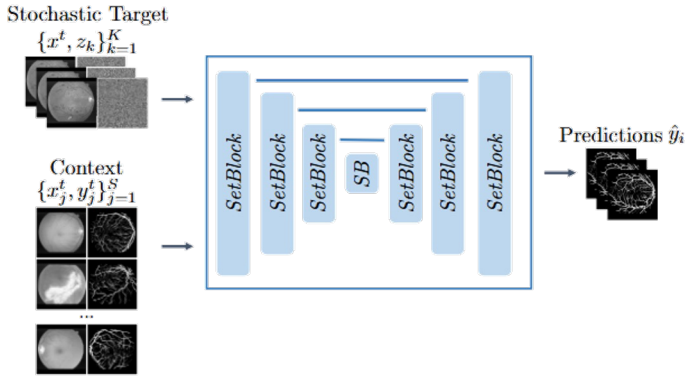


The UniverSeg family



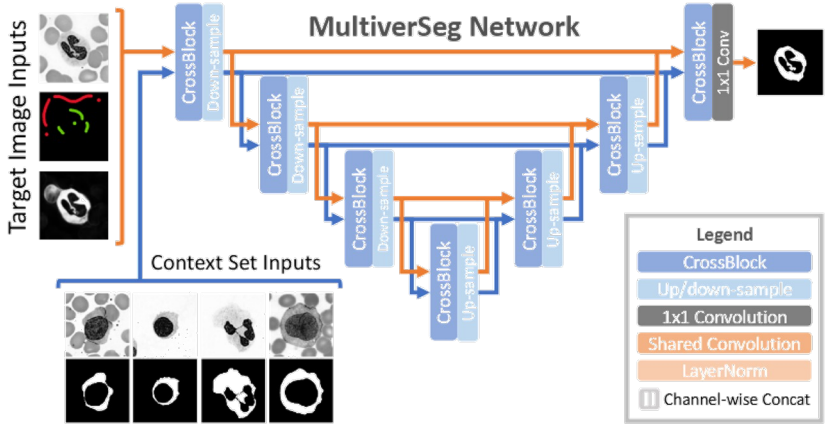
UniverSeg, *Butoi et al.*
ICCV'23

- Multi-task, multi domain
- CrossBlock



Tyche, *Rakic et al.*
CVPR'24

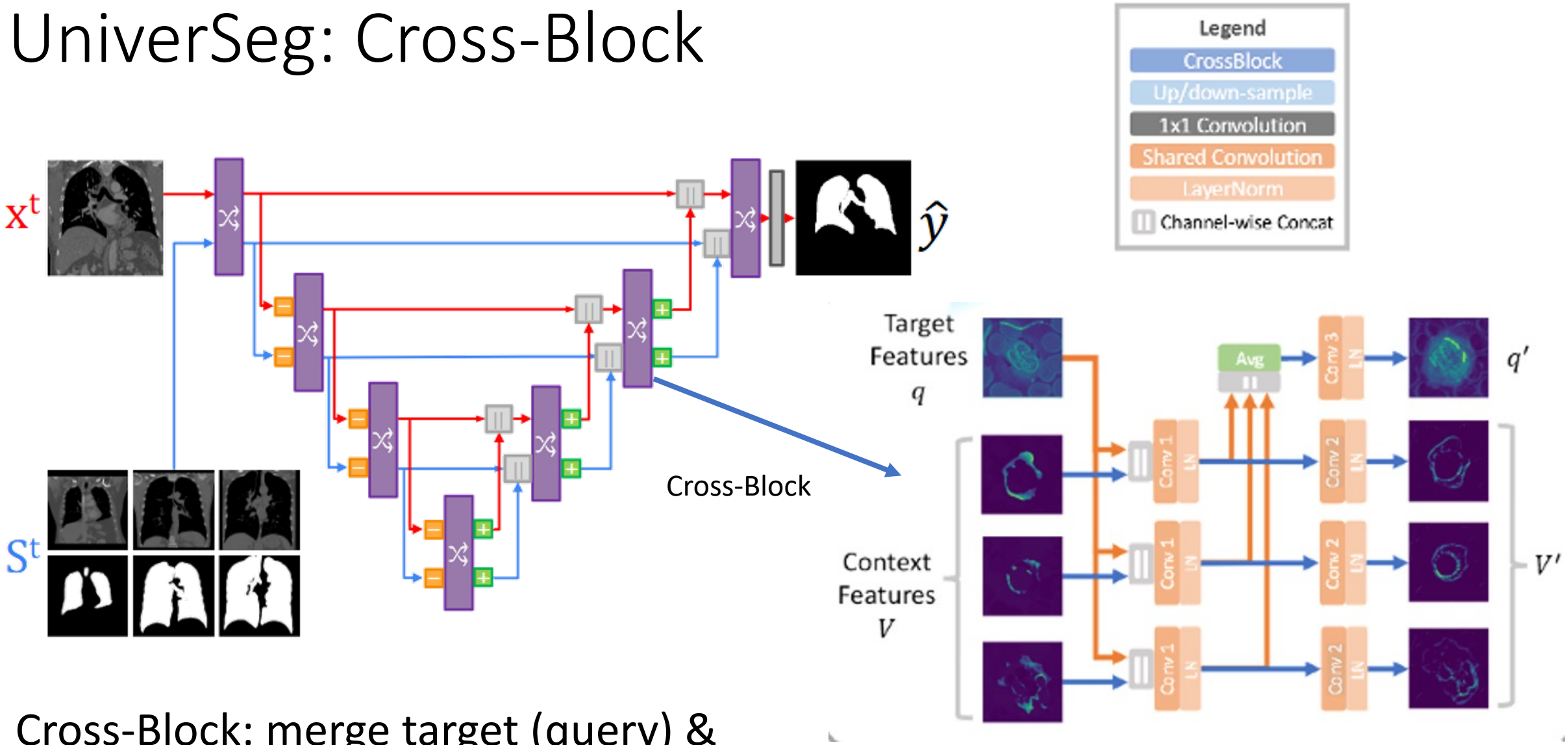
- Stochastic segmentation:
- SetBlock



MultiverSeg, *Wong et al.*
ICCV'25

- User interactions, e.g., scribbles
- Combines with context

UniverSeg: Cross-Block



- Cross-Block: merge target (query) & context(image + masks) features

UniverSeg training

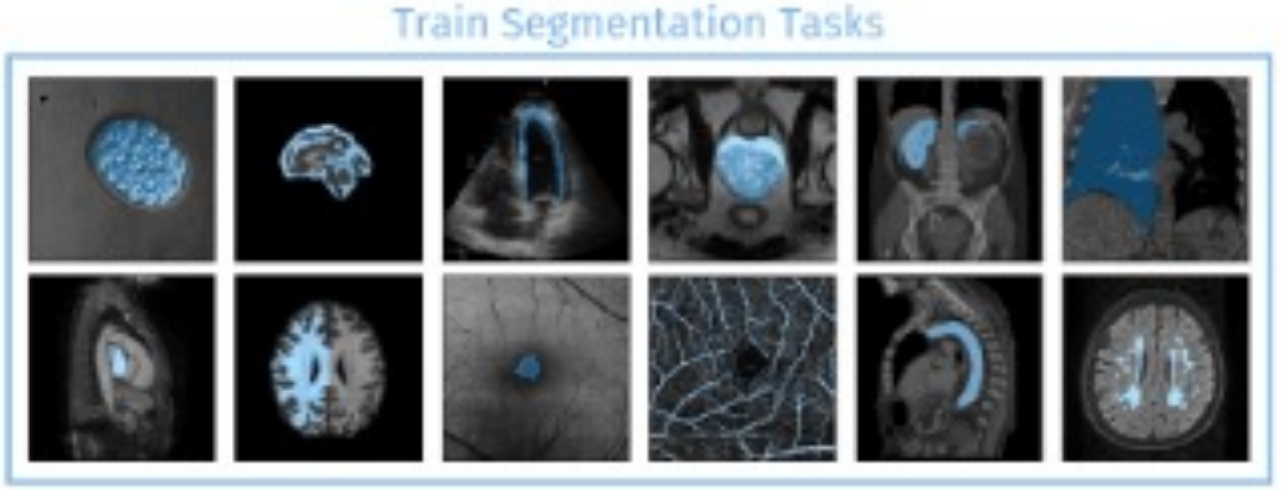
1. Multi-task training: different modalities & classes

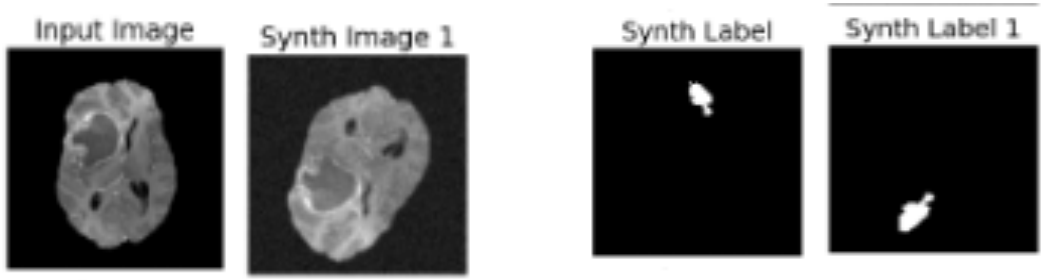
2. Synthetic task augmentation
   - Unsupervised masks, e.g., SAM

   $\Rightarrow$ Additional training signal

3. Test : OOD evaluation

   => zero-shot generalization





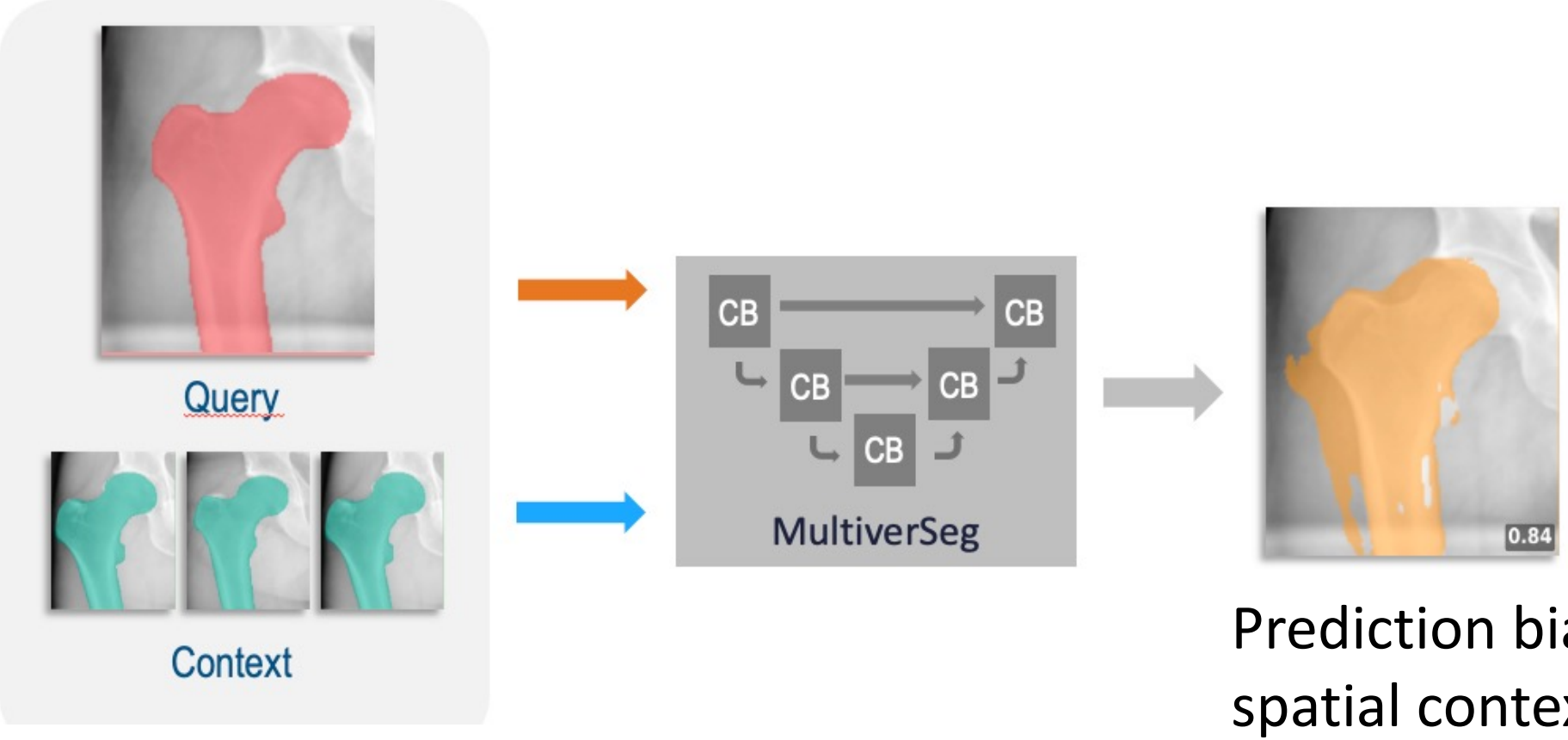
UniverSeg: strengths & limitations

✓ Adapt to distribution shift via few-shot ICL

- Crucial in medical images, variations: devices, anatomy, contrast product, etc

✗ Query/Context alignment assumption

✗ Context entanglement



Prediction biased by spatial context

Foundation models: risks and challenges

- Access to huge-scale datasets
  - Diverse, anonymized data
  - Pre-training on generalist data?
- Robustness and certification: uncertainty, OOD detection, stability, etc
  - A general issue in deep learning, exacerbated with general-purpose AI systems
  - Crucial and especially sensible in healthcare
- Adaptation: few-shot, zero-shot, test-time
- Explainability, interpretability: harder or easier?
- Ethical considerations
  - Biases and fairness/discriminability
  - Privacy, informed consent, transparency

Thank you for your attention!

Questions?