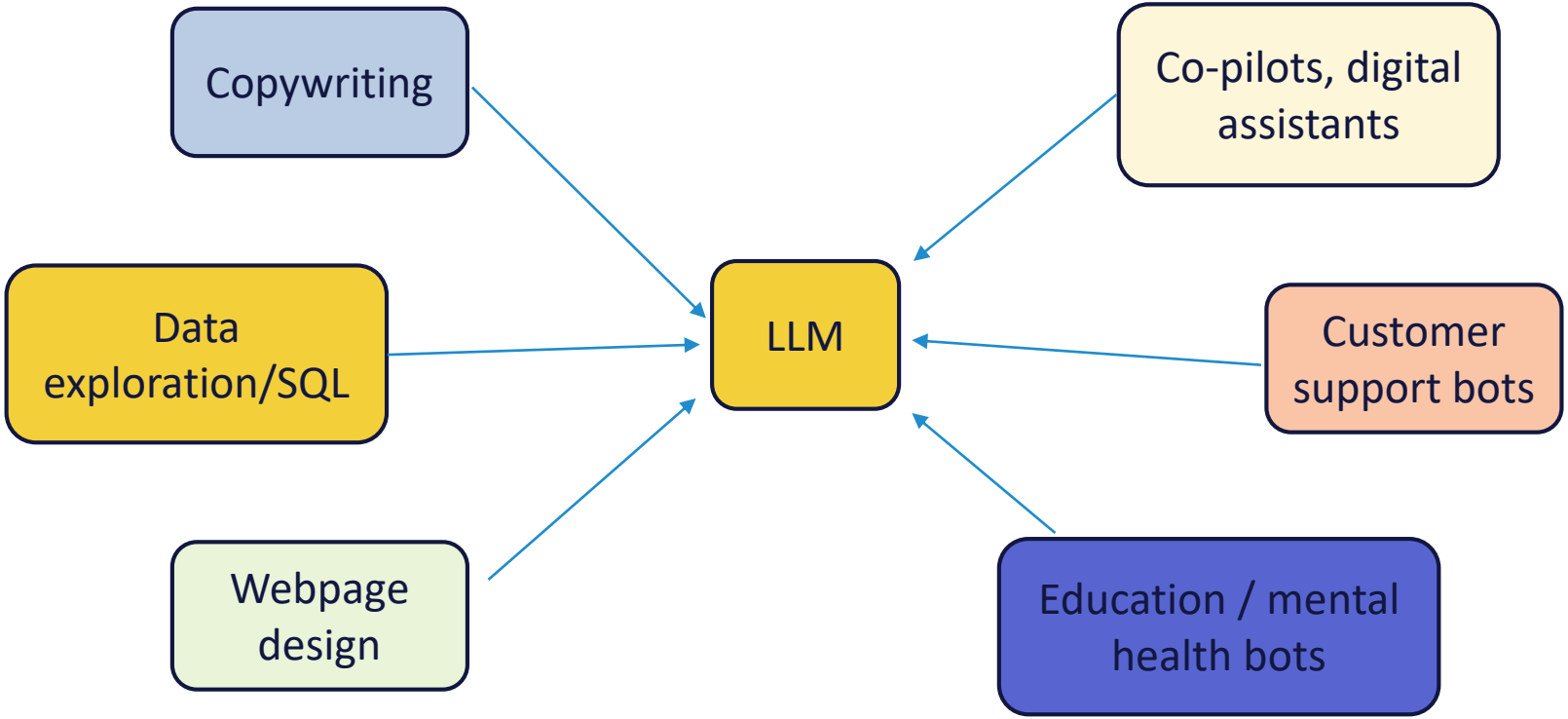
Unveiling the Power of Multi-Modal Large Language Models: Revolutionizing Perceptual AI

Istvan Fehervari

Director, Data and ML

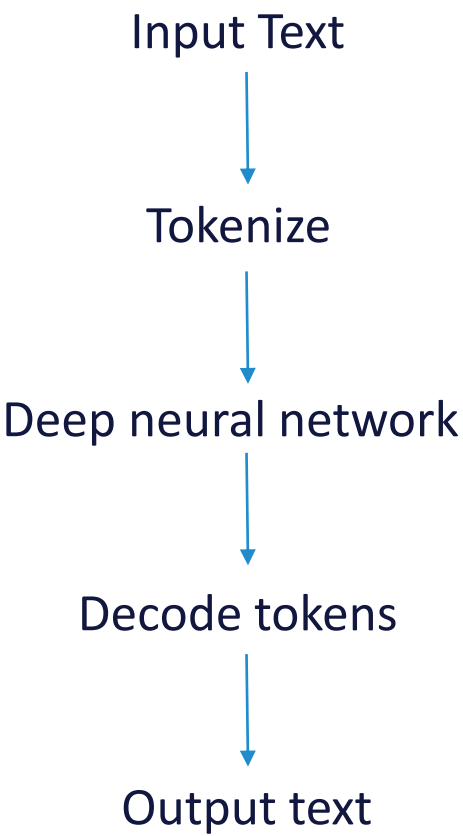
BenchSci

The era of large language models



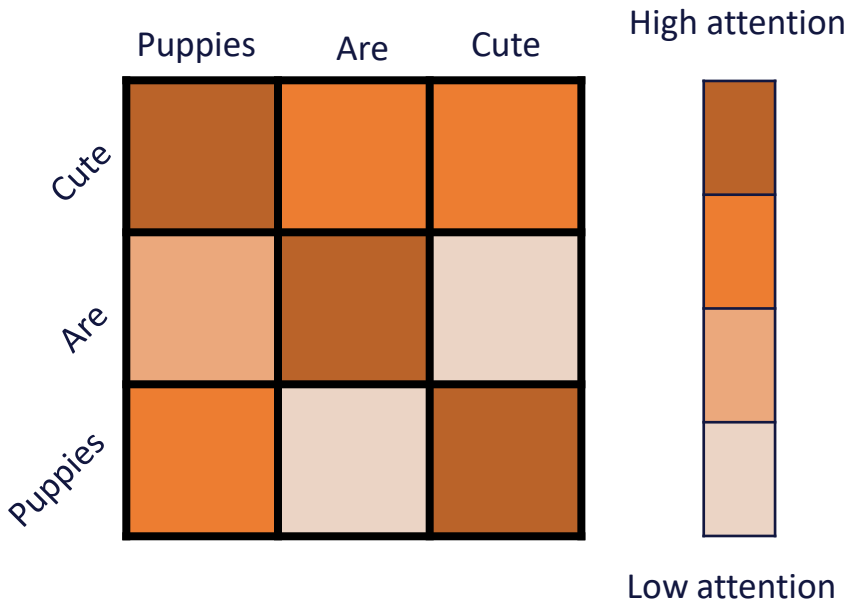
Large language models

- Revolution started in machine translation
- Context-sensitive next token prediction via attention
- Transformer blocks composed of layers of attention and feed-forward blocks
- Encoder-decoder architectures
- Intelligence emerges through scale



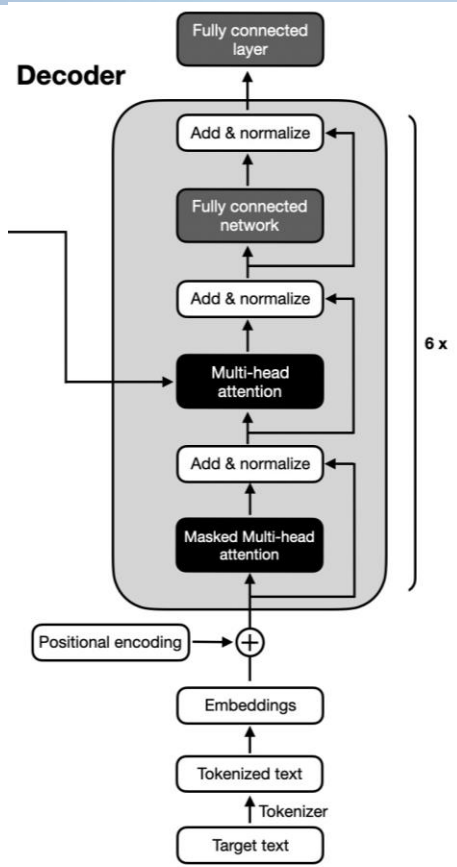
Why the attention mechanism is critical

- Enables learnable weights of content pieces for arbitrary context → better reasoning
- Works very well for ordered and unordered sets
- Works well with external contexts, e.g., cross-attention with vision
- All modern ML models leverage some form of attention



Building blocks of LLMs

- LLMs are composed of decoder blocks
- Inputs are all previous tokens – including predicted ones
- Decoder outputs a token distribution based on all previous tokens
- During generation we sample tokens from the output distribution
  - Temperature
  - Top-k / top-n



Training LLMs

- Supervised training
  - Input – output pairs (e.g., translation)
  - Fine-tune on specific task
- Self-supervised training
  - Next/masked token prediction – needs a large body of data
- Reinforcement learning human feedback (RLHF)
  - For instruction tuning, use human ranking to learn a reward function

Main foundational open LLMs

- LLaMA (2023/2) - 7B / 13B / 33B / 65B
- Falcon (2023/5) - 7B / 40B / 180B
- LLaMA2 (2023/6) - 7B / 13B / 70B
- Mistral (2023/9) - 7B (based on LLaMA2)
- Vicuna / Alpaca (based on LLaMA)
- Phi-2 (2023/12) - 2.7B
- Mixtral (2023/12) - 8x7B

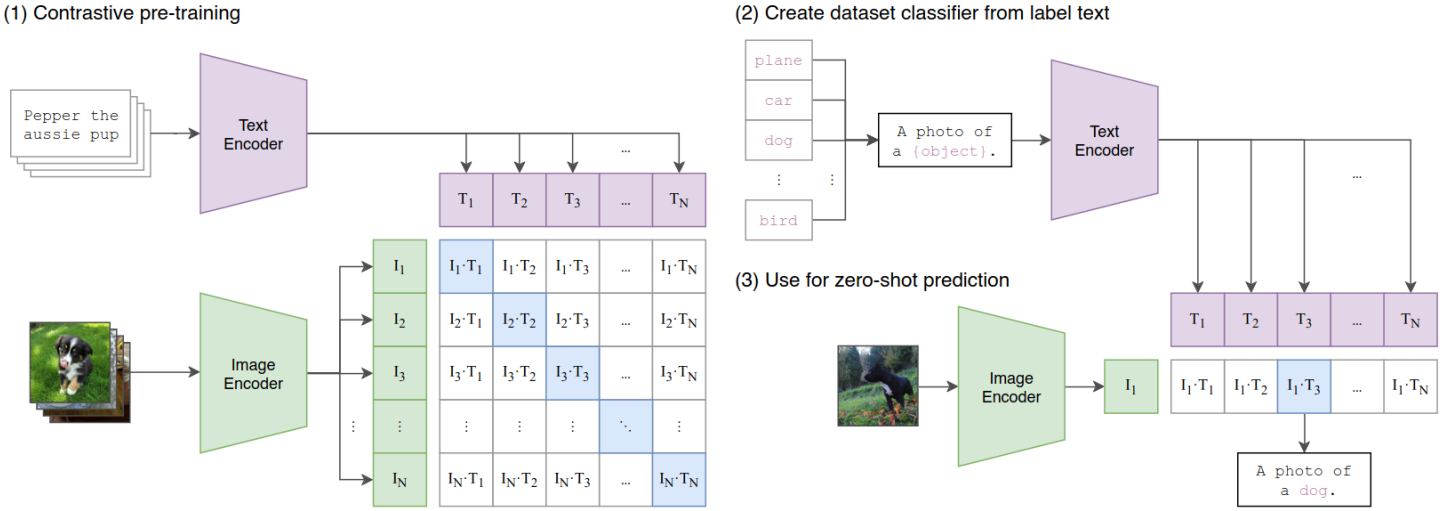
Perception via Language

Rise of a new dataset

- Annotated class labels are expensive → captions are abundant
- Era of natural language supervision
- WebImageText dataset: 400 million images with text captions
  - Created with web scraping
  - Query words are composed of all words occurring at least 100 times on Wikipedia

CLIP: combining language and vision

- Predicting captions directly does not scale well
- Instead, predict how well a text description and an image “fit together”
- First example of prompt engineering in vision



Object detection with CLIP

Object detection with Grounded DINO

- Open-vocabulary detection
- Text backbone is a pretrained transformer like BERT
- Text-image and image-text cross-attention at several stages

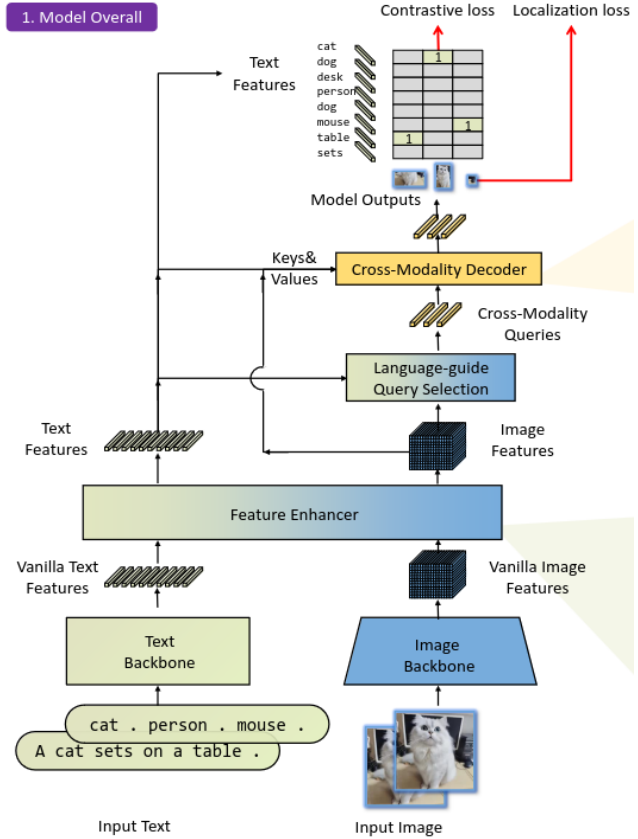


Image segmentation with CLIP

Segmentation with Grounded SAM

- Open-vocabulary segmentation
- Detect boxes with Grounded DINO → Predict mask with SAM

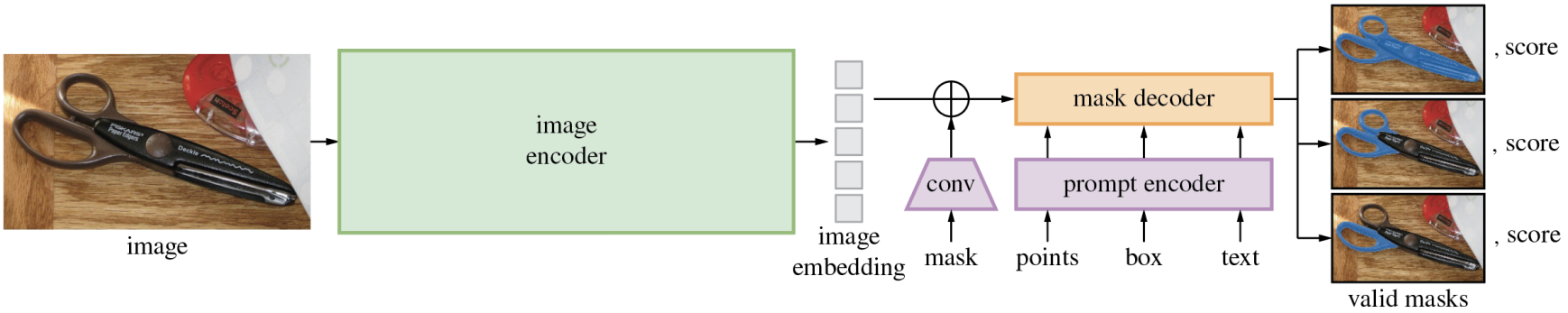
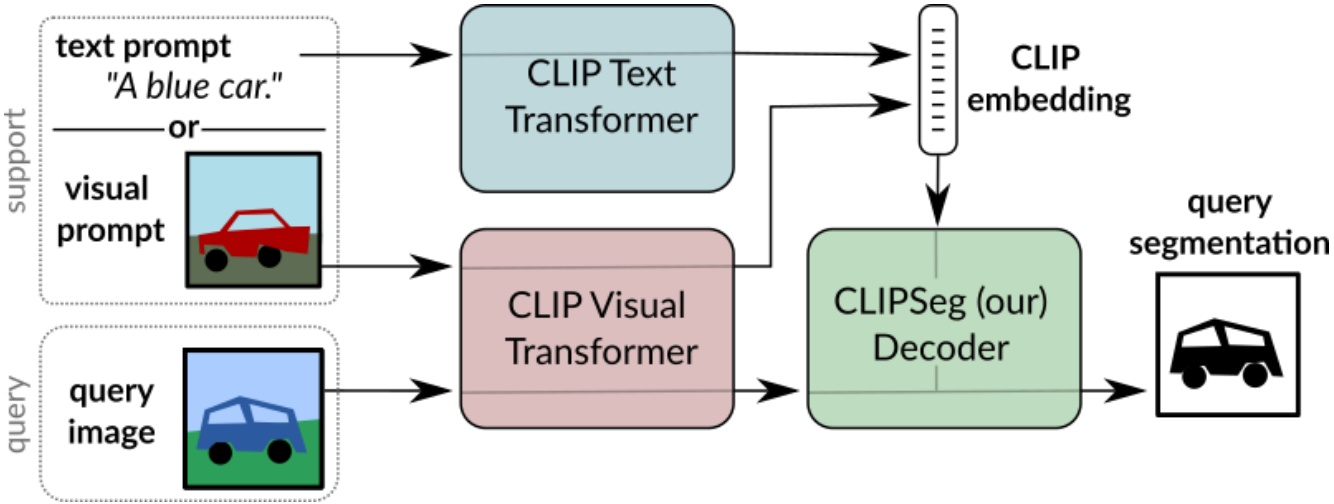


Image segmentation with CLIP

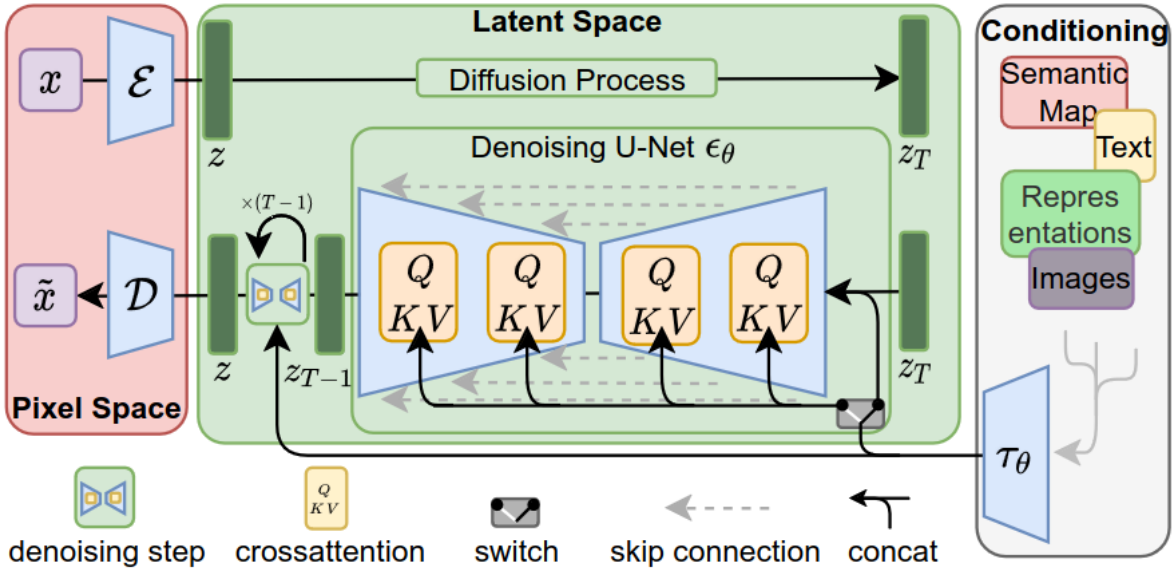
- CLIPSeg – Image segmentation with prompts



Lüddecke et al. - Image Segmentation Using Text and Image Prompts, CVPR 2022

Image generation with CLIP

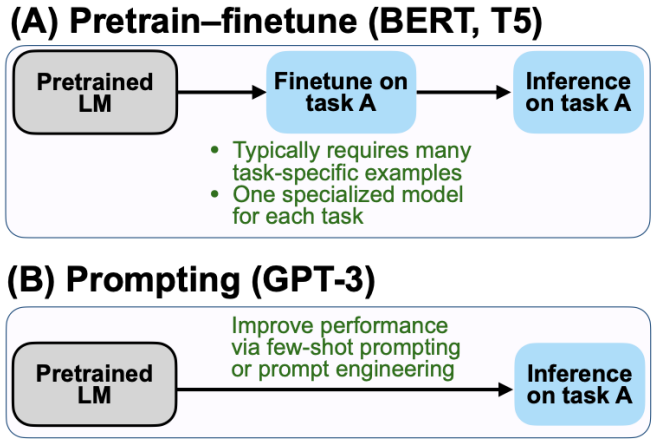
- Stable Diffusion uses CLIP text embeddings



LLMs with Vision

Learning paradigms for (V)LLMs

- We want our models to **reason** over visual input
- What data is needed?



Training V-LLMs

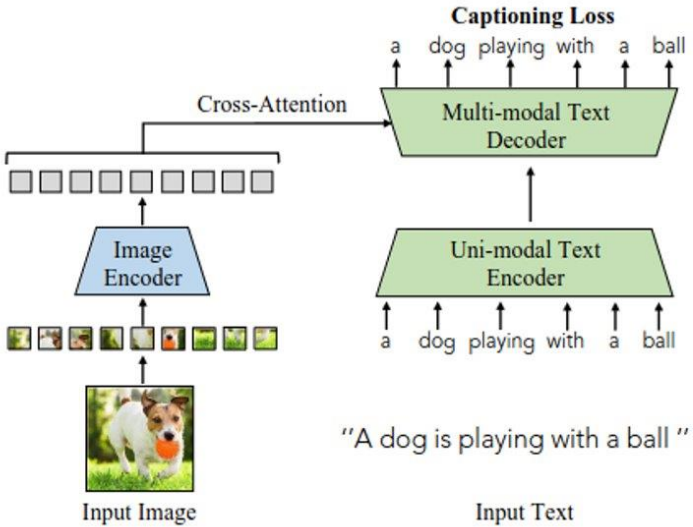
Dataset building via

1. Image captioning (ideally with bounding box ground truth)
2. Visual QA datasets
3. Synthetic: create (2) from (1)

Can be done manually or LLM-assisted

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

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



Learning paradigms for (V)LLMs

- (Multi-modal) in-context learning (e.g., Otter)
  - Inject demonstration set to context
  - Requires large context
  - Can be used to teach LLMs to use external tools
- (Multi-modal) chain-of-thought (e.g., ScienceQA)
  - Immediate reasoning steps for superior output
  - Adaptive or pre-defined chain configuration
  - Chain construction: infilling or predicting

<BOS> Below are some examples and an instruction that describes a task. Write a response that appropriately completes the request

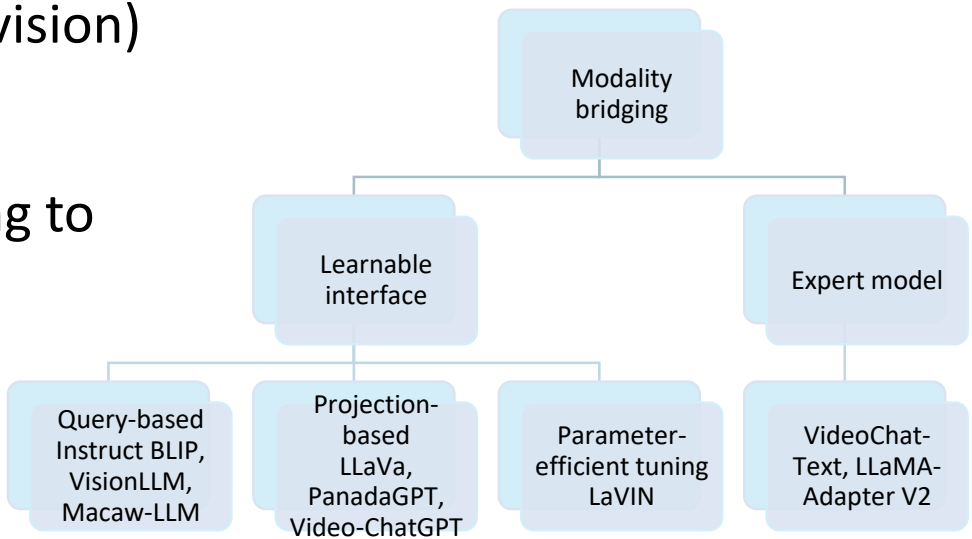
Instruction: {instruction}
Image: <image>
Response: {response}

Image: <image>
Response: {response}

Image: <image>
Response: <EOS>

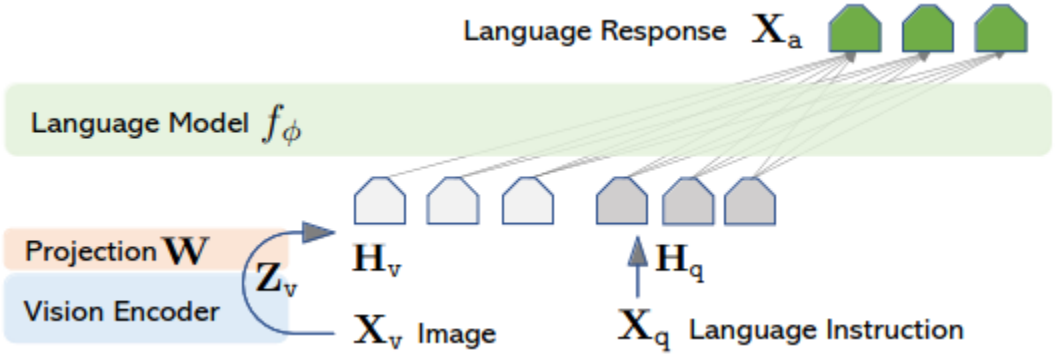
LLMs with vision capabilities

- Learnable interface between modalities
- Expert model translating (e.g., vision) into text
  - Special tokens/function calling to access aux models



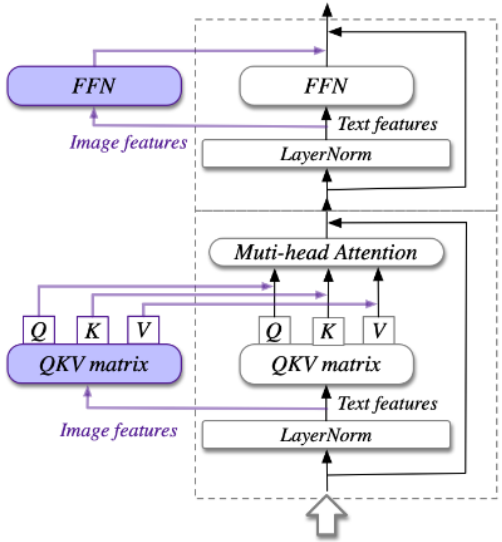
Modality bridging with shallow alignment

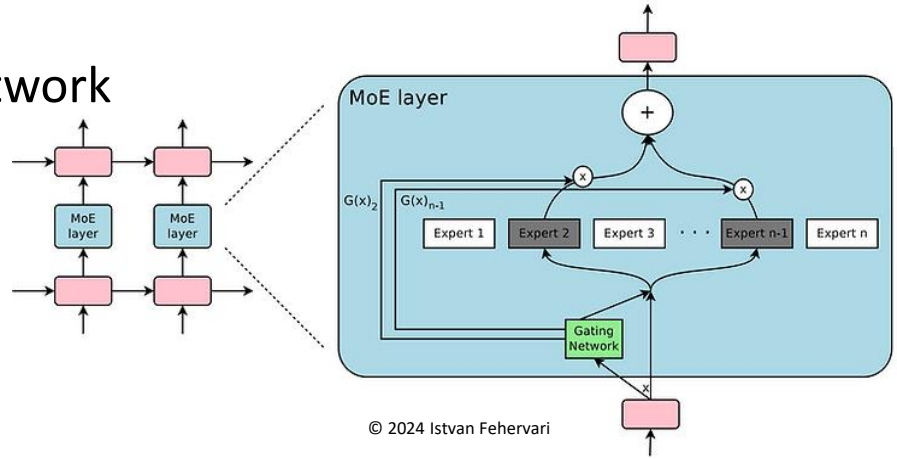
- Use CLIP to map vision and language tokens to the same latent space (shallow alignment) – LLaVa-1.5
- Keep LLM and image encoder frozen – only train a shallow projection layer



Deeper alignment: Mixture of experts with vision

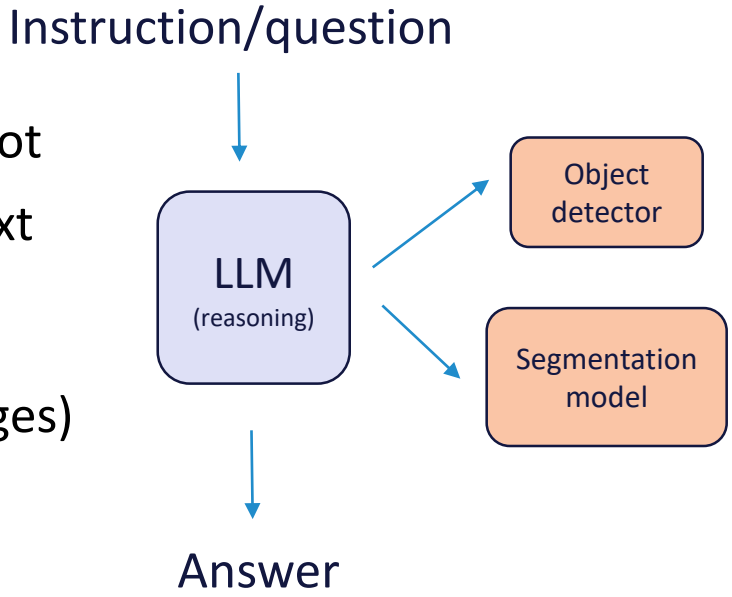
- Visual Experts – Mixture of experts with vision, e.g., CogVLM
- Experts are separate feedforward layers
- Only a few experts are activated during inference
- Learn a gating network





LLM-aided visual reasoning

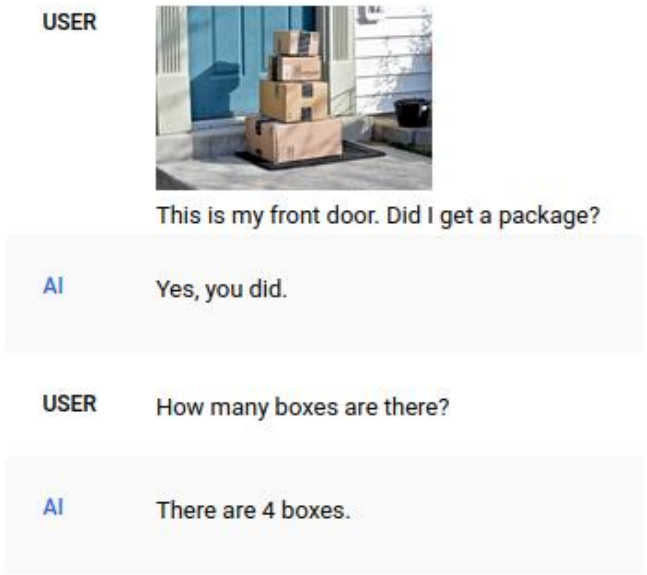
- LLM function calling
  - Controller – task planning
  - Decision maker – summarize, continue or not
  - Semantic refiner – generate text wrt. context
- Strong generalization
- Emergent ability (e.g., understand meme images)
- Better control



Vision-language on the Edge

Applications

- Programming → natural language instructions
  - Training free solutions
  - Shorter time-to-market
  - Short lead time to adapt to changing environments



Applications

- Control: more natural, frictionless UX
  - Voice or chat to control/monitor devices/networks
  - Answer usability questions (no more manuals)
  - Personalized onboarding to new devices
- Feedback:
  - Output is interpreted without human-in-the-loop (e.g., alarm systems)

Challenges

- LLMs need lot of resources (compute, memory)
  - Visual input, CoT requires larger context
  - Latency is still an issue on the edge
- Output control of LLMs is still unsolved, prone to hallucinations
- AI safety – bias is an unsolved issue
- AI alignment is an upcoming field of research

Conclusions

- Language as control brought tremendous improvements
- LLMs can operate very well with visual signals
- Future products will be more user-friendly, more natural
- Faster time to market, better adaptability both tech and business
- Performance on the edge today is a challenge but will be solved
- AI safety / alignment is the new challenge without a clear answer in sight

Questions?

Resources

- Yin et al. - A Survey on Multimodal Large Language Models
- Zhang et al. - Vision-Language Models for Vision Tasks: A Survey
- Yin et al. - A Survey on Multimodal Large Language Models
- Lüddecke et al. - Image Segmentation Using Text and Image Prompts
- Liu et al. - Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection
- Kirillov et al. - Segment Anything
- Wang et al. - CogVLM: Visual Expert for Pretrained Language Models
- Liu et al. - LLaVA: Large Language and Vision Assistant
- Li et al. - Otter: A Multi-Modal Model with In-Context Instruction Tuning