



ISCAS Tutorial



New Era of Artificial Intelligence: Unleashing the Power of Large Models in Visual Applications



Jiaying Liu



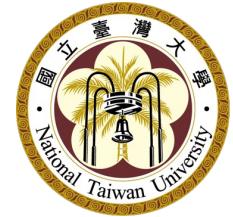
Wen-Huang Cheng



Shuai Yang

May 19, 2024 @ Singapore

Agenda



- **Introduction**
 - 13:30 – 13:40 (Wen-Huang)
- **Part 1: Specializing Large Models for Domain-Specific Vision Tasks**
 - 13:40 – 14:15 (Wen-Huang)
- **Part 2: AIGC for Image and Video Generation (I): Basic**
 - 14:15 – 15:00 (Shuai)
- **Coffee Break**
 - 15:00 – 15:30
- **Part 2: AIGC for Image and Video Generation (II): Method**
 - 15:30 – 16:30 (Shuai)
- **Part 3: Employing Diffusion Models for Low-level Vision**
 - 16:30 – 17:00 (Jiaying)



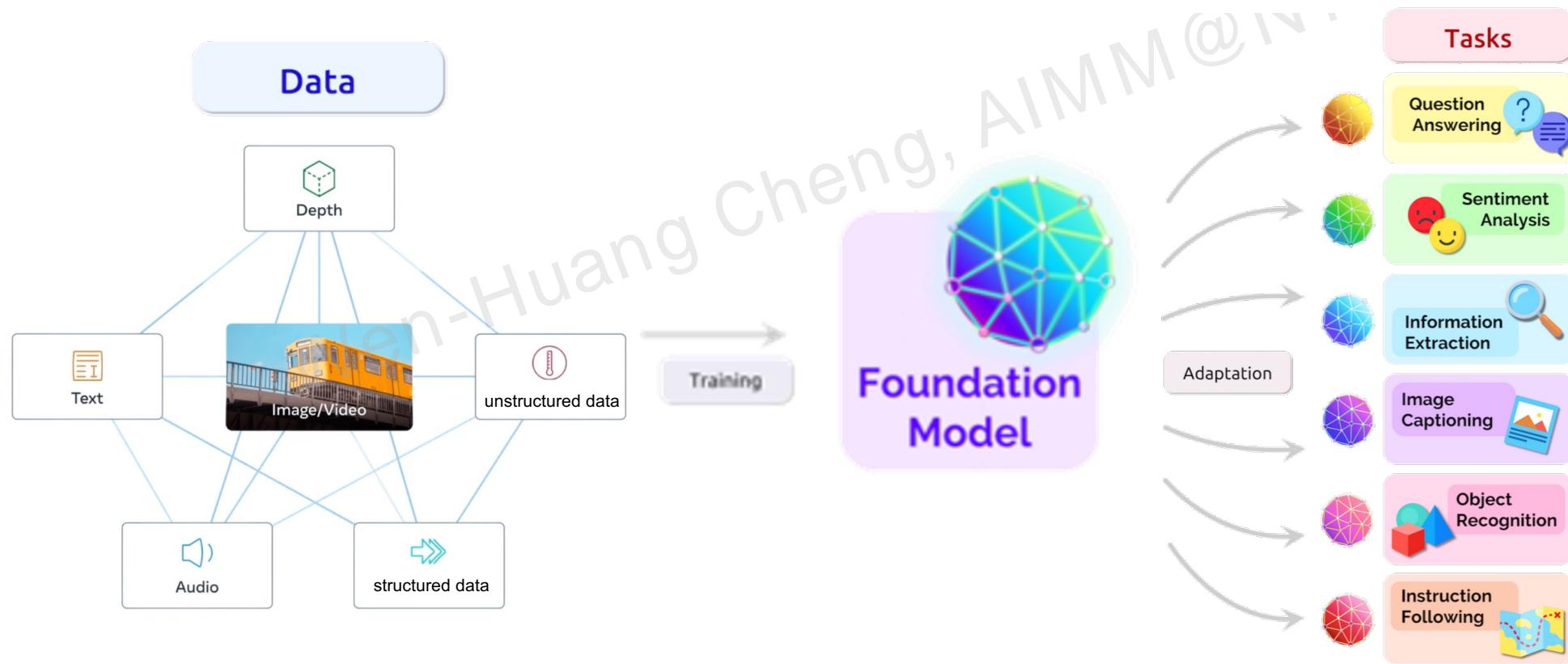
Introduction

**Wen-Huang Cheng,
National Taiwan University**



Foundation Model

- A foundation model is a large deep learning model that is trained on broad data such that it can be applied across a wide range of use cases.



Foundation Models in Visual Applications



- Foundation models, particularly Large Vision Models (LVMs), open up a world of possibilities for visual applications.

Image Captioning



“A city view showcasing a blend of modern high-rise buildings and historic architecture.”

Visual Question Answering



Q: “What vehicles are these?”
A: “Yellow taxi”

Foundation Models in Visual Applications



- Foundation models, particularly Large Vision Models (LVMs), open up a world of possibilities for visual applications.

Image Generation



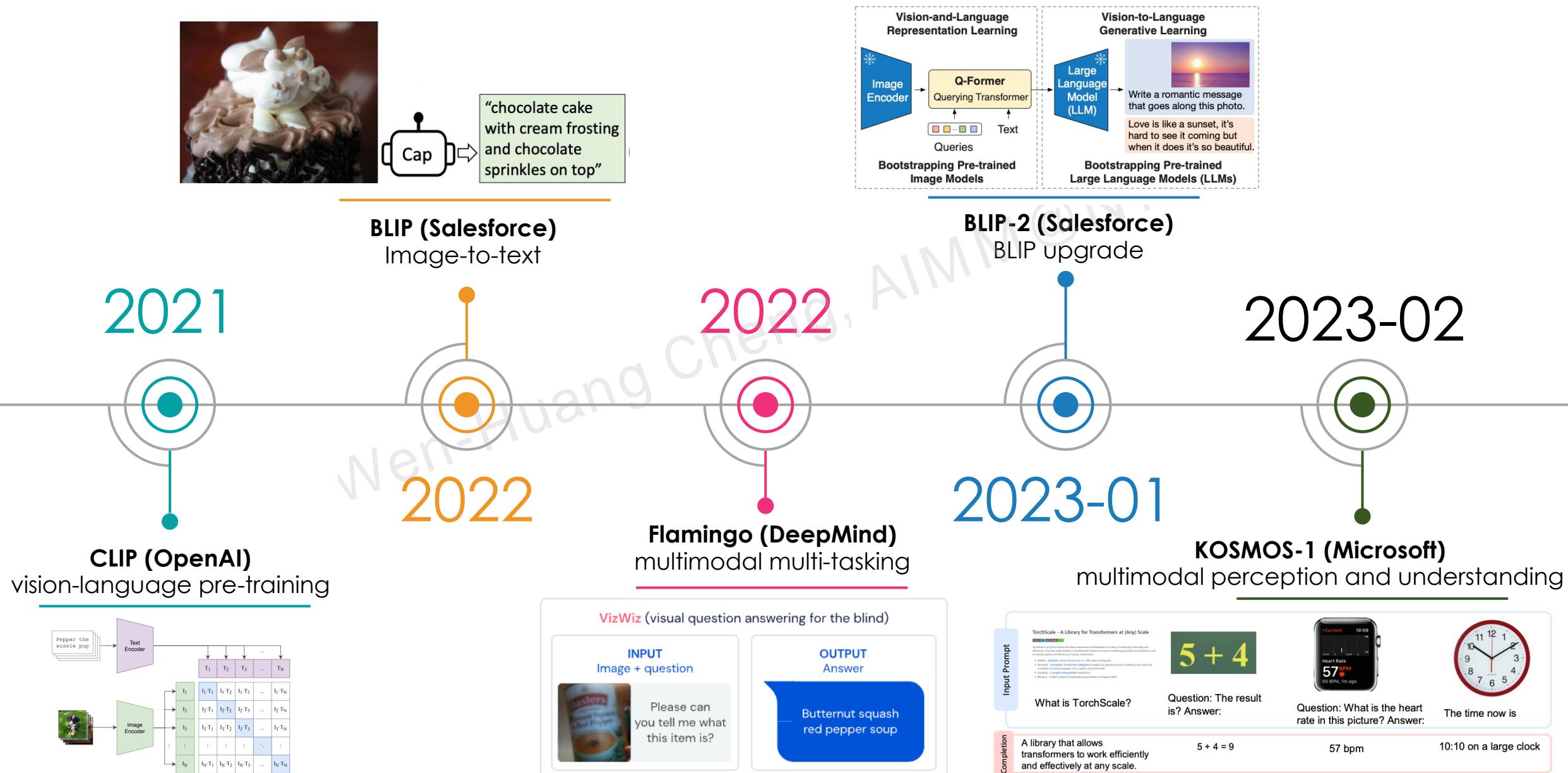
Prompt: “Tokyo, Shinjuku”

Video Generation

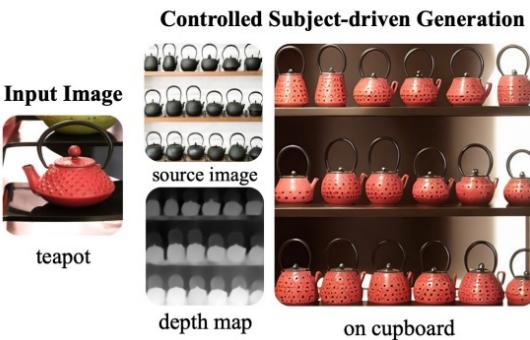


Prompt: “a scuba diver discovers a hidden futuristic shipwreck, with cybernetic marine life and advanced alien technology”

LVMs since 2023 Review: A Timeline



LVMs since 2023 Review: A Timeline



Prompt:
This is the front-view image of your car. Please describe the current traffic scenario you're in.



GPT-4V:
You're on a city street with multiple lanes. There are parked cars on both sides of the road. Ahead, there's moderate traffic with vehicles moving in both directions. There's also a pedestrian crossing zone visible. Buildings and trees line the street, and there are various street signs along the way.

Prompt:
This is the front-view image of your car. Please describe the current traffic scenario you're in.



GPT-4V:
You are on a road during twilight or early evening. The road has a pedestrian crossing ahead. There's a vehicle with its tail lights on, indicating it's either stationary or moving away from you. There are traffic signs on both sides of the road, and the area appears to be moderately lit. Trees and other structures are visible in the background.

2023-05



InstructBLIP (Salesforce)
general-purpose multi-models



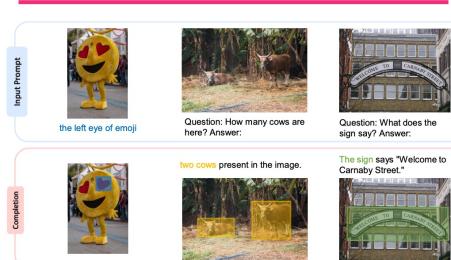
BLIP-Diffusion (Salesforce)
controllable generation

2023-06



2023-05

KOSMOS-2 (Microsoft)
KOSMOS-1 upgrade

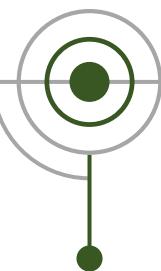


GPT-4V (OpenAI)
GPT-4 with vision

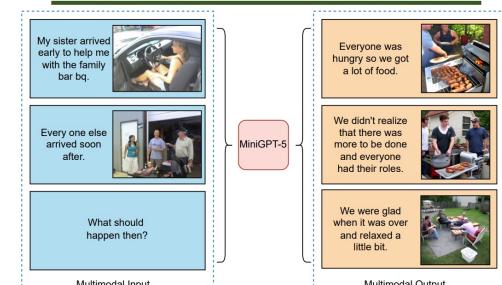


2023-09

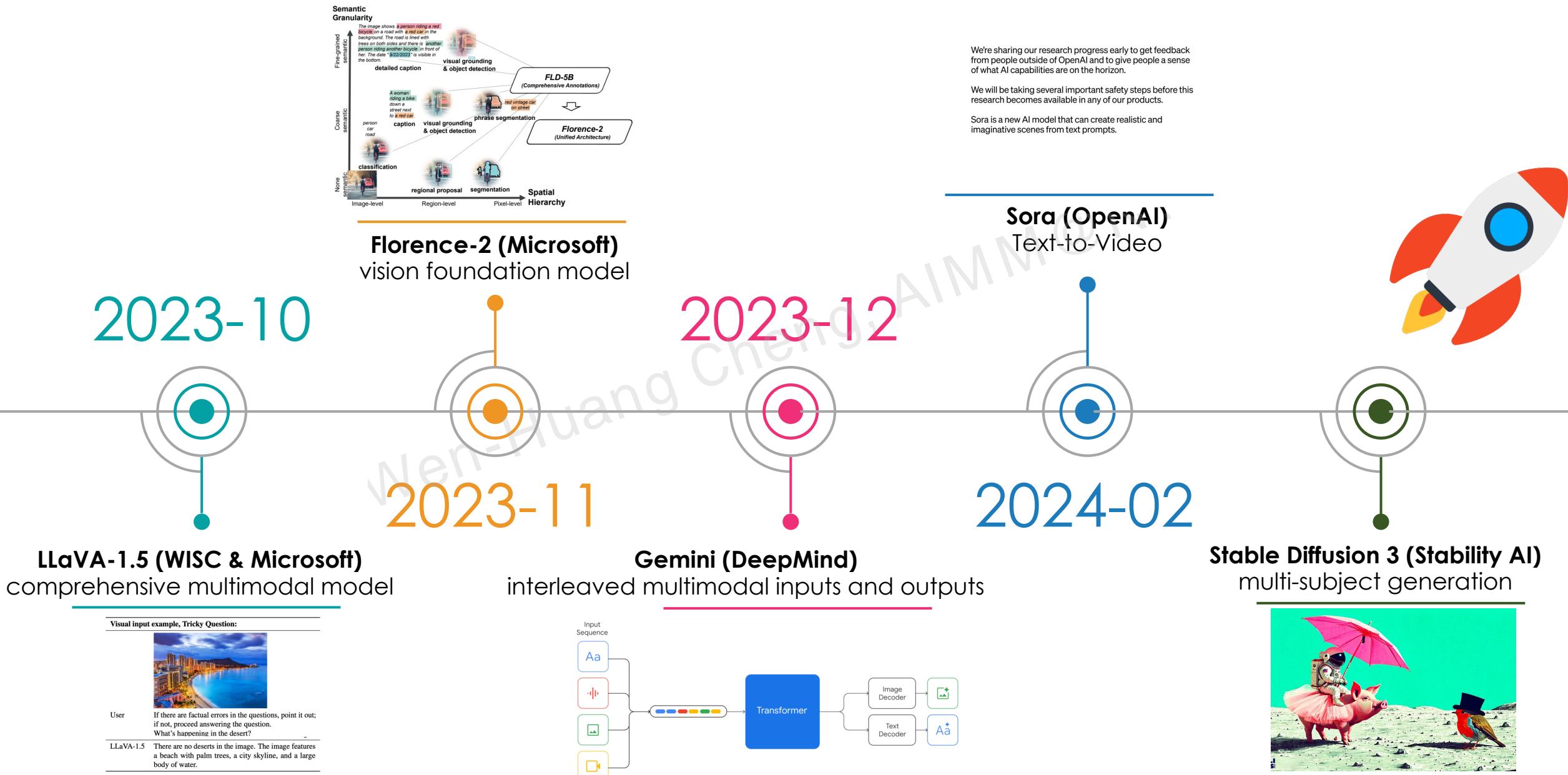
2023-10



MiniGPT-5 (UCSC)
multimodal inputs and outputs



LVMs since 2023 Review: A Timeline

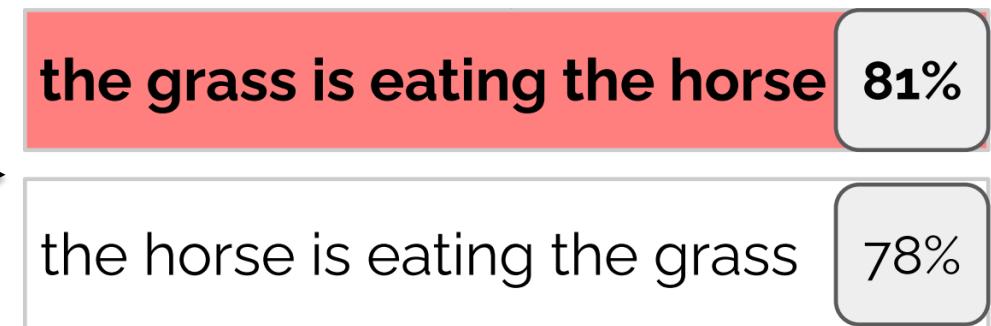




Challenges still persist with LVMs

- Compositionality issues

- The encoding of compositional relationships between objects and their attributes in text or images may not always be accurate.



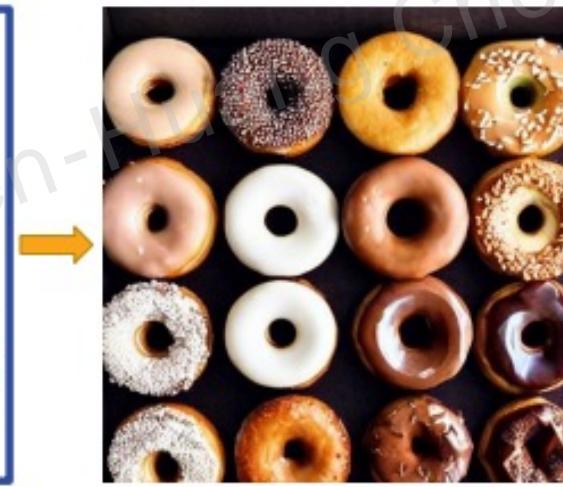


Challenges still persist with LVMs

- Inadequate detail control over generation
 - VLMs might overlook certain details and rather follow the visual or linguistic training prior.

A box contains ten donuts with varying types of glazes and toppings. Large square in the top red donut. Large square in the top right brown glazed chocolate donut.

Input Condition



Generated Image



A lion with a crown and flowers, the crown on the **bottom**, flowers on the **top**.



Organization of this tutorial

- Unleashing the Power of Large Models in Visual Applications

Part 1
**Specializing Large
Models for Domain-
Specific Vision Tasks**



Wen-Huang Cheng

Part 2
**AIGC for Image
and Video
Generation**



Shuai Yang

Part 3
**Employing Diffusion
Models for Low-level
Vision**



Jiaying Liu



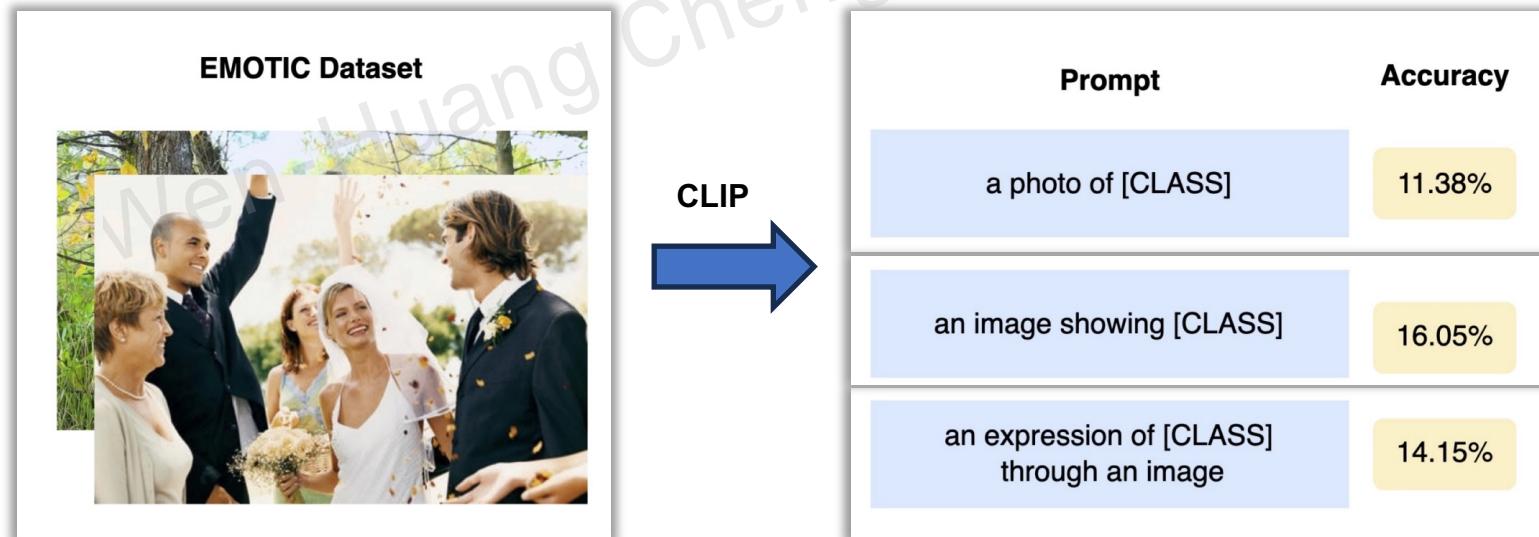
Specializing Large Models for Domain-Specific Vision Tasks

Wen-Huang Cheng
National Taiwan University



Limitations of Generic LVMs for Domain-Specific Vision Tasks (DSVT)

- Generic LVMs excel in open-world vision tasks, such as classification, detection, segmentation, and captioning. However, they are still in its infancy when it comes to specific domains.
 - A motivating example: Visual emotion understanding





Limitations of Generic LVMs for Domain-Specific Vision Tasks (DSVT)

- Generic LVMs excel in open-world vision tasks, such as classification, detection, segmentation, and captioning. However, they are still in its infancy when it comes to specific domains.
 - A motivating example: Visual emotion understanding



"Do people in the image feel fear."

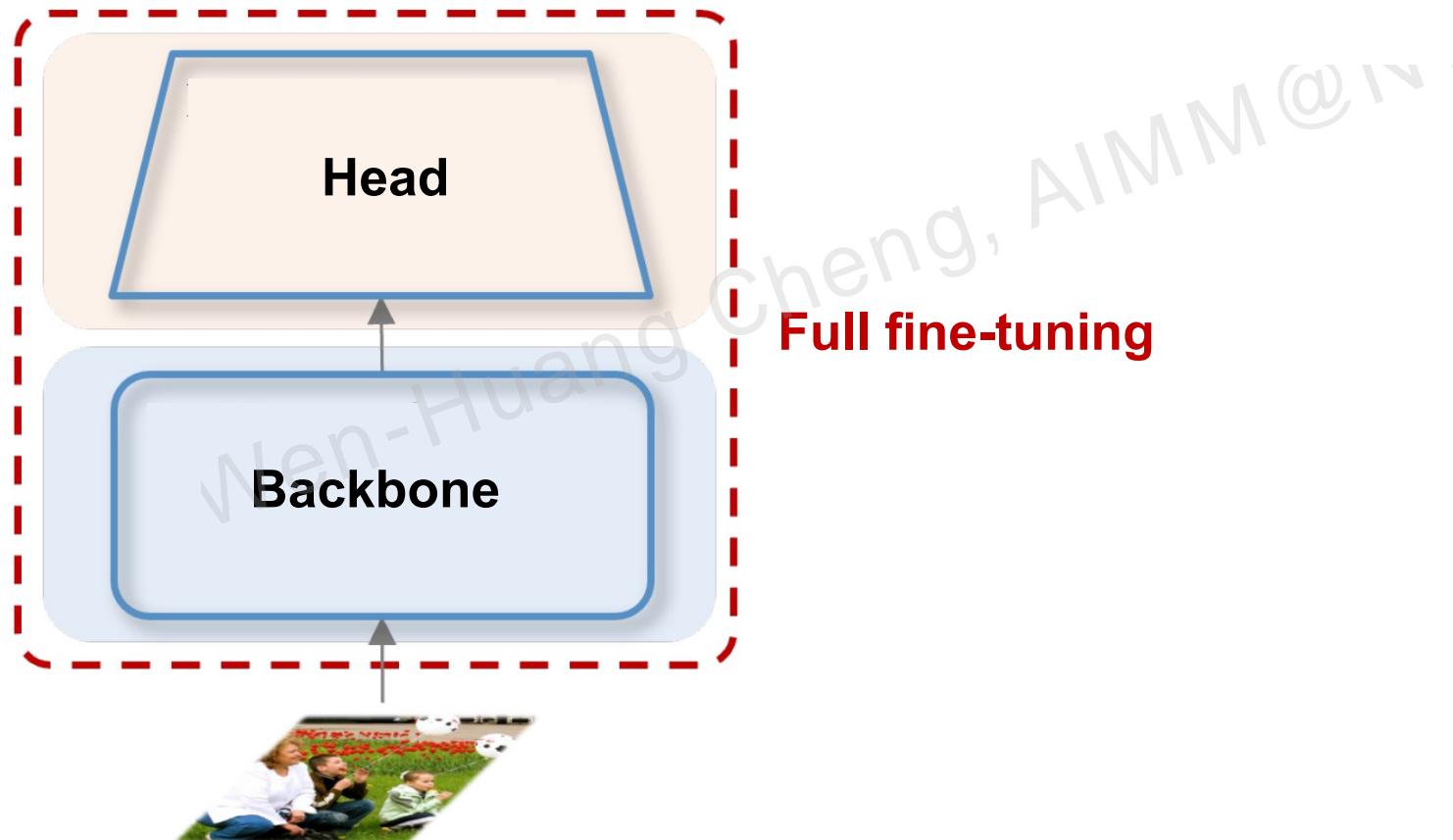
Gemini

Yes. The person in the image is a soccer player. He is smiling and lying on the grass. He is probably **happy** because he scored a goal or his team won the game.



Strategies for DSVT: Fine-tuning

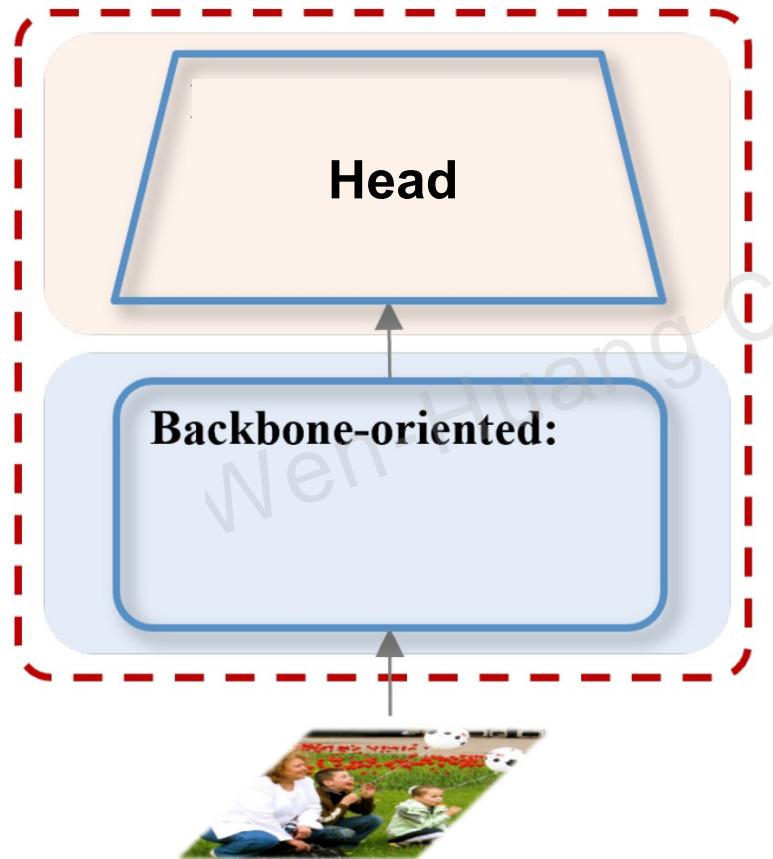
- Adapting LVMs to new tasks





Strategies for DSVT: Fine-tuning

- Adapting LVMs to new tasks

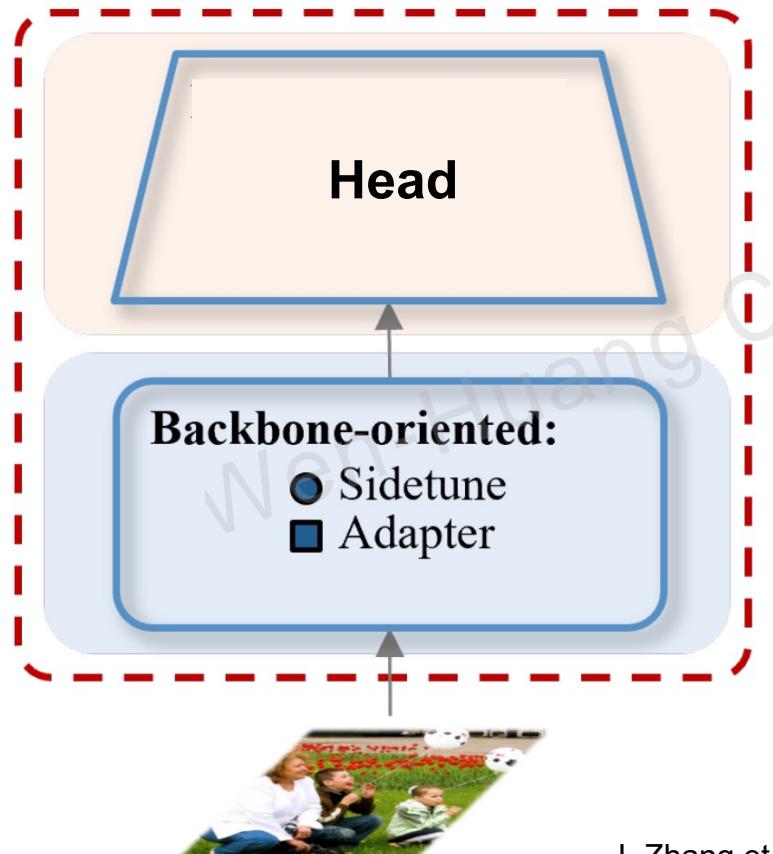


An alternative approach is to fine-tune only a subset of the parameters



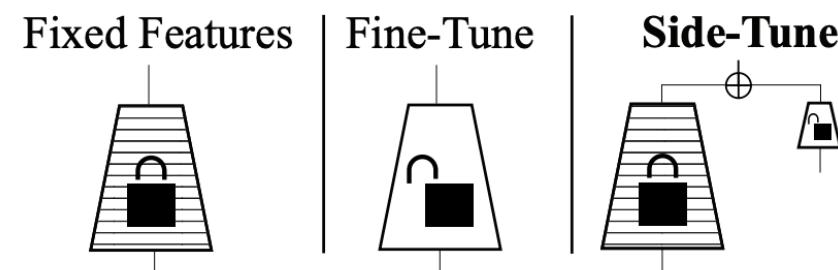
Strategies for DSVT: Fine-tuning

- Adapting LVMs to new tasks



An alternative approach is to fine-tune only a subset of the parameters

Example: Side-tuning adapts the pre-trained network by training a lightweight conditioned “side” network that is fused with the (unchanged) pre-trained network using summation.

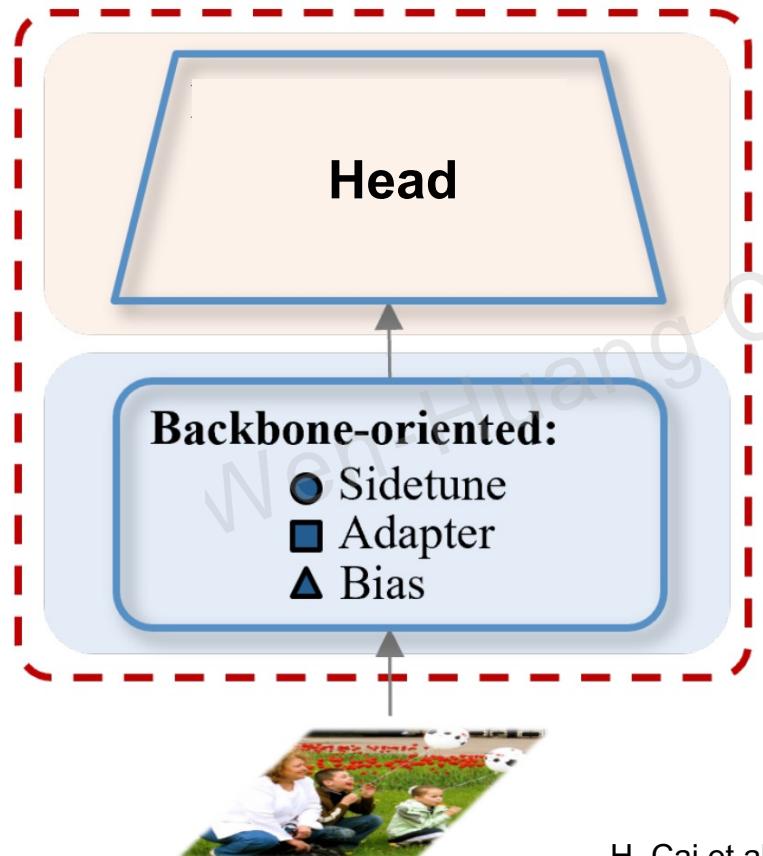


J. Zhang et al., “Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks,” ECCV 2020.



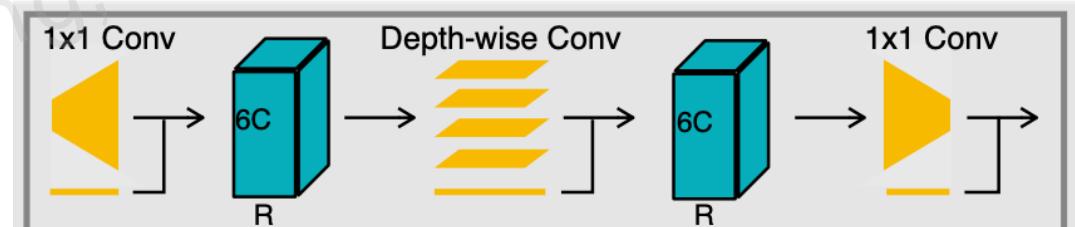
Strategies for DSVT: Fine-tuning

- Adapting LVMs to new tasks

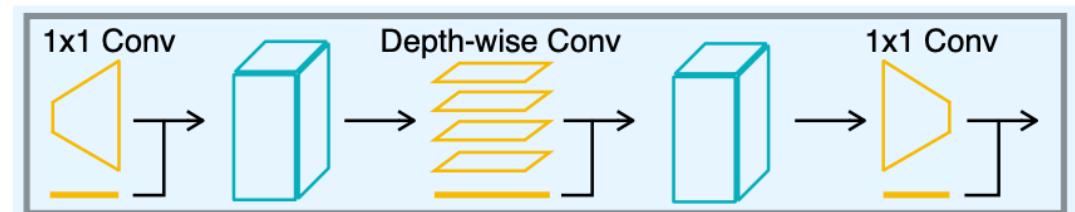


An alternative approach is to fine-tune only a subset of the parameters

Example: only the bias terms



(a) Fine-tune the full network (Conventional)



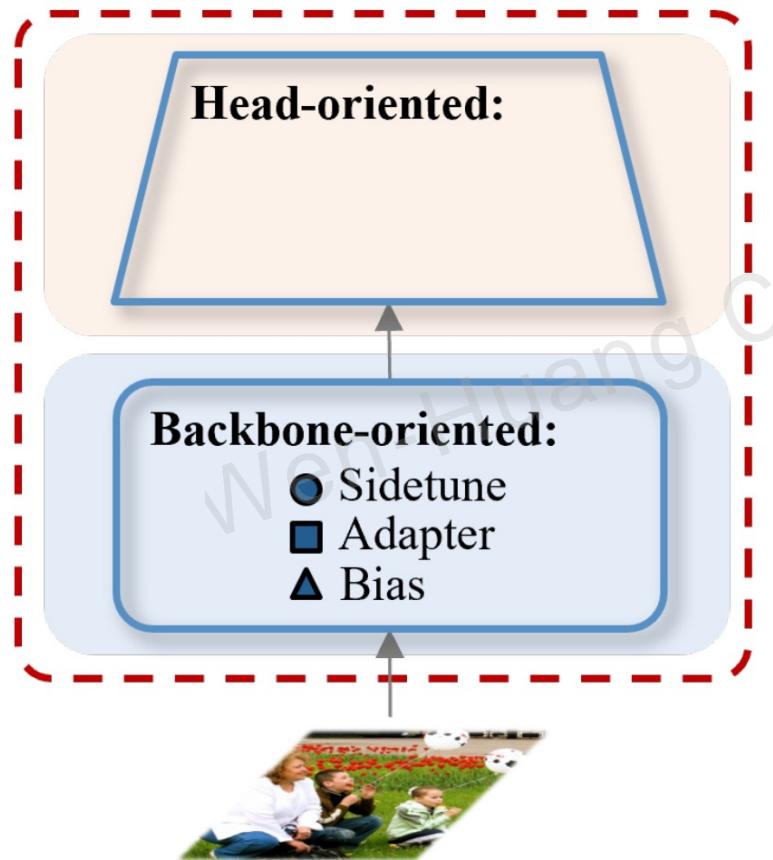
(b) Fine-tune bias only

H. Cai et al., "TinyTL: Reduce memory, not parameters for efficient on-device learning," NeurIPS 2020.



Strategies for DSVT: Fine-tuning

- Adapting LVMs to new tasks

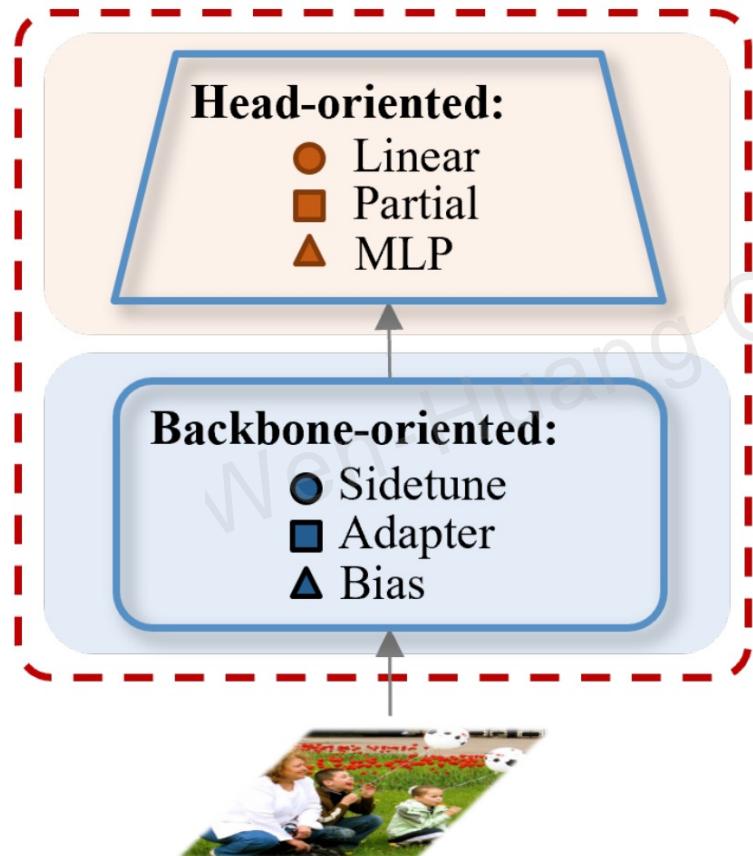


An alternative approach is to fine-tune only a subset of the parameters



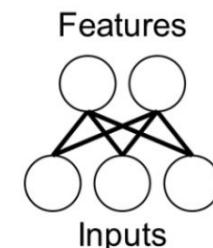
Strategies for DSVT: Fine-tuning

- Adapting LVMs to new tasks

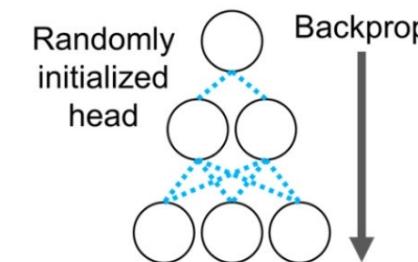


An alternative approach is to fine-tune only a subset of the parameters

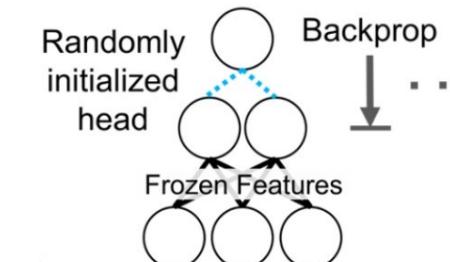
Pretraining



(a) Fine-tuning



(b) Linear probing

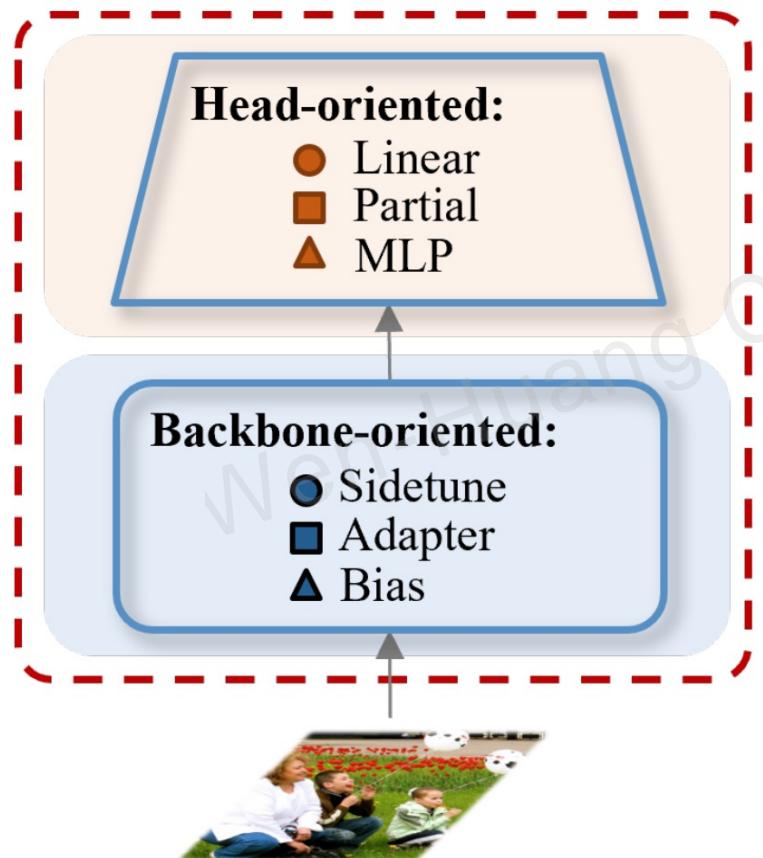


X. Chen, "An empirical study of training self-supervised vision transformers," ICCV 2021.
M. Jia, "Exploring Visual Engagement Signals for Representation Learning," ICCV 2021.



Strategies for DSVT: Fine-tuning

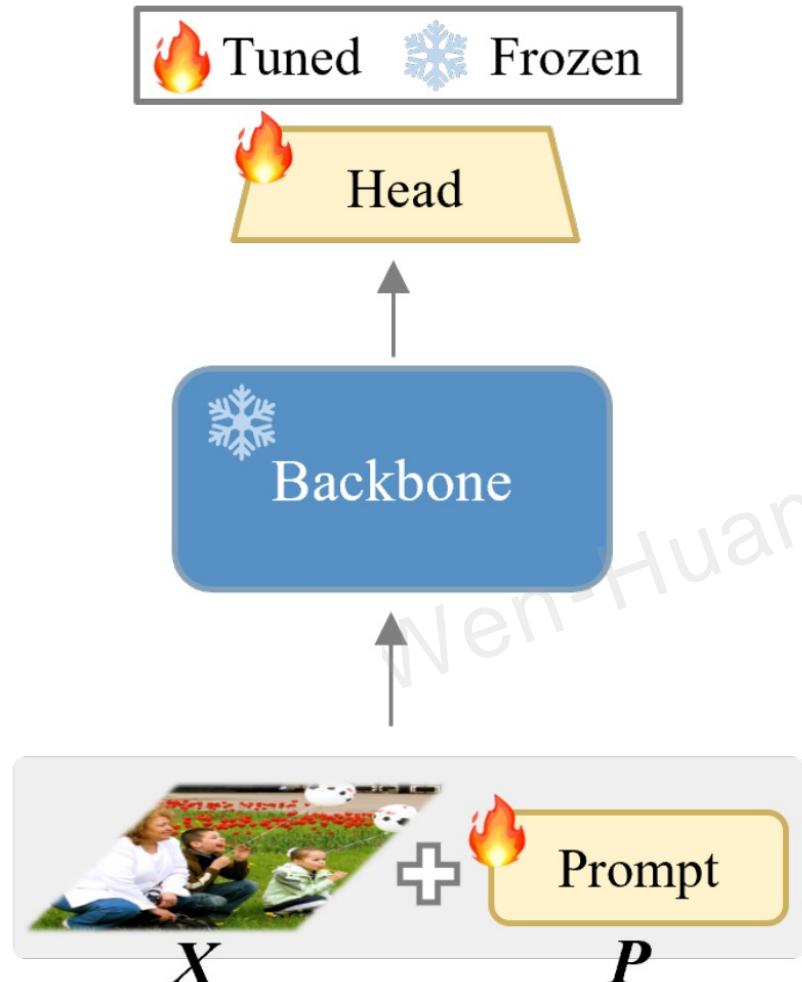
- Adapting LVMs to new tasks



Fine-tuning often requires maintaining separate copies of the parameters for each distinct task, posing challenges in storage and deployment.



Strategies for DSVT: Visual-Prompt Tuning



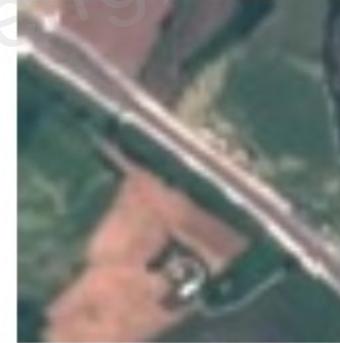
- Employing a minimal fraction of trainable parameters in the input space while maintaining a frozen backbone model

M. Jie et al, "Visual Prompt Tuning," ECCV 2022.



Strategies for DSVT: Visual-Prompt Tuning

- Identifying the right prompt is a non-trivial task (**prompt engineering**)
 - A slight change in wording could make a huge difference in performance.

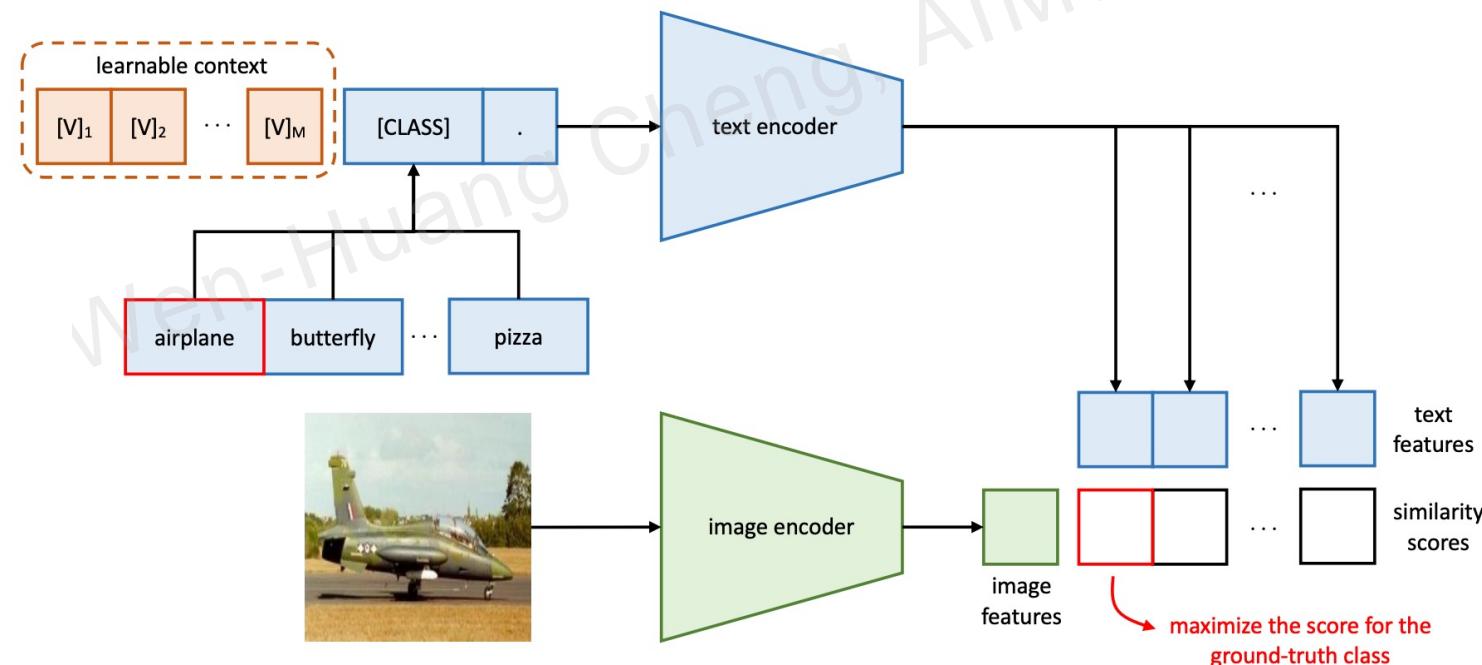
Caltech101	Prompt	Accuracy	EuroSAT	Prompt	Accuracy
	a [CLASS].	82.68		a photo of a [CLASS].	24.17
	a photo of [CLASS].	80.81		a satellite photo of [CLASS].	37.46
	a photo of a [CLASS].	86.29		a centered satellite photo of [CLASS].	37.56

K. Zhou et al., "Learning to Prompt for Vision-Language Models," IJCV 2022.



Strategies for DSVT: Visual-Prompt Tuning

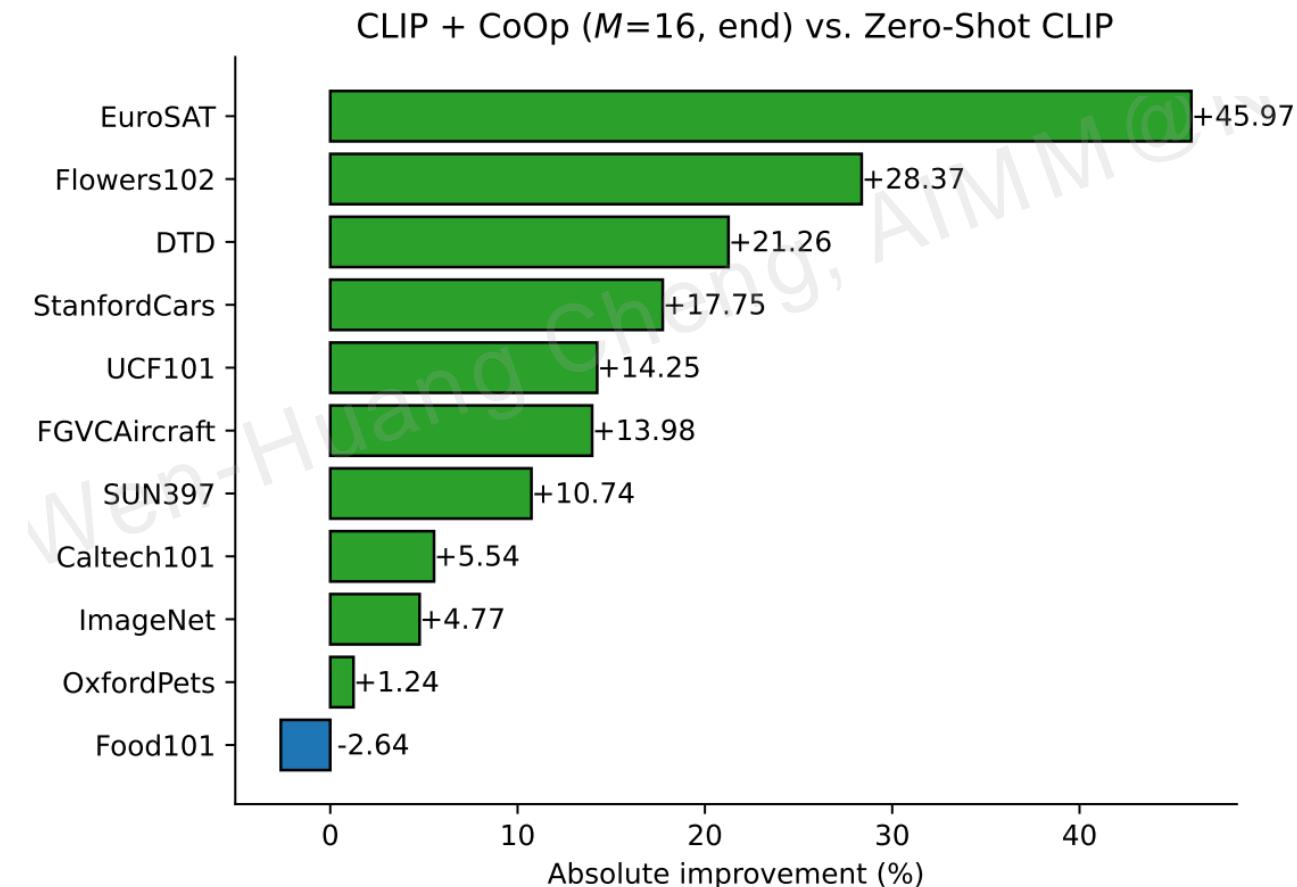
- **Context Optimization (CoOp)** is a technique to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss.





Strategies for DSVT: Visual-Prompt Tuning

- Comparison with hand-crafted prompts



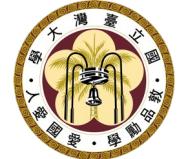


Strategies for DSVT: Visual-Prompt Tuning

• Pitfalls with CoOp

- The learned context is not generalizable to wider unseen classes within the same task (i.e., recognizing scenes). Such a problem is caused by CoOp's **static** nature: the context, which is fixed once learned, is optimized only for a specific set of (training) classes.

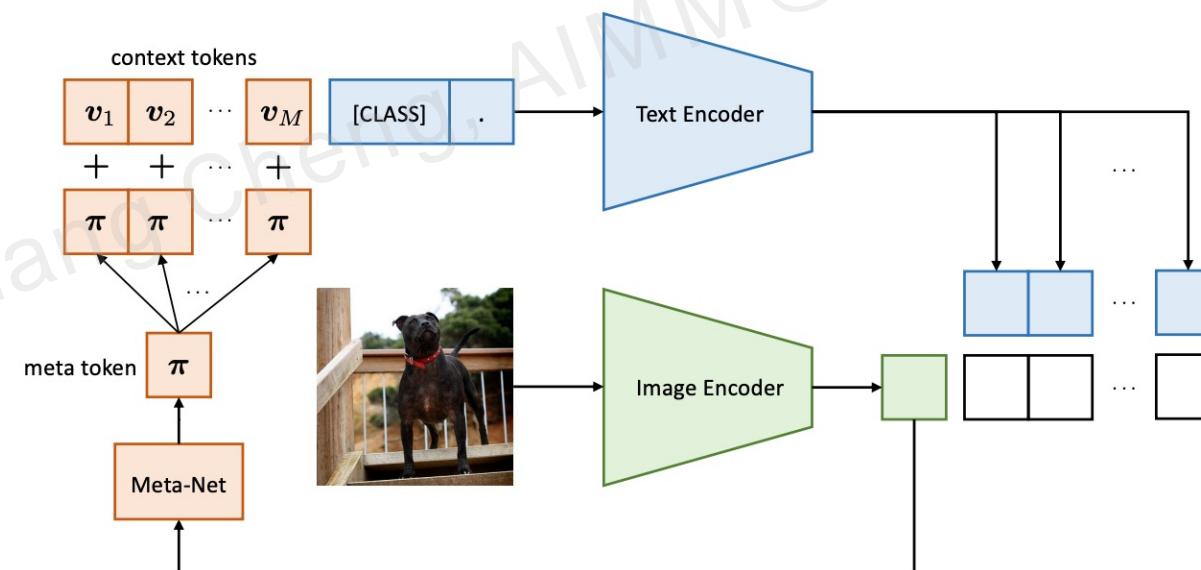




Strategies for DSVT: Visual-Prompt Tuning

- **Conditional Context Optimization (CoCoOp)** extends CoOp to further learn a lightweight neural network to generate for each image an input-conditional token.

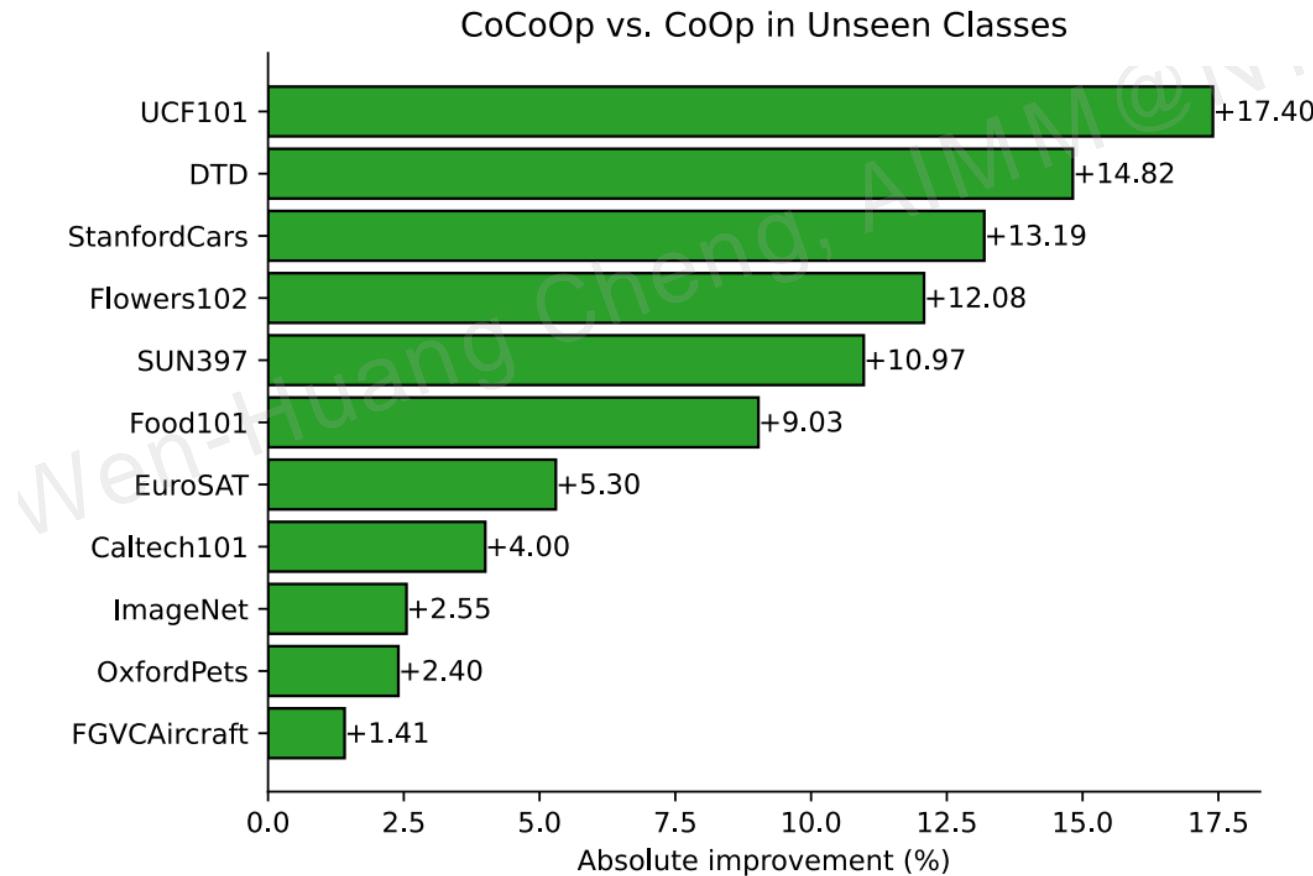
Meta-Net: A lightweight neural network that generates for each image an input-conditional token, which is built with a two-layer bottleneck structure (Linear-ReLU-Linear).





Strategies for DSVT: Visual-Prompt Tuning

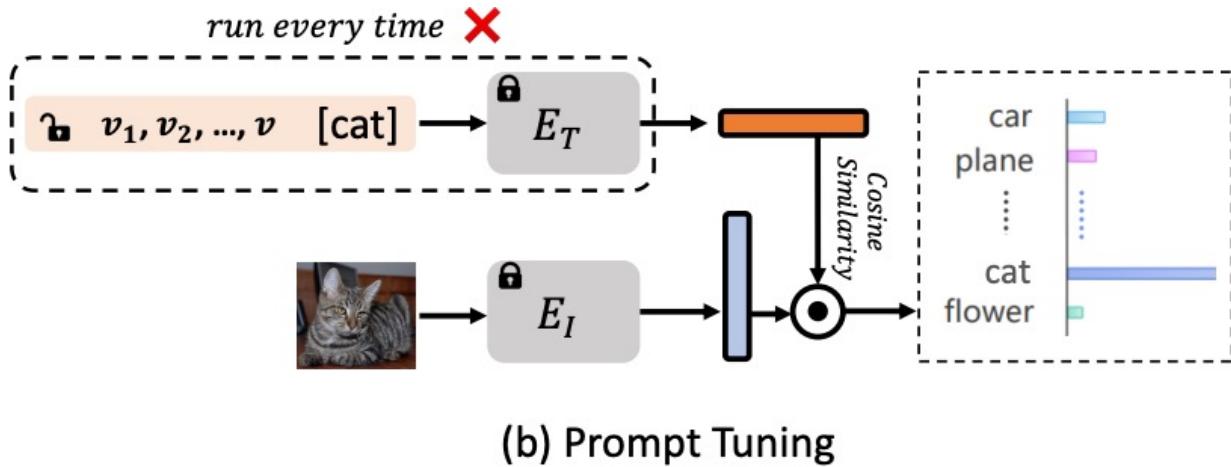
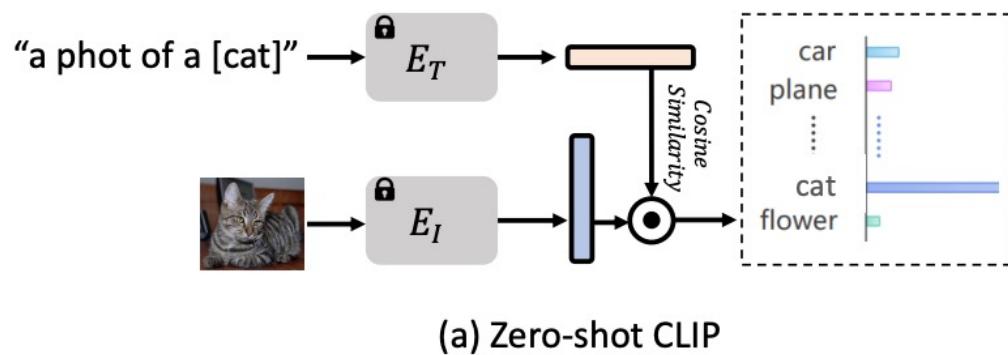
- Improvement of CoCoOp over CoOp





Strategies for DSVT: Visual-Prompt Tuning

- Observations on pitfalls of existing tuning paradigms

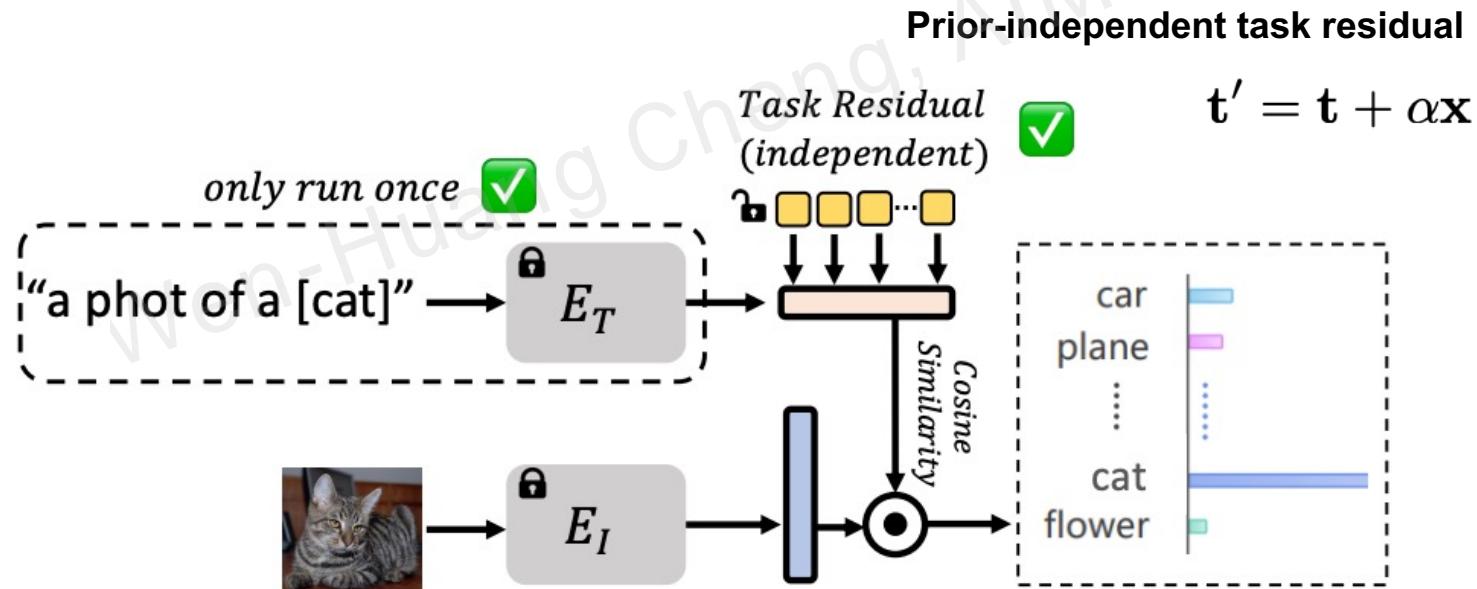


Lack of guarantees of prior knowledge preservation
in prompt tuning



Strategies for DSVT: Visual-Prompt Tuning

- Task Residual Tuning (TaskRes) performs directly on the text-based classifier and explicitly decouples the old knowledge of the pretrained models and the new knowledge for a target task.

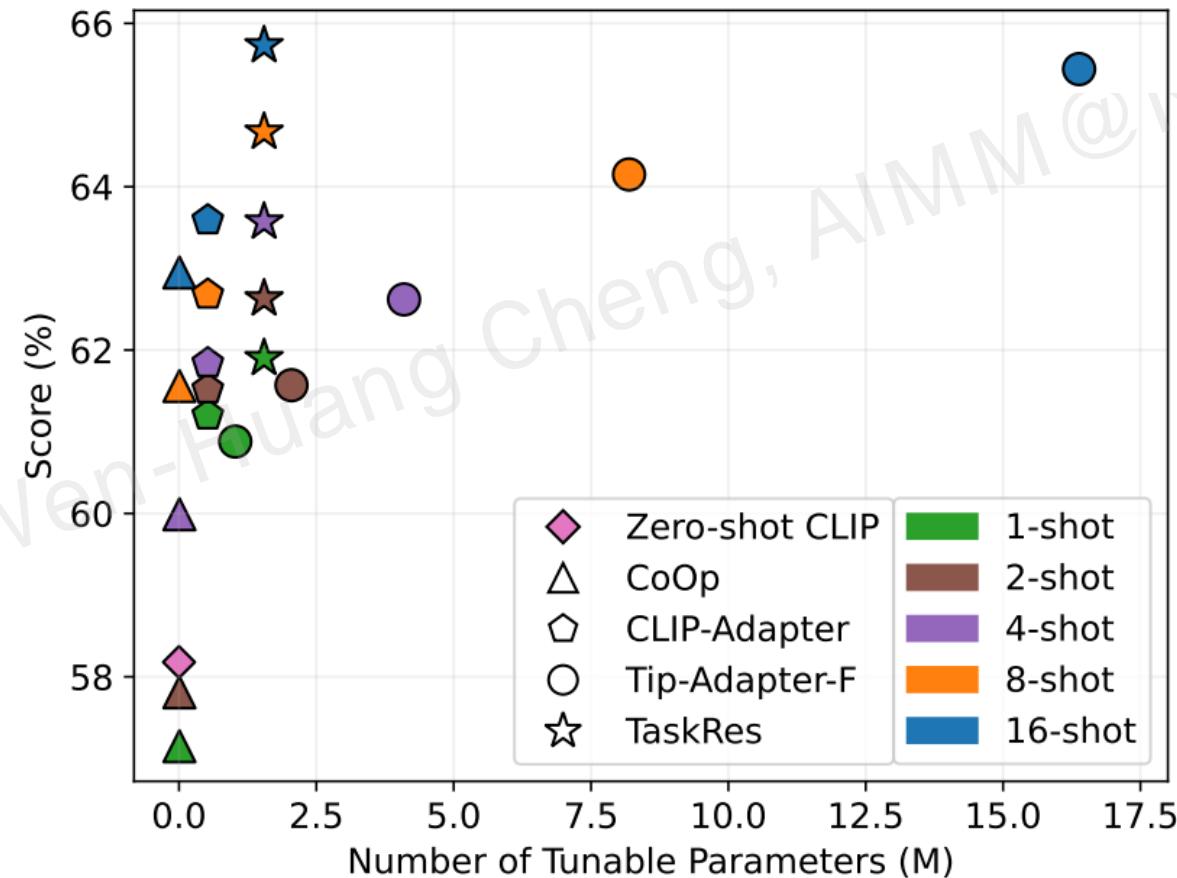


T. Yu et al., “Task Residual for Tuning Vision-Language Models,” CVPR 2023.



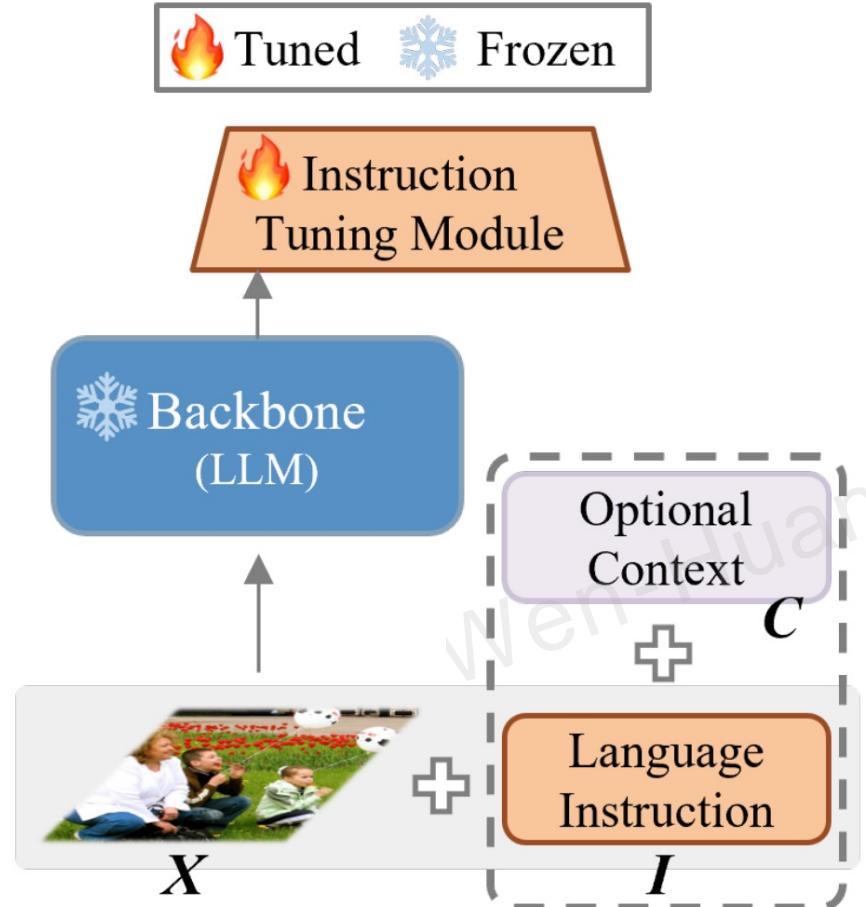
Strategies for DSVT: Visual-Prompt Tuning

- Performance comparison between TaskRes and existing tuning approaches





Strategies for DSVT: Visual Instruction Tuning

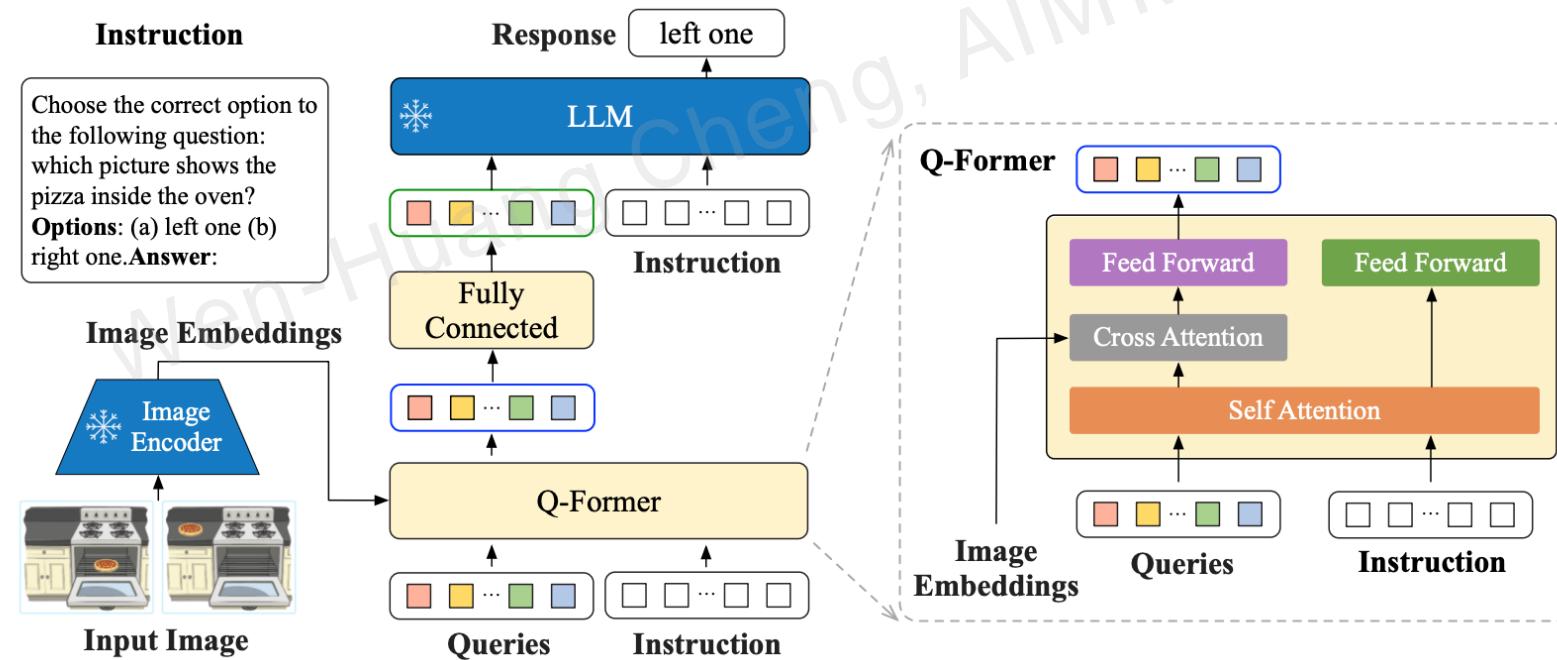


- **Visual Prompt Tuning** focuses on optimizing LLMs using a small set of parameters, whereas **Visual Instruction Tuning (VIT)** aims to improve the model's comprehension of instructions, thereby addressing the model's shortcomings in specific domains.

Strategies for DSVT: Visual Instruction Tuning



- The idea is to extract instruction-aware visual features from the output embeddings of the frozen image encoder, and feeds the visual features as soft prompt input to the frozen LLM.





[Input Image]



[Input Instruction] Do you know who drew this painting?

[Model Response]

InstructBLIP



Leonardo da Vinci.

LLaVA

The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

MiniGPT-4

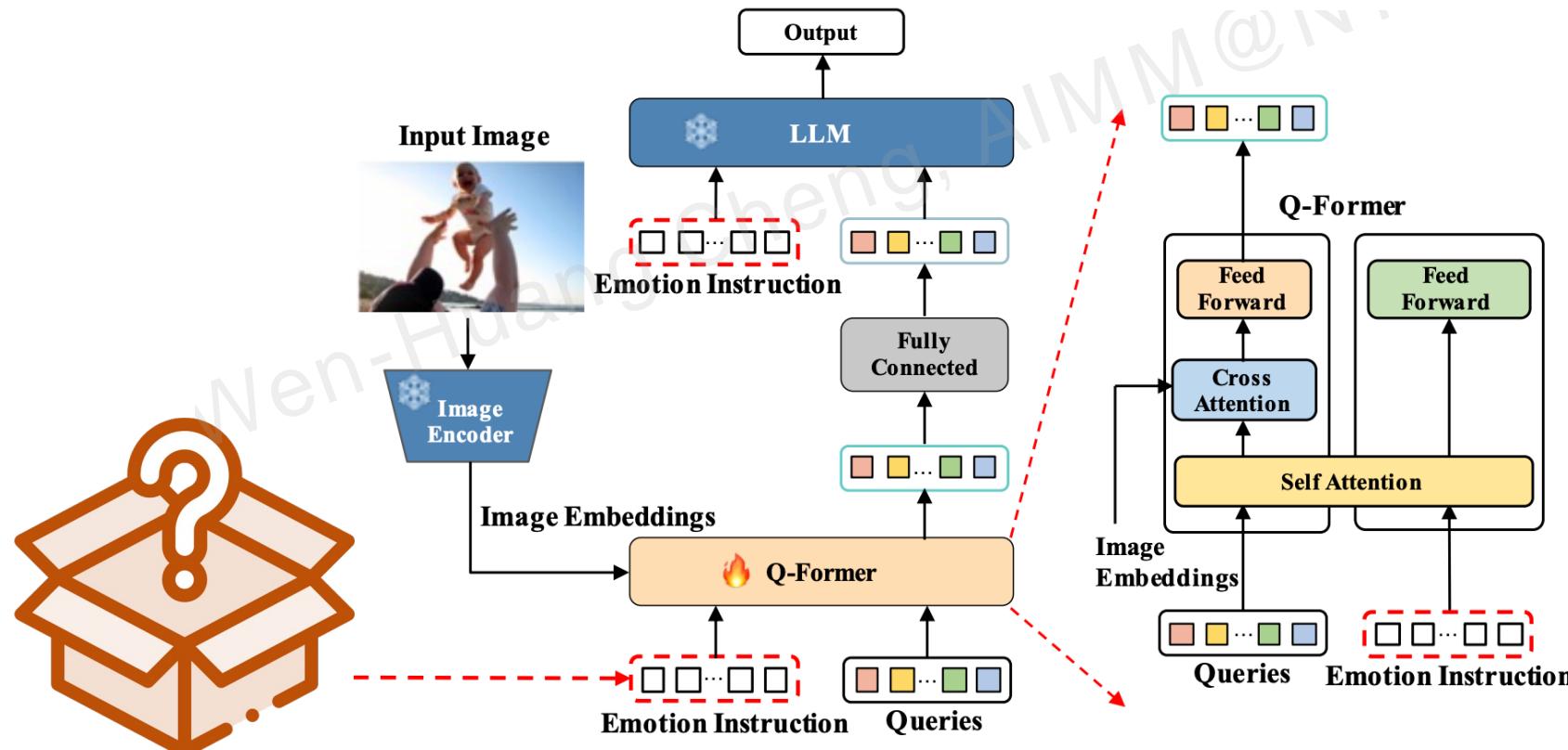
The painting is the Mona Lisa, also known as the Mona Lisa, by Leonardo da Vinci. It is one of the most famous paintings in the world and is known for its enigmatic smile and the mysterious identity of the sitter. The painting is housed in the Louvre Museum in Paris, France.



A Case Study: Visual Emotion Understanding

- Emotion Visual Instruction Tuning

- The goal is to align a LVM's existing knowledge with the emotion understanding domain



H. Xie et al, "EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning," CVPR 2024.



Visual Instruction Data Generation

- Instructions that are specifically tailored to dynamically match visual content are preferred over generic template-based and one-size-fits-all approaches.
 - (1) Existing LLMs can be used to create a pipeline for self-generating visual instructions.
 - (2) Visual attributes across multiple levels can be exploited:



Context type 1: Caption:

Two young women dancing in a kitchen.

Context type 2: Attribute:

<u>Emotion</u> : contentment	<u>Human action</u> : stretching arm
<u>Brightness</u> : 0.7	<u>Object</u> : ['Cupboard', 'Food', 'Countertop']
<u>Colorfulness</u> : 0.4	<u>Facial expression</u> : happy

J. Yang et al., "Emoset: A Large-scale Visual Emotion Dataset with Rich Attributes," CVPR 2023.

H. Xie et al., "EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning," CVPR 2024.



Our Emotion Instruction Data Sample



Context type 1: Caption:

Two young women dancing in a kitchen.

Context type 2: Attribute:

Emotion: contentment Human action: stretching arm
Brightness: 0.7 Object: ['Cupboard', 'Food', 'Countertop']
Colorfulness: 0.4 Facial expression: happy

Categorical

Question: Please select the emotion closest to the image from the following options: amusement, anger, awe, contentment, disgust, excitement, fear and sadness.

Answer: Predict emotion: contentment.

Conversation

Basic Interaction

Question: Can you describe the activity happening in the image?

Answer: The image depicts two young women enjoying themselves by dancing in a kitchen. One of the women is seen stretching her arm, most likely as part of their dance routine.

Conversation

Advanced Interaction

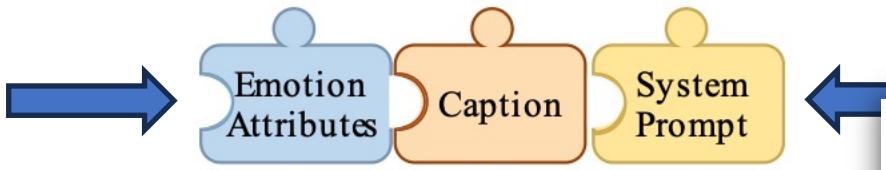
Question: Based on the given emotion of contentment and the activities in the image, can you guess about the possible context or story of this scenario?

Answer: These women are good friends or roommates, spending quality time together. The act of dancing, especially in an unconventional place like a kitchen, suggests they are comfortable with each other and are enjoying a break from routine.

Reasoning

Question: Based on the setting and activities of the two young women, what might have prompted this impromptu dance session in the middle of a kitchen?

Answer: The impromptu dance could have been prompted by a favorite song playing, or simple spontaneous fun amidst preparing a meal or organizing the kitchen.



Emotion Instruction

Categorical

Conversation

Basic Interaction

Advanced Interaction

Reasoning

You are an AI visual assistant, and you are seeing a single image. What you see are provided with one caption and some emotion related attributes, describing the same image you are looking at. Answer all questions as you are seeing the image. The range of brightness is from 0 (darkest) to 1 (brightest), and the range of colorfulness is from 0 (black-and-white) to 1 (the most colorful).

Design two questions for a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

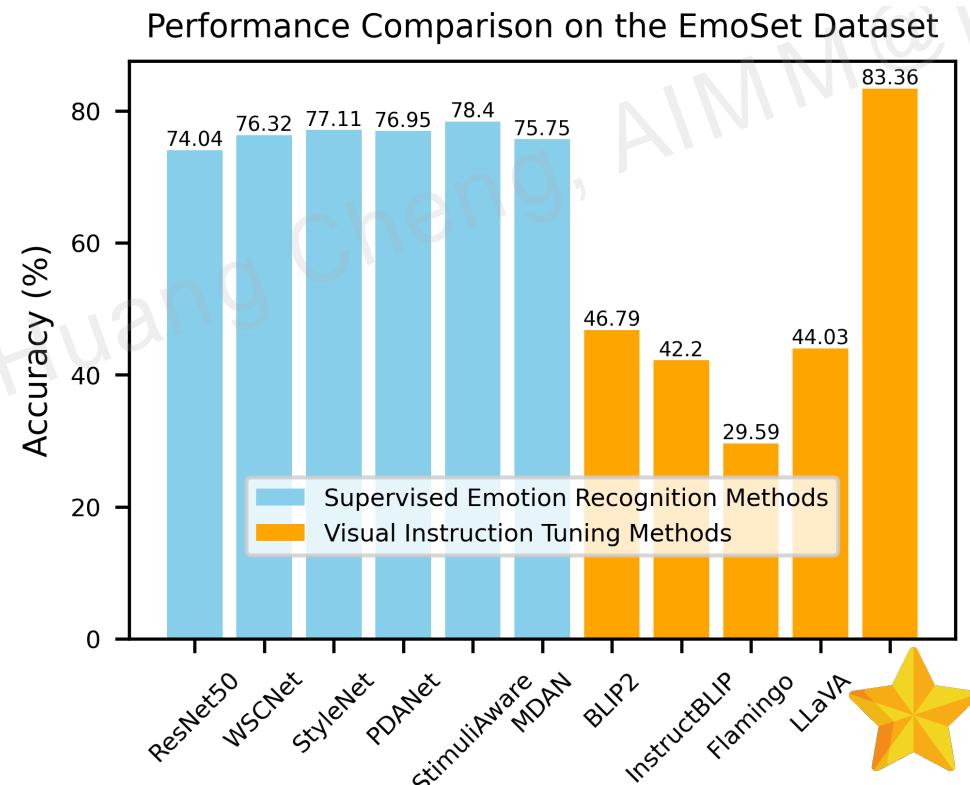
Include questions asking about the visual content of the image, including the object types, object actions, relationship among objects, etc. Only include questions that have definite answers: (1) one can see the content in the image that the question asks about and can answer confidently; (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently. Please answer with the format Question: Answer:

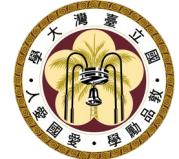
Also include one complex question that is relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.



Strategies for DSVT: Visual Instruction Tuning

- **Visual Instruction Tuning with image-wise instructions needs less than 50% of the usual training data and outperforms state-of-the-art visual emotion recognition models.**





Open Problems: Hallucination

- Hallucination refers to the misalignment between factual visual content and corresponding textual generation.

Judgement Hallucination Examples



Is there a cat in the image? 

Yes, there is a cat in the image. 

Are there four birds in the image? 

Yes, there are four birds. 

Is the blue bird on the left of the yellow bird? 

No, the blue bird is on the right of the yellow one. 

H. Liu, "A Survey on Hallucination in Large Vision-Language Models," arXiv 2024.



Open Problems: Hallucination

- Hallucination refers to the misalignment between factual visual content and corresponding textual generation.

Description Hallucination Examples



Describe the image as detail as possible.

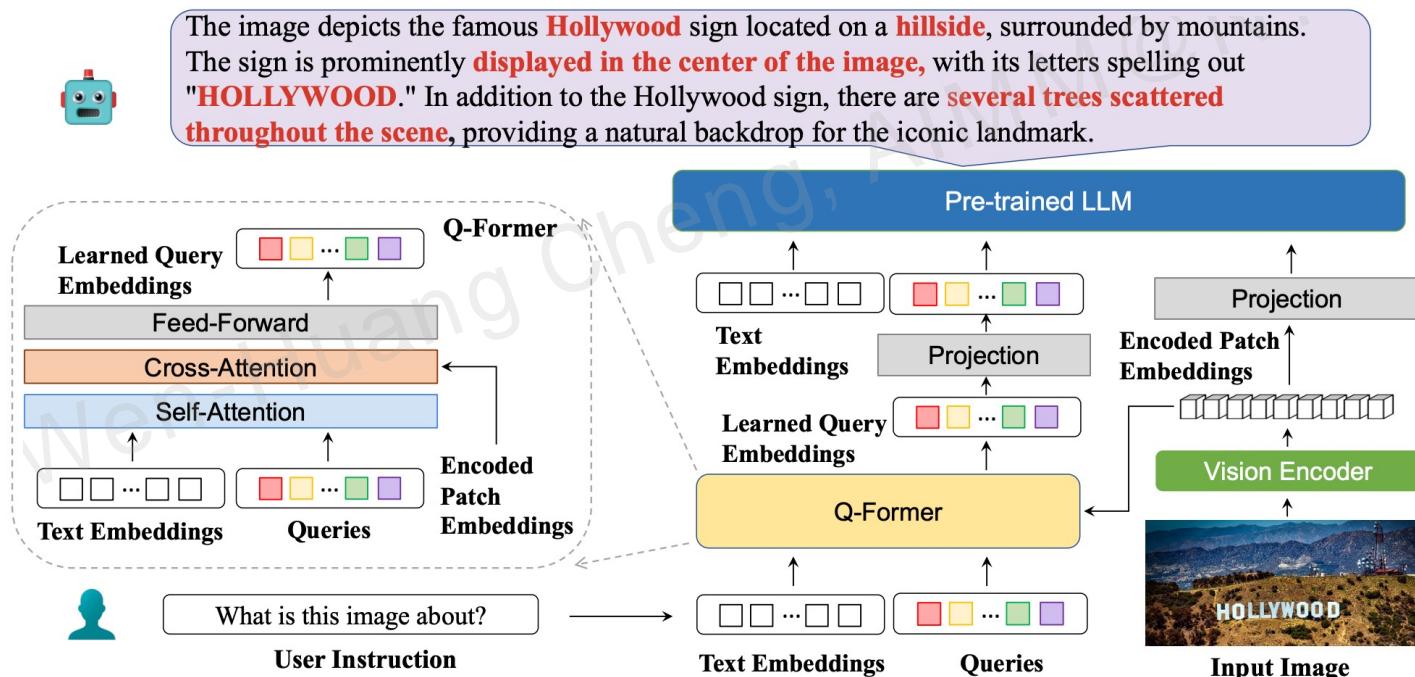


The picture shows a long-haired man in a suit sitting on the steps in the city, checking his watch. Beside him are two green cups, a laptop, and some scattered documents, with a bicycle parked in front of him. A small dog on the step is curiously observing him.



Open Problems: Long-context Modeling

- How to scale LVMs to extremely long inputs (e.g., text-rich instructions) with bounded memory and computation?



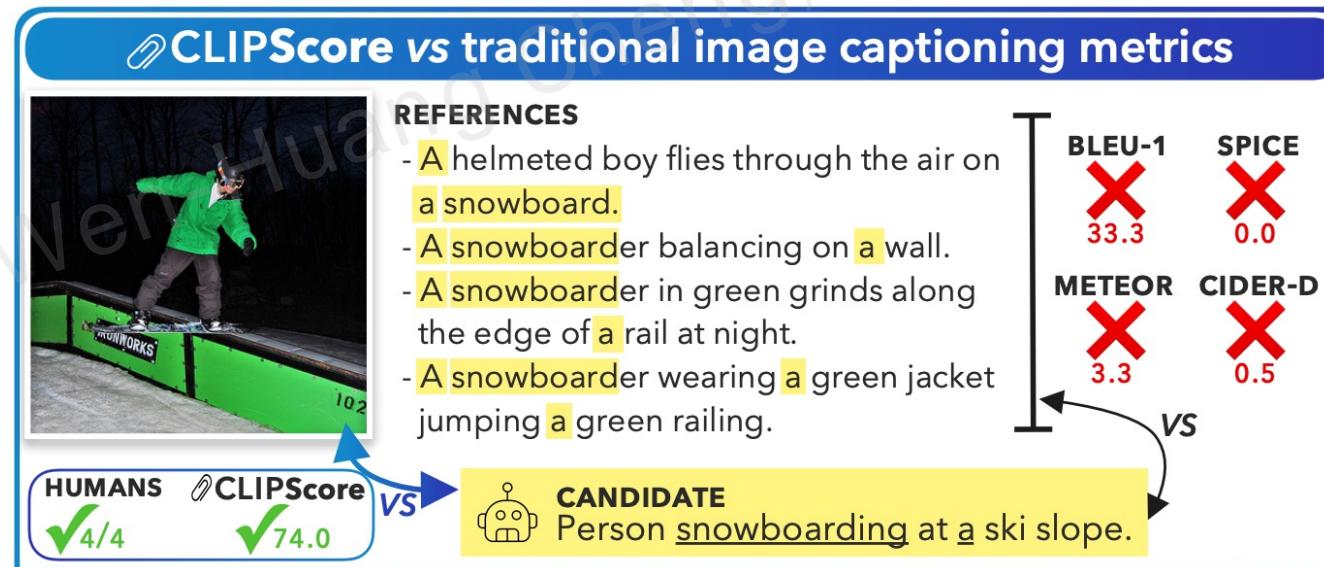
W. Hu et al., "BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions," AAAI 2024.

T. Munkhdalai et al., "Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention," arXiv 2024.



Open Problems: Robust Automatic Evaluation

- Evaluating the quality of output answers is difficult!
- Existing metrics rely on word-level similarity measurements (e.g., BLEU) between a candidate and the reference.
- Recently, reference-free metrics have been proposed.



J. Hessel et al., "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," EMNLP 2021.



Thank you for your attention!