Topics:

- Vision-Language Models

CS 4644-DL / 76T KIRA

Slides by Chenyuan Li et al., Azade Farshad and Mei Sun, Claude

Visual Math and Text Reasoning

What is “a” in the first equation?
What is y in the last equation?

5 + a = 7	40 + y = 45
10 + y = 15	39 + a = 41
20 + a = 22	24 + y = 29
48 + a = 50	15 + a =17
10 + a = 12	34 + y = 39

The value of “a” in the first equation “5 + a = 7” is 2.

The value of “y” in the last equation “34 + y = 39” is 5.

Visual-Conditioned Joke/Meme

What is funny about it?

The humor in the comic may come from the situation, the characters, or the dialogue. The person is feeling unwell and is asked if they are pregnant or have the coronavirus, which is a current and relevant topic. The humor may come from the absurdity of the situation or the unexpectedness of the question.

Spatial/Coordinate Understanding

Where is the frisbee?
What is the relative position of the frisbee and the man?

The frisbee is located at coordinates (63,119,133,177).

The frisbee is located above the man in the image. The man is falling on the ground while the frisbee is in the air. The frisbee is also closer to the top of the image than the man..

Visual Planning and Prediction

What should I do after kneading the dough, before dividing it into pieces? For how long?

After kneading the dough, you should put it in a warm place before dividing it into pieces, according to the recipe in the image.

You should put the dough in a warm place for 1 hour.

Multi-Image Reasoning

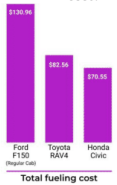
How much in total did I spend on travel?
How much in total did I pay for taxes?

You spent $1299.90 on airfare and $43.83 on an Uber ride for a total of $1343.73 on travel.

You paid $323.23 in total for taxes. This includes $321.90 for taxes and fees for the flight purchase and $1.33 for sales tax for the Windmill Family Restaurant.

Multi-Hop Document Understanding

Excluding the regular cab, what is the average total fueling cost?



The average total fueling cost excluding the Ford F150 (Regular Cab) is $76.55. This is calculated by taking the sum of the total fueling cost of Toyota RAV4 and Honda Civic and dividing it by 2. ($82.56 + $70.55) / 2 = $76.55.

Open-World Concept Understanding

What is the brand of these drink? Which one is unsweetened?

The brand of these drinks is “BETTER THAN MILK”. The unsweetened one is “BETTER THAN MILK organic OAT DRINK”.

Who are the celebrities in this image?

Keira Knightley and Natalie Portman are the celebrities in this image.

Is it safe to eat this mushroom?

Morel mushrooms are generally considered safe to eat when properly cooked.

Video Summarization/Event Localization

Can you breakdown this video tutorial to step-by-step instructions, with start and end time?

Here's a step-by-step breakdown of the tutorial with start and end time.

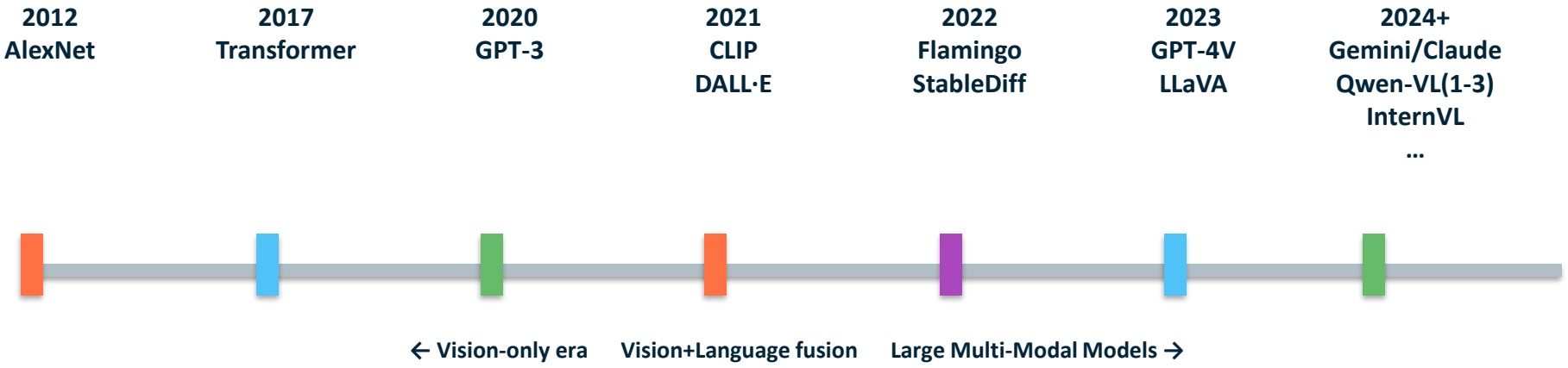
... Add Lettuce Add Tomato Add Bacon ...

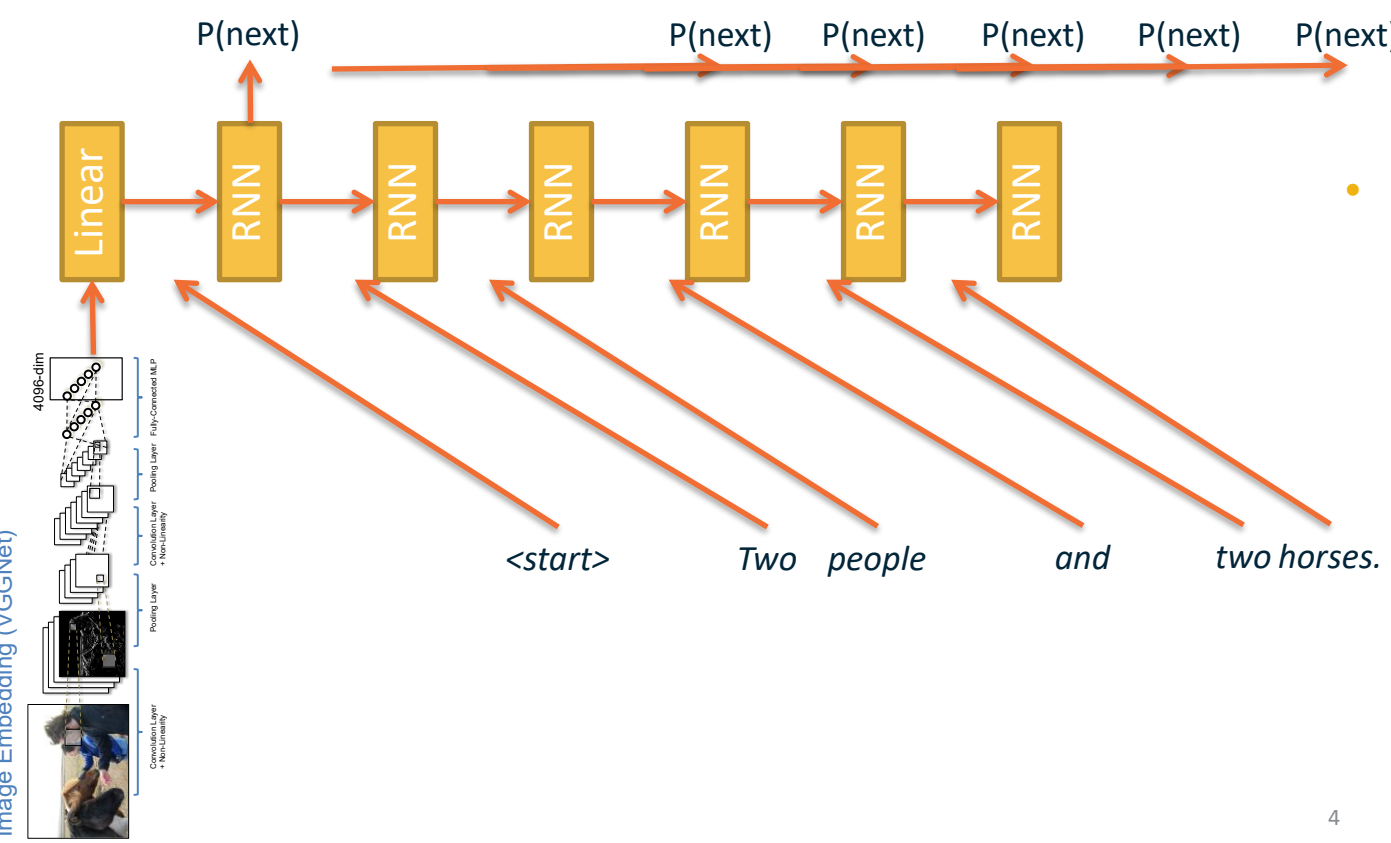
Yang et al., MM-ReAct MM-ReAct: Prompting ChatGPT for Multimodal Reasoning and Action

Motivation: Vision-Language Models

Georgia Tech

History and Evolution of Multi-Modal AI





- **Features**:
  - Pre-trained visual encoder (vector)
  - Linear projection layer to map to “captionable” space
  - Vector serves as initial input to RNN

The Early Days: Neural Image Captioning

Georgia Tech

Unimodal Models Recap

Language Models

- Transformer-based auto-regressive LLMs
- BERT (masked language modeling)
- GPT series (causal language models)
- Input: tokenized text sequences
- Strong: reasoning, generation, QA
- Weak: cannot perceive visual input

Vision Models

- CNNs: local feature extraction (ResNet)
- ViT: image split into 16×16 patches
- DINO, MAE: self-supervised vision
- Input: pixel grids → patch embeddings
- Strong: object recognition, detection
- Weak: lack semantic language grounding

?

How should we encode this (representations?)?

How will they be learned?

VLMs: Considerations

Georgia Tech

How/what should we train?

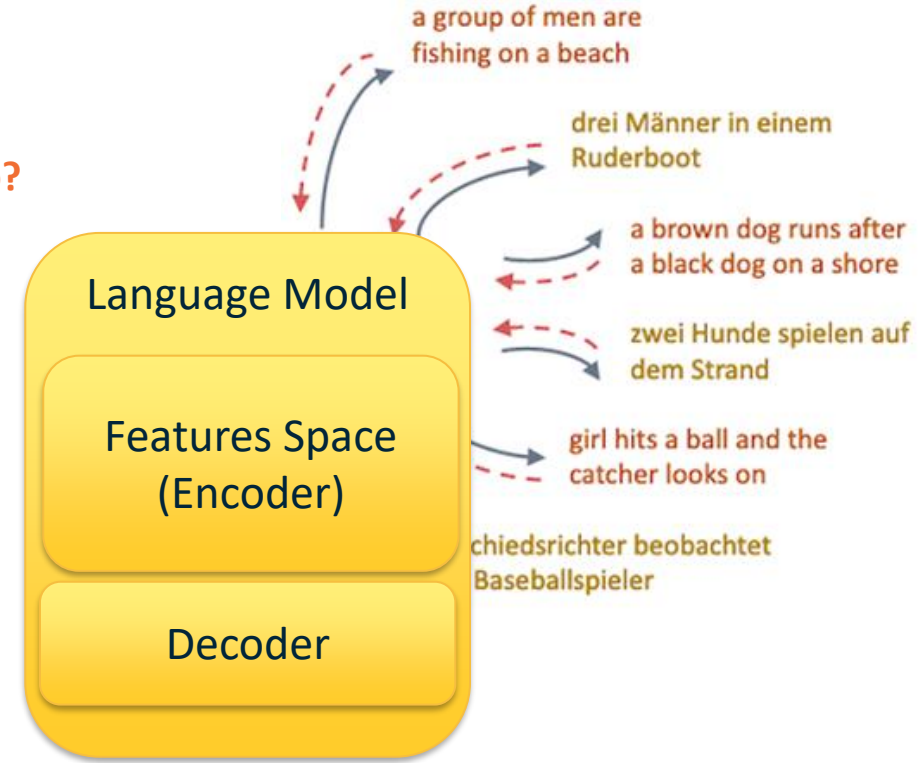
Using what data?

What tasks can we do?

Visual Feature Space

?

What should the interface be?

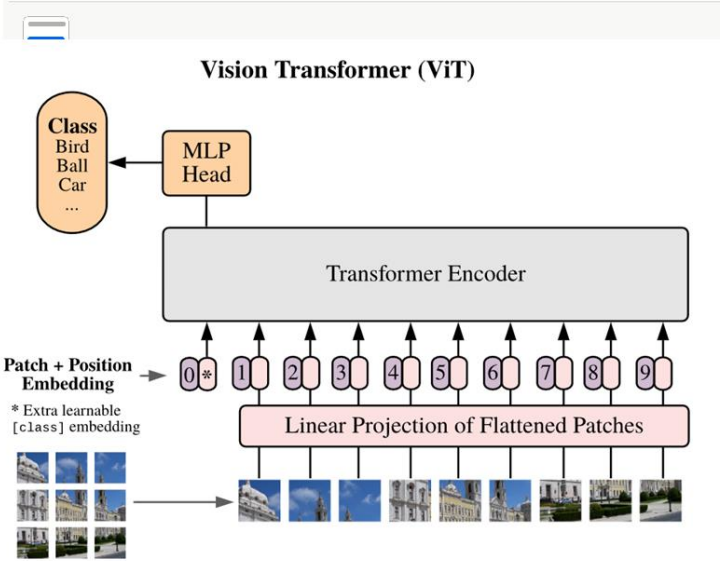


VLMs: Considerations

Georgia Tech

Potential ways of representing an image?

- Image encoder
  - Any architecture: ResNet, Vision transform (ViT)
  - Randomly initialized, SL/SSL pre-trained
    - DINOv1-3, MAE, etc.



Slide by Vicente Ordóñez

Georgia Tech

Architecture Overview: Taxonomy

- **Dual-encoder**: separate encoders, shared embedding space (CLIP)
  - → Trained with contrastive loss; no generation capability by itself
- **Encoder-decoder:** encoder processes image, decoder generates text (Flamingo, CogVLM)
  - → More parameter-efficient cross-modal interaction via cross-attention layers
- **Decoder-only**: image tokens prepended to text tokens in single LLM (LLaVA, GPT-4V)
  - → Simplest approach; leverages full LLM capability; most common in 2023–2024

Key design choice: freeze vs. fine-tune the vision encoder

→ Frozen ViT: faster alignment, less forgetting; trained ViT: more flexible

Early Fusion vs. Late Fusion

Early Fusion

- Merge inputs at the token level
- Joint attention from layer 1
- Higher cross-modal information flow
- Expensive: $O((N_{text} + N_{img})^2)$
- Best for fine-grained grounding
- Example: Unified-IO, PaLI

Late Fusion

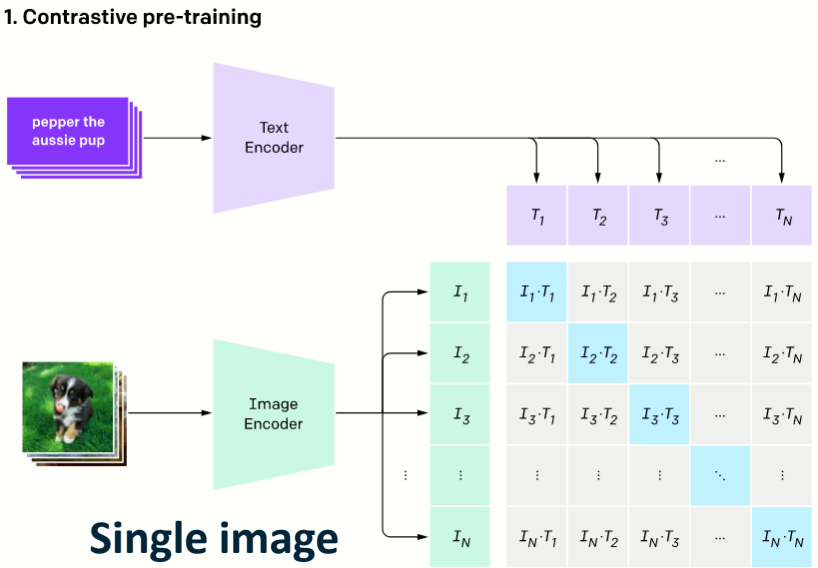
- Encode each modality separately
- Merge at final representation layer
- Lower computational cost
- Modular — swap encoders easily
- Best for retrieval, classification
- Example: CLIP, AudioCLIP

Method of alignment: Contrastive Learning

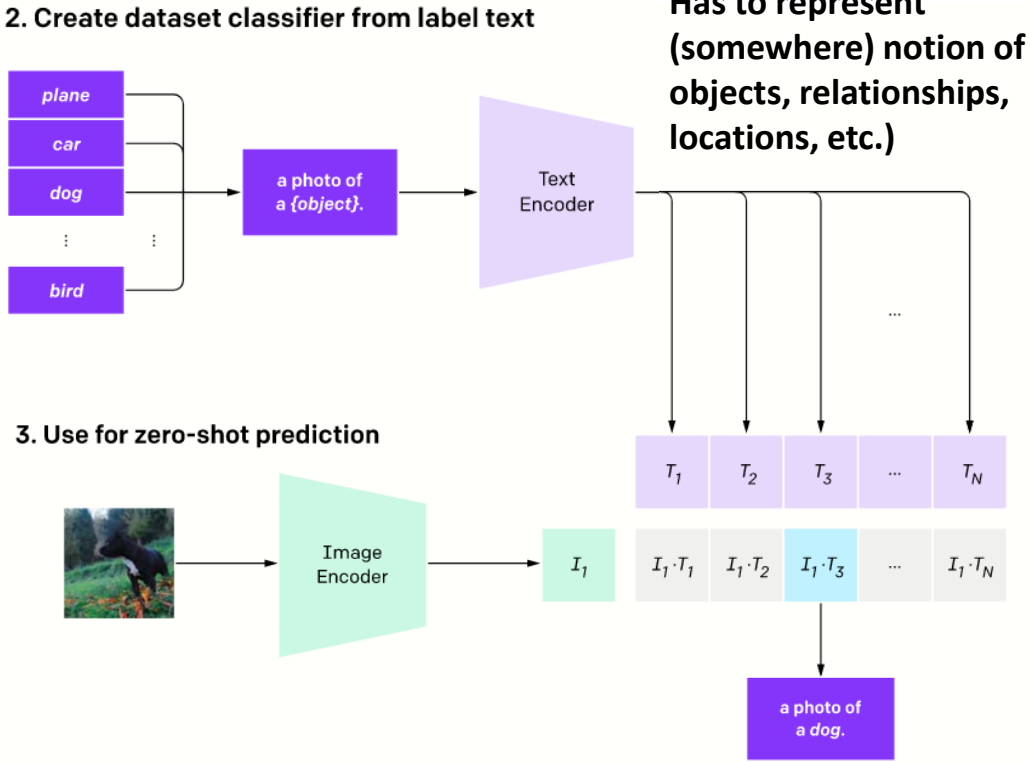
Data: 400M image-text pairs

Downside?

Coarse-grained. Has to represent (somewhere) notion of objects, relationships, locations, etc.)



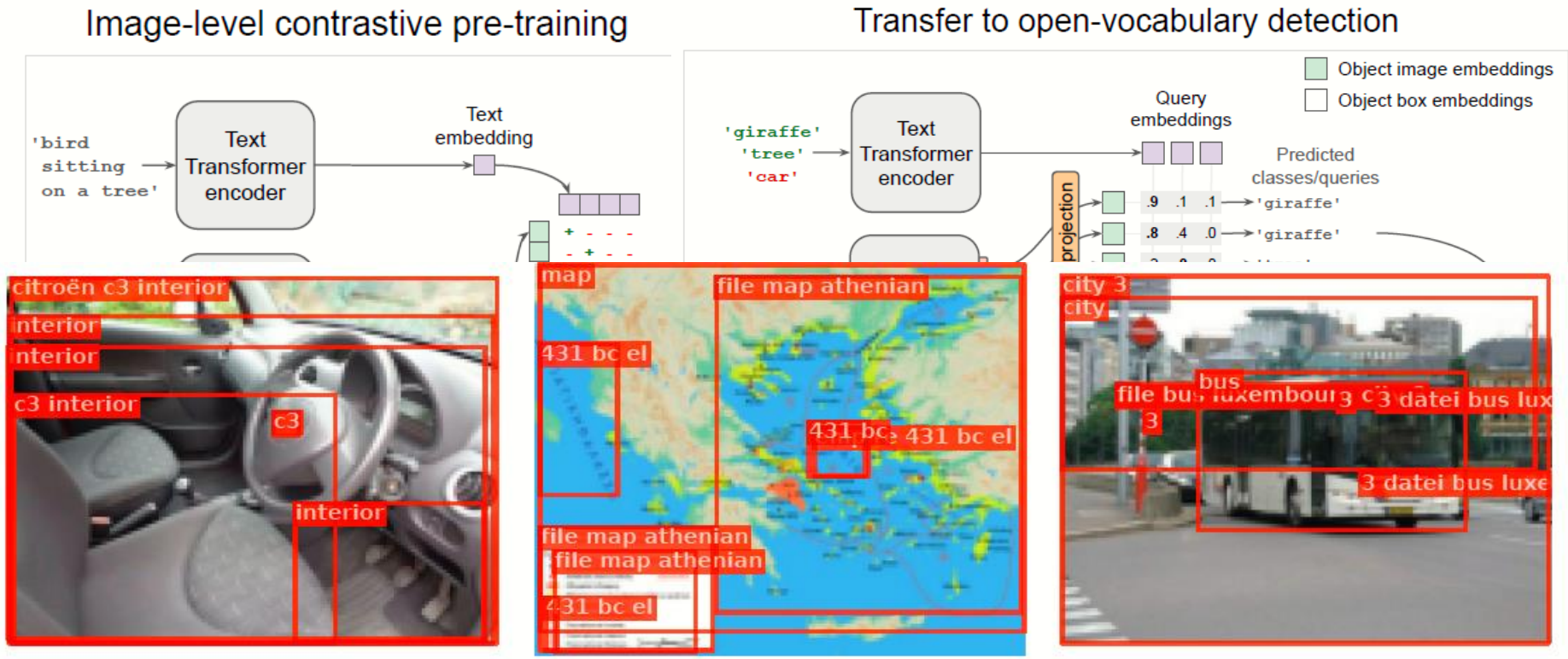
Single image Encoder (ResNet, ViT)



Radford et al., Learning Transferable Visual Models From Natural Language Supervision

CLIP: Learning More Aligned Represenations

Georgia Tech



Minderer et al., Simple Open-Vocabulary Object Detection with Vision Transformers
Minderer et al., Scaling Open-Vocabulary Object Detection

Open-Vocabulary Detection

Georgia Tech

How/what should we train?

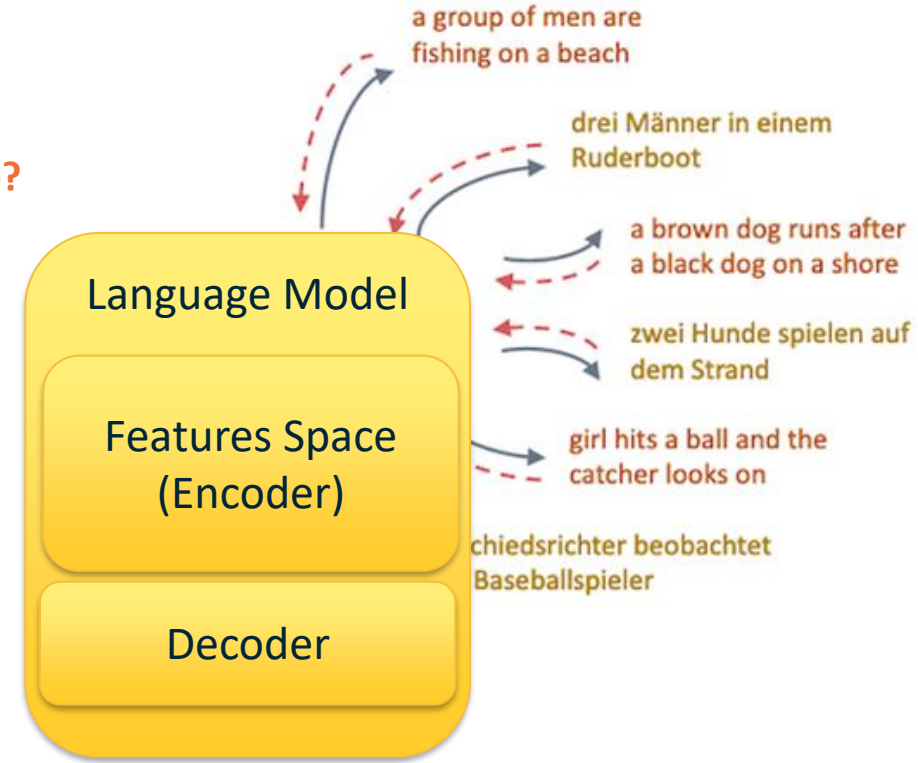
Using what data?

What tasks can we do?

Visual Feature Space

?

What should the interface be?



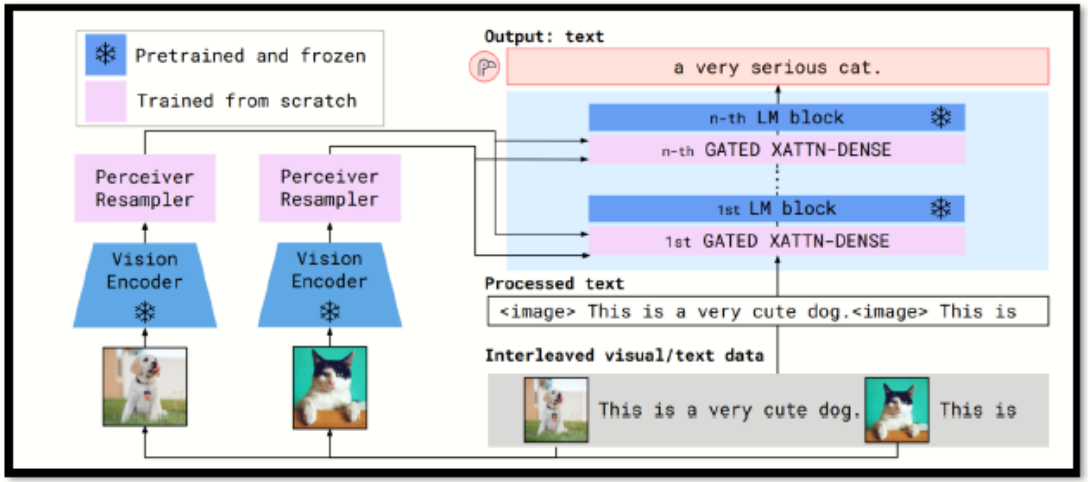
VLMs: Considerations

Georgia Tech

Embedding Space Alignment

- **Goal**: map visual / audio features into the LLM's token embedding space
  - → Spatial structure vs. token sequence — must bridge dimensionality and distribution
- **Linear Projector (MLP):** simple 1–2 layer MLP; used in LLaVA
  - → Fast, effective for high-res; maps ViT features → LLM hidden dim
- **Q-Former (BLIP-2)**: learnable query tokens attend to frozen image encoder
  - → 32–64 fixed queries collapse variable-length features into fixed-size representation
- **Perceiver Resampler (Flamingo)**: cross-attention from N latents to image features
  - → Key advantage: variable resolution images → fixed latent budget
- **LLM-side adapters:** lightweight LoRA layers in LLM learn to interpret visual tokens
  - → Reduces catastrophic forgetting of language knowledge during vision alignment
- **Joint training objective**: align visual features with corresponding text descriptions
  - → Image captioning loss, contrastive loss, or both during alignment pretraining

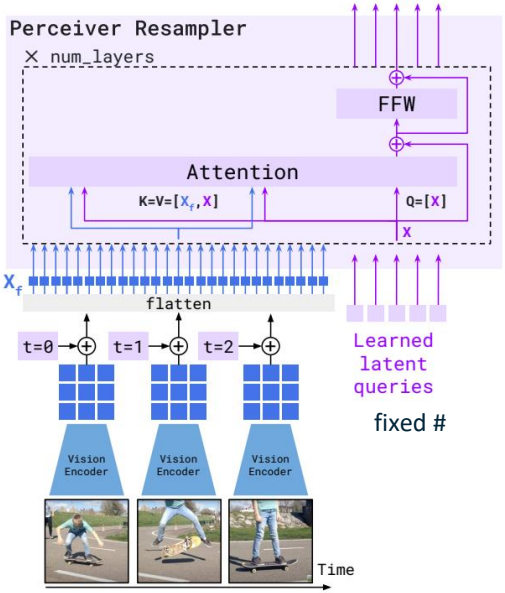
- Flamingo:



Language Model	Pre-trained: 70B Chinchilla
Connection Module	Perceiver Resampler Gated Cross-attention + Dense
Vision Encoder	Pre-trained: Nonrmalizer-Free ResNet (NFNet)

Flamingo VLM

- Model structure - **Supporting both images and videos**




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

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

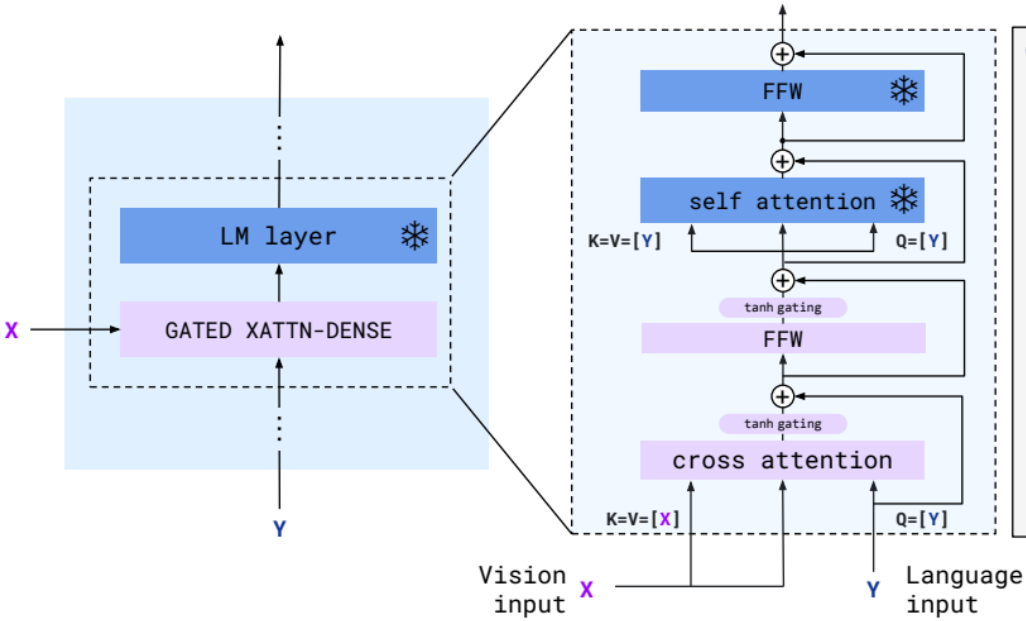
pseudo code

- Using pre-trained ResNet to get visual features X_f
- Compress the encode image into R tokens
- Core of this module : Attention .
  - Query: the learned latent token X
  - Key=Value: the concatenation of X_f, and the learned latent token X
  - Better performance by concatenating keys and values obtained from latent
- If the input is video
  - X_f will add time embeddings

Maps **a variable size grid of visual features** from the Vision Encoder to **a fixed number of output token** (5 in the figure.)

Flamingo VLM

- Model structure - **The interaction with image/video and text**




```
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
```

A **Gated Cross attention** mechanism is proposed to fuse images and text.

Flamingo VLM

- Model structure - **Obtaining multimodal dataset to induce good generalist capabilities**

Image-Text Pairs dataset
[N=1, T=1, H, W, C]

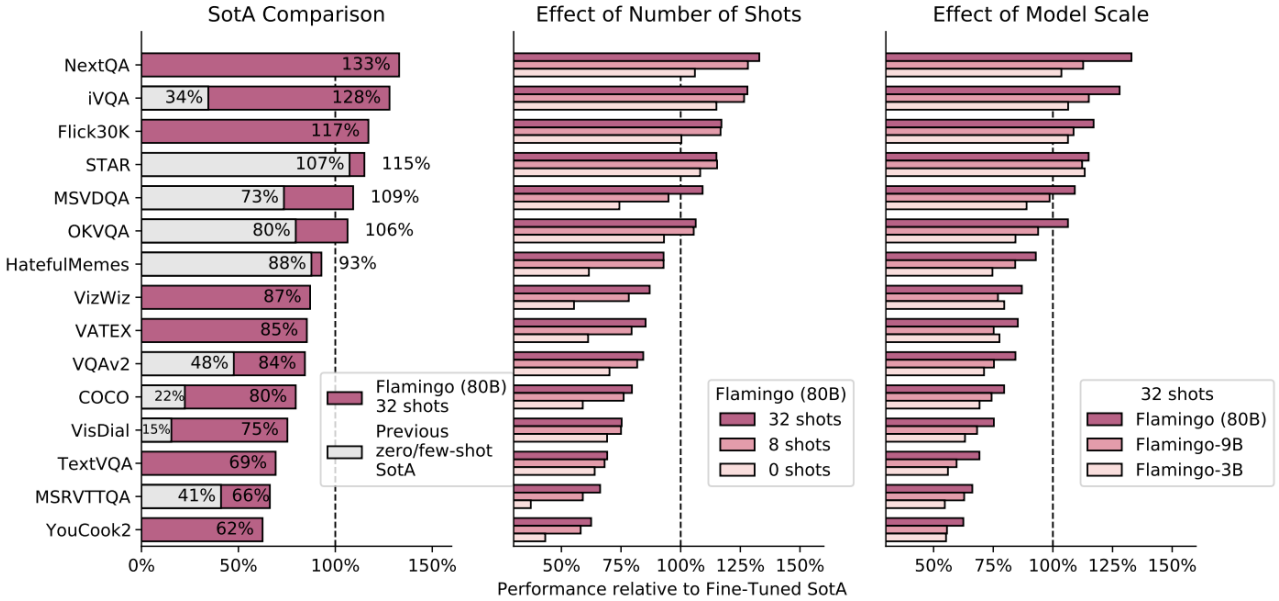
Video-Text Pairs dataset
[N=1, T>1, H, W, C]

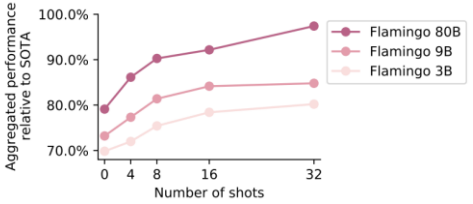
Multi-Modal Massive Web (M3W) dataset
[N>1, T=1, H, W, C]

- M3W: Scrapping 43 million webpages from the Internet
- Training on a mixture of vision and language datasets
  - M3W(185M images+ 182G text)
  - ALIGN(1.8B images with alt-text)
  - LTIP (312M images/text)
  - VTP(27M short video/text)

Flamingo VLM

- **Result: Overview of the results of the Flamingo models**



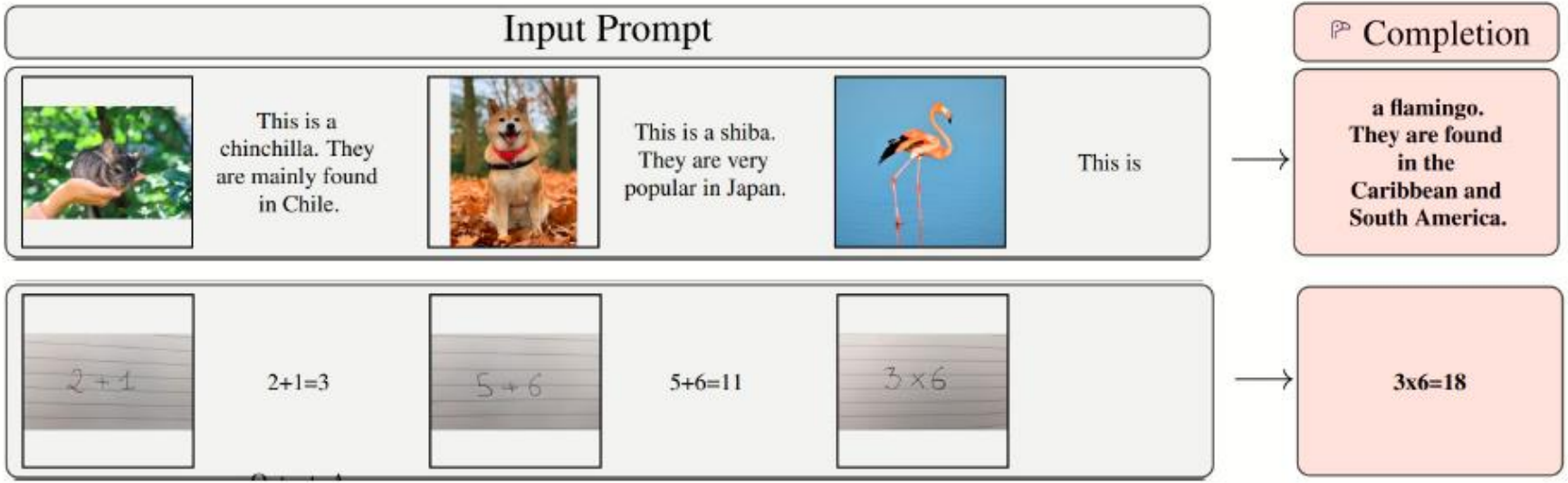


- Larger model sizes and more few-shot examples lead to better performance

- Performance of Flamingo model using different numbers of shots and of different sizes,(without fine-tuned) in comparison with SoTA fine-tuned baseline.

- Flamingo: Multimodal In-Context-Learning

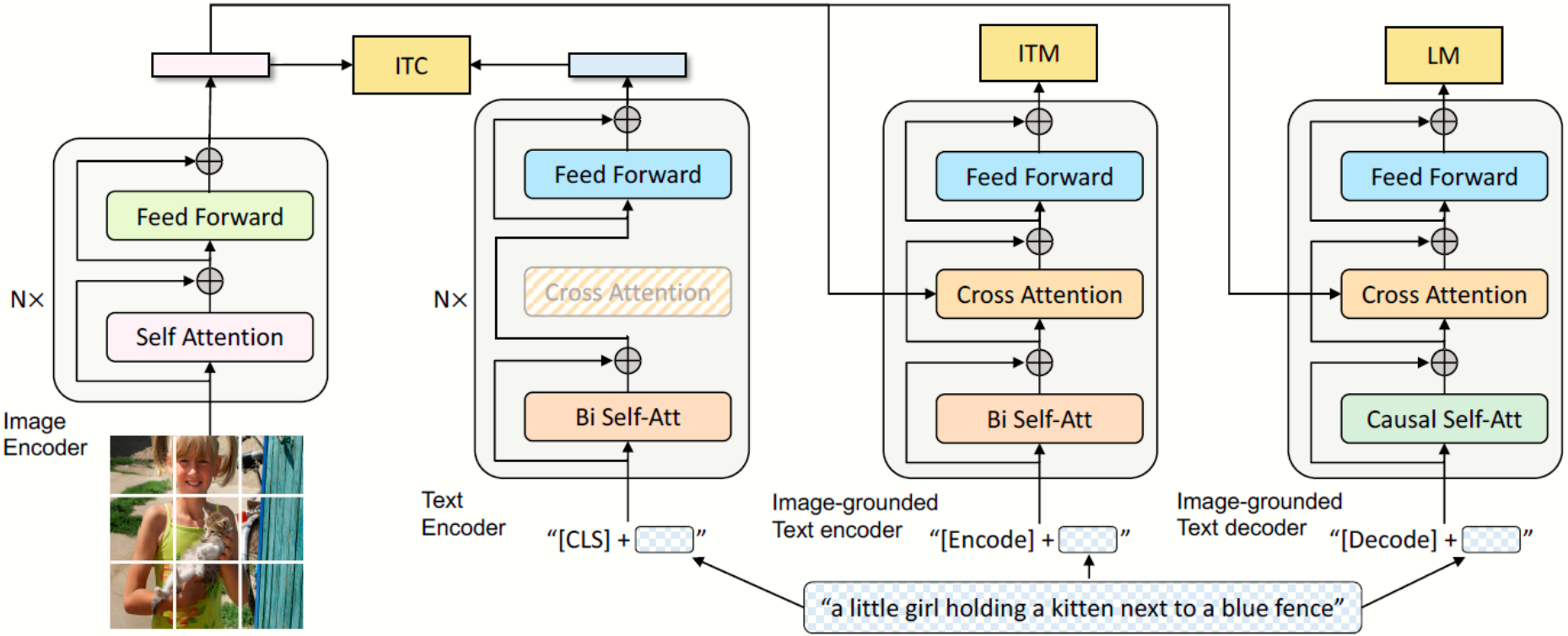
Emerging Property



Flamingo In-Context

Slide by Vicente Ordóñez

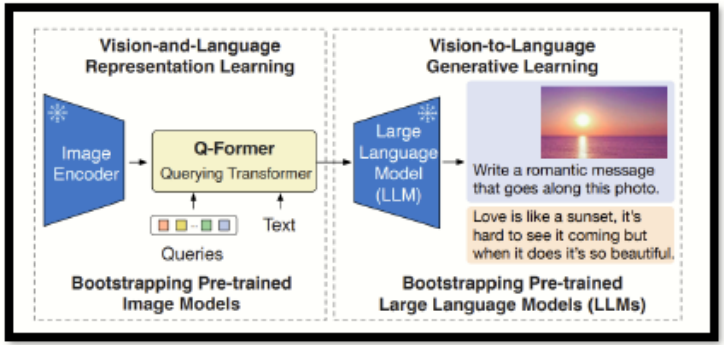
Georgia Tech



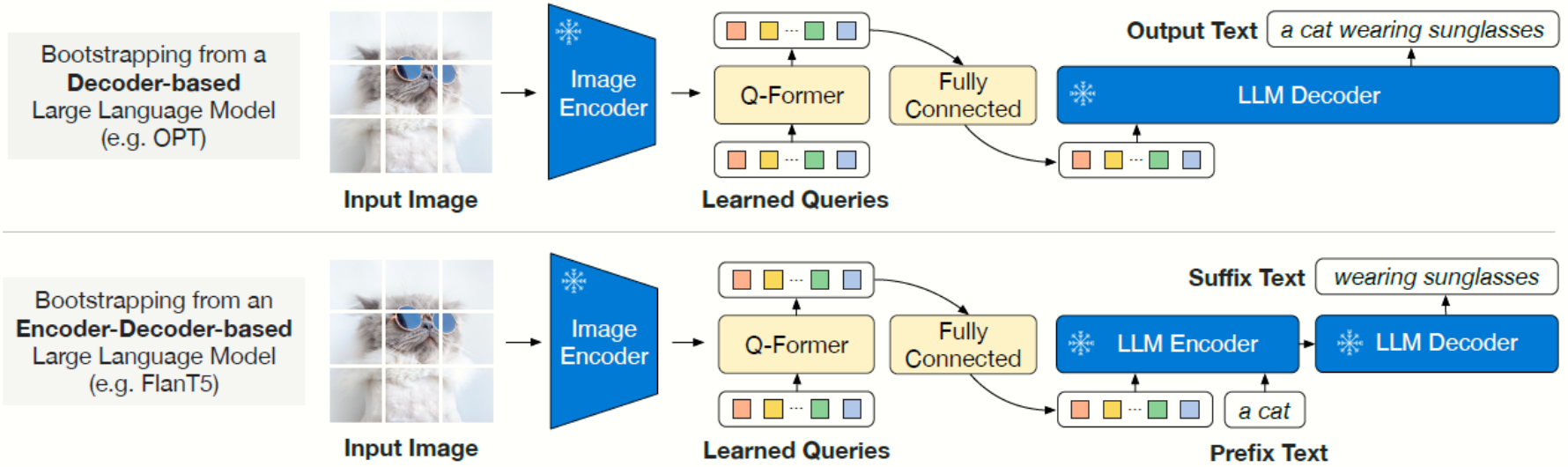
Li et al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

BLIP

• BLIP2



Language Model	Pre-trained: FLAN-T5/OPT
Connection Module	Q-Former: Lightweight Querying Transformer
Vision Encoder	Contrastive pre-trained: EVA/CLIP



Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

BLIP-2

Georgia Tech

Models	#Trainable Params	#Total Params	VQAv2 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
VL-T5$_{\text{no-vqa}}$	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	**50.6**	-
BLIP-2 ViT-L OPT$_{2.7B}$	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT$_{2.7B}$	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT$_{6.7B}$	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5$_{XL}$	103M	3.4B	62.6	62.3	39.4	44.4
BLIP-2 ViT-g FlanT5$_{XL}$	107M	4.1B	63.1	63.0	40.7	44.2
BLIP-2 ViT-g FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**	45.9	**44.7**

Table 2. Comparison with state-of-the-art methods on zero-shot visual question answering.

Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Visual Question Answering Examples

Georgia Tech

Recap on Language Modeling: Large Language Models (LLM)

GPT-2 → GPT-3 → ChatGPT / InstructGPT → GPT-4

What's new?

GPT-3	ChatGPT / InstructGPT	GPT-4
In-context-learning	In-context-learning	In-context-learning
Chain-of-thoughts (CoT)	Chain-of-thoughts (CoT)	Chain-of-thoughts (CoT)
	Instruction-Following	**Instruction-Following**
		Multimodal Input with image

Multimodal Space

Flamingo
BLIP2
GIT
...

Gap?
Instruction-Following
→ Alignment Research

Multimodal GPT-4

GPT4-V Gap

Slide by Chunyuan Li

Georgia Tech

Instruction Tuning

	Input →	Output
Translation	*Hello, Vancouver*	*你好，温哥华*
Summarization	*CVPR is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. This year, CVPR will be single track such that everyone (with full passport registration) can attend everything.*	*CVPR: top computer vision event, single-track, accessible to all.*

- Task instructions are implicit.
- Individual models are trained, or multi-tasking without specifying the instructions
- Hard to generalize to new tasks in zero-shot

Instruction Tuning

Instruction	Input	Output
Translate English into Simplified Chinese	*Hello, Vancouver*	*你好，温哥华*
Summarize in just 10 words to make the message even more brief and easier to remember.	*CVPR is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. This year, CVPR will be single track such that everyone (with full passport registration) can attend everything.*	*CVPR: top computer vision event, single-track, accessible to all.*

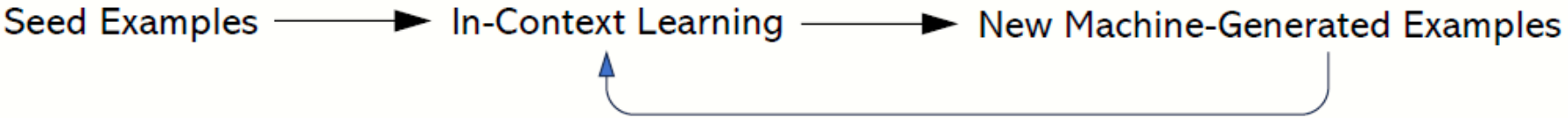
- Task instructions are explicit, expressed in natural language
- One single model is trained, multi-tasking with specified instructions
- Natural and easy to generalize to new tasks in zero-shot

How to collect a diverse set of high-quality instructions and their responses?

- ❑ Human-Human: Collected from humans with high cost
- ❑ Human-Machine: A Strong LLM Teacher such as GPT3 and GPT4

translation example *summarization example*

Please generate new instructions that meet the requirements: ....



Getting the Data: Self-Instruct Tuning

Instruction Tuning with Open-Source LLMs

Self-Instruct with Strong Teacher LLMs & Mixed Human Data

	LLaMA	Alpaca	Vicuna	GPT4-Alpaca	…	Tulu
Data Source		GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)	…	Mixed Data
Instruction-following Data (#Turns)	None	52K	500K (~150K conversions)	52K	…	

Getting the Data – Language Examples

Slide by Chunyuan Li

Georgia Tech

Visual Instruction Tuning with GPT-4

https://llava-vl.github.io/

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee (* Equal contribution)

Self-Instruct with Strong Teacher LLMs

	LLaMA	Alpaca	Vicuna
Teacher		GPT-3.5	ShareGPT (Human & GPT)
Instruction-following Data	None	52K	700K (70 conversions)

But No Teacher is available on multiGPT4?

	GPT-4-LLM	LLaVA
Teacher	GPT-4 (text-only)	GPT-4 (text-only)

- 158K multimodal instruction following data (First & High Quality)

→ Multimodal Chatbot

Large Language and Vision Assistant

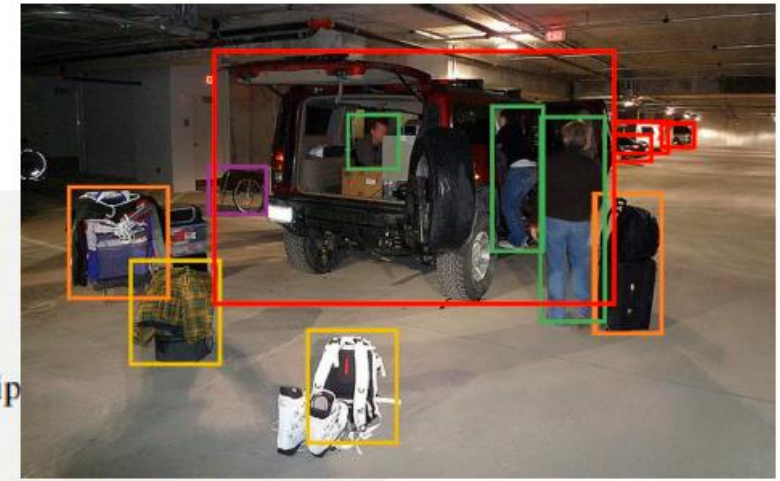
21

GPT-assisted Visual Instruction Data Generation

Downside?

- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples

→ Text-only GPT-4



Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Data – Using Captions

Slide by Chunyuan Li

Georgia Tech

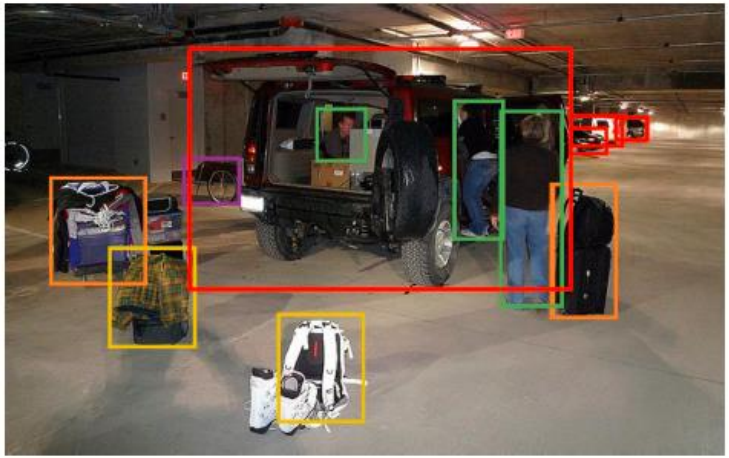
GPT-assisted Visual Instruction Data Generation

Three type of instruction-following responses

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

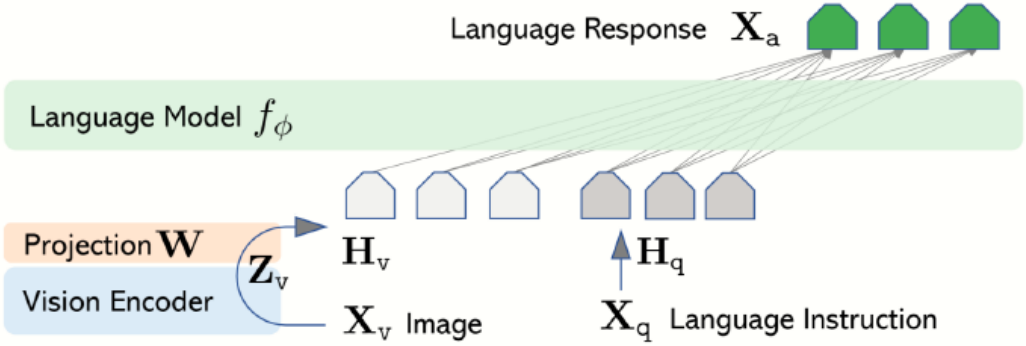
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.



LLaVA: Large Language-and-Vision Assistant

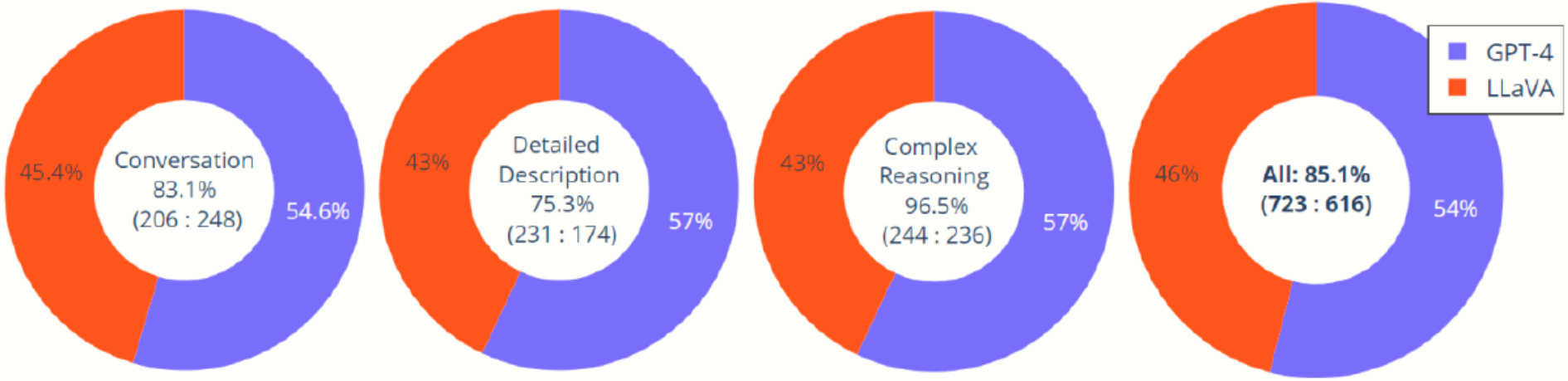
❑ Architecture



❑ Two-stage Training

- **Stage 1: Pre-training for Feature Alignment**.
  Only the projection matrix is updated, based on a subset of CC3M.
- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are updated
  - **Visual Chat**: Our generated multimodal instruction data for daily user-oriented applications.
  - **Science QA**: Multimodal reasoning dataset for the science domain.

Visual Chat: Towards building multimodal GPT-4 level chatbot



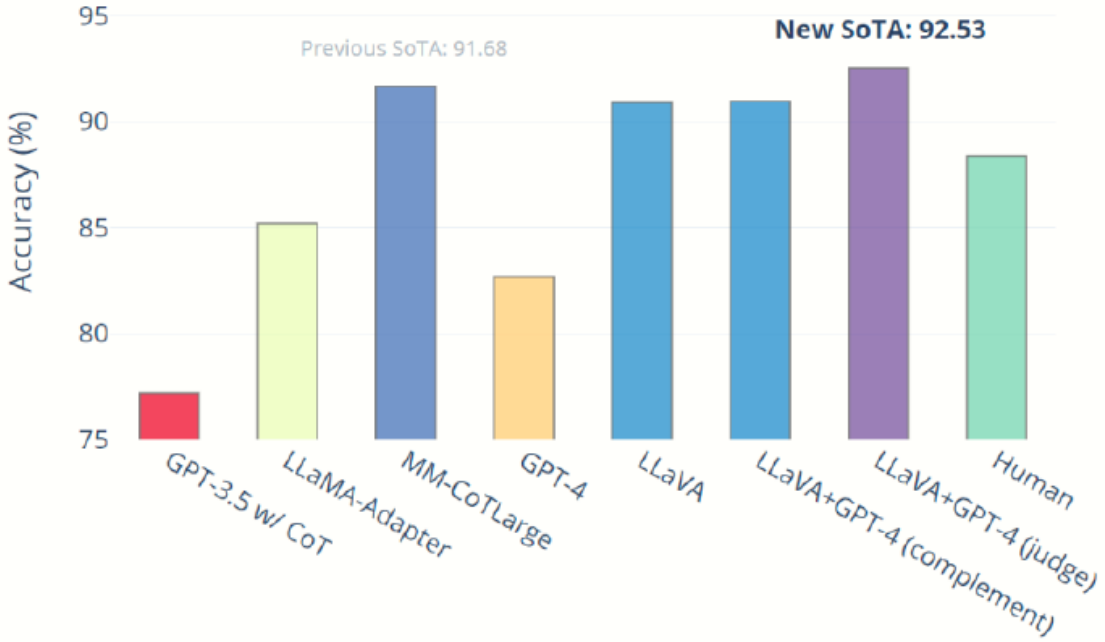
An evaluation dataset with 30 unseen images, 90 new language-image instructions

Overall, LLaVA achieves 85.1% relative score compared with GPT-4

Results

Science QA: New SoTA with the synergy of LLaVA with GPT-4

- LLaVA alones achieve 90.92%
- We use the text-only GPT-4 as the juedge, to predict the final answer based on its own previous answers and the LLaVA answers.
- This ``GPT-4 as juedge'' scheme yields a new SOTA 92.53%
- GPT-4 is an effective model ensemble method



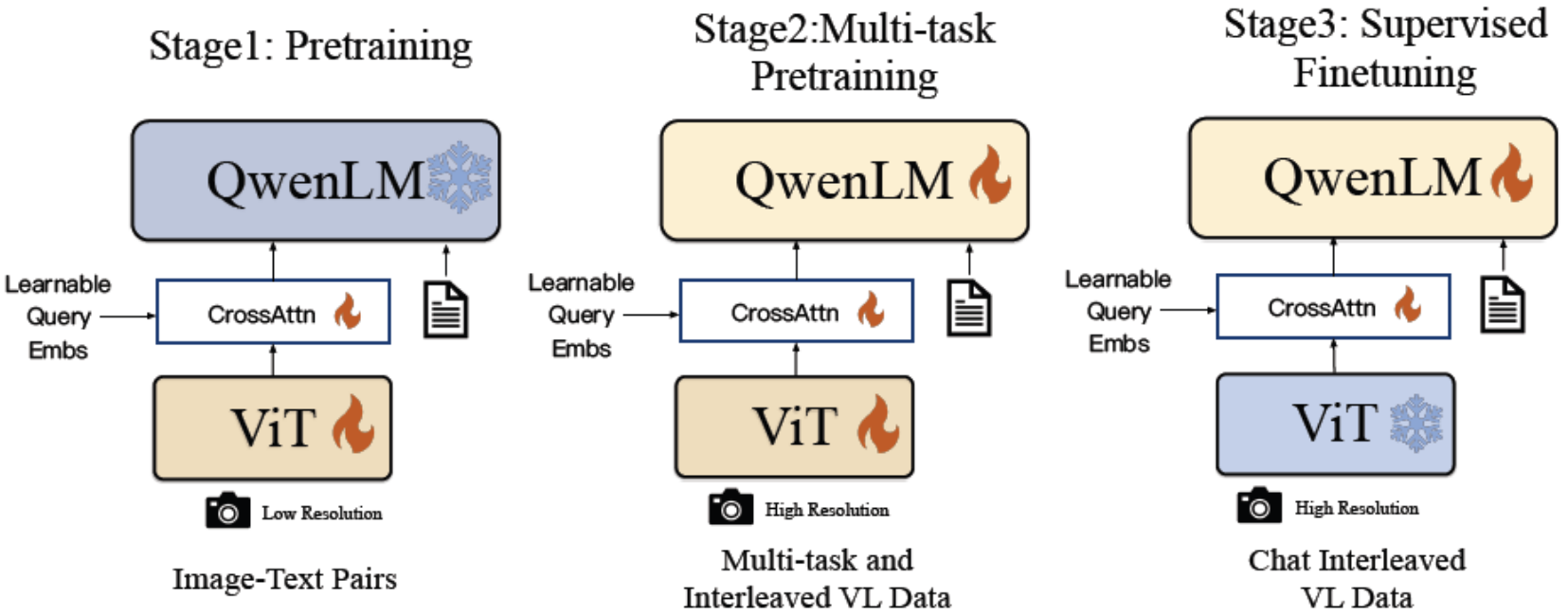
Results - Science

Slide by Chunyuan Li

Georgia Tech

A number of aspects improved over time:

- Scale (of course)
- Multi-state training
  - pre-training, multi-modal alignment, SFT/Instruction Tuning, RLHF, Post-Training/Reasoning
- Diversity of data
- Some architectural (hard to tell what matters, no ablations):
  - Multi-resolution
  - Long context
  - Temporal position embeddings
  - Synthetic data
- Huge diversity of evaluation benchmarks across many tasks

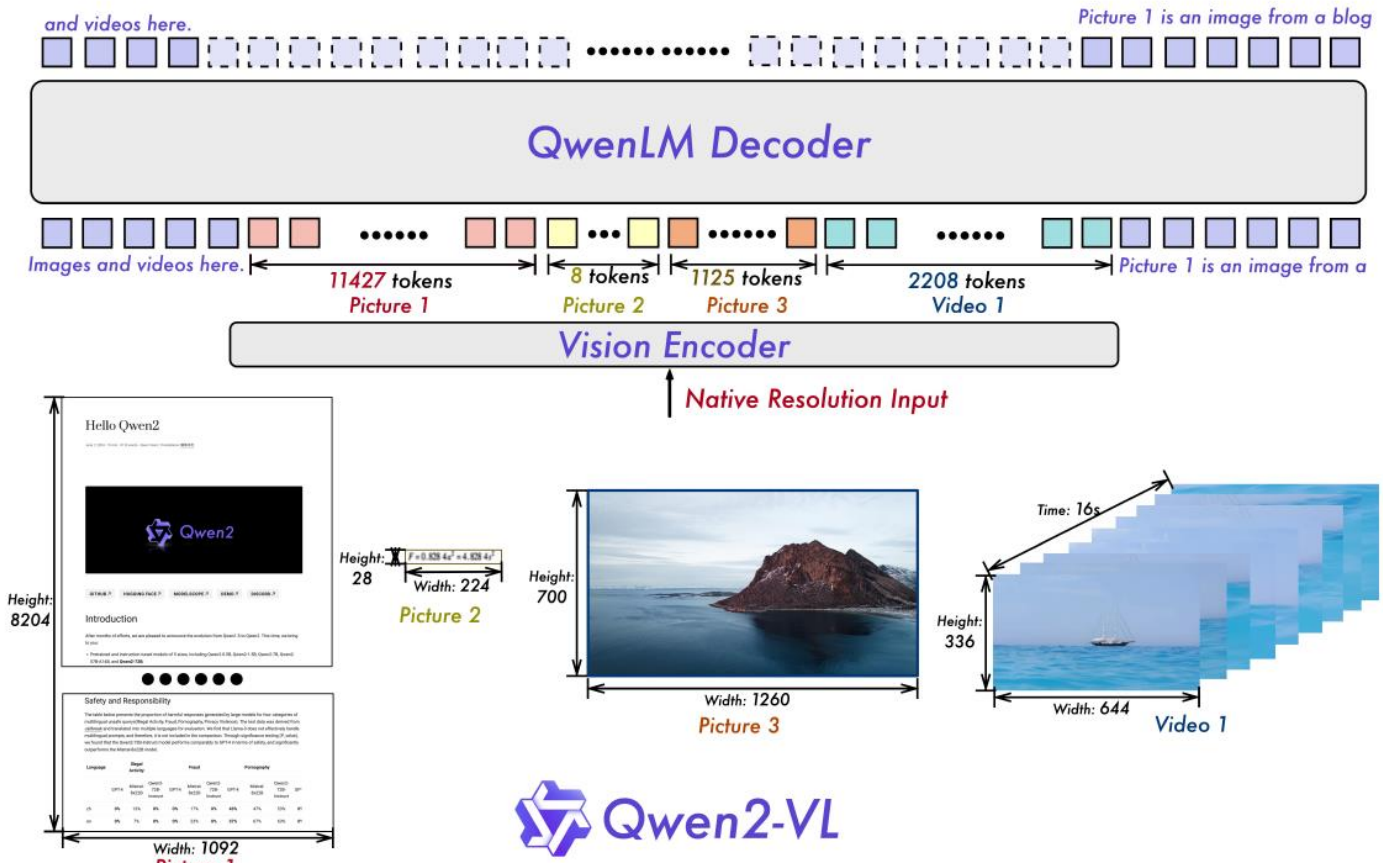


Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond

Qwen-VL Training

Slide by Chunyuan Li

Georgia Tech

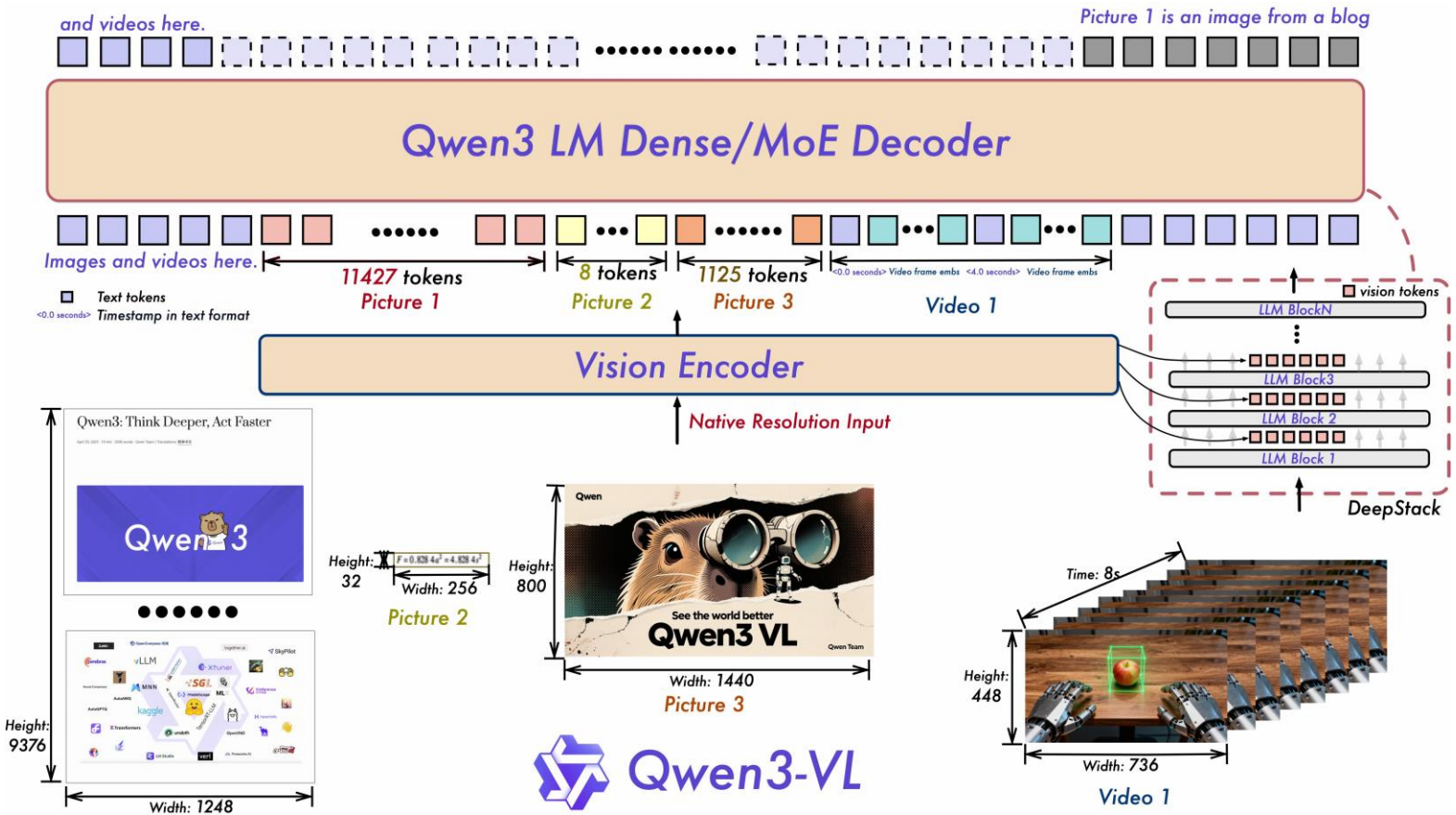


Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution

Qwen-VL Training

Slide by Chunyuan Li

Georgia Tech



https://github.com/QwenLM/Qwen3-VL?tab=readme-ov-file

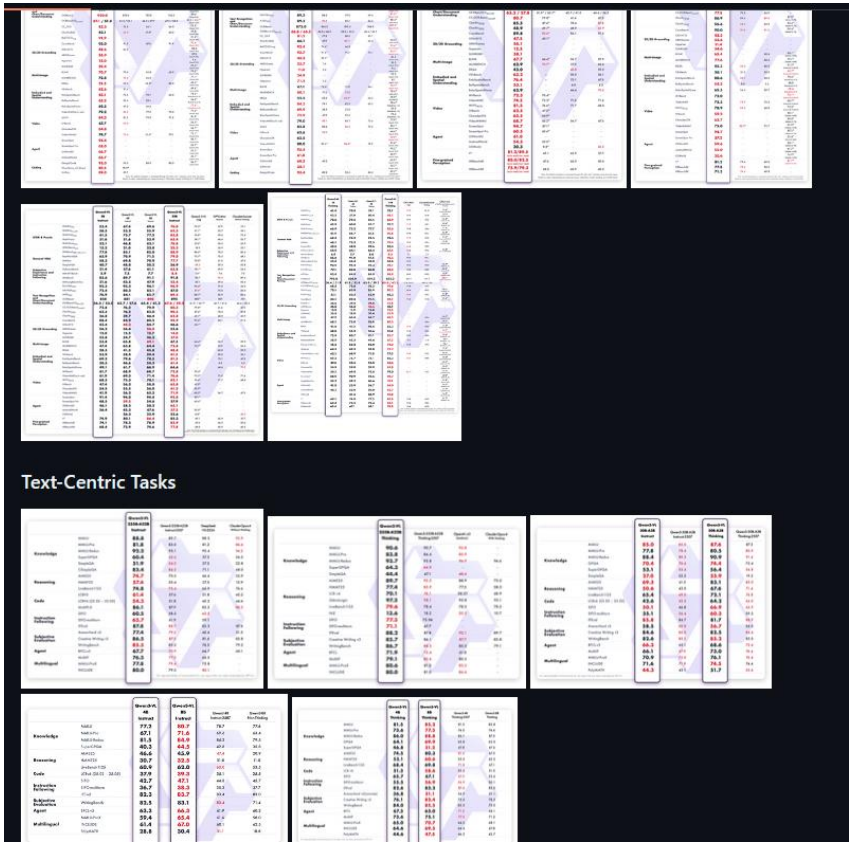
State of Art

Slide by Chunyuan Li

Georgia Tech

- **Key Enhancements:**
- **Visual Agent**: Operates PC/mobile GUIs—recognizes elements, understands functions, invokes tools, completes tasks.
- **Visual Coding Boost**: Generates Draw.io/HTML/CSS/JS from images/videos.
- **Advanced Spatial Perception**: Judges object positions, viewpoints, and occlusions; provides stronger 2D grounding and enables 3D grounding for spatial reasoning and embodied AI.
- **Long Context & Video Understanding**: Native 256K context, expandable to 1M; handles books and hours-long video with full recall and second-level indexing.
- **Enhanced Multimodal Reasoning**: Excels in STEM/Math—causal analysis and logical, evidence-based answers.
- **Upgraded Visual Recognition**: Broader, higher-quality pretraining is able to "recognize everything"—celebrities, anime, products, landmarks, flora/fauna, etc.
- **Expanded OCR**: Supports 32 languages (up from 10); robust in low light, blur, and tilt; better with rare/ancient characters and jargon; improved long-document structure parsing.
- **Text Understanding on par with pure LLMs**: Seamless text–vision fusion for lossless, unified comprehension.

https://github.com/QwenLM/Qwen3-VL?tab=readme-ov-file



Text-Centric Tasks

		Qwen3-VL 30B-A3B Thinking	GPT5-Mini High	Claude4-Sonnet Thinking	Other Best of Similar-sized Open-source Models
STEM & Puzzle	$MMMU_{VAL}$	76.0	79.0	74.4*	75.6* (Internvl3.5-30A3)
	$MMMU_{Pro\_full}$	63.0	67.3	61.6	57.1* (GLM-4.1V-9B)
	$MathVista_{mini}$	81.9	79.1	74.6	81.8* (MiMo-VL 7B)
	MathVision	65.7	71.9	62.1	60.4* (MiMo-VL 7B)
	$MathVerse_{mini}$	79.6	78.8	68.6	71.5* (MiMo-VL 7B)
General VQA	$MMBench_{DEV\_EN\_V1.1}$	88.9	86.8	82.2	85.8* (GLM-4.1V-9B)
	RealWorldQA	77.4	79.0	69.8	72.3* (Internvl3.5-30A3)
	MMStar	75.5	74.1	69.4	72.9* (GLM-4.1V-9B)
	SimpleVQA	54.3	56.8	53.3	–
Subjective Experience and Instruction Following	HallusionBench	66.0	63.2	59.2	53.8* (Internvl3.5-30A3)
	MM-MT-Bench	7.9	7.7	7.9	–
	MIABench	91.6	92.0	90.4	–
Text Recognition and Chart/Document Understanding	$DocVQA_{test}$	95.0	90.0	92.0	95.7* (MiMo-VL 7B)
	$InfoVQA_{test}$	86.0	78.0	58.0	88.0* (MiMo-VL 7B)
	$AI2D_{test}$	86.9	88.2	83.0	87.9* (GLM-4.1V-9B)
	OCRBench	839.0	821.0	739.0	880.0* (Internvl3.5-30A3)
	$OCRBenchV2_{en/zh}$	62.6 / 60.4	52.6 / 45.1	44.9 / 39.4	–
	$CC\text{-}OCR\text{-}Bench_{overall}$	77.8	70.8	66.9	–
	$CharXiv_{(DQ)}$	86.9	89.4	89.5	87.0* (MiMo-VL 7B)
	$CharXiv_{(RQ)}$	56.6	68.6	63.3	56.5* (MiMo-VL 7B)
2D/3D Grounding	CountBench	90.0	91.0	91.0	90.4* (MiMo-VL 7B)
	ODinW13	42.3	–	–	–
	ARKitScenes	55.6	–	–	–
	Hypersim	11.4	–	–	–
	SUNRGBD	34.6	–	–	–
Multi-Image	BLINK	65.4	–	60.4	65.1* (GLM-4.1V-9B)
	MUIRBENCH	77.6	–	68.6	74.7* (GLM-4.1V-9B)
Embodied and Spatial Understanding	ERQA	45.3	54.0	46.0	41.5* (Internvl3.5-30A3)
	VSI-Bench	56.1	31.5	33.3	63.7* (Internvl3.5-30A3)
	EmbSpatialBench	80.6	80.7	68.5	78.6* (RoboBrain 2.0)
	RefSpatialBench	54.2	9.0	3.6	54.0 (RoboBrain 2.0)
	RoboSpatialHome	65.5	54.3	69.7	72.4 (RoboBrain 2.0)
Video	MVBench	72.0	–	–	72.1* (Internvl3.5-30A3)
	VideoMME	73.3	78.9	72.3	68.7* (Internvl3.5-30A3)
	$MLVU_{MCQ}$	78.9	83.3	68.8	73.0* (Internvl3.5-30A3)
	LVBench	59.2	–	–	45.1* (GLM-4.1V-9B)

https://github.com/QwenLM/Qwen3-VL?tab=readme-ov-file

Evaluation

Slide by Chunyuan Li

Georgia Tech

- Simplify design, scale data and training
- From captioning → reasoning + action
- From static images → long-context multimodal streams
- From human data → synthetic, model-generated
- From SFT-only → full RL-style optimization pipelines

- Some unified auto-regressive + diffusion but early

Hallucination in Multi-Modal Models

- Hallucination: model generates plausible-sounding but incorrect visual claims
  - → More dangerous than text hallucination: used in medical / legal / safety settings
- Object hallucination: inventing objects not present in image
  - → 'I see a dog, cat, and elephant' — image has only dog and cat
- Attribute hallucination: correct object, wrong color/size/position
  - → 'The red car' when car is clearly blue — visual encoder not attended properly
- Causes: over-reliance on language priors; insufficient visual grounding
  - → Language model learned 'people in kitchens usually have knives' → hallucinates
- Mitigation: RLHF, DPO on factual preference data; contrastive decoding
  - → Contrastive decoding: penalize tokens that increase when visual context removed
- Evaluation: CHAIR (caption hallucination), HallusionBench, POPE
  - → POPE: 3K binary yes/no questions about object presence; adversarial negatives

Open Problems and Research Directions

- Compositionality: VLMs fail at 'red cube left of blue sphere above green cylinder'
  - → WinoGround benchmark: 72% chance accuracy; models struggle with spatial composition
- Long visual context: processing hour-long videos or 1000-page documents efficiently
  - → KV cache compression, token merging (ToMe), hierarchical encoders
- Efficient architectures: reduce quadratic attention cost for long multi-modal inputs
  - → Linear attention, state space models (Mamba) for visual sequence modeling
- Multi-modal reasoning chains: Chain-of-Thought for visual arithmetic and geometry
  - → ScienceQA + CoT: 'Step 1: identify the triangle. Step 2: apply Pythagoras...'
- Grounded generation: produce images/audio consistent with textual specifications
  - → Stable Diffusion / DALL-E 3 with enhanced text-image alignment
- Safety and alignment: preventing misuse in deepfakes, misinformation, surveillance
  - → C2PA watermarking, model fingerprinting, usage policies

- Vision+Language (and multi-modal) are hot!
- **Why?**
  - Align various interface modalities
  - Leverage more data (all modalities)
  - Physical world inherently multi-modal
- **Tasks**:
  - Image ⇔ language
  - Visual question answering
    - + Interaction
  - Embodied AI
- **Large number of design choices!**
  - Vision encoding?
  - Method of alignment?
  - Method of fusion?
  - Grounding?

Resources:
https://www.youtube.com/@VLPTutorial

Summary

Georgia Tech