Cross-Modal Knowledge Transfer

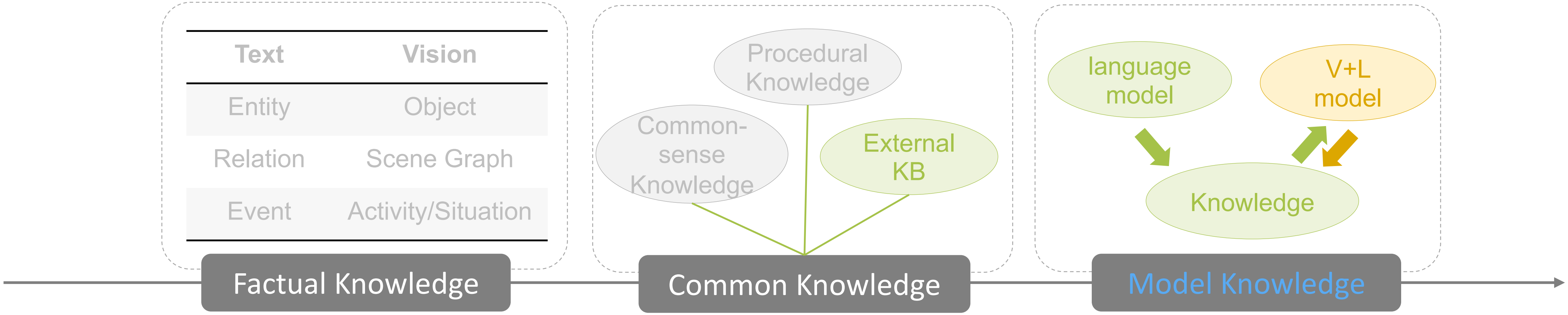
Knowledge-Driven Vision-Language Pretraining (Part V)

Jie Lei
Meta AI
jielei@meta.com

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
COLUMBIA UNIVERSITY
UNC
Meta AI

Overview

Compared to raw data, knowledge is **important and useful information.**

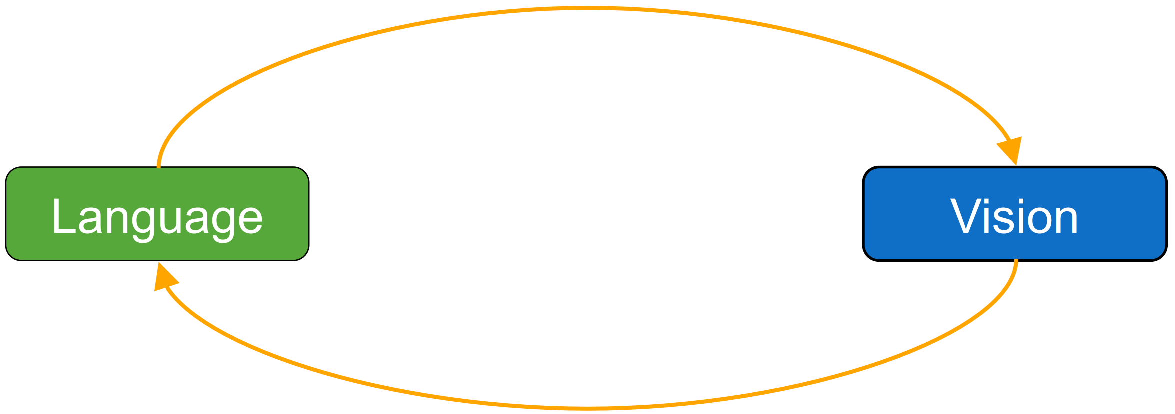


Text	Vision
Entity	Object
Relation	Scene Graph
Event	Activity/Situation

Agenda

Part 1. Language knowledge helps learn better vision models

- Pure vision tasks: object detection, image classification, etc.
- Multimodal tasks: VQA, video captioning, etc.

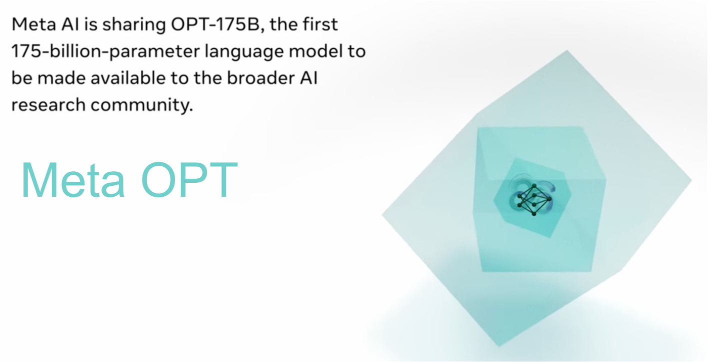


Part 2. Vision knowledge helps learn better language models

- Human learn language by connecting the words to their visual appearance in the surrounding world.

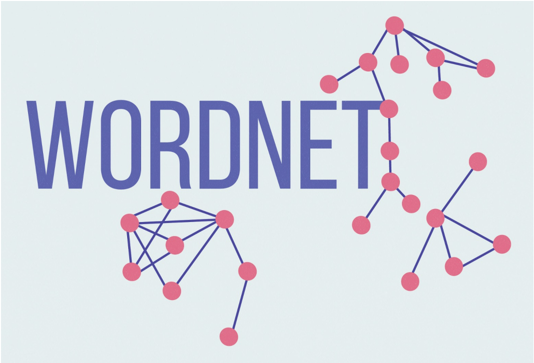
Part 1. Language → Vision

- Implicit knowledge from pre-trained Language Models (LM)



- Explicit knowledge from human curated sources (e.g., wiki) or model generated knowledge (e.g., GPT-3 generated category definitions)

Wiktionary
The free dictionary



- Concept name: snowberg
- Def_wik: None
- GPT3 Query:

Please explain the concept according to the context.

===

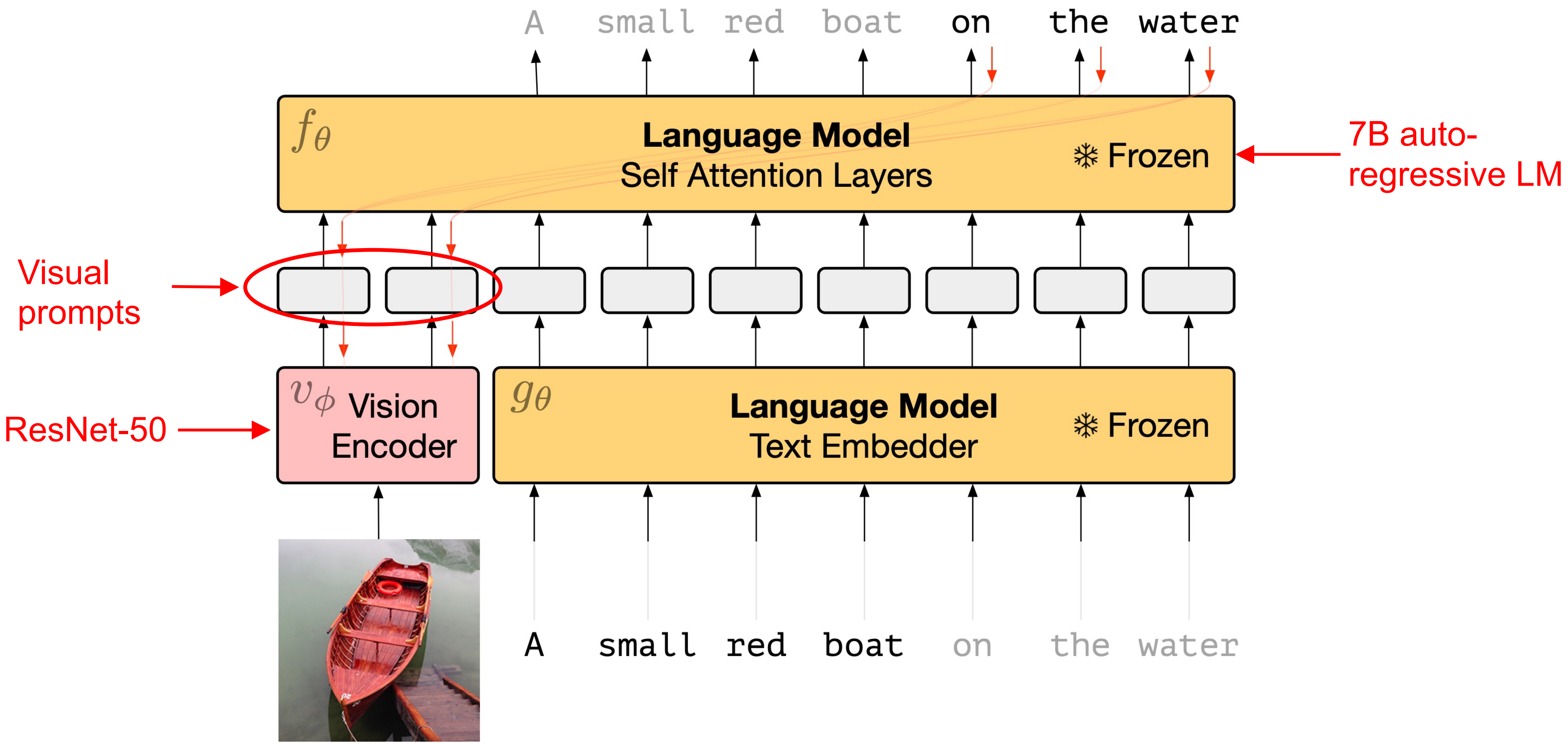
Q: ship

A: A water-borne vessel generally larger than a boat.

Part 1.1 Implicit Knowledge from Language Models

Frozen

- Preserve LM ability by freezing it during cross-modal model training.
- Gradient: frozen LM → vision encoder



Multimodal Few-Shot Learning with Frozen Language Models, Tsimpoukelli et al., NeurIPS 2021

Frozen

- Few-shot multimodal in-context learning after trained on 3M image-text pairs.

Wiki knowledge

- Reasonably good zero/few-shot performance, but still underperform SOTA: limited multimodal data? (3M); LM is relatively small? (7B)

VQAv2

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	29.5	35.7	38.2	✗
Frozen scratch	0.0	0.0	0.0	✗
Frozen finetuned	24.0	28.2	29.2	✗
Frozen train-blind	26.2	33.5	33.3	✗
Frozen VQA	48.4	–	–	✓
Frozen VQA-blind	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

OKVQA

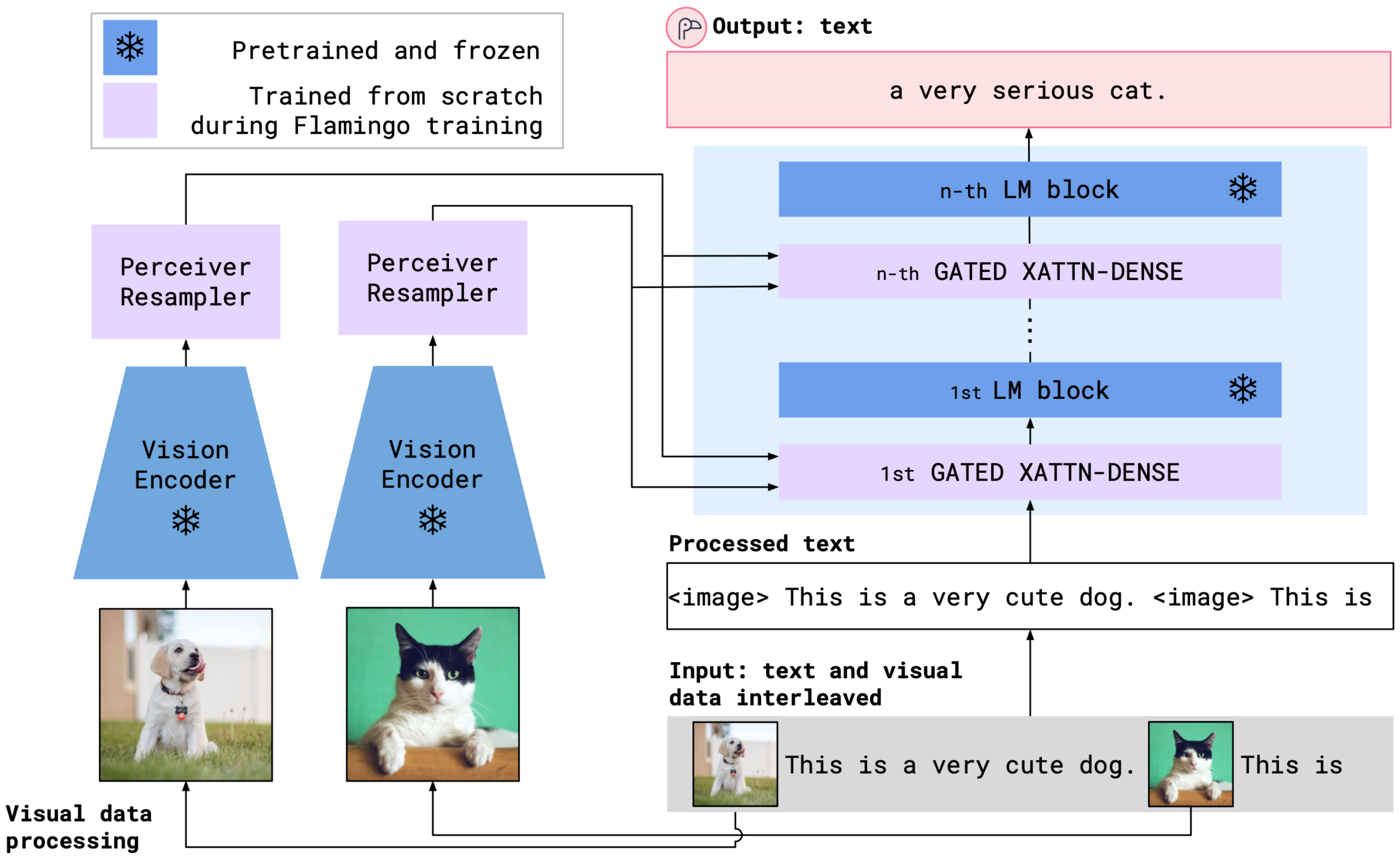
n-shot Acc.	n=0	n=1	n=4	τ
Frozen	5.9	9.7	12.6	✗
Frozen 400mLM	4.0	5.9	6.6	✗
Frozen finetuned	4.2	4.1	4.6	✗
Frozen train-blind	3.3	7.2	0.0	✗
Frozen VQA	19.6	–	–	✗
Frozen VQA-blind	12.5	–	–	✗
MAVEx [42]	39.4	–	–	✓

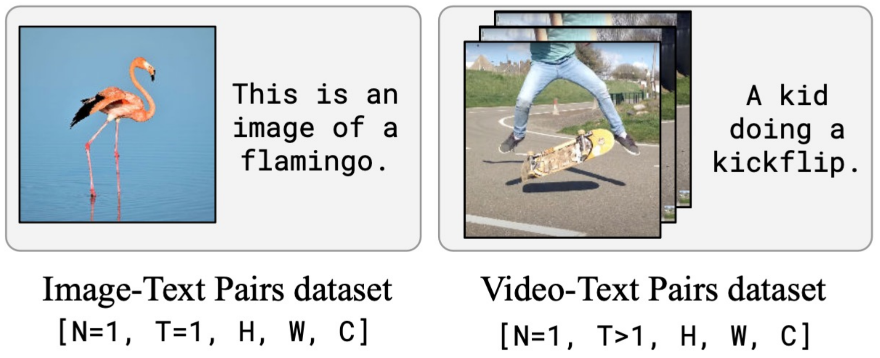
Large gap w/ SOTA

Multimodal Few-Shot Learning with Frozen Language Models, Tsimpoukelli et al., NeurIPS 2021

Flamingo

- A frozen 70B pre-trained LM + a frozen pre-trained ResNet.
- Trained w/ image/video-text pairs, along with interleaved image-text data (M3W), which is important for in-context learning.





ALIGN: 1.8B +
LTIP: 312M images

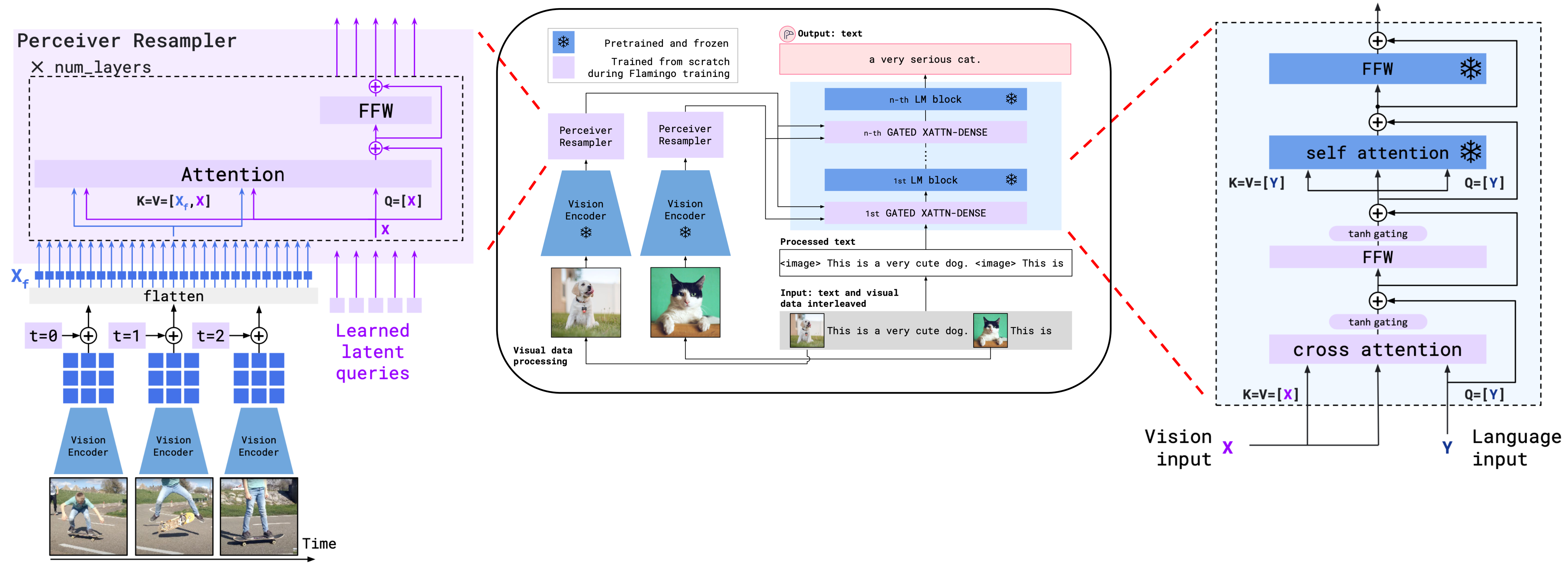
VTP: 27M videos

M3W: 43M webpages (185M images)

Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac et al, arXiv 2022

Flamingo

- Perceiver resampler (left): map variable sized visual inputs to fixed length visual tokens.
- Gated x-attn (right): bridge vision and language inputs & better preserve info in pre-trained LM. The tanh gate with zero initialization makes sure a smooth transition from a text-only LM to a visual-language model.



Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac et al, arXiv 2022

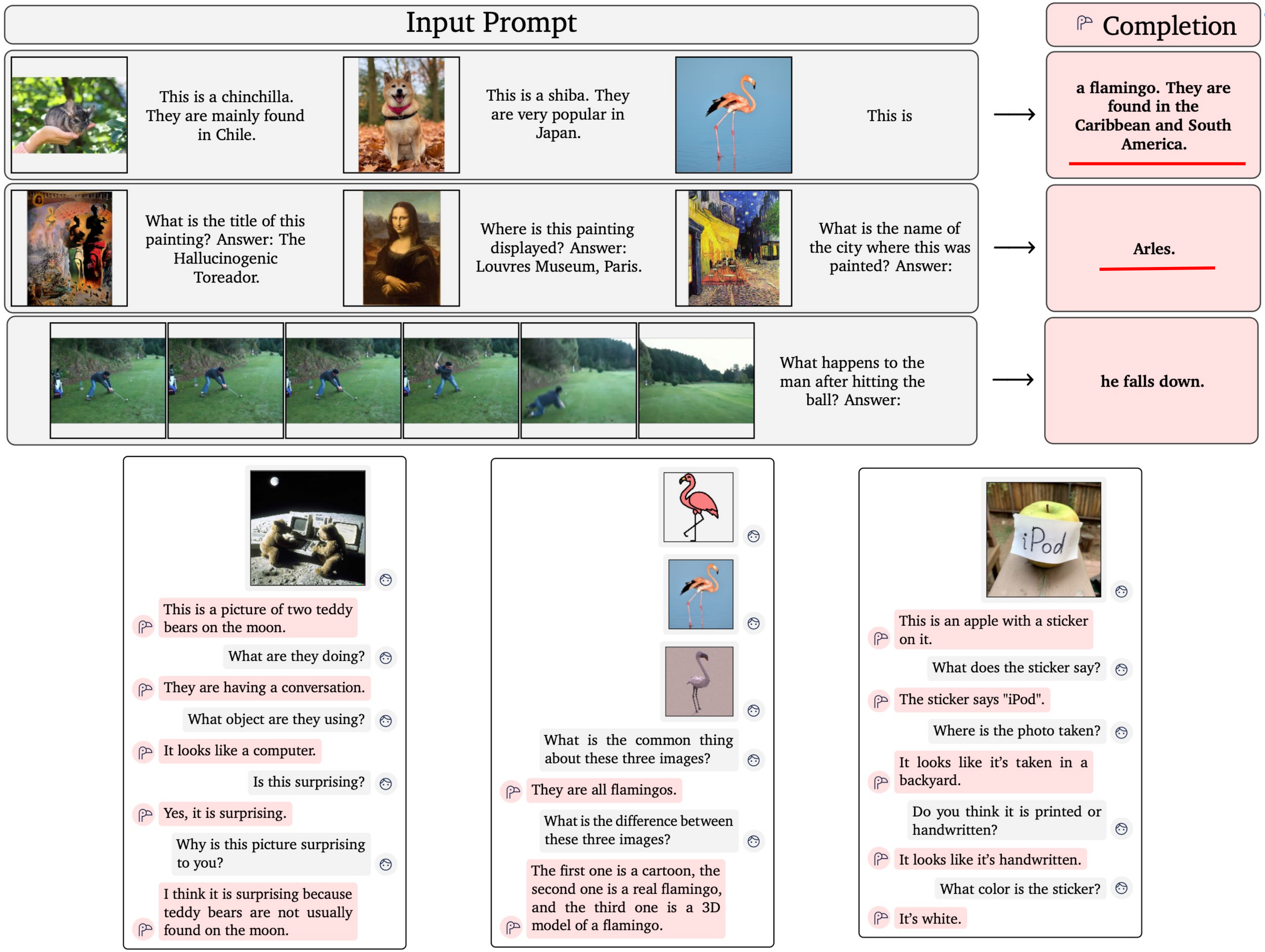
Flamingo

Image Captioning

Image QA

Video QA

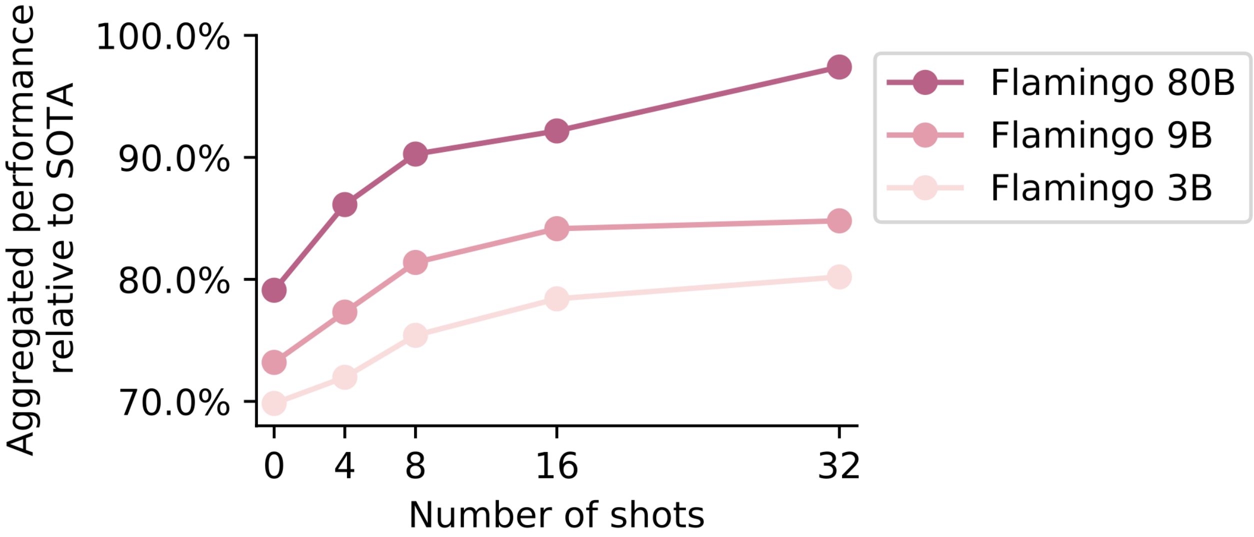
Visual Dialogue

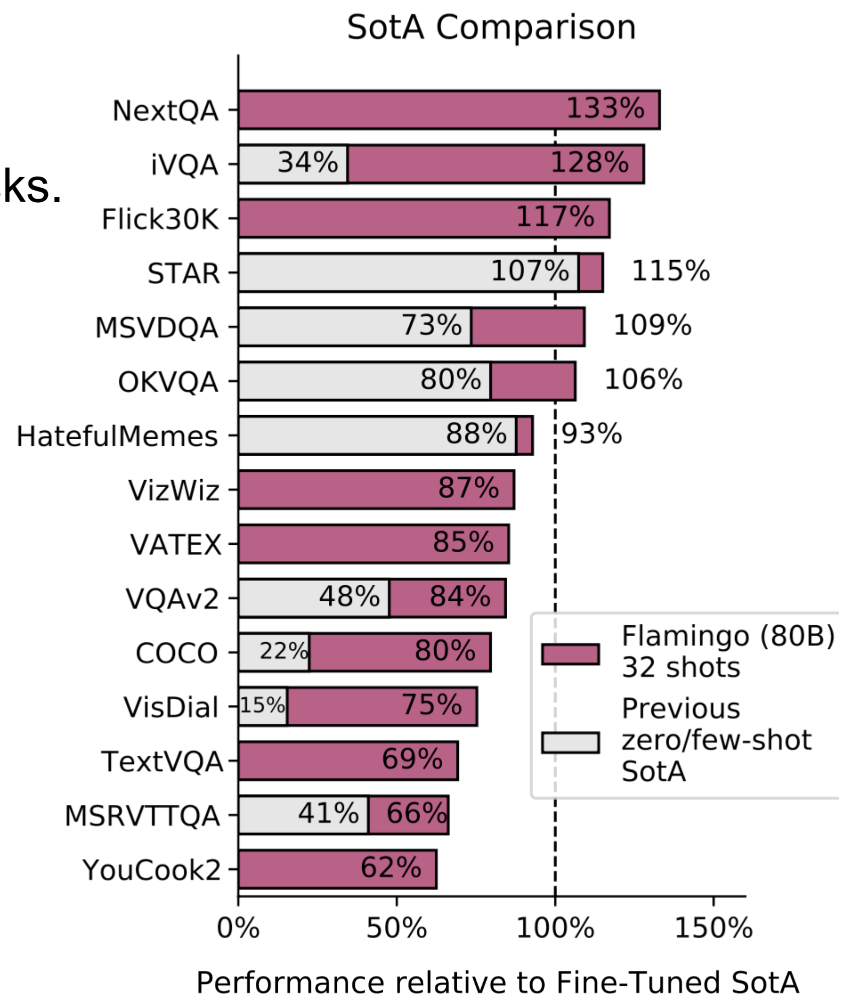


Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac et al, arXiv 2022

Flamingo

- Left: larger model works better; more in-context examples helps.
- Right: thanks to larger model and more training data, he model achieves comparable or better results than SOTA on multiple tasks.



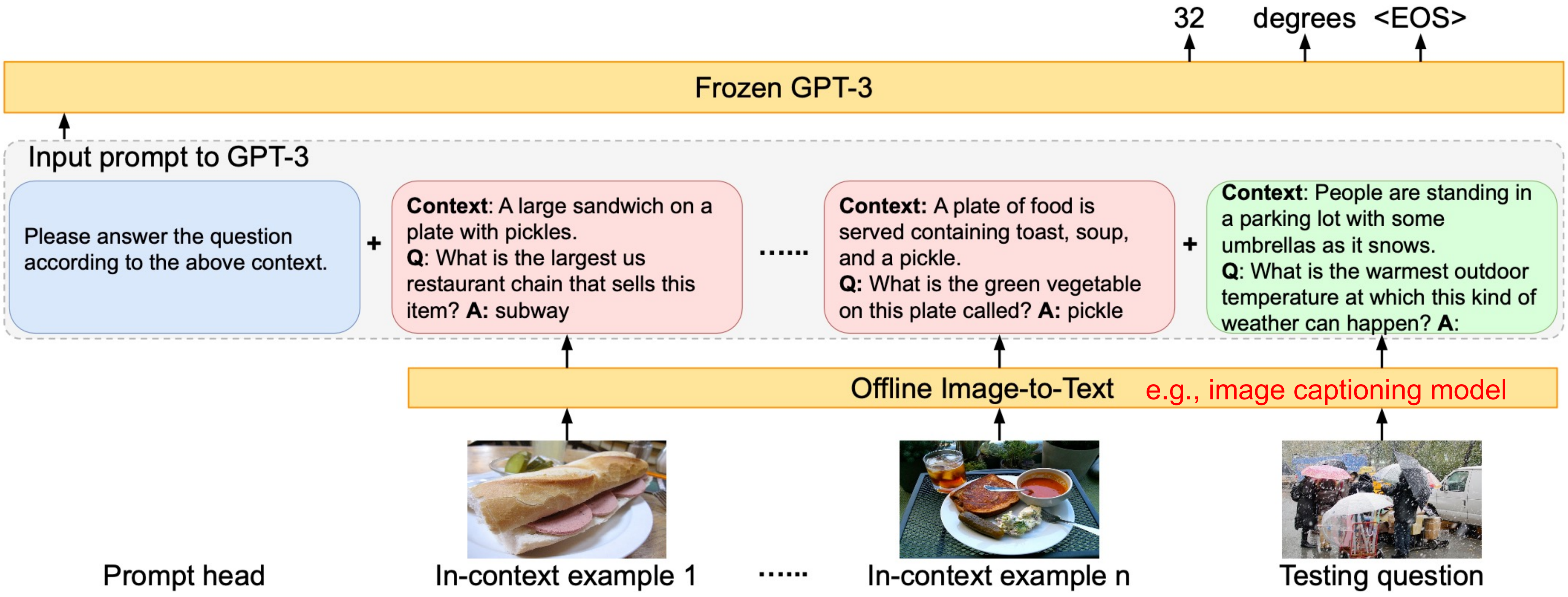


Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac et al, arXiv 2022

The method discussed above all require additional multi-modal pre-training, however, it is very expensive for LLMs. Is there an alternative way to utilize knowledge in LLMs?

PICa for few-shot knowledge-based VQA

- Summarize image info in text form with an image-to-text model, and prompt GPT-3 to get an answer.
  - Image QA problem is converted into a text QA problem.
  - Implicit GPT-3 knowledge <-> previous approaches explicitly query external knowledge
  - Few-shot w/o parameter update.



An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, Yang et al, AAAI 2022

PICa for few-shot knowledge-based VQA

- Works better than fine-tuned models that use explicit wiki knowledge.
- A core issue: image-to-text model is not perfect, it will cause information loss.

OKVQA

Method	Image Repr.	Knowledge Resources	Few-shot	Accuracy
MUTAN+AN (Ben-Younes et al. 2017)	Feature Emb.	Wikipedia	✗	27.8
Mucko (Zhu et al. 2020)	Feature Emb.	Dense Captions	✗	29.2
ConceptBert (Garderes et al. 2020)	Feature Emb.	ConceptNet	✗	33.7
ViLBERT (Lu et al. 2019)	Feature Emb.	None	✗	35.2
KRISP (Marino et al. 2021)	Feature Emb.	Wikipedia + ConceptNet	✗	38.9
MAVEx (Wu et al. 2021)	Feature Emb.	Wikipedia + ConceptNet + Google Images	✗	39.4
Frozen (Tsimpoukelli et al. 2021)	Feature Emb.	Language Model (7B)	✓	12.6
PICa-Base	Caption	GPT-3 (175B)	✓	42.0
PICa-Base	Caption+Tags	GPT-3 (175B)	✓	43.3
PICa-Full	Caption	GPT-3 (175B)	✓	46.9
PICa-Full	Caption+Tags	GPT-3 (175B)	✓	**48.0**

(e) What color is the man's jacket?
Context: A man flying through the air while riding a snowboard.
Answer: black
GT Answer: ['red', 'red', 'red', 'orange', 'red', 'red', 'red', 'red', 'red', 'red']

Acc.: 0.0

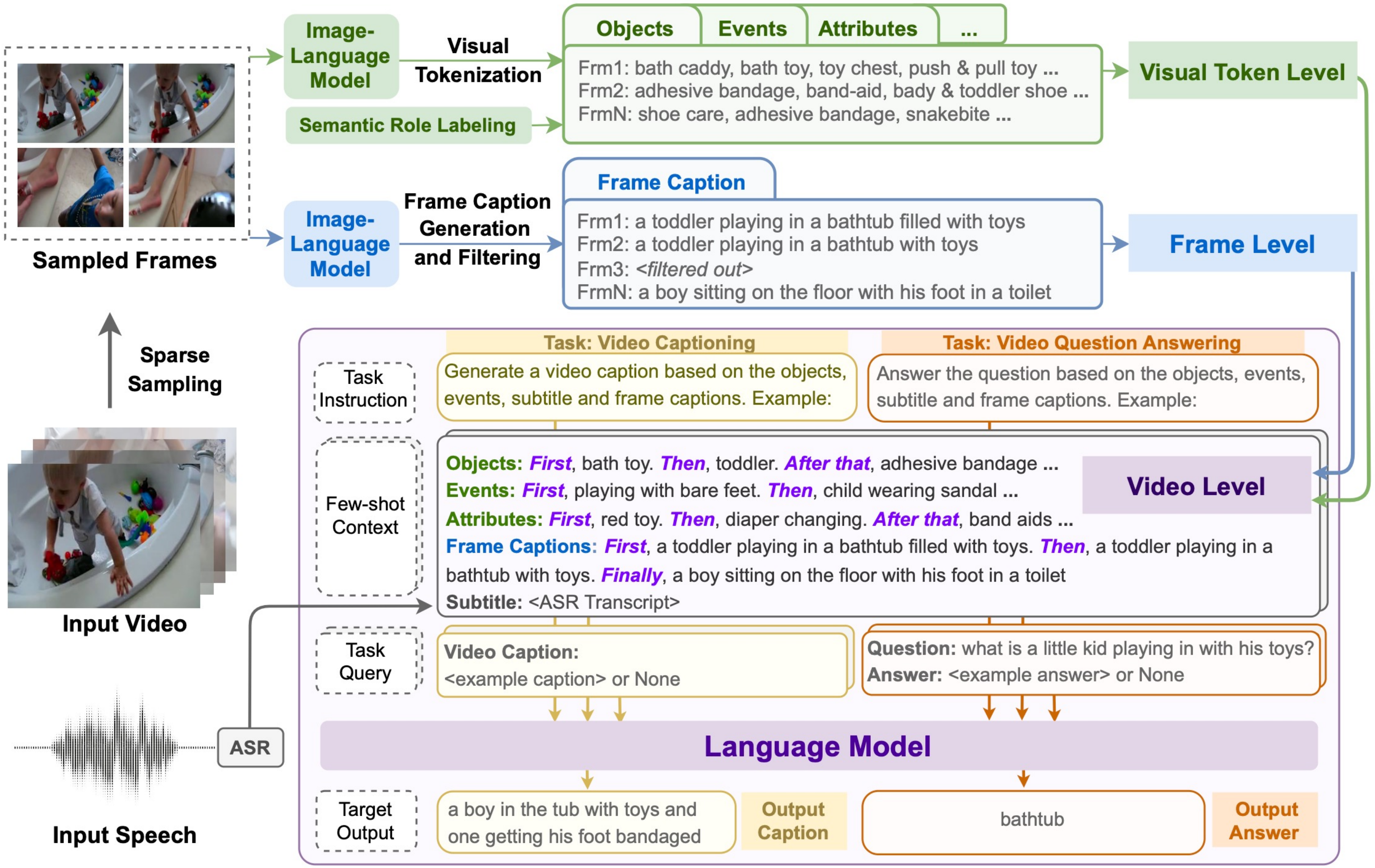
(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3
GT Answer: ['6', '6', '8', '6', '8', '6', '6', '7', '8', '7']

Acc.: 0.0

An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, Yang et al, AAAI 2022

VidIL: LLM video + language learning

- Generate frame-level info at various granularity, and put them in a temporal aware prompt for LLM.

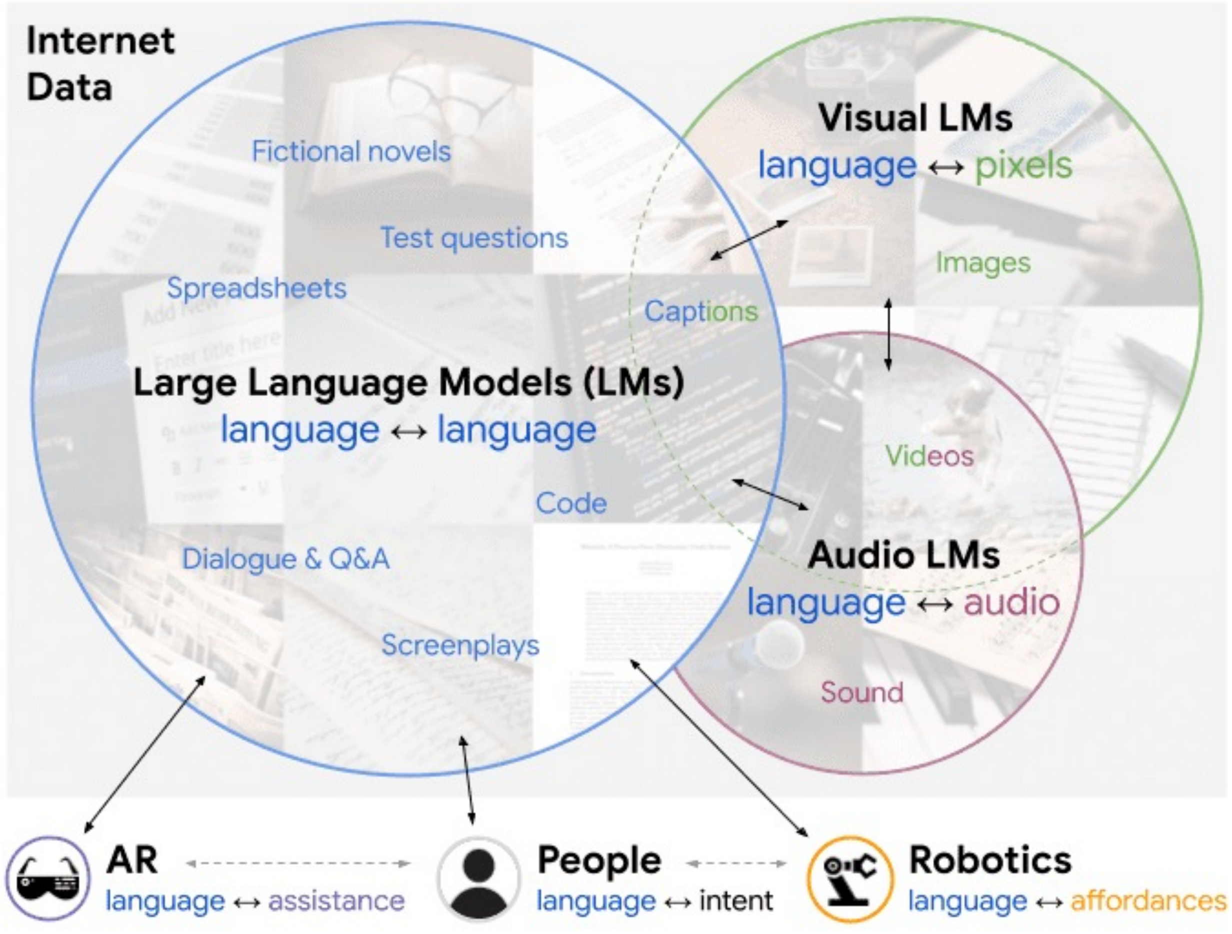


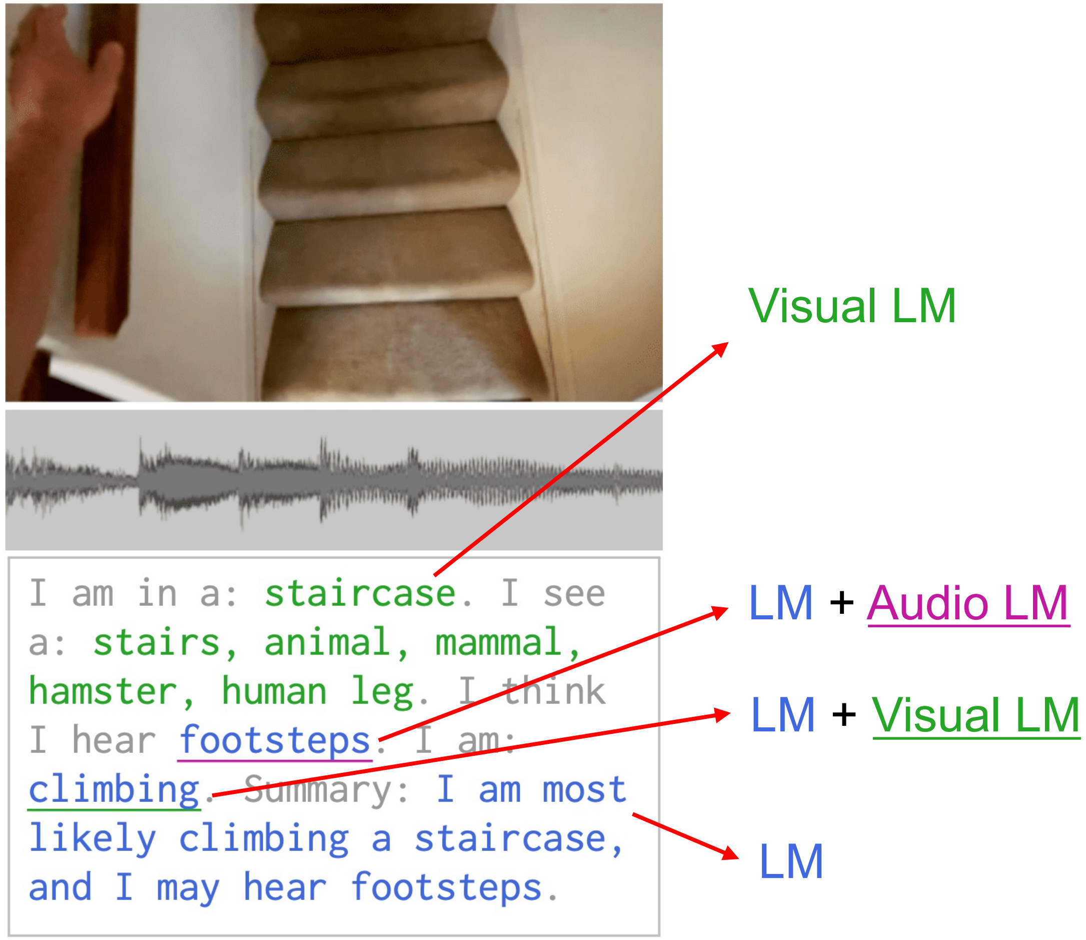
Temporal-Aware Few-shot Prompt

Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners, Wang et al, NeurIPS 2022

Socratic: Composing Multi-modality w/ LLM

- A modular framework in which multiple pretrained models may be composed zero-shot through language without training.





Summarize ego-centric videos.

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Zeng et al, arXiv 2022

Socratic: Composing Multi-modality w/ LLM

- The model works well on image-text tasks such as image captioning, and video-text tasks such as text-to-video retrieval. It can also parse & generate robot instructions from free form human language.

COCO captions

Method	BLEU-4	METEOR	CIDEr	SPICE	ROUGE-L
*ClipCap [45]	40.7	30.4	152.4	25.2	60.9
†MAGIC [61]	11.4	16.4	56.2	11.3	39.0
ZeroCap [62]	0.0	8.8	18.0	5.6	18.3
SMs 0-shot (ours)	6.9	15.0	44.5	10.1	34.1
SMs 3-shot (ours)	**18.3**	**18.8**	**76.3**	**14.8**	**43.7**

*finetuned on full training set with image-text pairs.
†finetuned on unpaired training set, zero-shot on image-text pairs.

MSRVTT retrieval

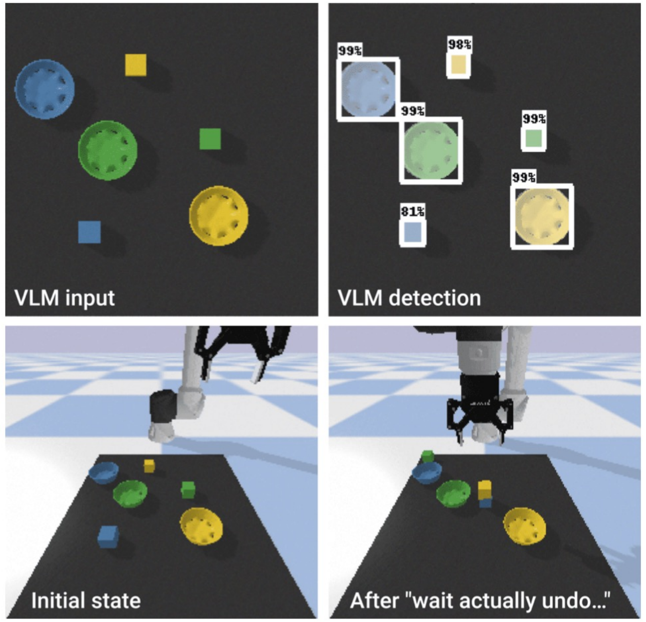
		MSR-VTT Full				
Category	Method	R@1↑	R@5↑	R@10↑	MdR↓	Audio
Finetuned	JEMC [70]	12.5	32.1	42.4	16.0	yes
	Collab. Experts [55]	15.6	40.9	55.2	8.3	yes
	CLIP2Video [71]	**54.6**	**82.1**	**90.8**	**1.0**	no
Zero-shot	CLIP via [67]	40.3	69.7	79.2	**2.0**	no
	SMs (ours)	**44.7**	**71.2**	**80.0**	**2.0**	yes

Visual LM

User Instruction

LM

```
objects = ["green block", "blue block", "yellow block", "green
bowl", "blue bowl", "yellow bowl"]
# move all the blocks to different corners.
Step 1. robot.pick_and_place("green block", "top left corner")
Step 2. robot.pick_and_place("blue block", "top right corner")
Step 3. robot.pick_and_place("yellow block", "bottom left corner")
# now move the blue block to the middle.
Step 1. robot.pick_and_place("blue block", "middle")
# stack the blocks on top of each other.
Step 1. robot.pick_and_place("yellow block", "blue block")
Step 2. robot.pick_and_place("green block", "yellow block")
# wait actually undo that last step.
Step 1. robot.pick_and_place("green block", "top left corner")
# put the yellow block in the bowl you think it best fits.
Step 1. robot.pick_and_place("yellow block", "yellow bowl")
# ok now sort the remaining blocks in the same way.
Step 1. robot.pick_and_place("blue block", "blue bowl")
Step 2. robot.pick_and_place("green block", "green bowl")
```



Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Zeng et al, arXiv 2022

LLM for ZS multi-modal learning: Pros/Cons

Pros

- It provides an efficient way to utilize foundation models of different modalities, no extra training required.
- The approaches are modular: new modules can be seamlessly plugged into the framework.

Cons

- Modality specific models are not perfect, there will be info loss when converted into text.
  - The lower performance vs. e2e trained Flamingo model might partly due to this info loss.

(e) What color is the man's jacket?
Context: A man flying through the air while riding a snowboard.
Answer: black
GT Answer: ['red', 'red', 'red', 'orange', 'red', 'red', 'red', 'red', 'red', 'red']

Acc.: 0.0

(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3
GT Answer: ['6', '6', '8', '6', '8', '6', '6', '7', '8', '7']

Acc.: 0.0

Failure cases from the PICa model.

An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, Yang et al, AAAI 2022

The use of implicit knowledge from pre-trained LMs shows strong zero-shot performance for multi-modal tasks, however, they are hard to interpret. Is there a more interpretable way of using language knowledge?

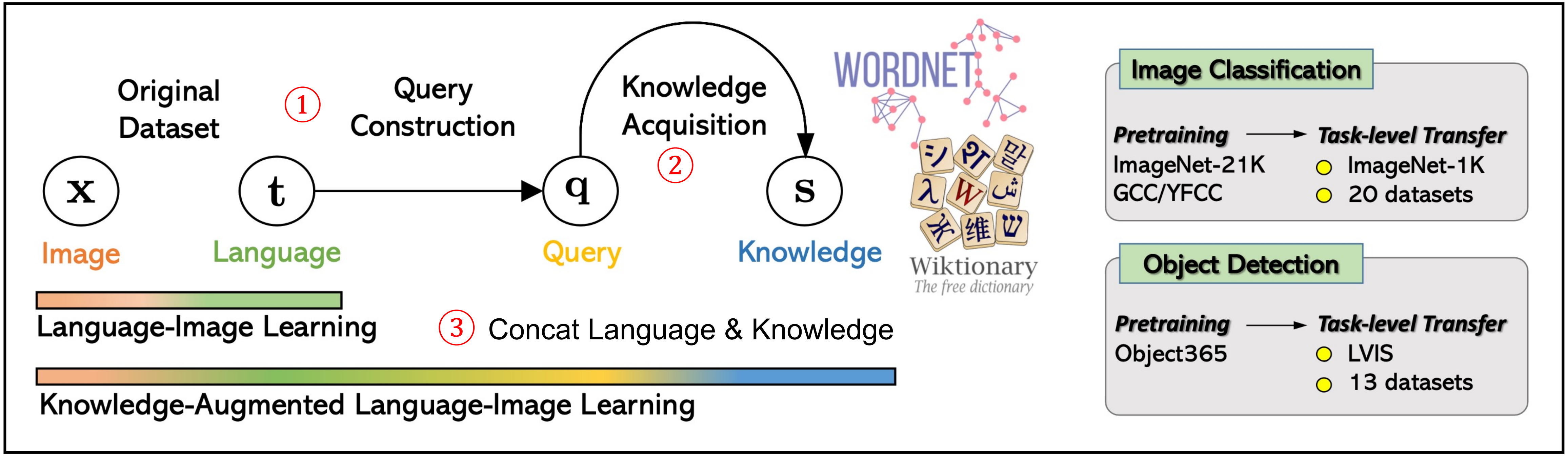
Part 1.2 Explicit Knowledge from Language

K-LITE

- External knowledge is useful to help the model understand rare concepts.

Sashimi

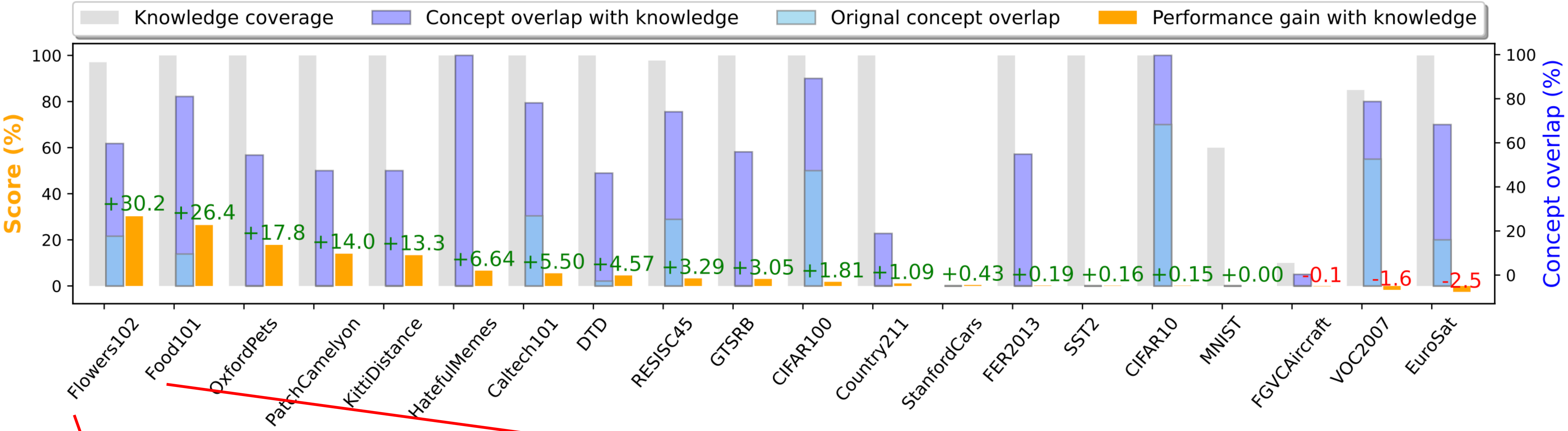
A dish consisting of ***thin slices*** or pieces of ***raw fish or meat***.



K-LITE: Learning Transferable Visual Models with External Knowledge, Shen et al., NeurIPS 2022

K-LITE

- Orange: knowledge improves zero-shot performance on 16/20 image classification datasets.



✓ **English marigold**: Any of the Old World plants, of the genus Calendula, with orange, yellow or reddish flowers.

✗ **Wallflower**: Any of several short-lived herbs or shrubs of the Erysimum genus with bright yellow to red flowers.

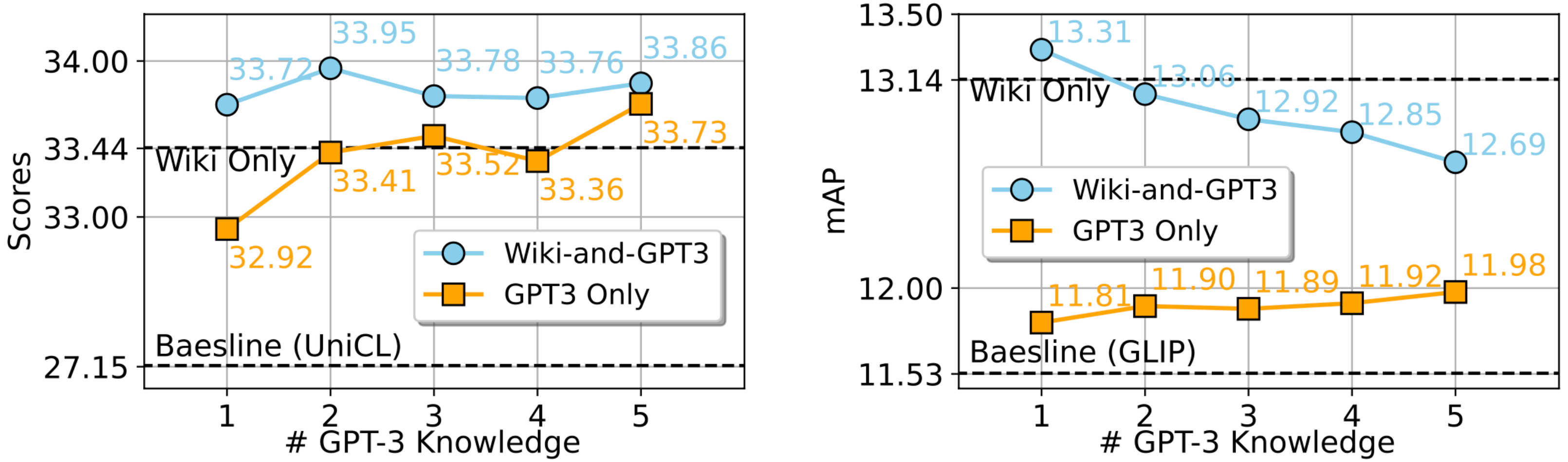
✓ **Lobster bisque:** A thick creamy soup made from fish, shellfish, meat or vegetables.

✗ **Hot and sour soup**: Any one of several soups, served in various Asian cuisines, which are both spicy and sour

ELEVATER

- Same K-LITE model, but with GPT-3 knowledge
- GPT-3 knowledge improves ZS image classification and object detection. More is better.
- GPT-3 + wiki is often better for image classification, but not for object detection.

❑ Concept name: snowberg
Def_wik: None
GPT3 Query:
Please explain the concept according to the context.
===
Q: ship
A: A water-borne vessel generally larger than a boat.
===
Q: storage tank
A: A closed container for liquids or gases.
===
Q: snowberg
A:
GPT3 Answer: A large mass of ice floating in the sea.



(a) Image classification (b) Object detection

Zero-shot performance

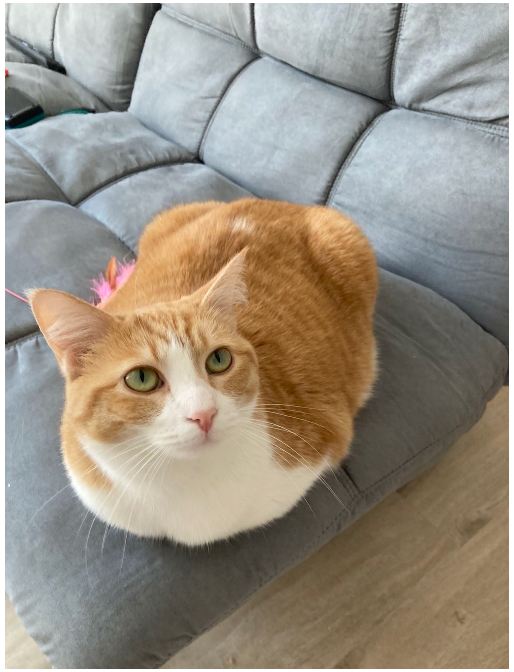
K-LITE: Learning Transferable Visual Models with External Knowledge, Shen et al., NeurIPS 2022
ELEVATER: A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models, NeurIPS 2022

Vision → Language

Could vision knowledge help learn language?

Could vision knowledge help learn language?

- Visual pointing is an essential step for most children to learn meanings of words [Bloom 2002].

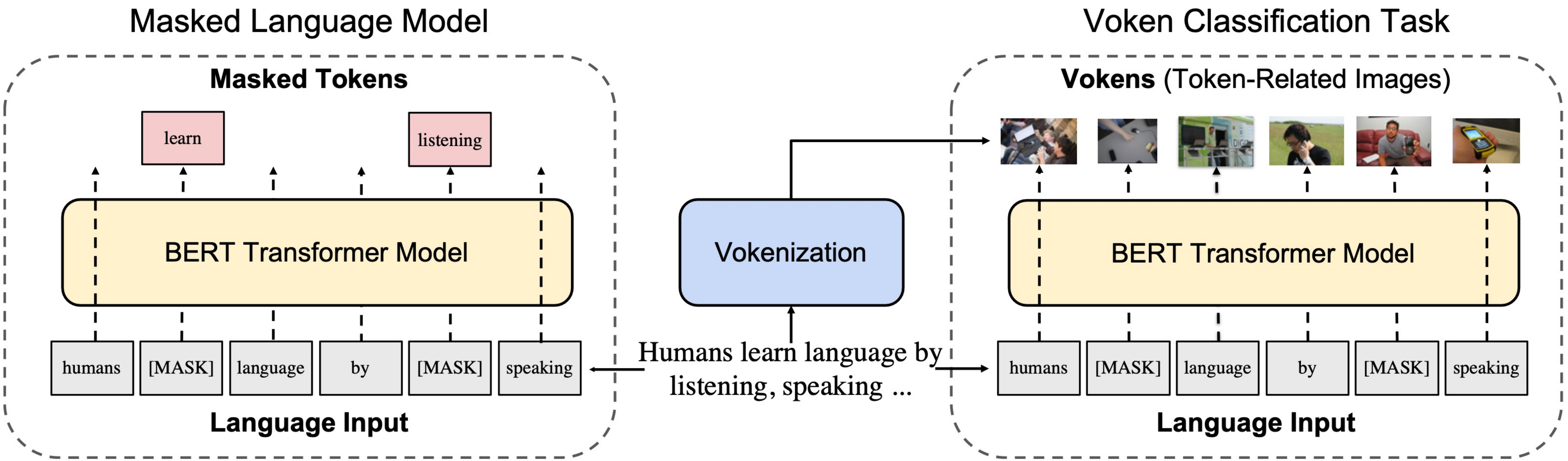


Look! This is a “cat”!

“ButterCup”, cat photo credit to Xiaoyu Xiang
How children learn the meanings of words. Paul Bloom. 2002. MIT press.
Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision, Tan et al., EMNLP 2020

Vokenization: LM w/ Vision Supervision

- Besides standard Masked Language Modeling (MLM), the LM is also trained w/ a voken classification task, by assigning each text token into one of the images (vokens) in the pool.
- Vokens are pre-defined, and are obtained by using a pre-trained image-text retrieval model



Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision, Tan et al., EMNLP 2020

Vokenization: LM w/ Vision Supervision

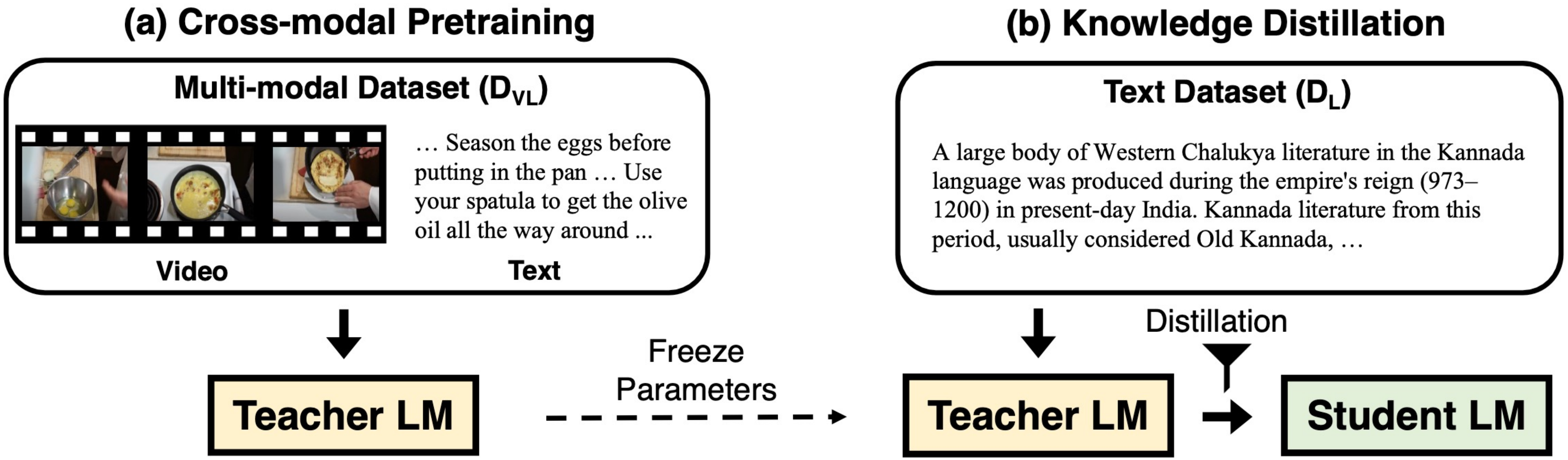
- Voken classification task improves LM performance on a wide range of pure-language tasks.
- This conclusion holds for both BERT and RoBERTa.

Method	SST-2	QNLI	QQP	MNLI	SQuAD v1.1	SQuAD v2.0	SWAG	Avg.
$\text{BERT}_{6L/512H}$	88.0	85.2	87.1	77.9	71.3/80.2	57.2/60.8	56.2	75.6
$\text{BERT}_{6L/512H}$ + Voken-cls	89.7	85.0	87.3	78.6	71.5/80.2	61.3/64.6	58.2	76.8
$\text{BERT}_{12L/768H}$	89.3	87.9	83.2	79.4	77.0/85.3	67.7/71.1	65.7	79.4
$\text{BERT}_{12L/768H}$ + Voken-cls	**92.2**	**88.6**	**88.6**	**82.6**	**78.8/86.7**	68.1/71.2	**70.6**	**82.1**
$\text{RoBERTa}_{6L/512H}$	87.8	82.4	85.2	73.1	50.9/61.9	49.6/52.7	55.1	70.2
$\text{RoBERTa}_{6L/512H}$ + Voken-cls	87.8	85.1	85.3	76.5	55.0/66.4	50.9/54.1	60.0	72.6
$\text{RoBERTa}_{12L/768H}$	89.2	87.5	86.2	79.0	70.2/79.9	59.2/63.1	65.2	77.6
$\text{RoBERTa}_{12L/768H}$ + Voken-cls	**90.5**	**89.2**	**87.8**	**81.0**	**73.0/82.5**	**65.9/69.3**	**70.4**	**80.6**

Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision, Tan et al., EMNLP 2020

VidLanKD: LM w/ Video-Distilled Knowledge

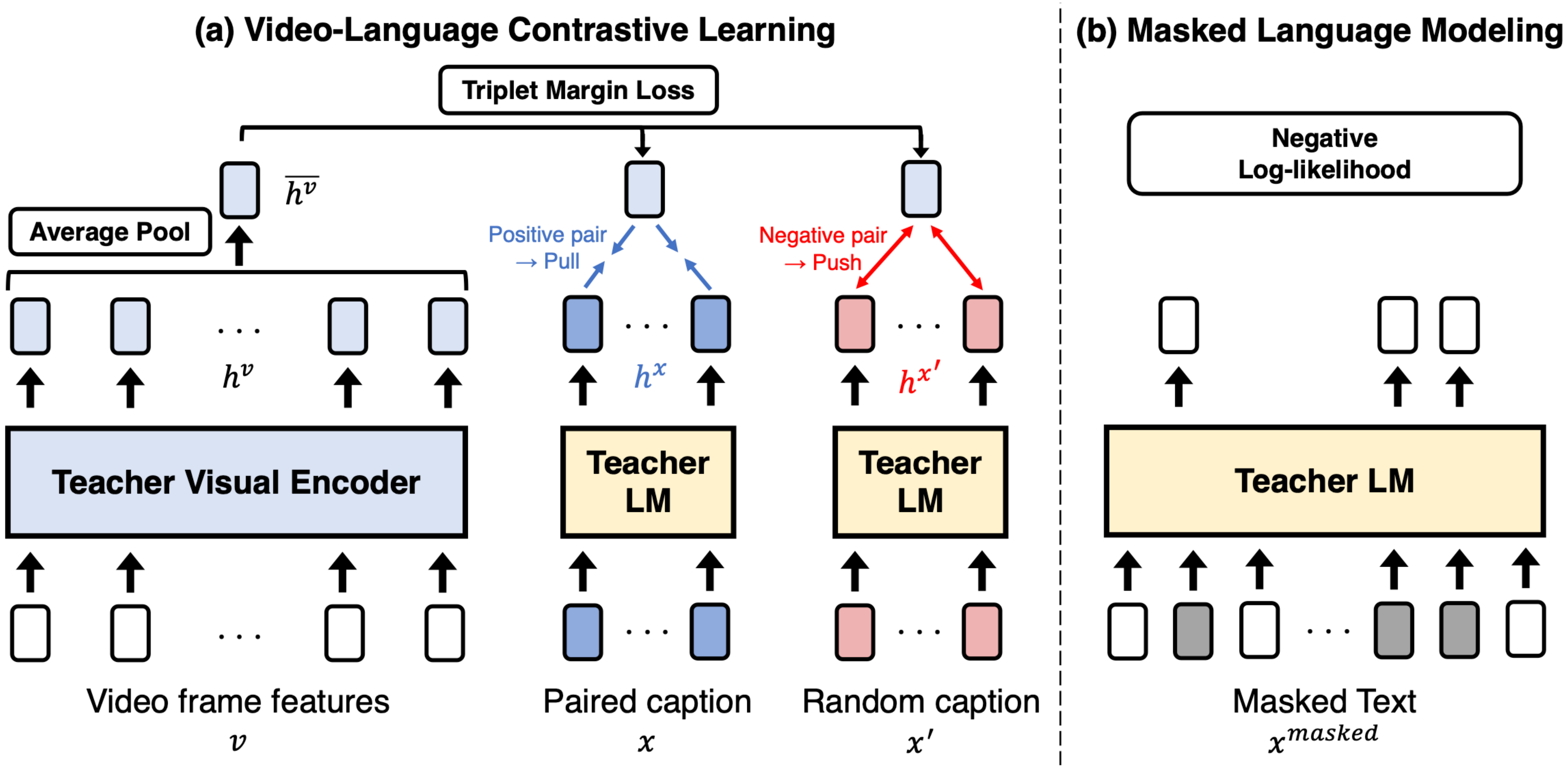
- Vokenization suffers from approximation error of using finite image labels + the lack of vocabulary diversity of a small image-text dataset (COCO).
- VidLanKD improves it by (1) using knowledge distillation instead of discrete vokenization to avoid approximation error; (2) using a large-scale video-language dataset HowTo100M.



VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer, Tang et al., NeurIPS 2021

VidLanKD: LM w/ Video-Distilled Knowledge

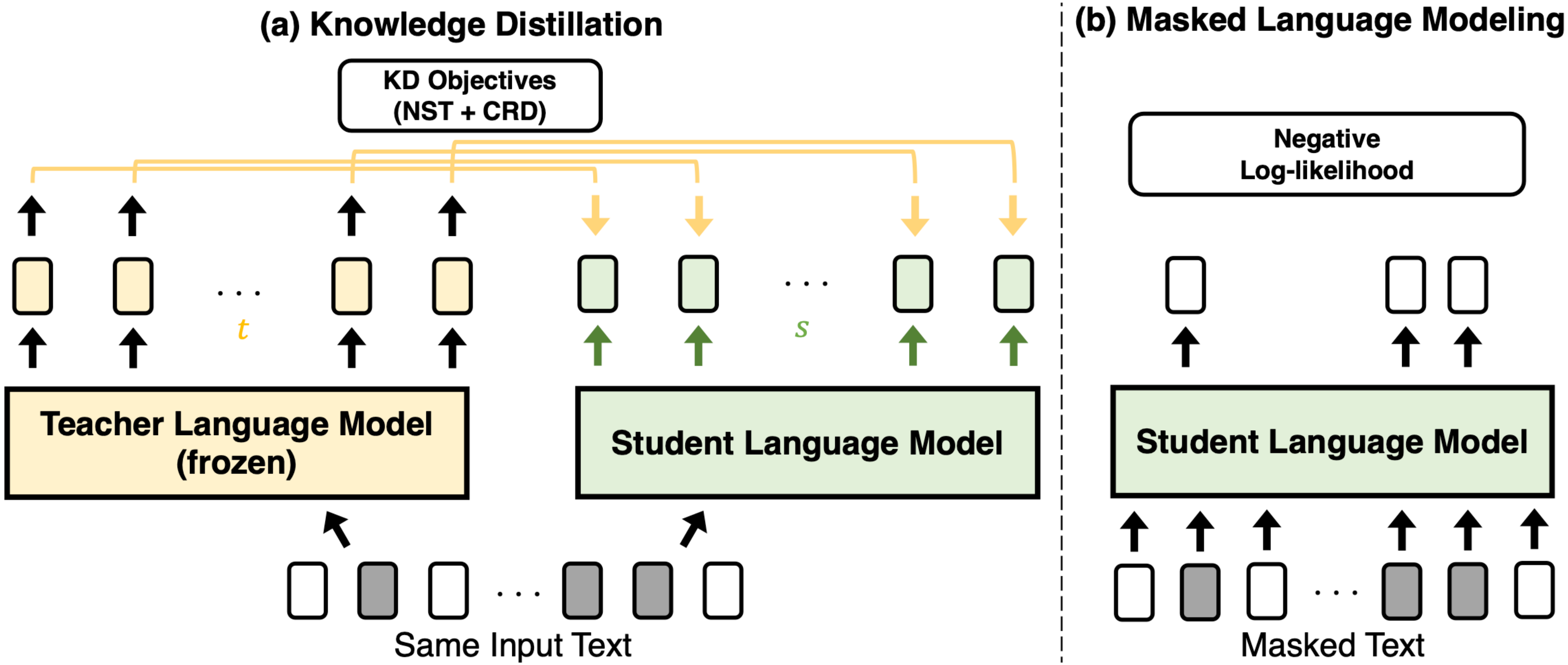
- The teacher LM is trained with (a) video-language contrastive learning; + (b) masked language modeling



VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer, Tang et al., NeurIPS 2021

VidLanKD: LM w/ Video-Distilled Knowledge

- The student LM is trained with (a) knowledge distillation; + (b) masked language modeling



VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer, Tang et al., NeurIPS 2021

VidLanKD: LM w/ Video-Distilled Knowledge

- Cross-modal KD (last 2 rows) achieves better performance than image vokenization.

	SST-2 Acc	QNLI Acc	QQP Acc	MNLI Acc	SQuAD v1.1 EM†	SQuAD v2.0 EM	SWAG Acc	Avg.
$BERT_{12L/768H}$ [68]	89.3	87.9	83.2	79.4	77.0	67.7	65.7	78.6
+ KD (Img-Voken) [68]	92.2	88.6	88.6	82.6	78.8	68.1	70.6	81.4
$BERT_{12L/768H}$	89.0	88.0	86.2	79.2	77.2	68.0	65.0	78.9
+ KD (Vid-Voken) w/ ResNet	93.4	89.2	88.7	83.0	78.9	68.7	70.0	81.7
+ KD (Vid-Voken) w/ CLIP	94.1	**89.8**	89.0	83.9	79.2	68.6	71.6	82.3
+ KD (NST+CRD) w/ ResNet	94.2	89.3	89.7	84.0	79.0	**68.9**	71.8	82.4
+ KD (NST+CRD) w/ CLIP	**94.5**	89.6	**89.8**	**84.2**	**79.6**	68.7	**72.0**	**82.6**

- Performance gain is mostly from knowledge, physical interaction, & temporal reasoning

	GLUE diagnostics				PIQA	TRACIE
	Lexicon	Predicate	Logic	Knowledge		
$BERT_{6L/512H}$	53.0	64.2	44.5	44.0	56.9	63.4
+ KD-NST	53.3 (+0.3)	63.7 (-0.5)	44.8 (+0.3)	48.6 (**+4.6**)	60.0 (**+3.1**)	66.7 (**+3.3**)

PIQA: QA w/ physical interactions + commonsense reasoning
TRACIE: a temporal reasoning benchmark

VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer, Tang et al., NeurIPS 2021

Take-way messages

L → V: Implicit Knowledge

- Training vision model w/ frozen LM
  - Preserves the in-context learning ability of LM
  - Larger LM is better, the same as pure language tasks
  - They are quite general and are applicable to a wide range of tasks
- Convert multimodal task as text task for LM
  - All of above.
  - Computation efficient: no finetuning is required
  - Inherently modular, easy to update individual modules
  - May suffer info loss when during the conversion to text

L → V: Explicit Knowledge

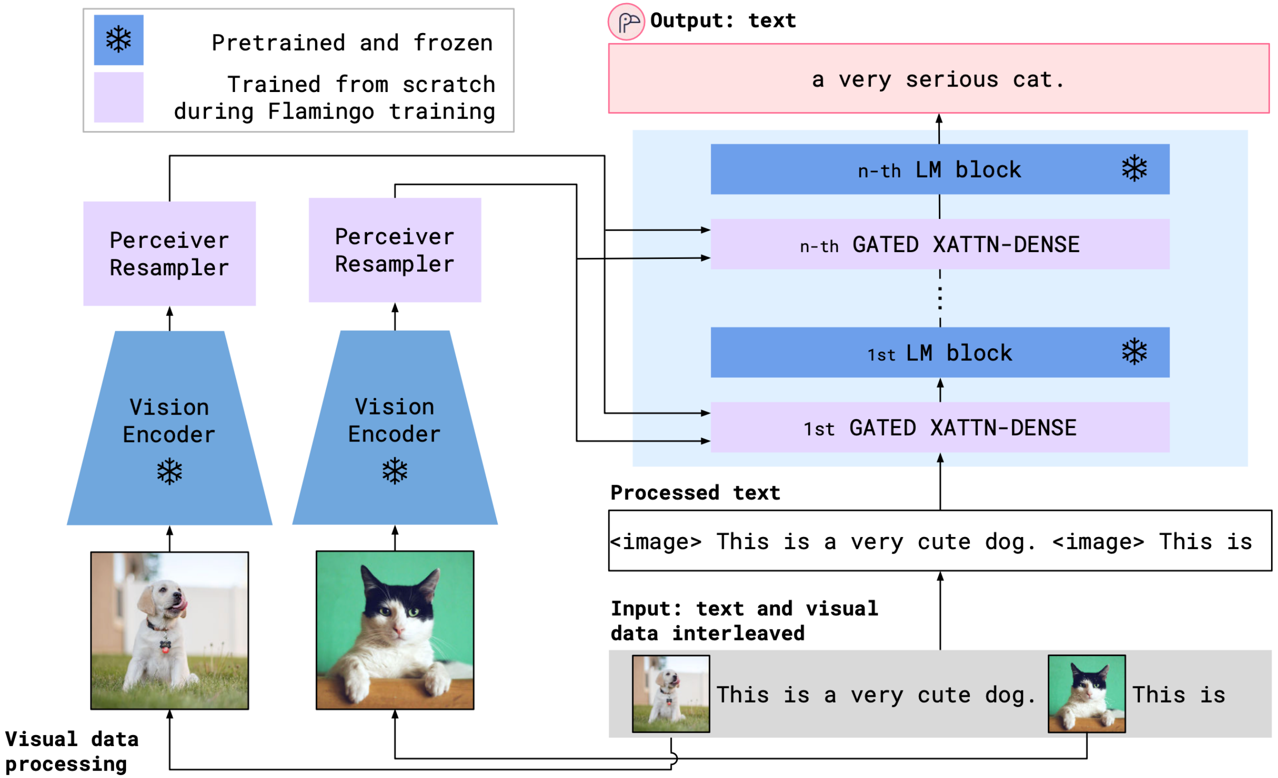
- Human curated (e.g., wordnet) or LLM (GPT-3) improves image classification and detection

V → L

- Vision knowledge via vokenization or distillation improves LMs, especially for physical and commonsense knowledge, and temporal reasoning.

Future Work

- Existing approaches using frozen LLM shows better performance, but they typically require full backpropagation through a LLM, which is very expensive.



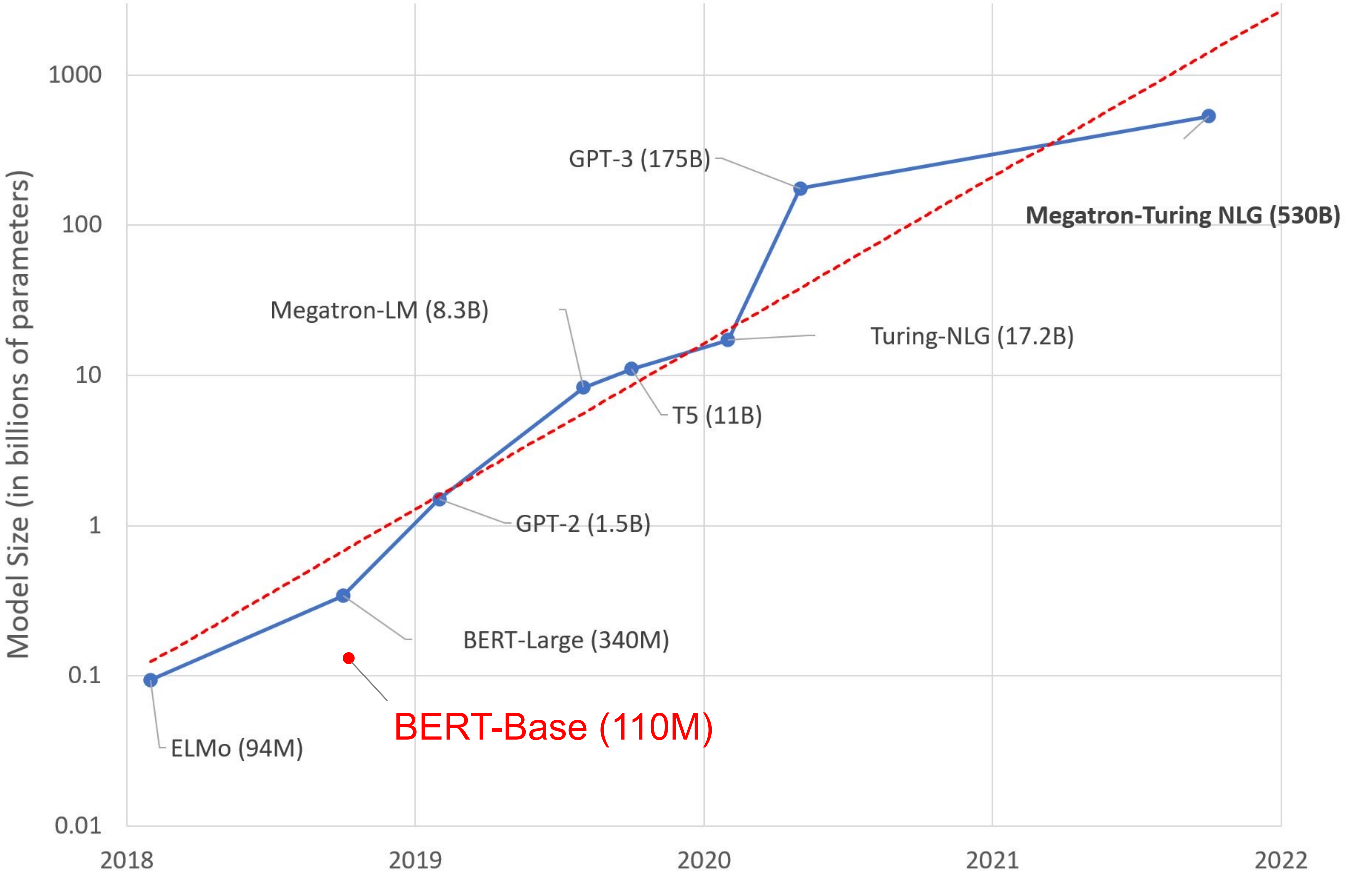
Flamingo, [Alayrac et al, arXiv 2022]

80B model: 1536 TPUv4 chips X 15 days.

- Full-backpropagation → Sparse backpropagation [Cheng et al, CVPR 2022]
- Deep fusion → Shallow fusion

Future Work

- Using vision (image or video) supervision has shown some early success.



- Bidirectional LM only, casual LM is not explored.
- Small model (up to 110M BERT-base), *vs.*, 175B GPT-3

- How about using other modalities (audio) as supervision?

Figure credit to Julien Simon, https://huggingface.co/blog/large-language-models

Thanks!