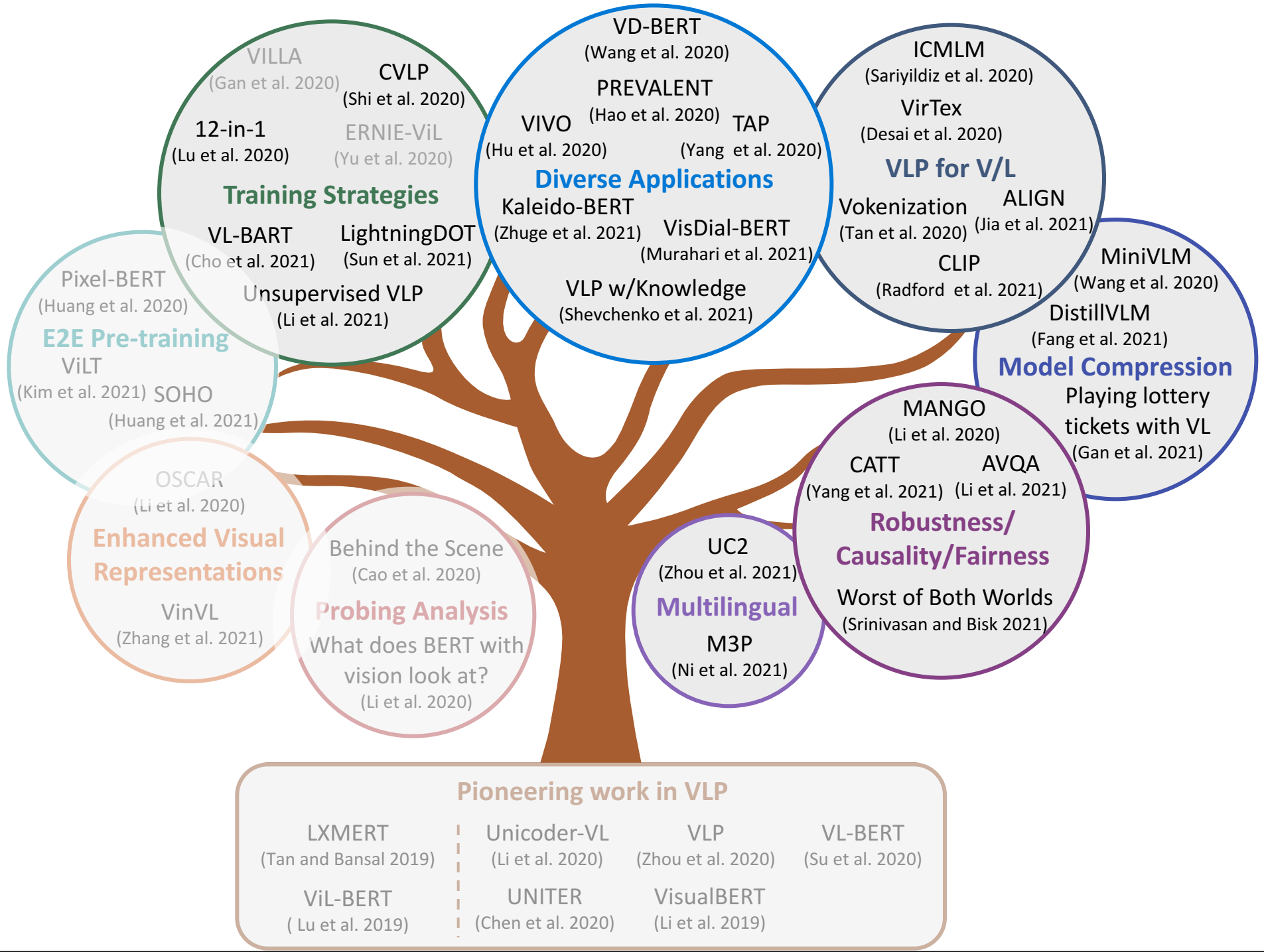
Vision-Language Pre-training Part II

Linjie Li

Researcher

Microsoft


VILLA
(Gan et al. 2020)
CVLP
(Shi et al. 2020)
12-in-1
(Lu et al. 2020)
ERNIE-ViL
(Yu et al. 2020)
Training Strategies
VL-BART
(Cho et al. 2021)
LightningDOT
(Sun et al. 2021)
Unsupervised VLP
(Li et al. 2021)
VD-BERT
(Wang et al. 2020)
PREVALENT
(Hao et al. 2020)
VIVO
(Hu et al. 2020)
TAP
(Yang et al. 2020)
Diverse Applications
Kaleido-BERT
(Zhuge et al. 2021)
VisDial-BERT
(Murahari et al. 2021)
VLP w/Knowledge
(Shevchenko et al. 2021)
ICMLM
(Sariyildiz et al. 2020)
VirTex
(Desai et al. 2020)
VLP for V/L
Vokenization
(Tan et al. 2020)
ALIGN
(Jia et al. 2021)
CLIP
(Radford et al. 2021)
MiniVLM
(Wang et al. 2020)
DistillVLM
(Fang et al. 2021)
Model Compression
Playing lottery tickets with VL
(Gan et al. 2021)
Pixel-BERT
(Huang et al. 2020)
E2E Pre-training
ViLT
(Kim et al. 2021)
SOHO
(Huang et al. 2021)
OSCAR
(Li et al. 2020)
Enhanced Visual Representations
VinVL
(Zhang et al. 2021)
Behind the Scene
(Cao et al. 2020)
Probing Analysis
What does BERT with vision look at?
(Li et al. 2020)
UC2
(Zhou et al. 2021)
Multilingual
M3P
(Ni et al. 2021)
MANGO
(Li et al. 2020)
CATT
(Yang et al. 2021)
AVQA
(Li et al. 2021)
Robustness/ Causality/Fairness
Worst of Both Worlds
(Srinivasan and Bisk 2021)
Pioneering work in VLP
LXMERT
(Tan and Bansal 2019)
Unicoder-VL
(Li et al. 2020)
VLP
(Zhou et al. 2020)
VL-BERT
(Su et al. 2020)
ViL-BERT
(Lu et al. 2019)
UNITER
(Chen et al. 2020)
VisualBERT
(Li et al. 2019)

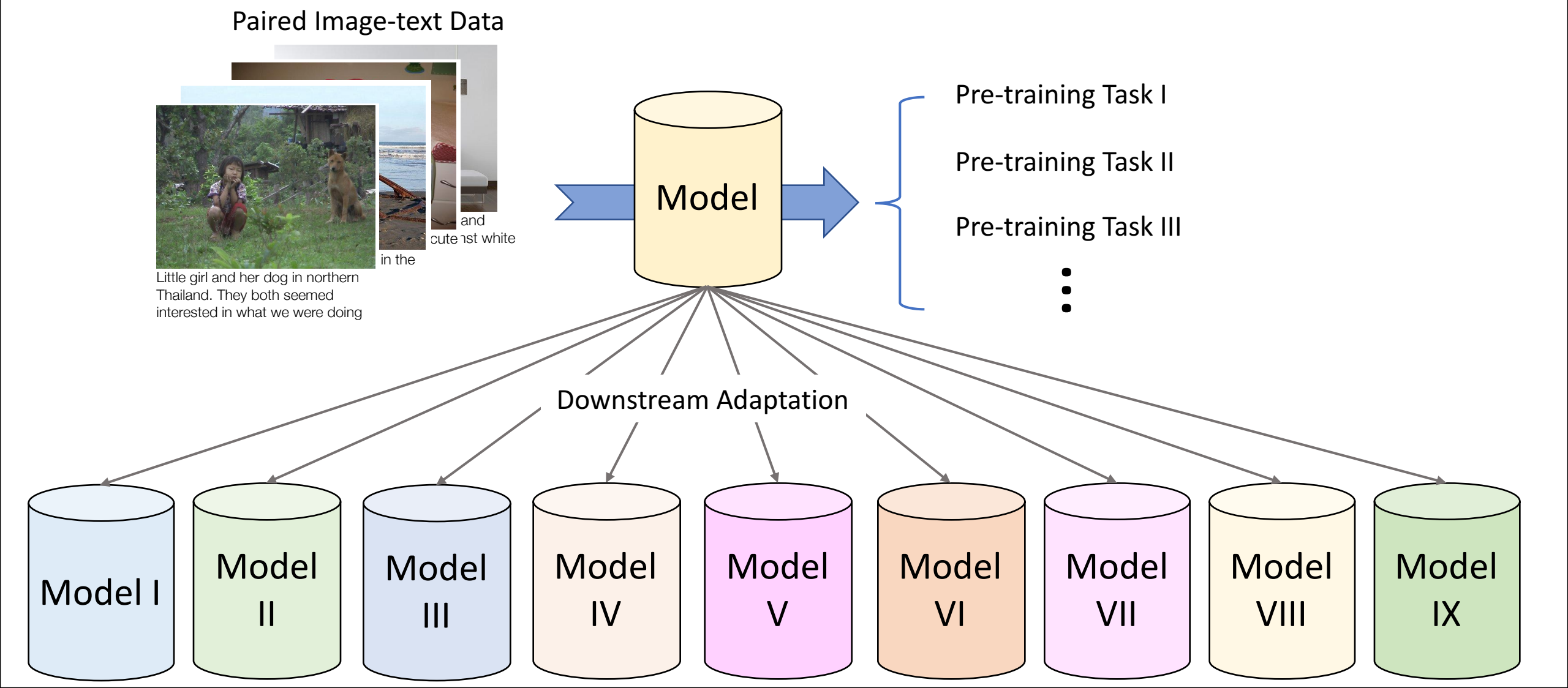
Agenda

- Advanced training strategies in VLP
- Diverse applications of VLP
- VL for V/L
- Compressing VLP models
- Robustness/causality/fairness of VLP models
- Multilingual VLP

Agenda

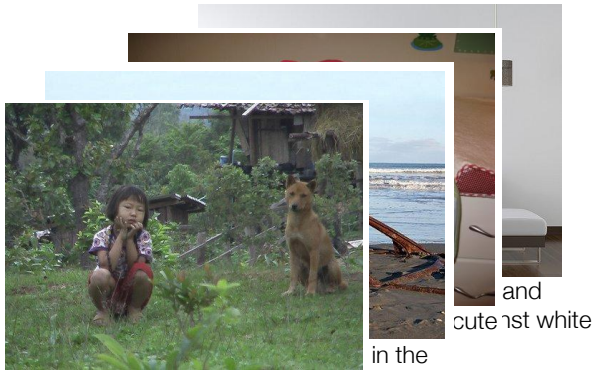
- Advanced training strategies in VLP
- Diverse applications of VLP
- VL for V/L
- Compressing VLP models
- Robustness/causality/fairness of VLP models
- Multilingual VLP

Training Strategies in VLP



Training Strategies in VLP

Paired Image-text Data



Q: Can we leverage un-paired data that widely exist on the web?

Model

Pre-training Task I : MLM

Pre-training Task II: MRM

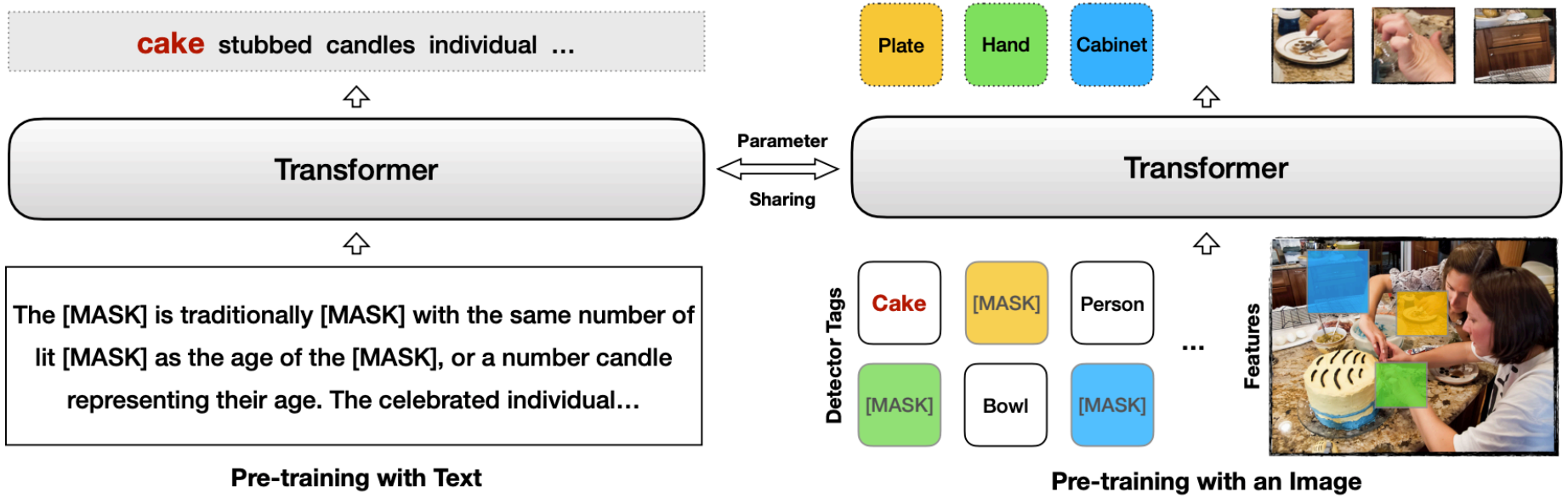
Pre-training Task III: ITM

Downstream Adaptation

Model I | Model II | Model III | Model IV | Model V | Model VI | Model VII | Model VIII | Model IX

Unsupervised VLP

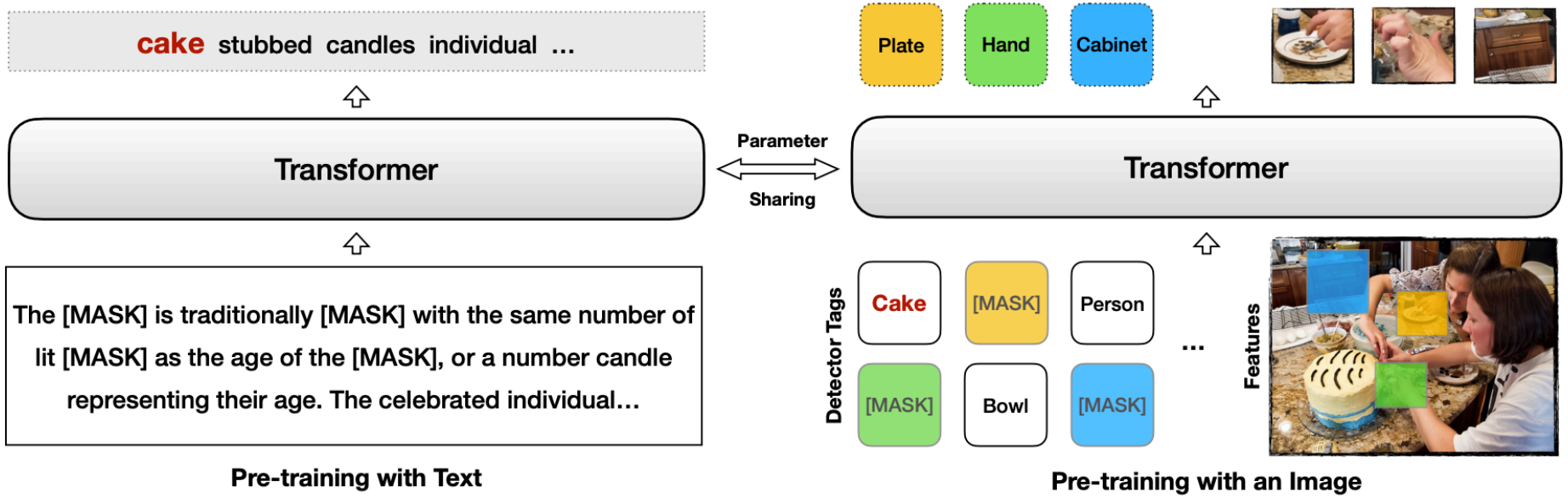
- Pre-training without paired image-text data



Unsupervised VLP (Li et al. 2021)

Unsupervised VLP

- Pre-training without paired image-text data

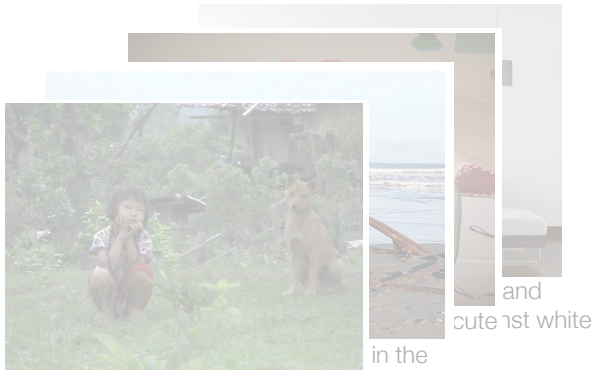


Model	#Image+Text	VQA	NLVR	Flickr30K	RefCOCO+
Supervised with paired image-text	5.5M	**70.87**	**73.69**	**79.80**	**72.54**
Unsupervised with unpaired image-text	5.5M	70.74 (-0.13)	71.38 (-2.31)	76.05 (-3.75)	71.91 (-0.63)

Unsupervised VLP (Li et al. 2021)

Training Strategies in VLP

Paired Image-text Data



Model

Q: Can we design better pre-training tasks?

Pre-training Task I : MLM

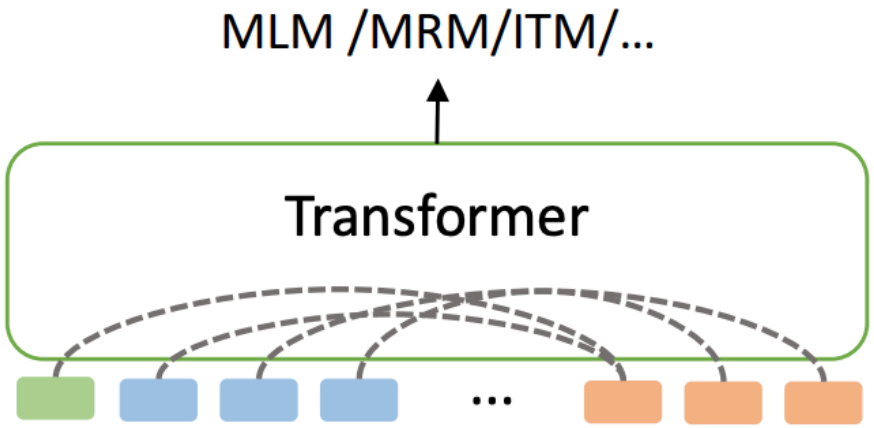
Pre-training Task II: MRM

Pre-training Task III: ITM

Downstream Adaptation

Model I | Model II | Model III | Model IV | Model V | Model VI | Model VII | Model VIII | Model IX

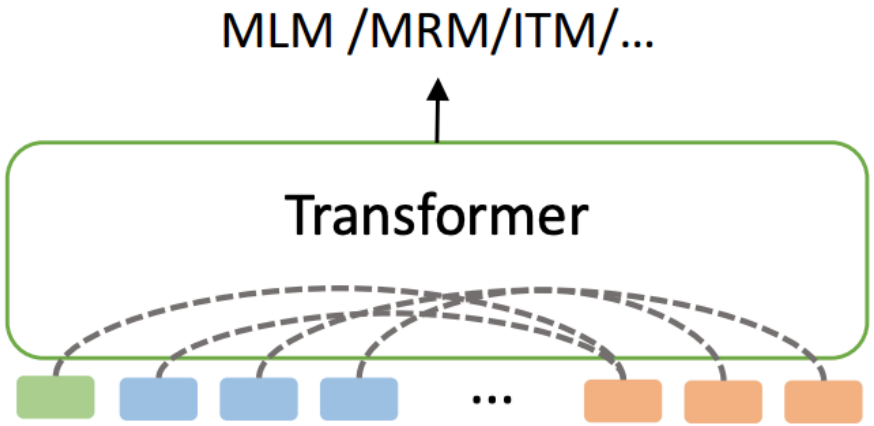
Contrastive Vision-Language Pre-training



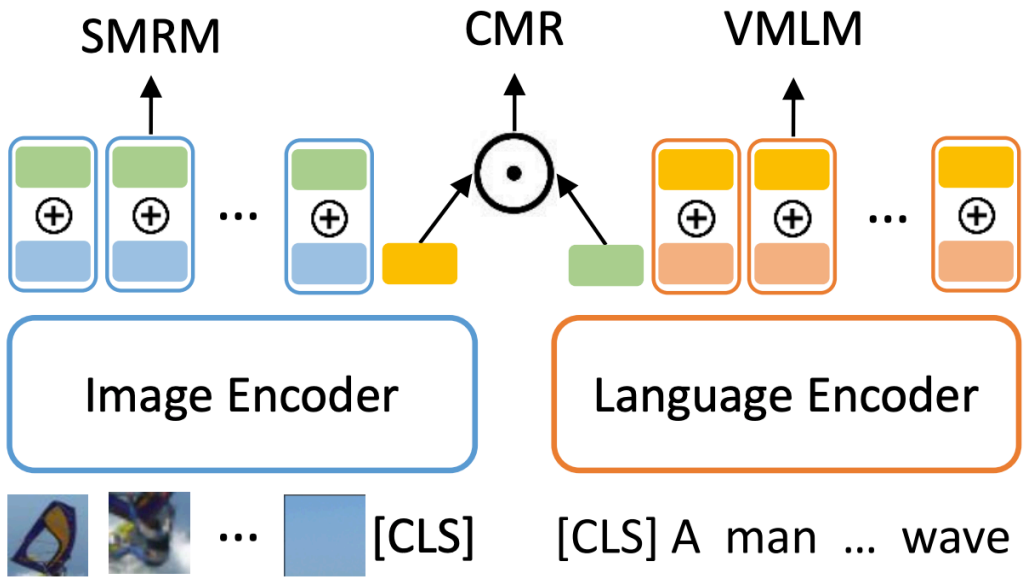
UNITER (Chen et al. 2020)

⊙ Dot Product ---- Cross Attention [CLS] Features Word Features Region Features

Contrastive Vision-Language Pre-training



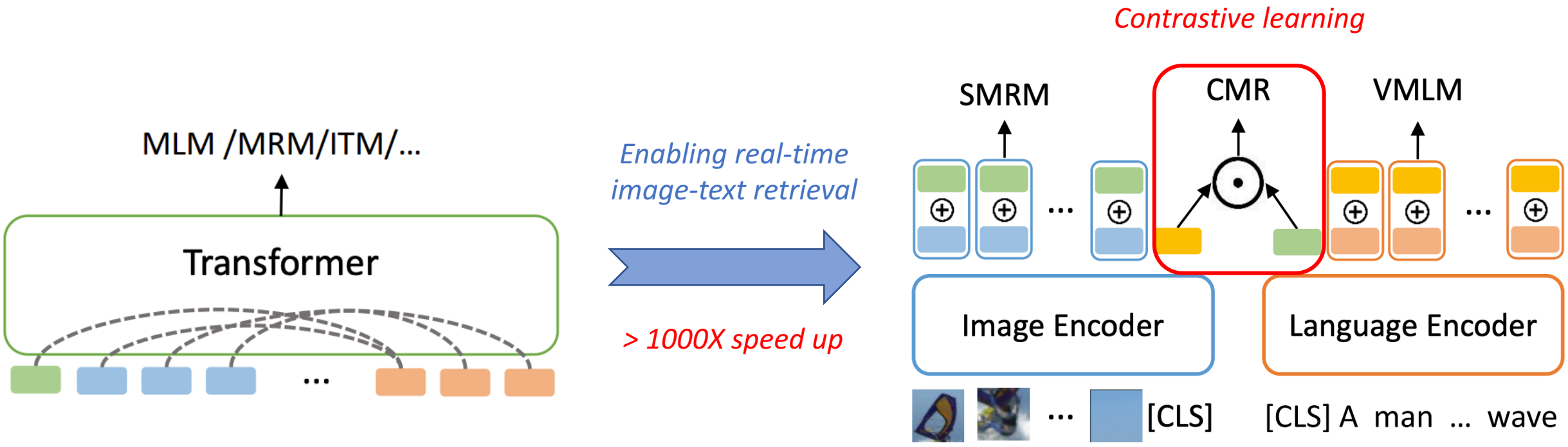
UNITER (Chen et al. 2020)



LighteningDOT (Sun et al. 2021)

Dot Product Cross Attention [CLS] Features Word Features Region Features

Contrastive Vision-Language Pre-training

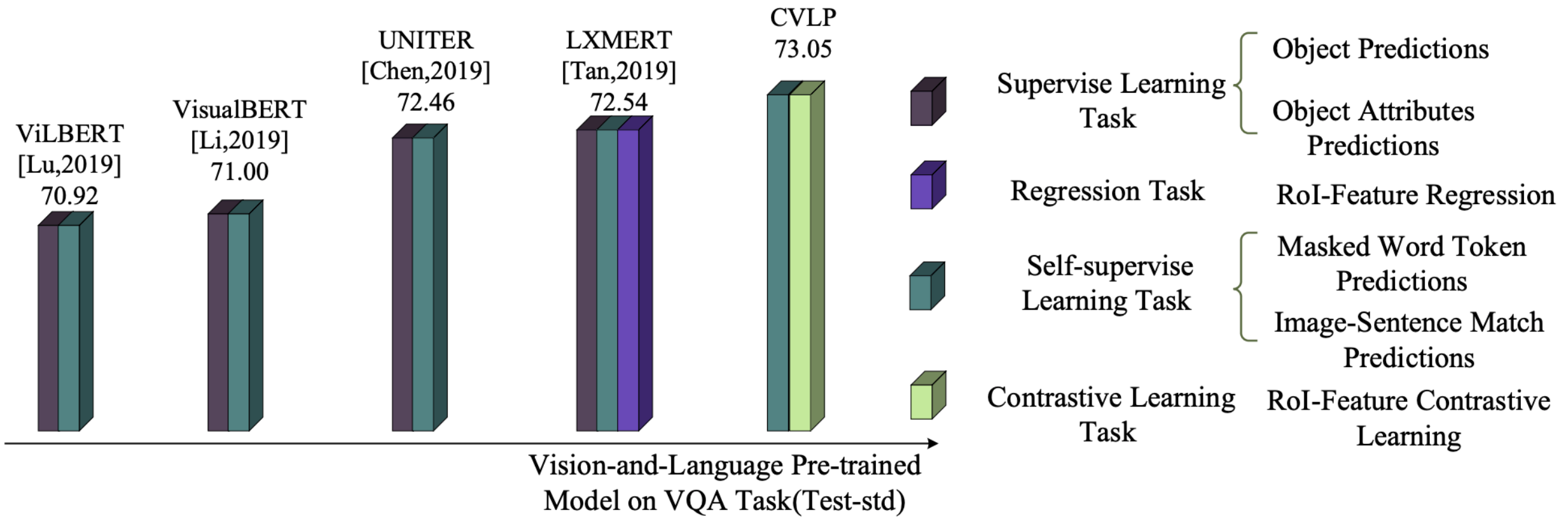


UNITER (Chen et al. 2020)

LighteningDOT (Sun et al. 2021)

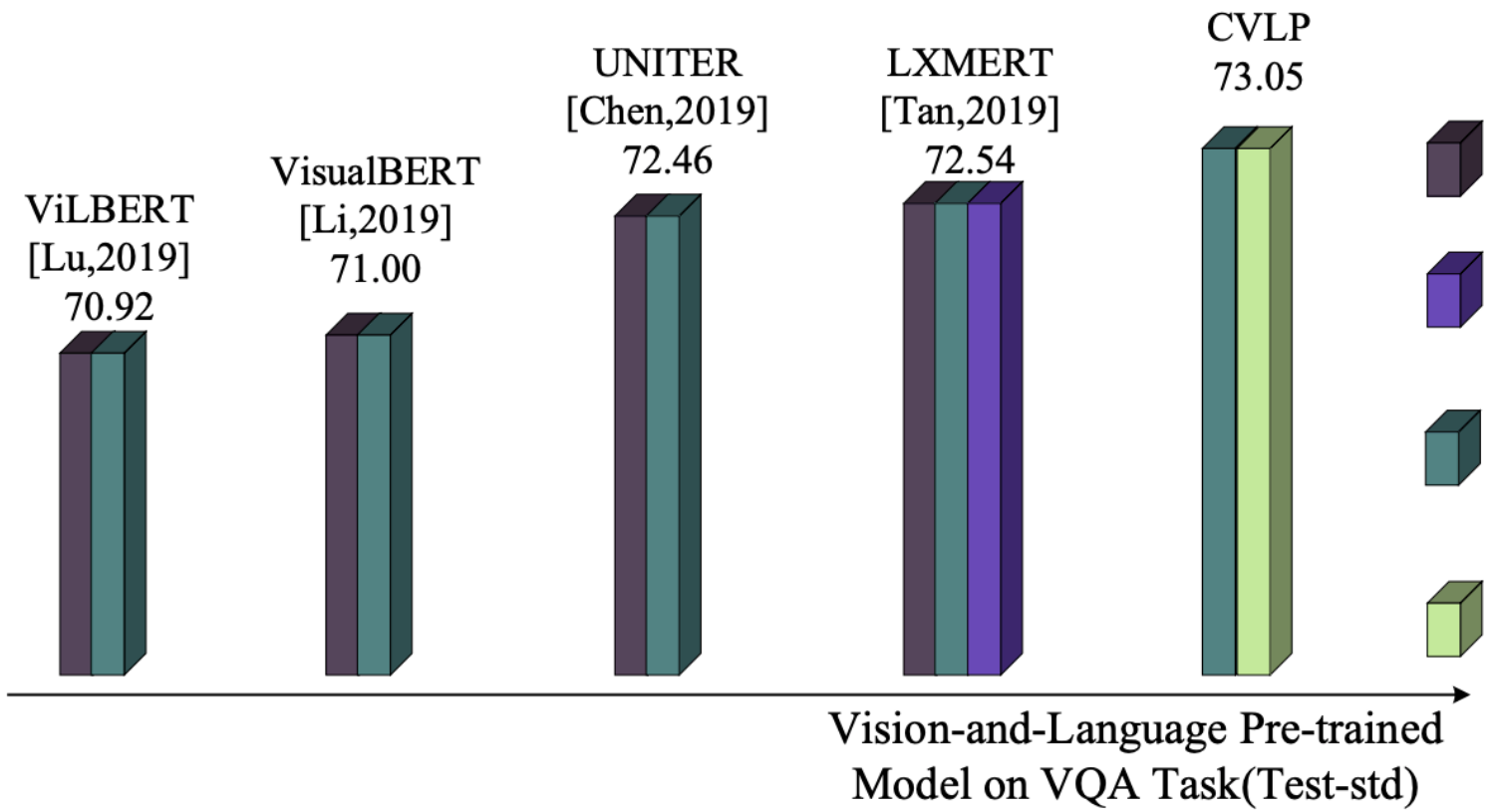
⊙ Dot Product ---- Cross Attention [CLS] Features Word Features Region Features

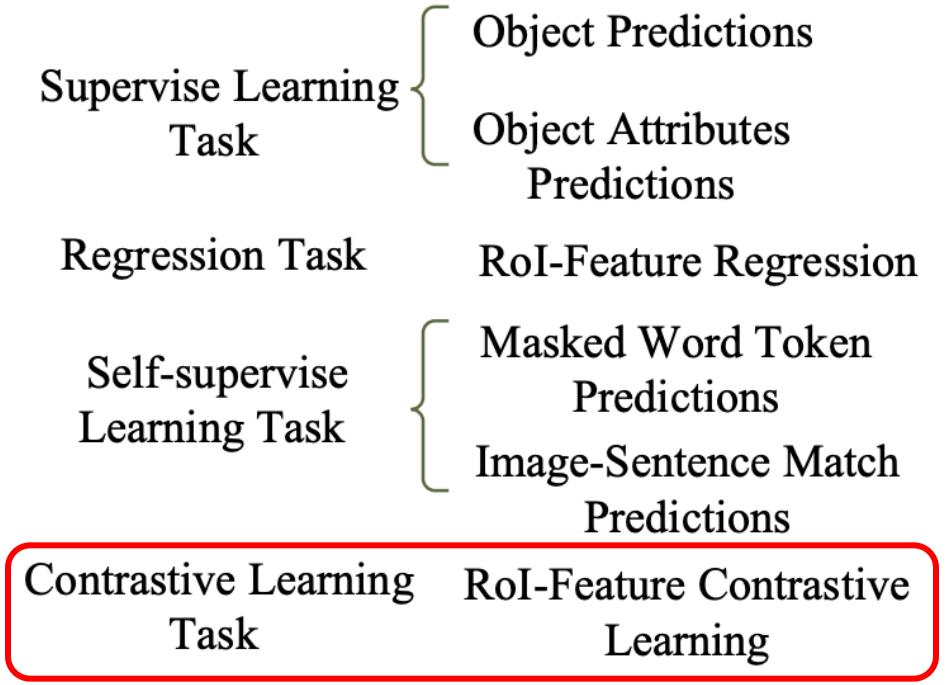
Contrastive Vision-Language Pre-training



CVLP (Shi et al. 2020)

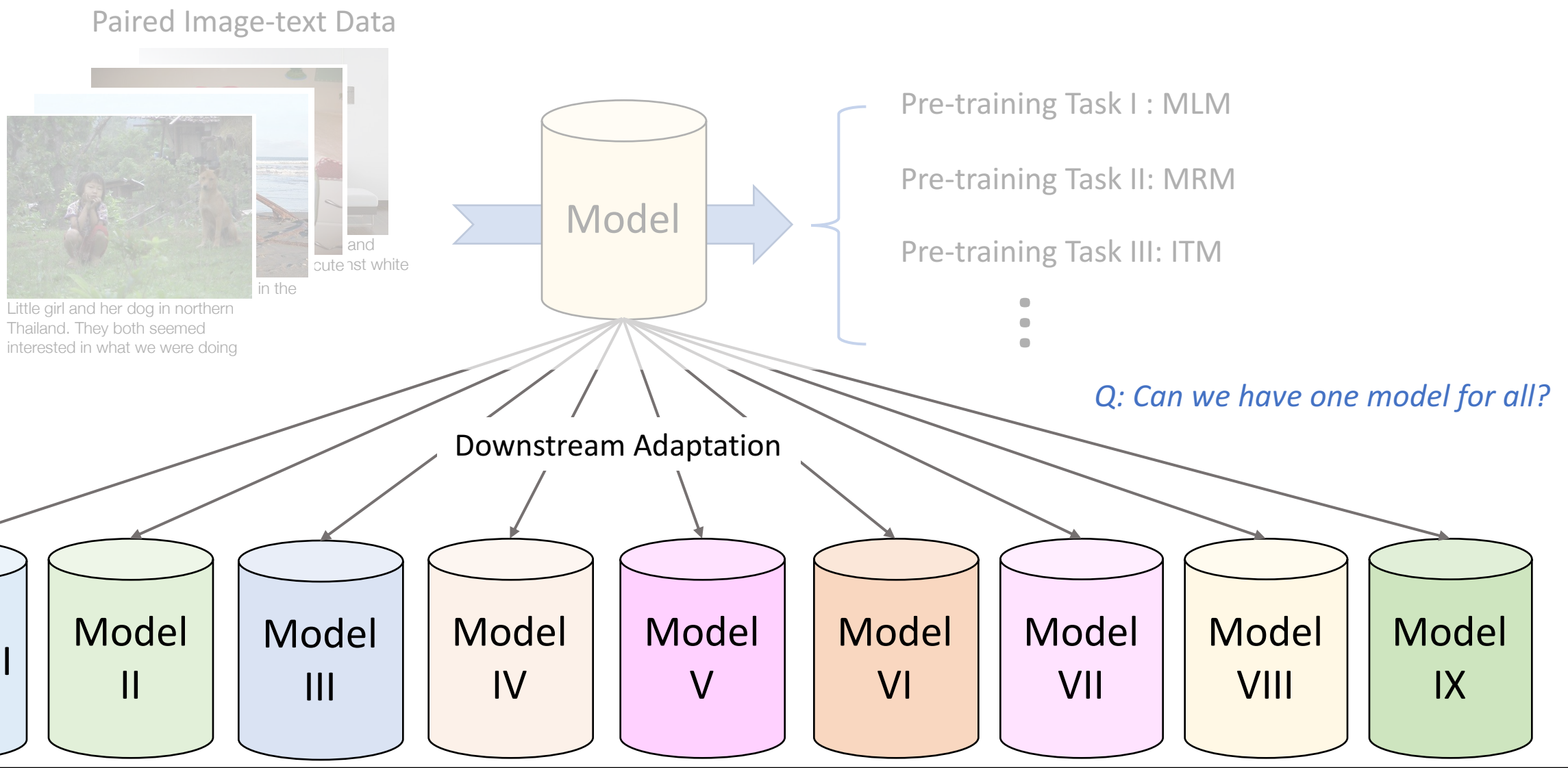
Contrastive Vision-Language Pre-training



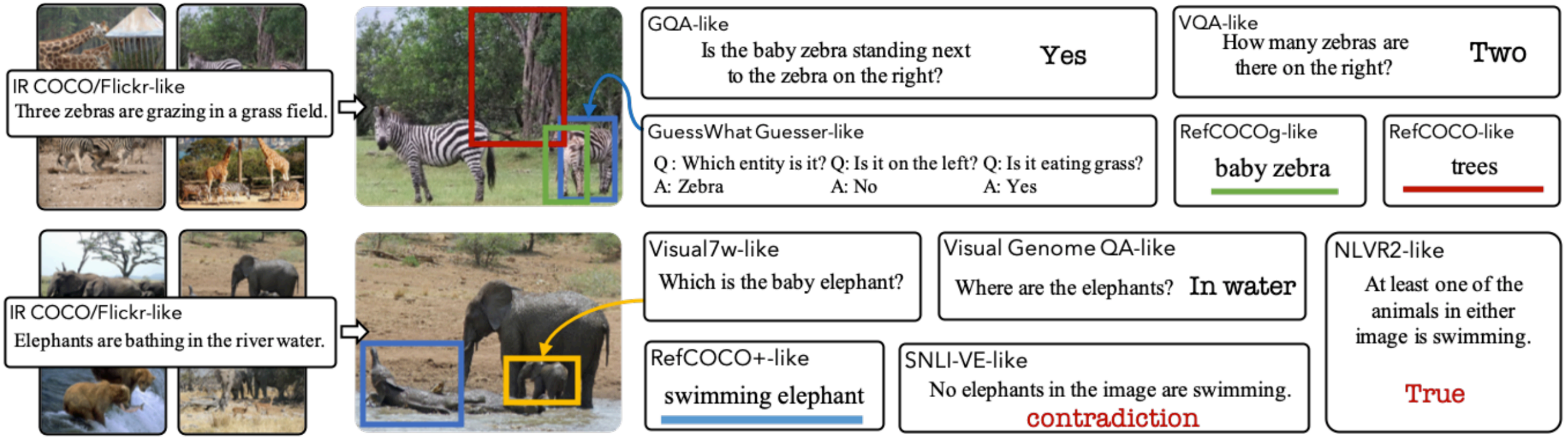


CVLP (Shi et al. 2020)

Training Strategies in VLP



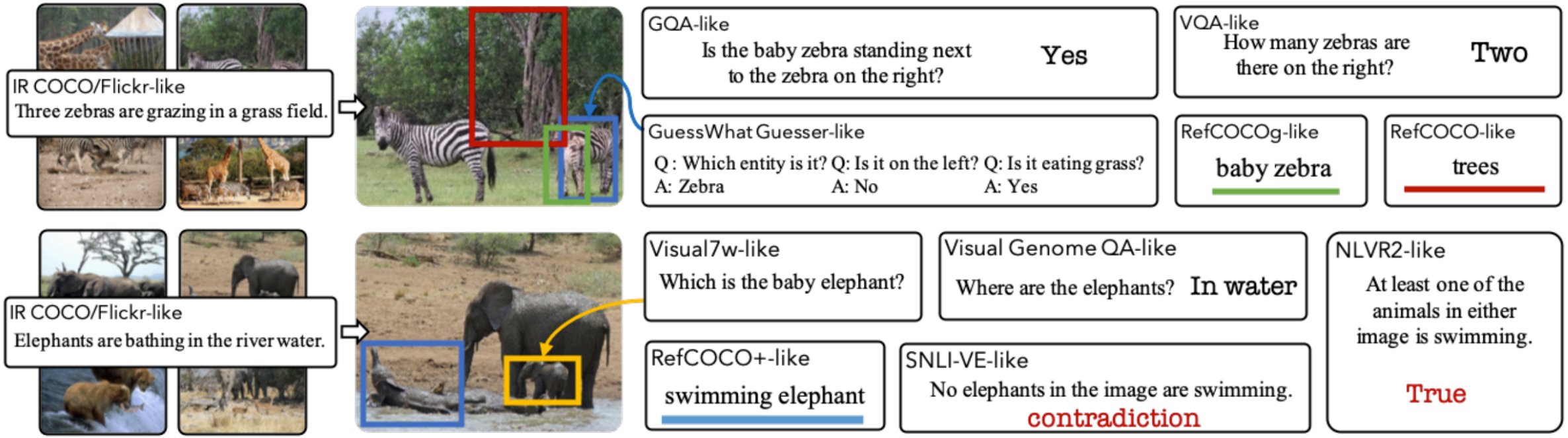
Multi-task Training



All these task require visually-grounded language understanding skills.

12-in-1 (Lu et al. 2020)

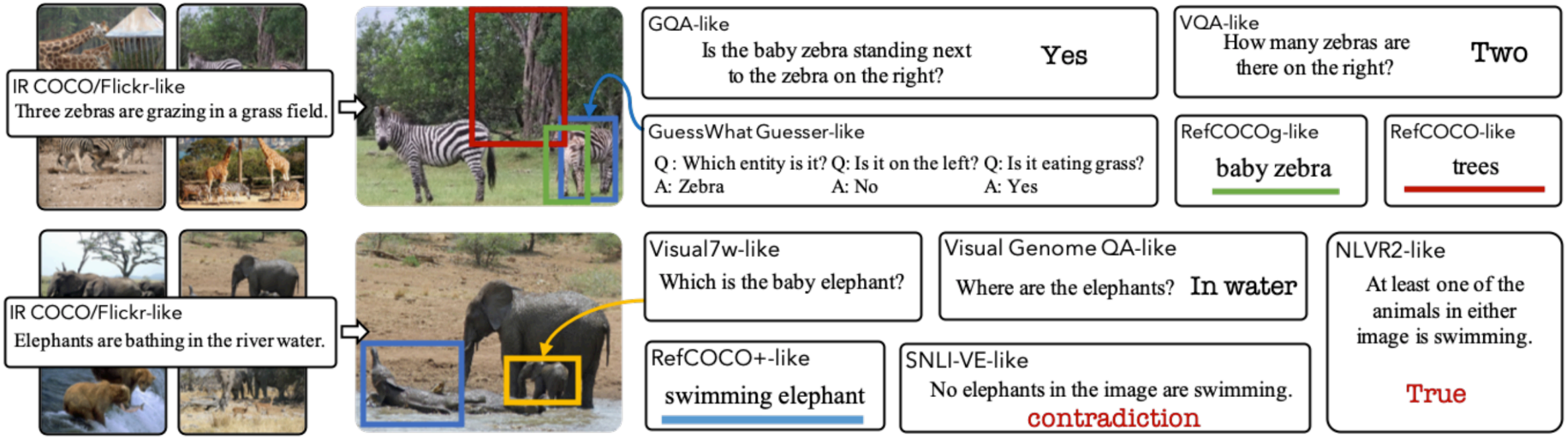
Multi-task Training



Model	# models	# parameters
Independently train for each task	12	12X270M = 3B

12-in-1 (Lu et al. 2020)

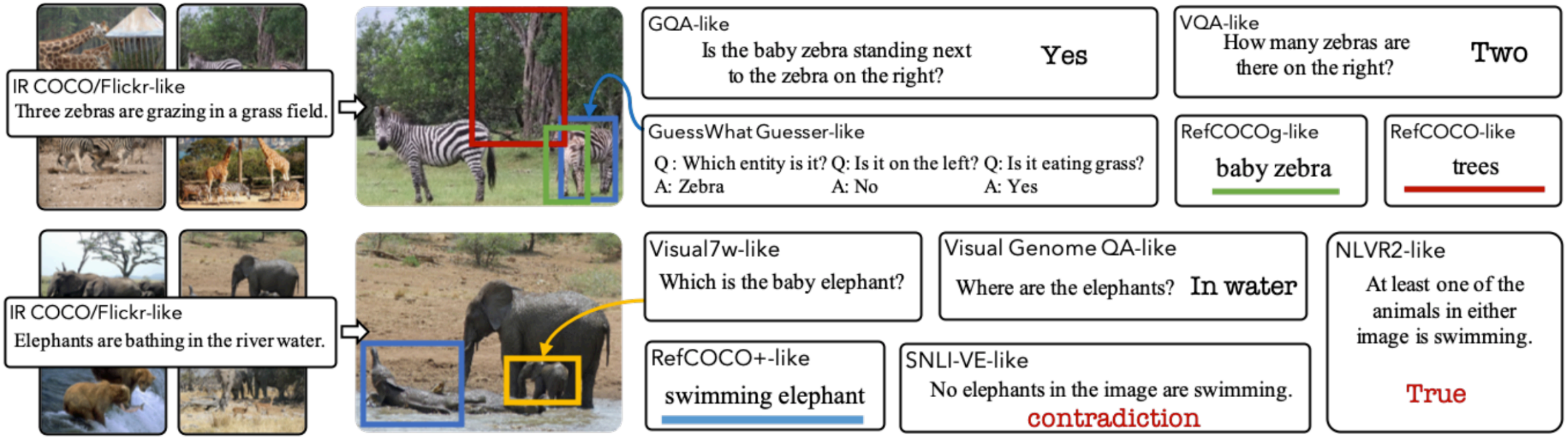
Multi-task training



Model	# models	# parameters
Independently train for each task	12	12X270M = 3B
Train all tasks together (12-in-1)	**1**	**270M**

12-in-1 (Lu et al. 2020)

Multi-task training

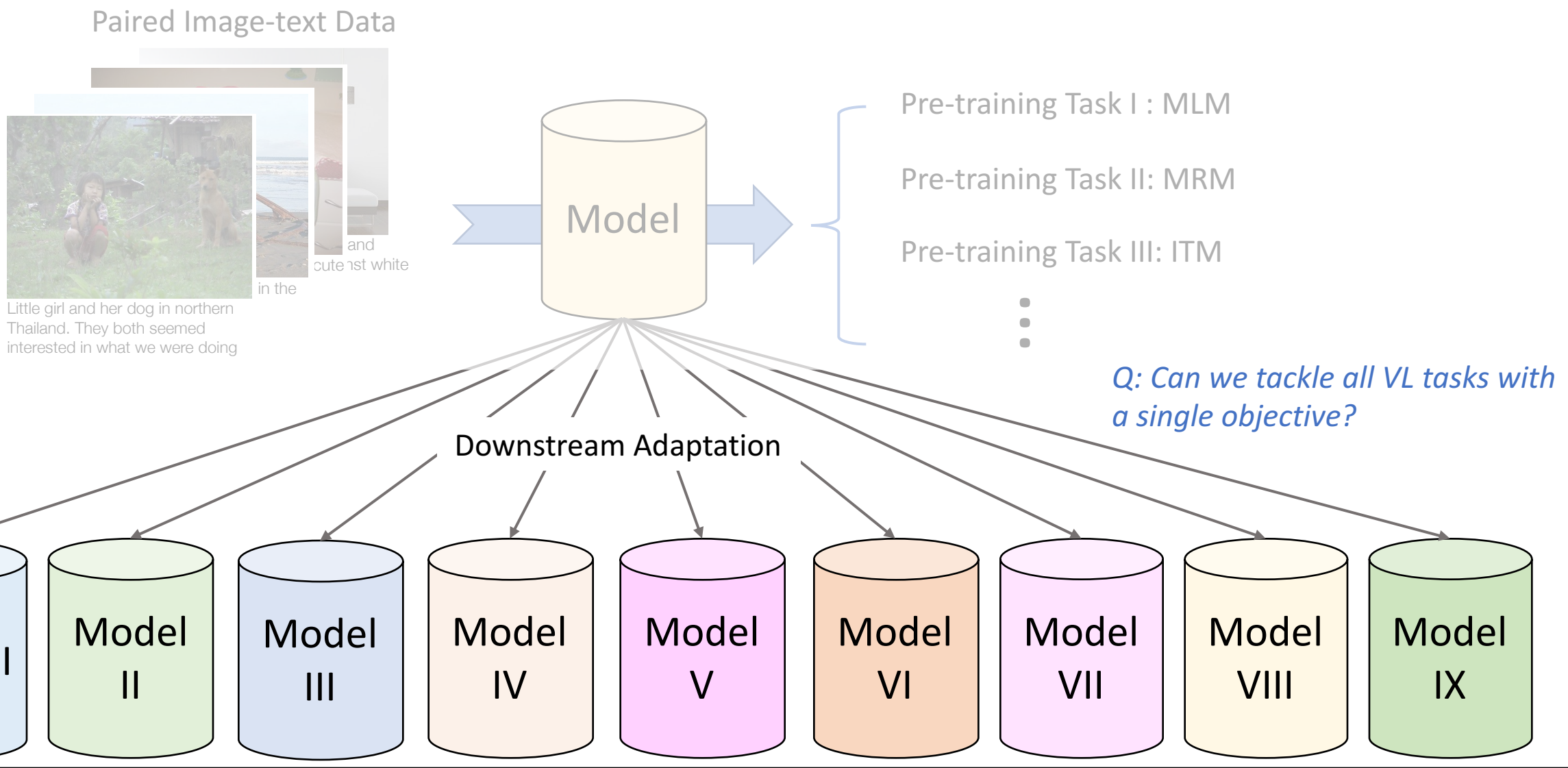


Model	# models	# parameters	Average Performance
Independently train for each task	12	12X270M = 3B	67.24
Train all tasks together (12-in-1)	**1**	**270M**	**69.08** (+1.84)

All these task share similar visually-grounded language understanding skills.

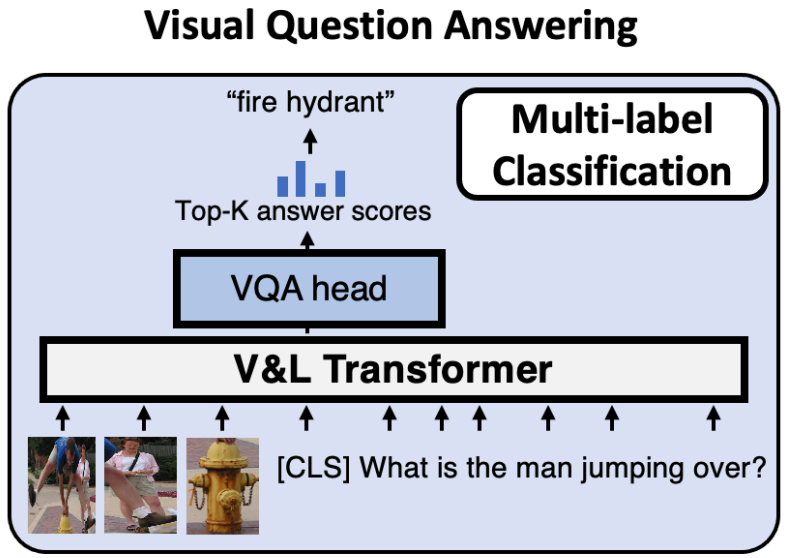
12-in-1 (Lu et al. 2020)

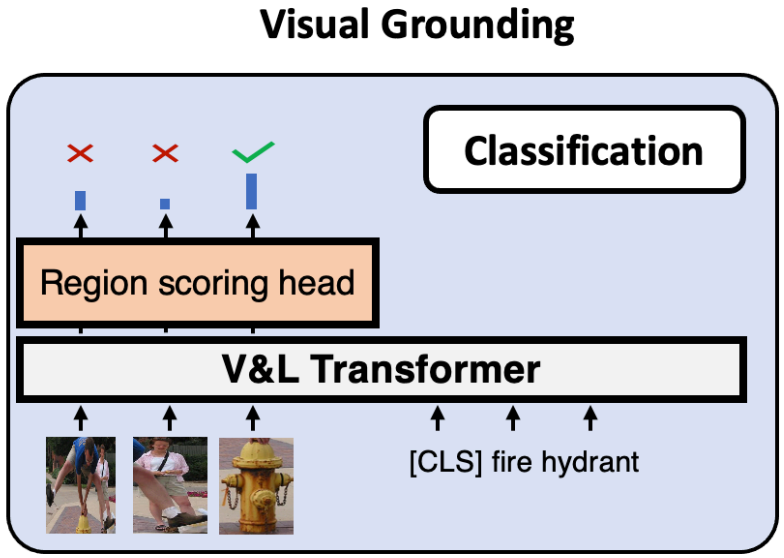
Training Strategies in VLP



Unifying VL Tasks via Text Generation

UNITER
(Chen et al. 2020)



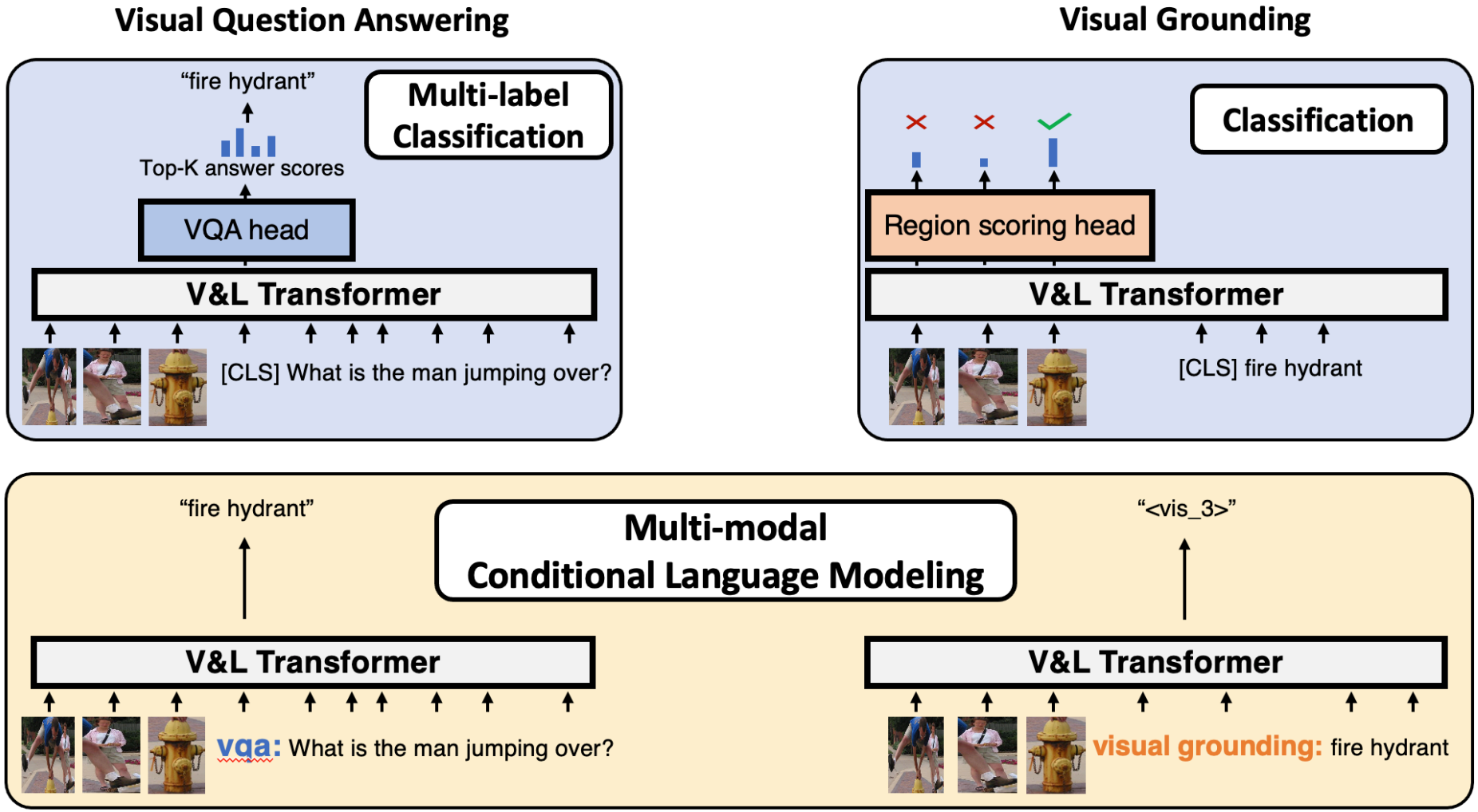


Credit to: Jaemin Cho

Unifying VL Tasks via Text Generation

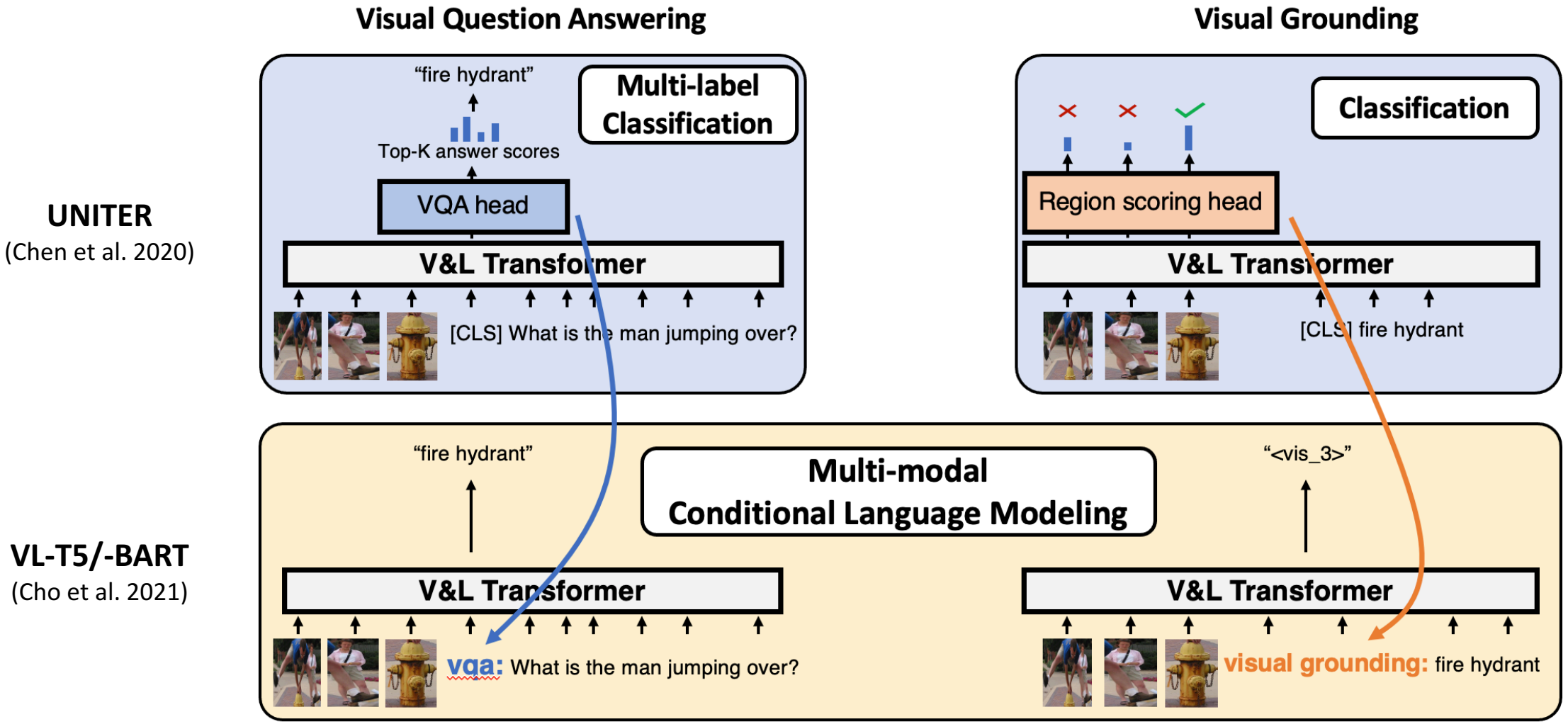
UNITER
(Chen et al. 2020)

VL-T5/-BART
(Cho et al. 2021)



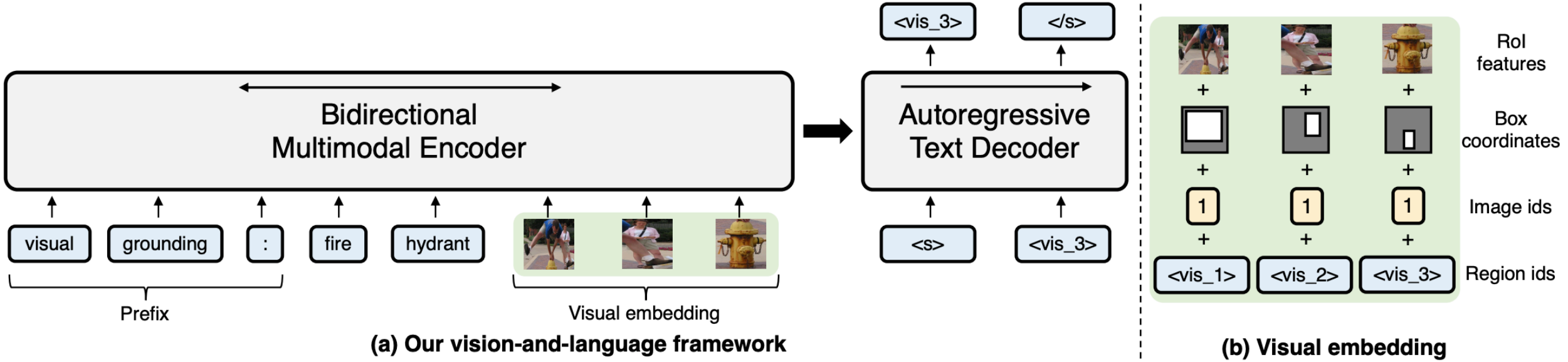
Credit to: Jaemin Cho

Unifying VL Tasks via Text Generation



Credit to: Jaemin Cho

Unifying VL Tasks via Text Generation

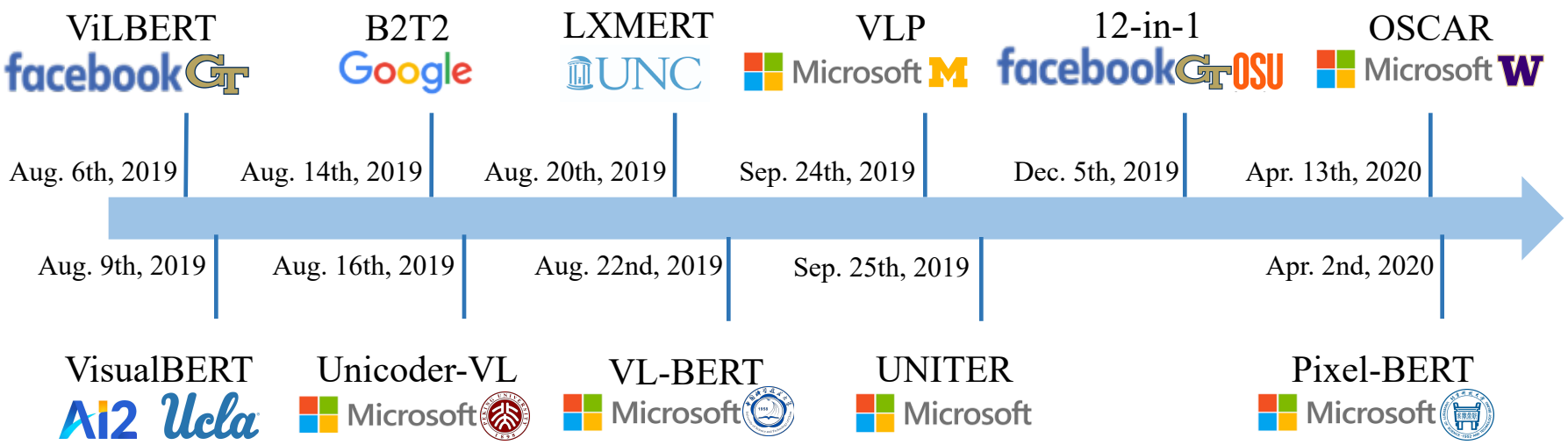


VL-T5/-BART (Cho et al. 2021)

Agenda

- Advanced training strategies in VLP
- Diverse applications of VLP
- VL for V/L
- Compressing VLP models
- Robustness/causality/fairness of VLP models
- Multilingual VLP

Great success of VLP models

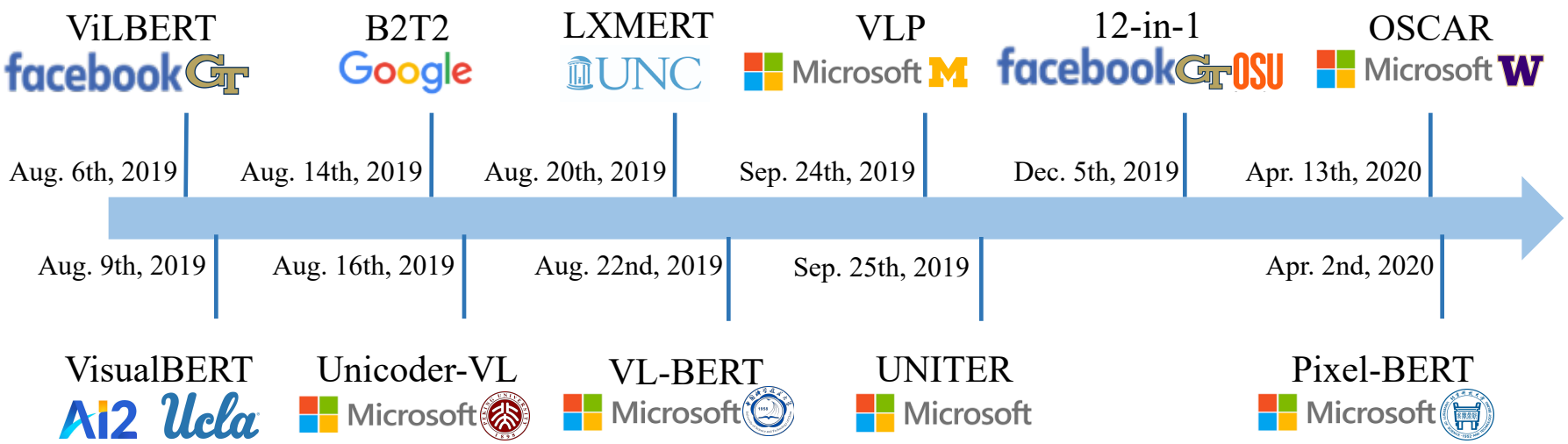


V+L Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

Can we apply VLP to other VL tasks?

Great success of VLP models



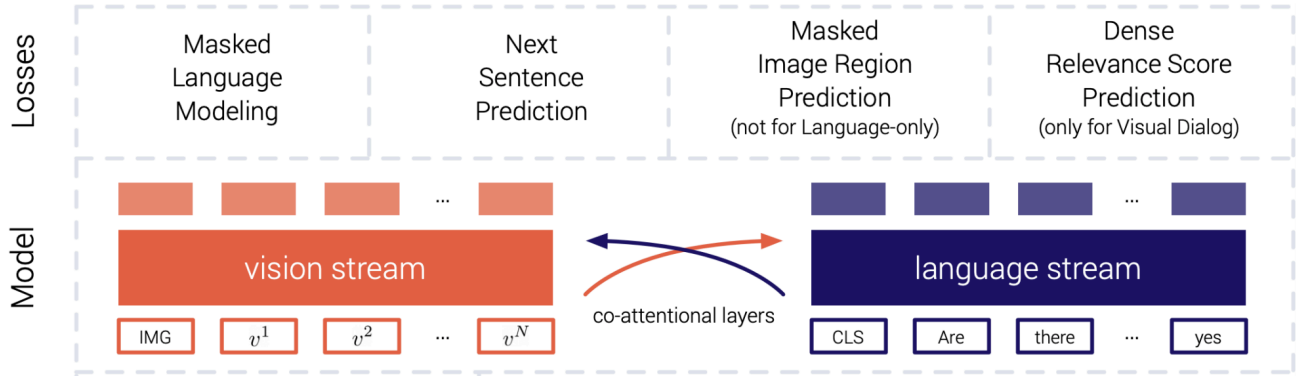
V+L Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

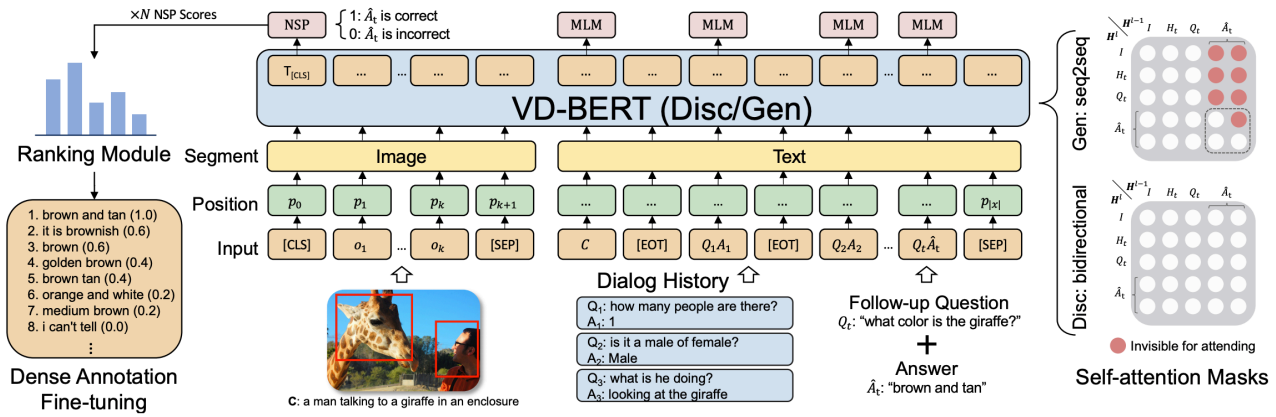
Can we apply VLP to other VL tasks? **YES!**

Diverse Applications of VLP

Visual Dialog (Murahari et al. 2021)



VisDial-BERT (Wang et al. 2020)

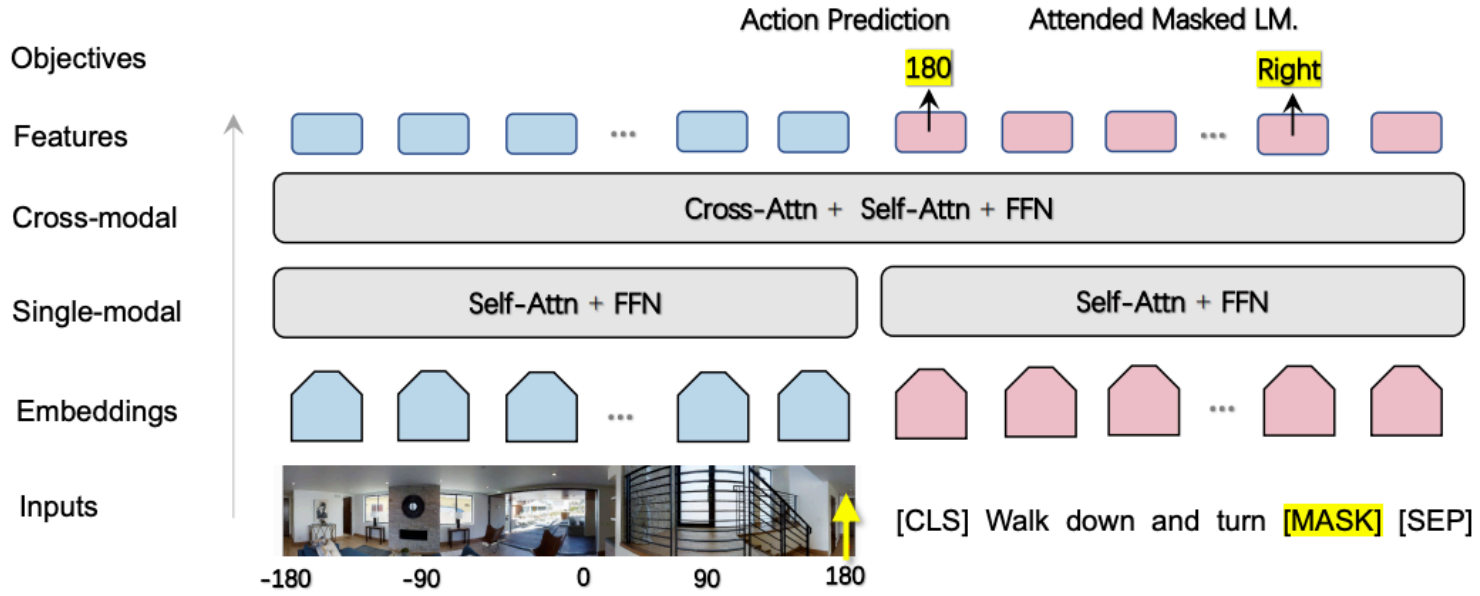


VD-BERT (Wang et al. 2020)

Diverse Applications of VLP

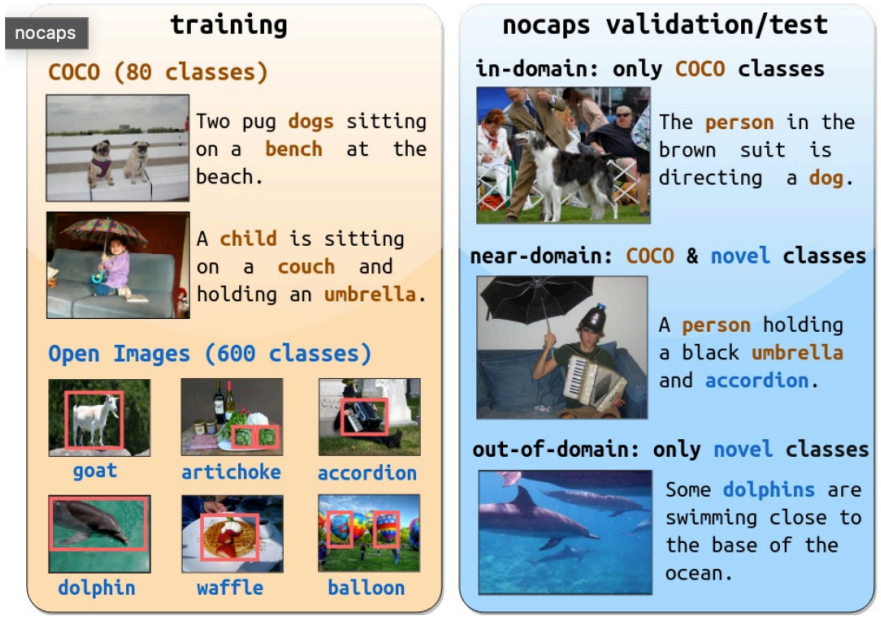
Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

Vision Language Navigation
(Anderson et al. 2017)

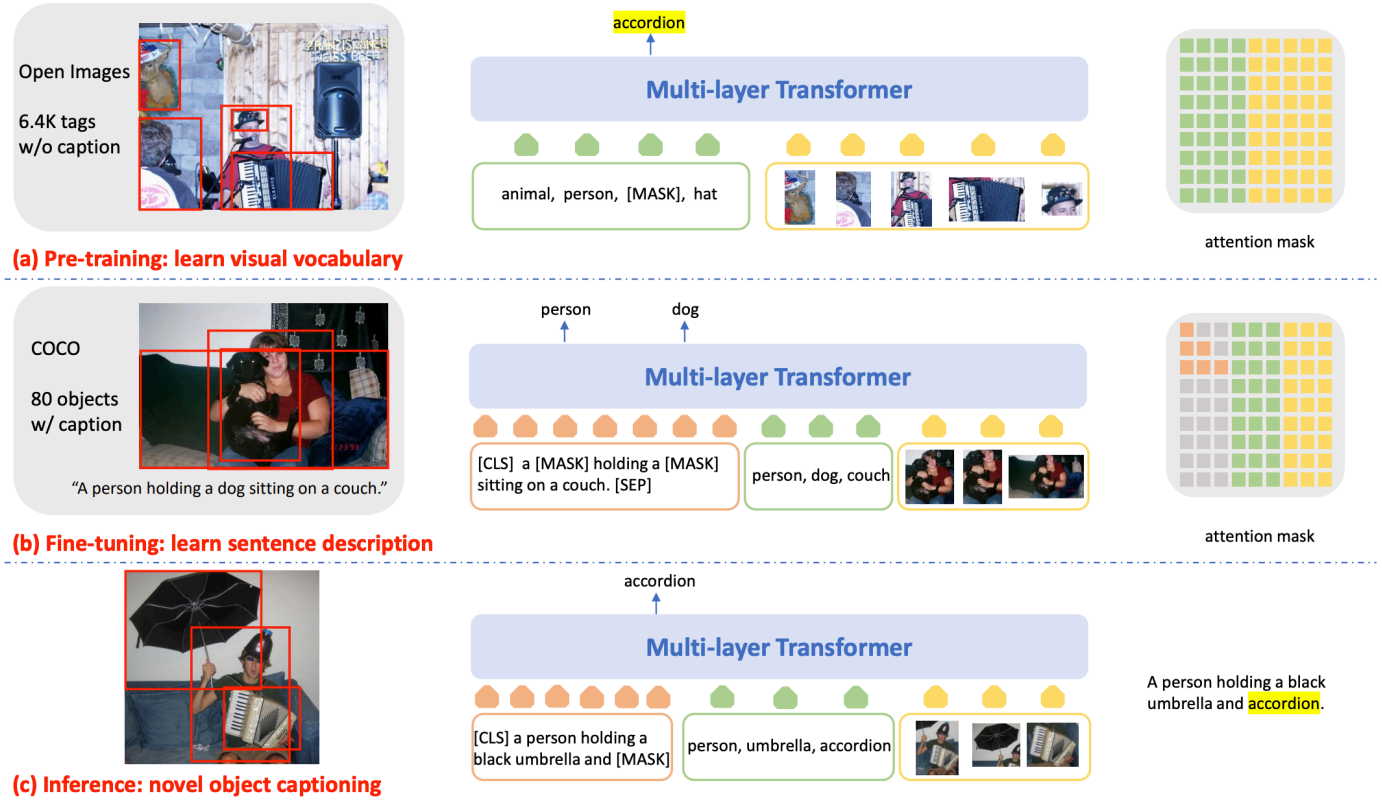


PREVALENT (Hao et al. 2020)

Diverse Applications of VLP



Novel Object Captioning
(Agrawal et al. 2019)



VIVO (Hu et al. 2020)

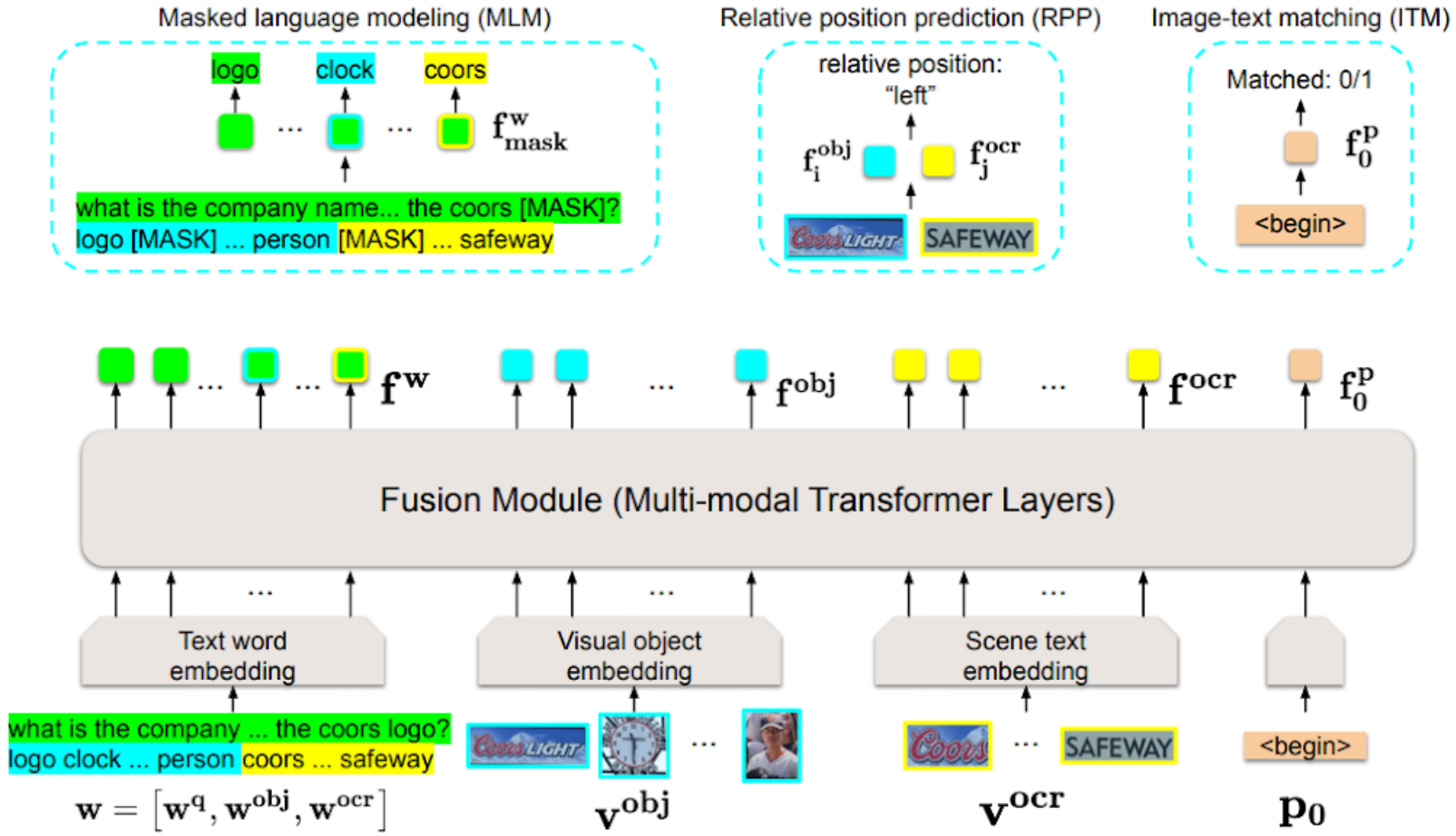
Diverse Applications of VLP

Question: what **number** is on the bike on the right? ---- A: the number is **317**

Text-VQA (Singh et al. 2019)

A group of motorcyclists with **number 317, 44, 30, 338, 598** racing outdoor.

Text-Captioning
(Sidorov et al. 2020)



TAP (Yang et al. 2021)

Diverse Applications of VLP

Q: Which American president is associated with the stuffed animal seen here?

A: Teddy Roosevelt

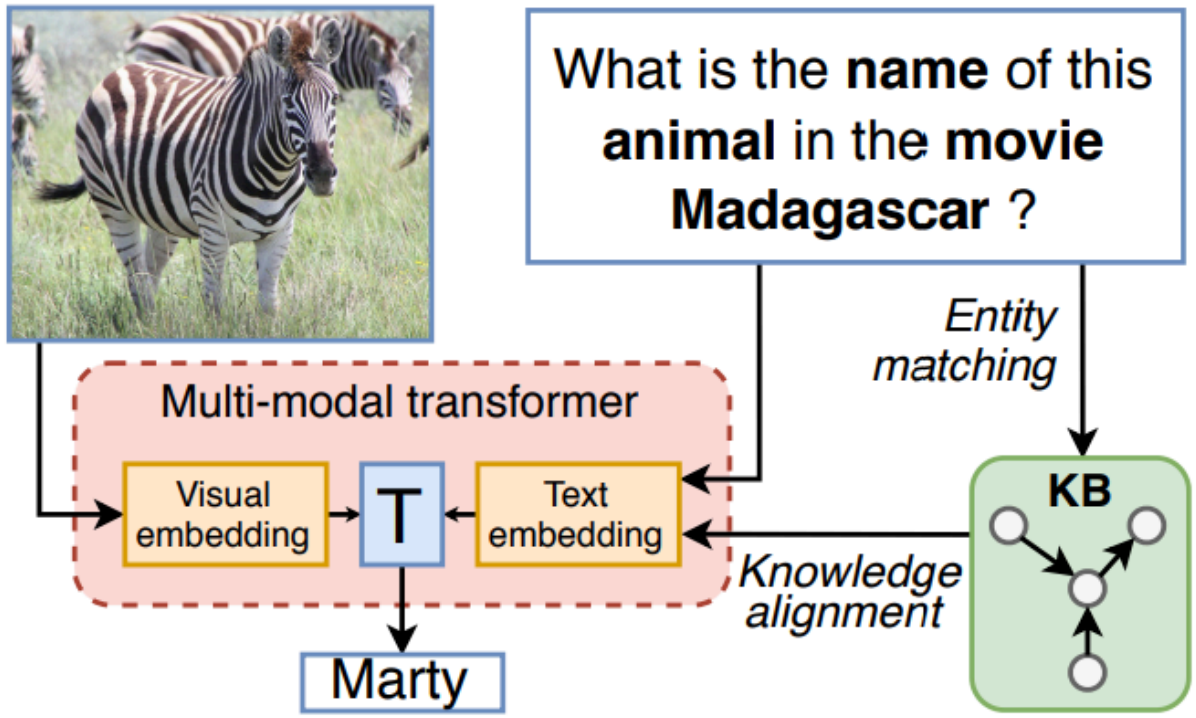
Outside Knowledge

Another lasting, popular legacy of Roosevelt is the stuffed toy bears—teddy bears—named after him following an incident on a hunting trip in Mississippi in 1902.

Developed apparently simultaneously by toymakers ... and named after President Theodore "Teddy" Roosevelt, the teddy bear became an iconic children's toy, celebrated in story, song, and film.

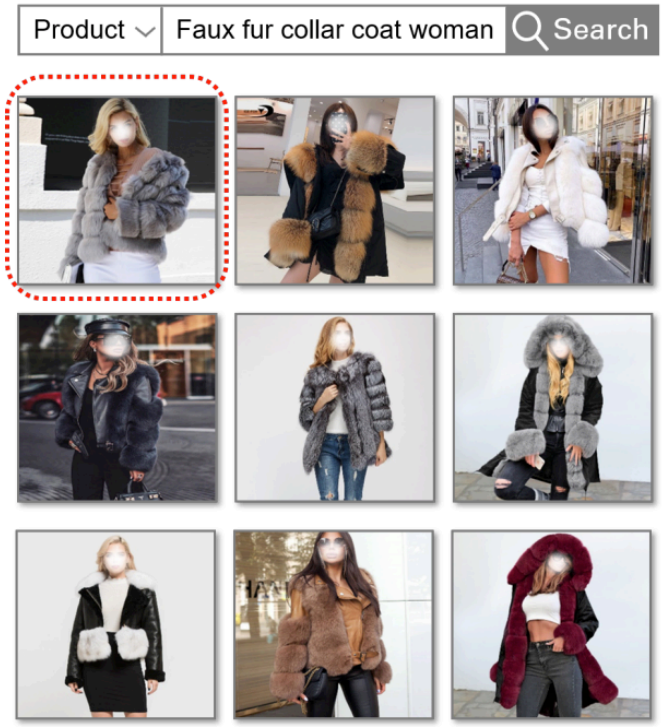
At the same time in the USA, Morris Michtom created the first teddy bear, after being inspired by a drawing of Theodore "Teddy" Roosevelt with a bear cub.

OK-VQA (Sidorov et al. 2019)

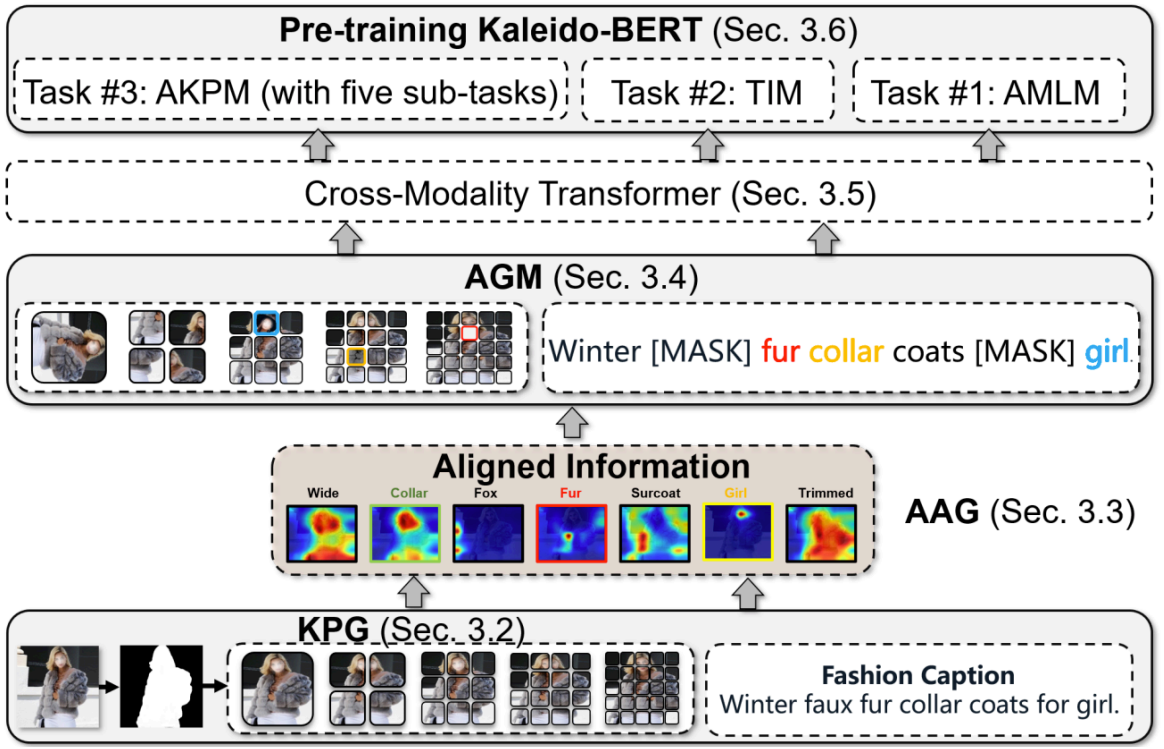


Reasoning over Vision and Language:
Exploring the Benefits of Supplemental Knowledge
(Shevchenko et al. 2021)

Diverse Applications of VLP



Application: Fashion Product Searching System



Kaleido-BERT (Zhuge et al. 2021)

TAP: Text-Aware Pre-training

- Towards VLP model that can read

Question: what **number** is on the bike on the right? ---- A: the number is **317**

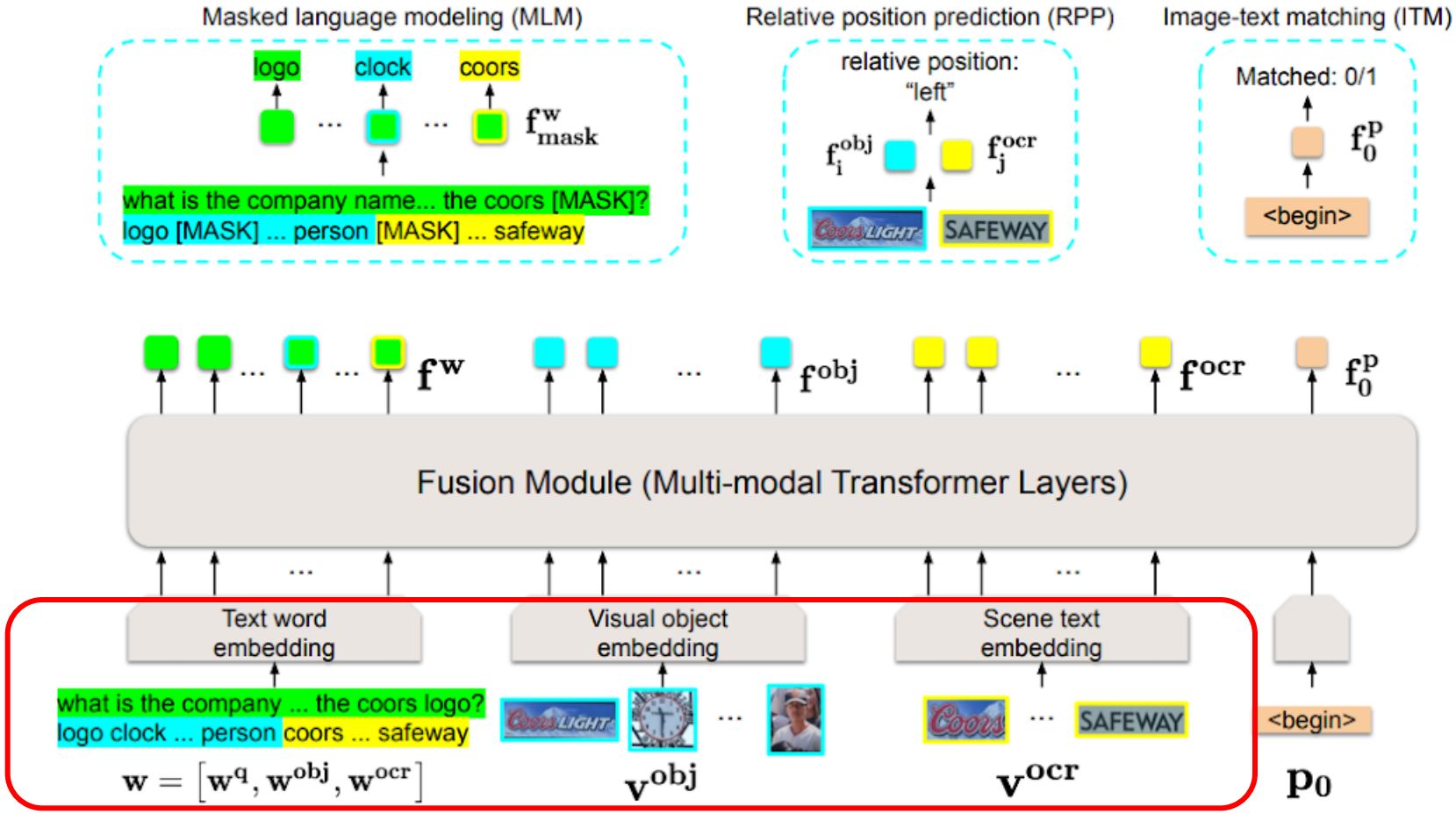
Text-VQA (Singh et al. 2019)

A group of motorcyclists with **number** **317, 44, 30, 338, 598** racing outdoor.

Text-Captioning (Sidorov et al. 2020)

Credit to: Zhengyuan Yang

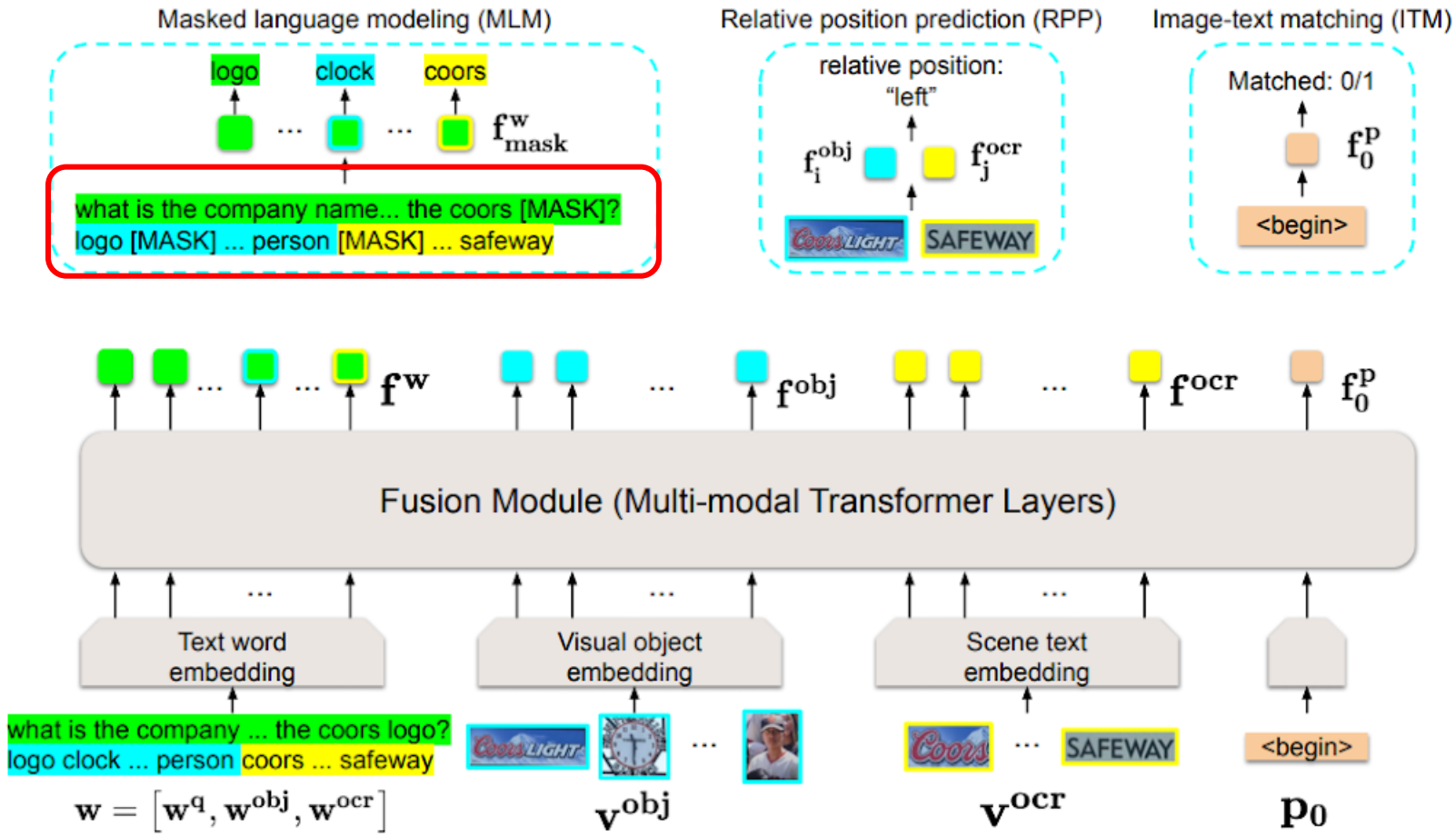
TAP: Text-Aware Pre-training



TAP (Yang et al. 2021)

Credit to: Zhengyuan Yang

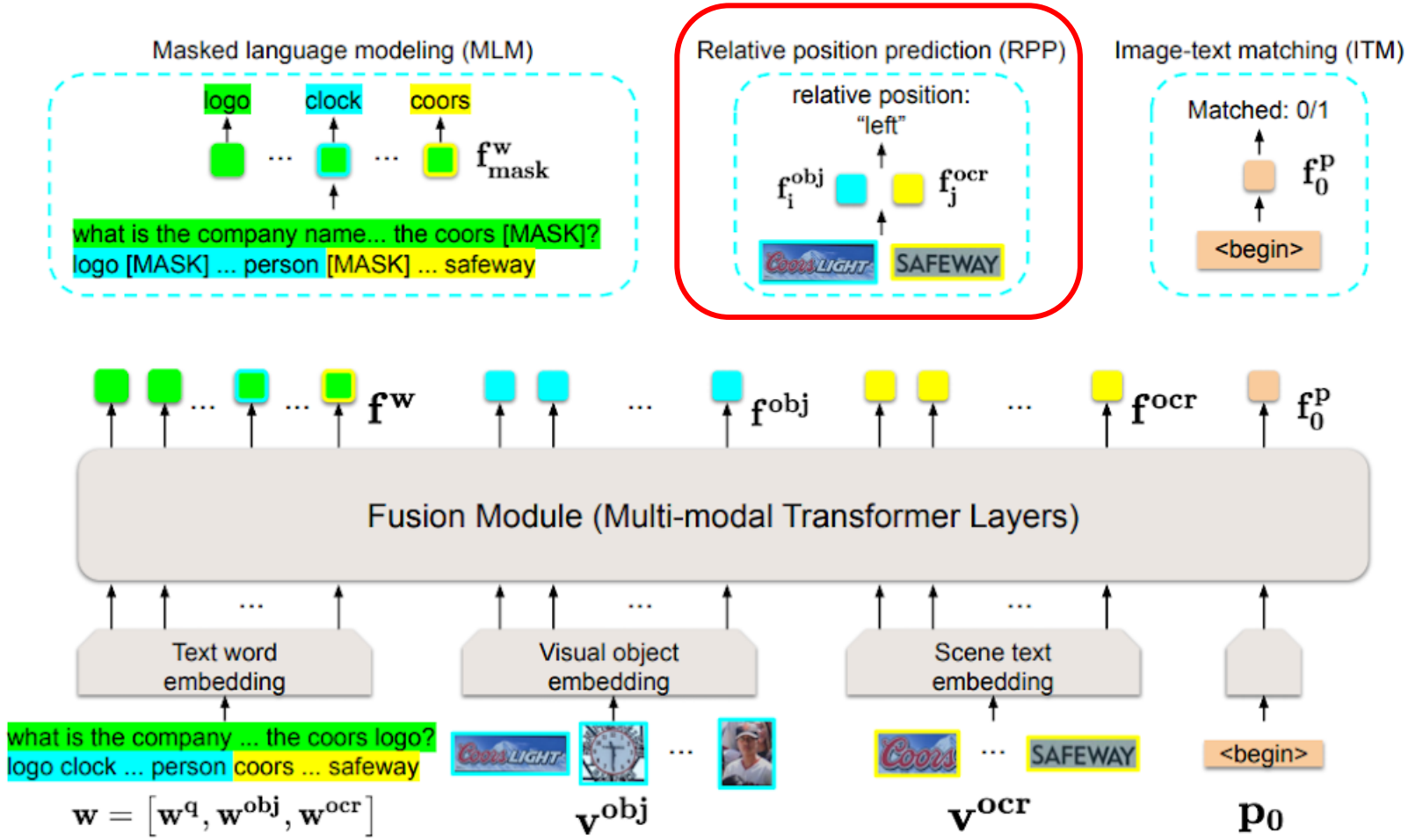
TAP: Text-Aware Pre-training



TAP (Yang et al. 2021)

Credit to: Zhengyuan Yang

TAP: Text-Aware Pre-training



TAP (Yang et al. 2021)

Credit to: Zhengyuan Yang

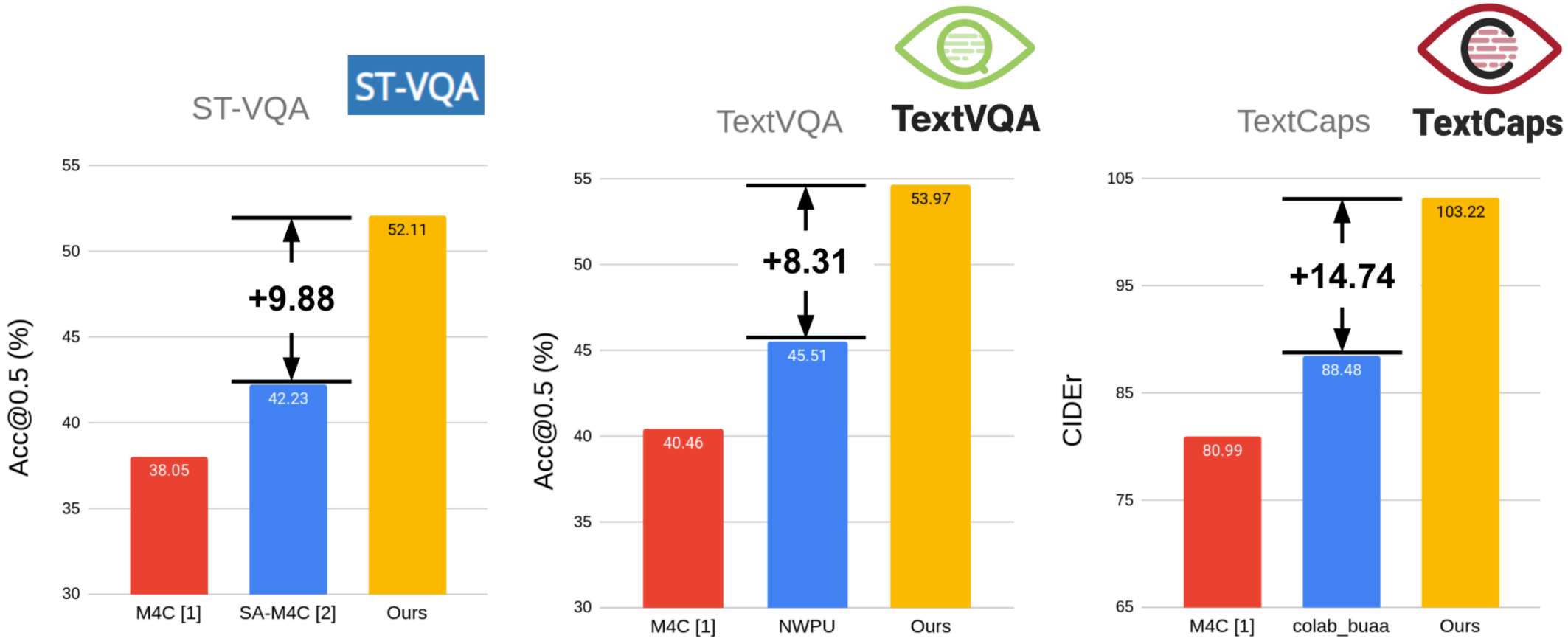
TAP: Text-Aware Pre-training

Dataset	TextVQA	ST-VQA	TextCaps	CC-OCR
Training images	**22K**	**19K**	**22K**	**1.37M**
Text	35K OCR-QA pairs	26K OCR-QA pairs	110K OCR-Caption	One caption per image
Image source	Open Image	ICDAR 2013/15, ImageNet, VizWiz, IIIT Scene Text Retrieval, Visual Genome, COCO-Text	TextVQA	Conceptual captions
# OCR	mean: 23.1, med:12	mean: 19.2, med:10	mean: 23.1, med:12	mean: 11.4, med: 6

TAP (Yang et al. 2021)

Credit to: Zhengyuan Yang

TAP: Text-Aware Pre-training



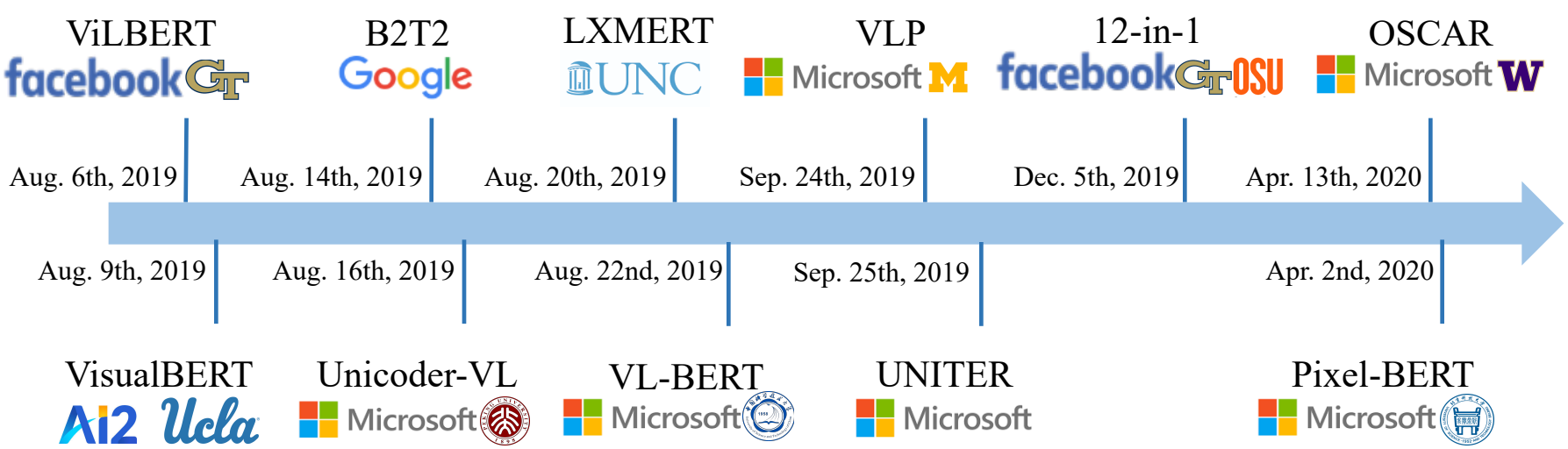
TAP (Yang et al. 2021)

Credit to: Zhengyuan Yang

Agenda

- Advanced training strategies in VLP
- Diverse applications of VLP
- VL for V/L
- Compressing VLP models
- Robustness/causality/fairness of VLP models
- Multilingual VLP

Great success of VLP models



V+L Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

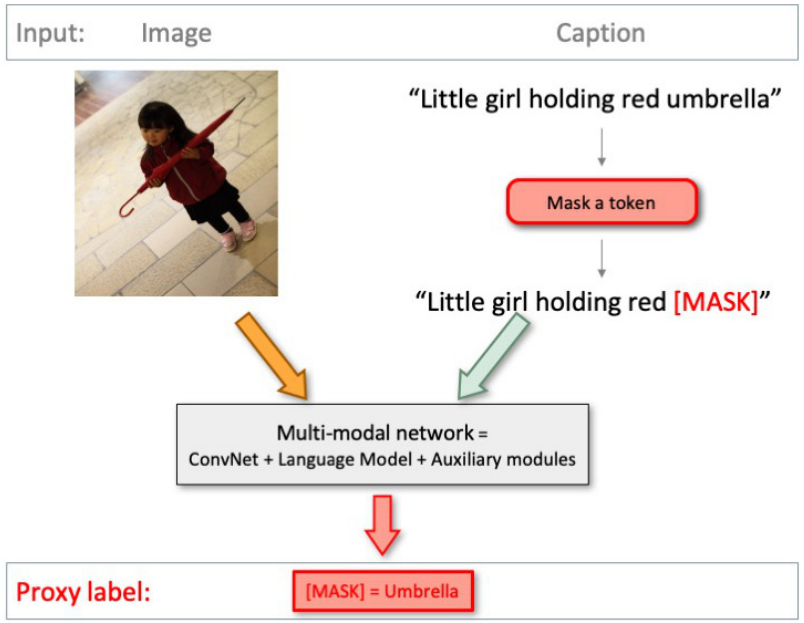
Can VLP help unimodal tasks?

VL for V/L

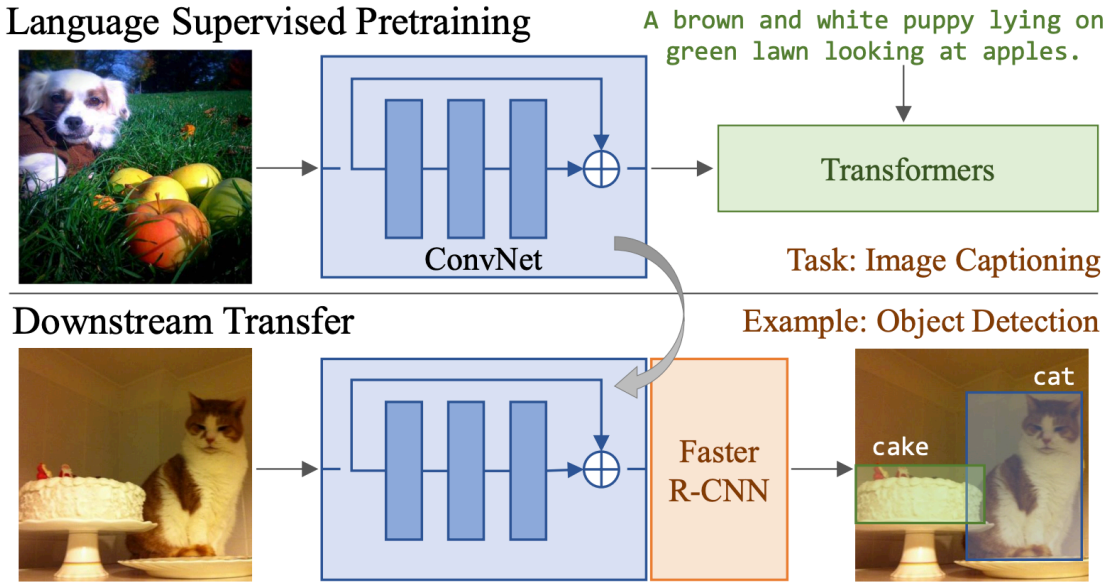
- *VL for V*: A scalable way to learn visual representations
  - SOTA computer vision models rely on carefully annotated labels/bounding boxes for learning
  - Self-supervised learning is scalable, but supervision signal is weak
  - Image-text pairs widely exist on the web

VL for V/L

- *VL for V*: A scalable way to learn visual representations
  - Early attempts on using human annotated image-text pairs: COCO, VG



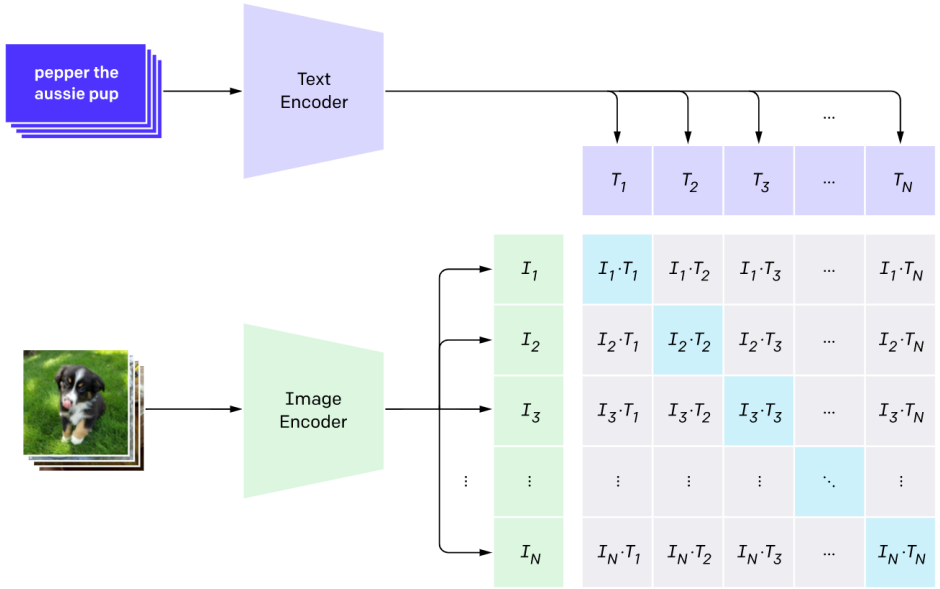
ICMLM (Sariyildiz et al. 2020)



VirTex (Desai and Johnson, 2020)

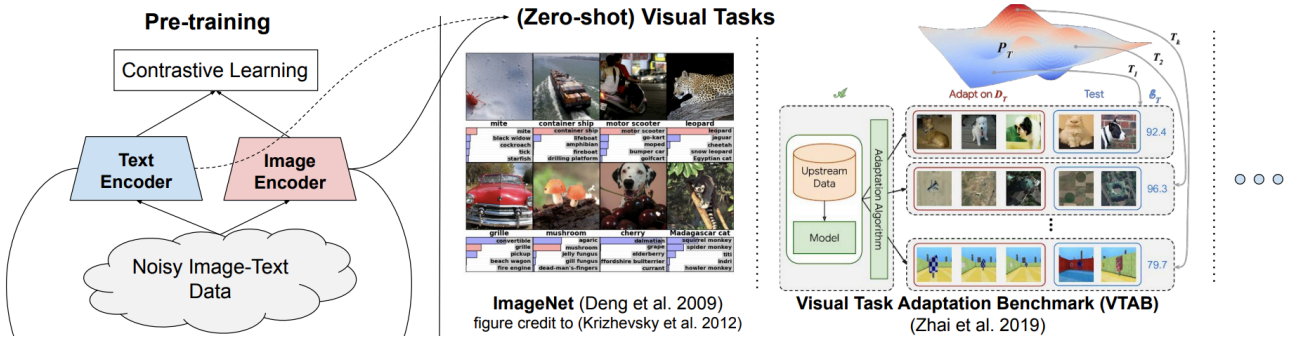
VL for V/L

- *VL for V*: A scalable way to learn visual representations
  - Scaling up to billions of web-crawled image alt-text data!



CLIP (Radford et al. 2021)

~400M image alt-text pairs!

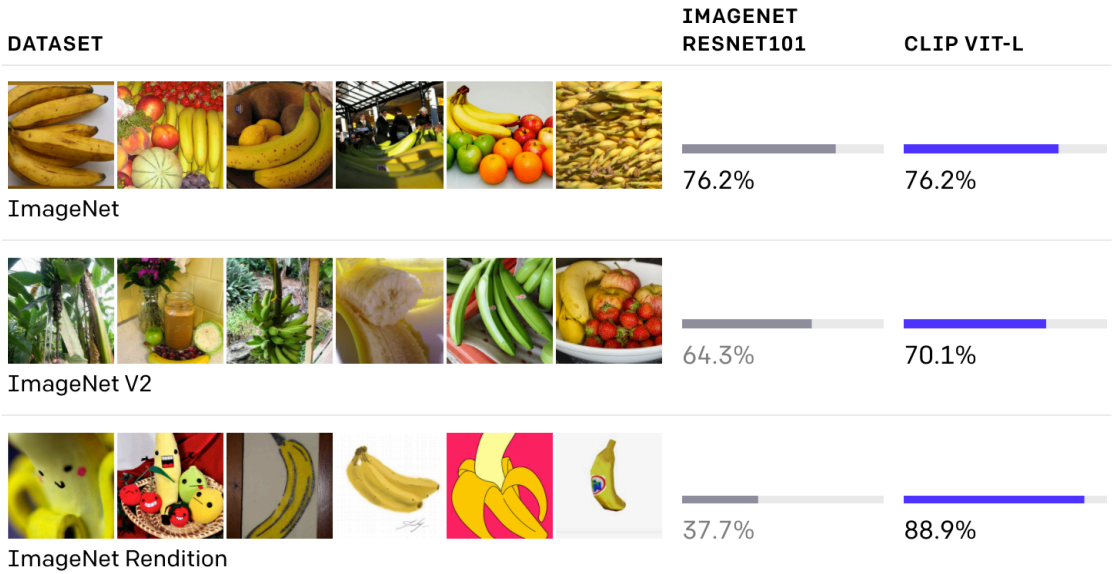


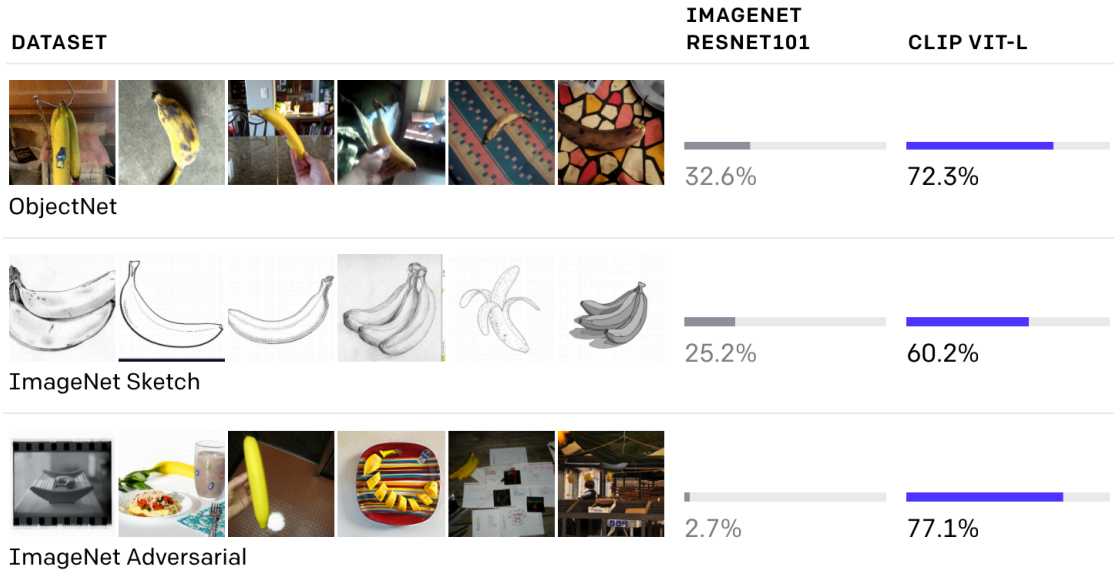
ALIGN (Jia et al., 2020)

Over one billion image alt-text pairs!!!

VL for V/L

- *VL for V*: A scalable way to learn visual representations
  - Scaling up to billions of web-crawled image alt-text data!
  - Achieving strong performance while closes the “robustness gap” by up to 75%

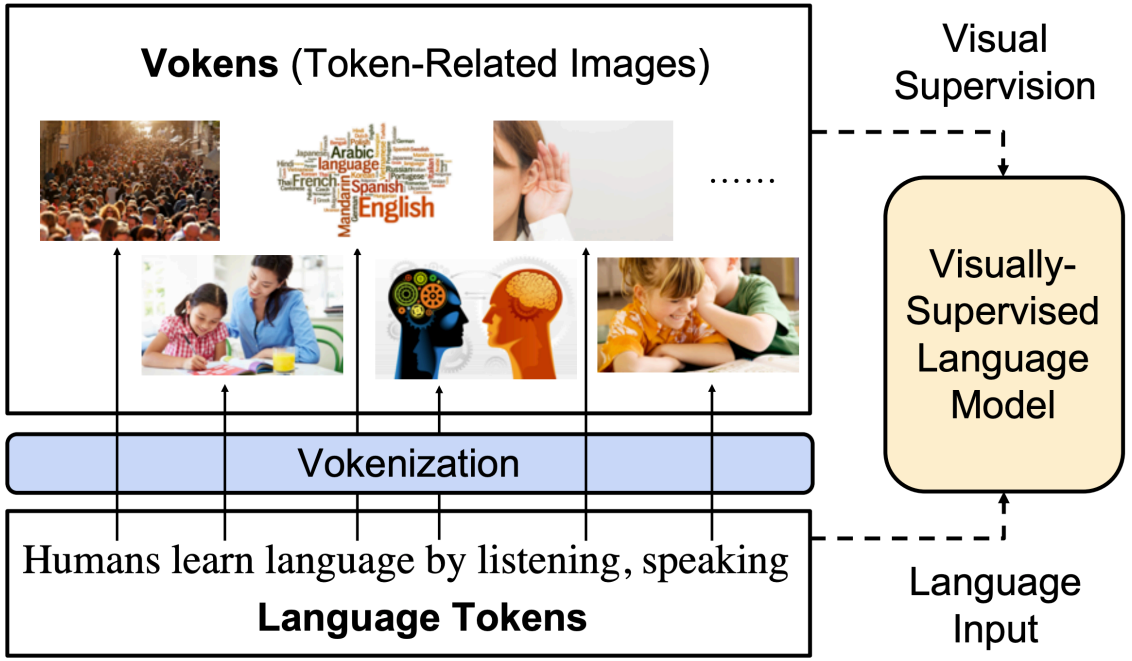




CLIP (Radford et al. 2021)

VL for V/L

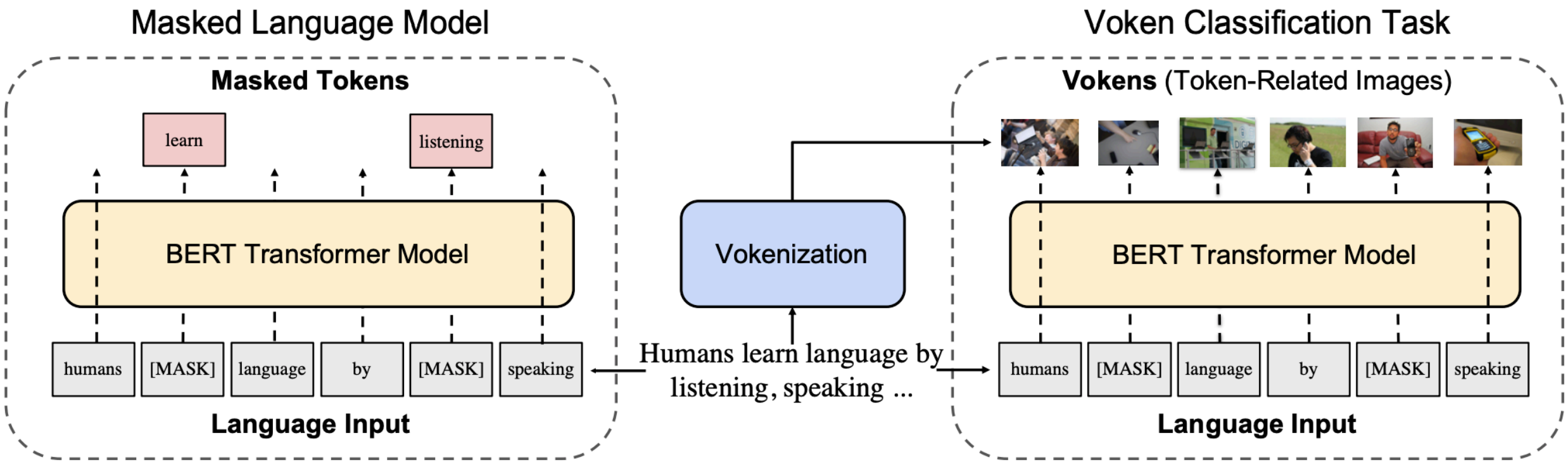
- *VL for L*: Great potential to enhance language representations
  - A picture is worth a thousand words



Vokenization (Tan and Bansal, 2020)

Vokenization

- Goal: improving language understanding with contextualized, visual-grounded supervision

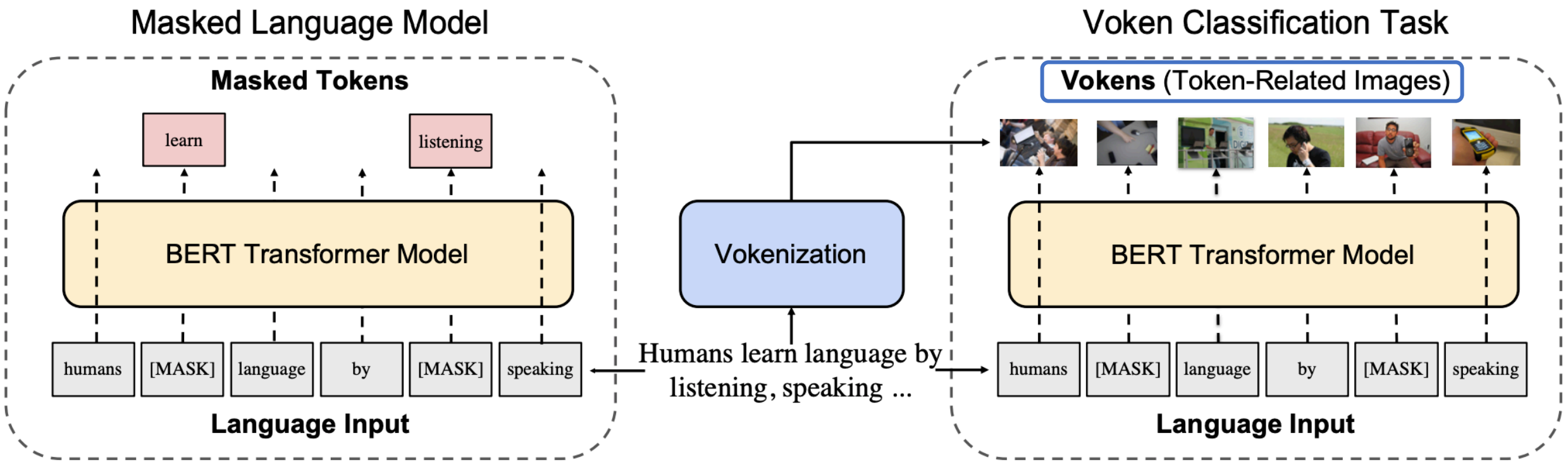


Vokenization (Tan and Bansal, 2020)

Vokenization

- Goal: Improving Language Understanding with Contextualized, Visual-Grounded Supervision

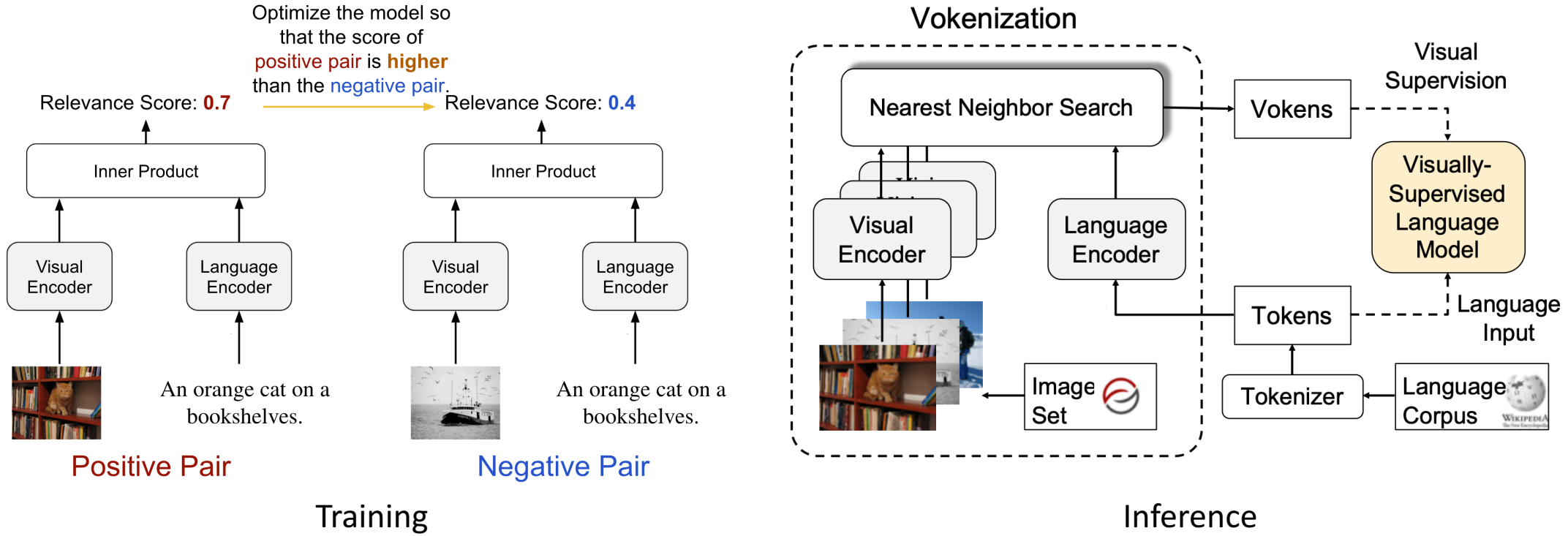
How to generate vokens?



Vokenization (Tan and Bansal, 2020)

Vokenization

- Vokenization process: assign each token with a relevant image



Vokenization (Tan and Bansal, 2020)

Vokenization

- Improve language understanding with related visual information

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**

+2.7% average improvement

Example 1: Humans learn language by listening, speaking, writing, reading

Learns token-image alignment

Vokenization (Tan and Bansal, 2020)

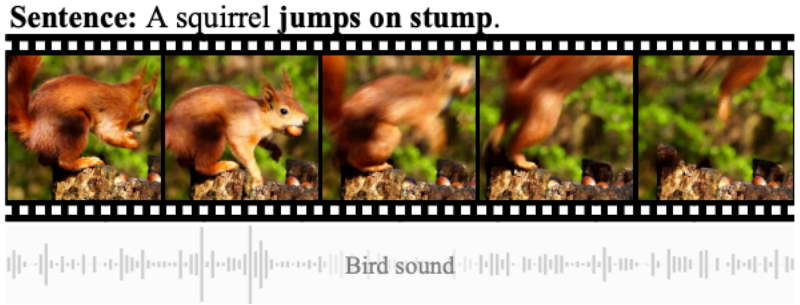
Video-Language for Language

Video-aided Unsupervised Grammar Induction

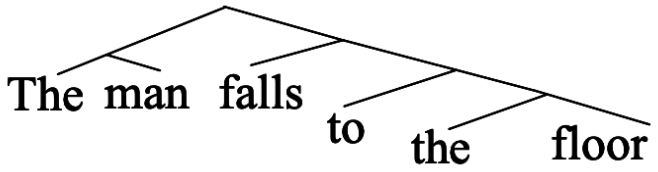
Songyang Zhang[1]*, **Linfeng Song**[2], **Lifeng Jin**[2], **Kun Xu**[2], **Dong Yu**[2] and **Jiebo Luo**[1]
[1]University of Rochester, Rochester, NY, USA
szhang83@ur.rochester.edu, jluo@cs.rochester.edu
[2]Tencent AI Lab, Bellevue, WA, USA
{lfsong,lifengjin,kxkunxu,dyu}@tencent.com

NAACL 2021 Best Long Paper

Pre-training



Downstream

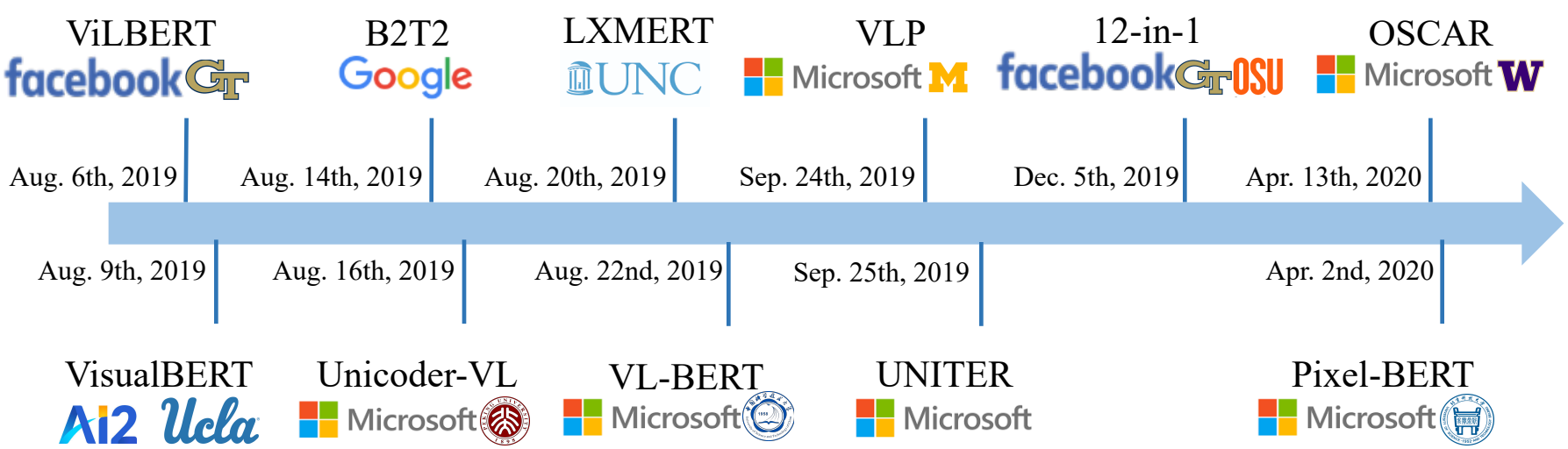


Parse tree for the sentence
"The man falls to the floor".

Agenda

- Advanced training strategies in VLP
- Diverse applications of VLP
- VL for V/L
- Compressing VLP models
- Robustness/causality/fairness of VLP models
- Multilingual VLP

Great success of VLP models



V+L Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

How about efficiency? Can we compress a large VLP model while preserving its performance and transferability?

Model Compression Technique

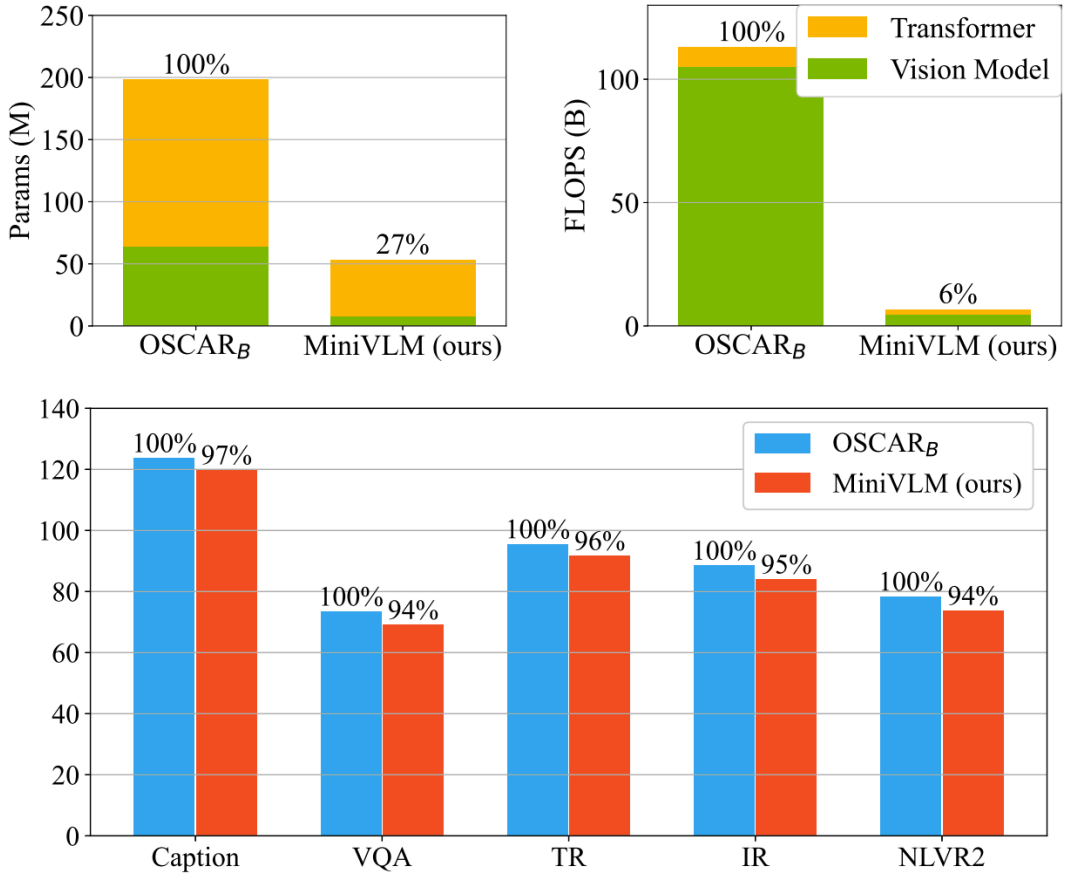
- Low-rank approximation and sparsity
- Neural Architecture Search
- Knowledge distillation
- Pruning
- Quantization

Compressing VLP Models

- Low-rank approximation and sparsity
- Neural Architecture Search
- Knowledge distillation
- Pruning
- Quantization

Compressing VLP Models via Distillation

- Large VLP Model
  - Region feature extractor
    - R101 (BUTD, 2017)
    - X152 (VinVL, 2021)
  - Transformer
    - $\text{BERT}_{\text{BASE}}$ (12/768/3072)
    - $\text{BERT}_{\text{LARGE}}$
- Compact VLP Model
  - Region feature extractor
    - Two-stage Efficient Extractor (TEE)
  - Transformer
    - MiniLM (12/384/1536)

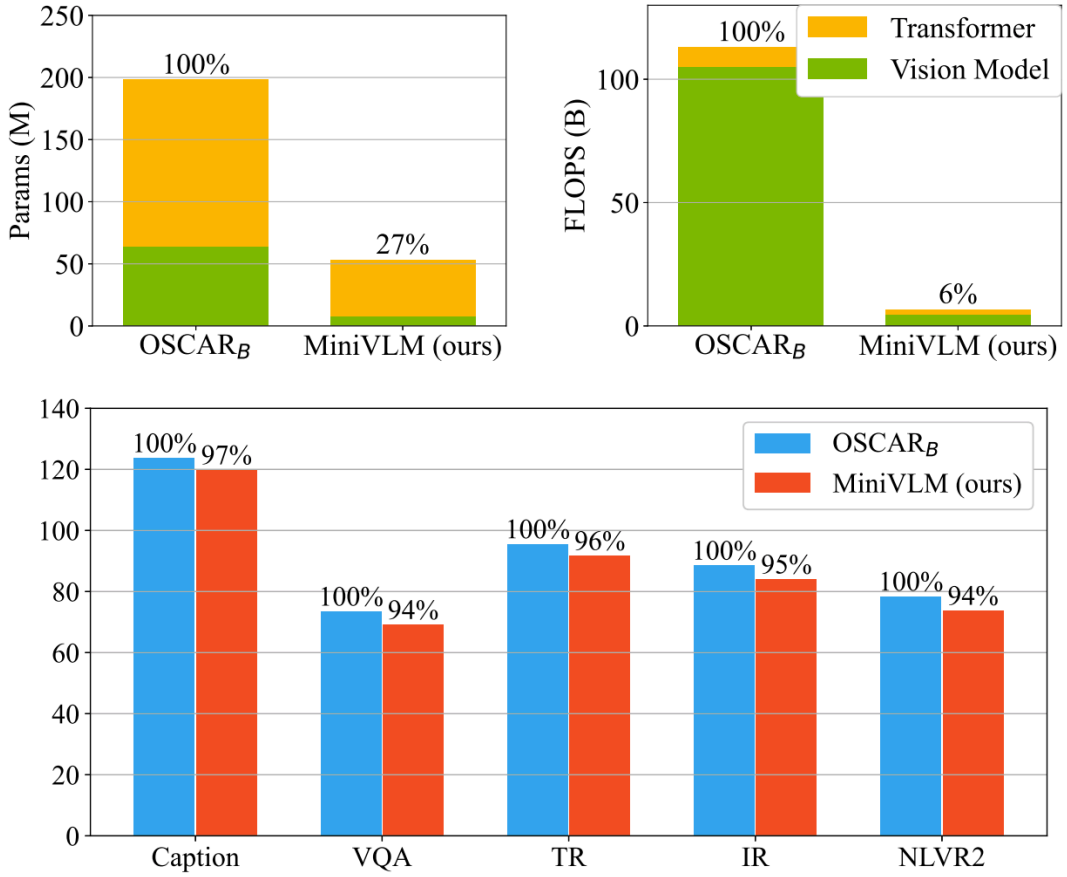


MiniVLM (Wang et al. 2020)

Credit to: Jianfeng wang

Compressing VLP Models via Distillation

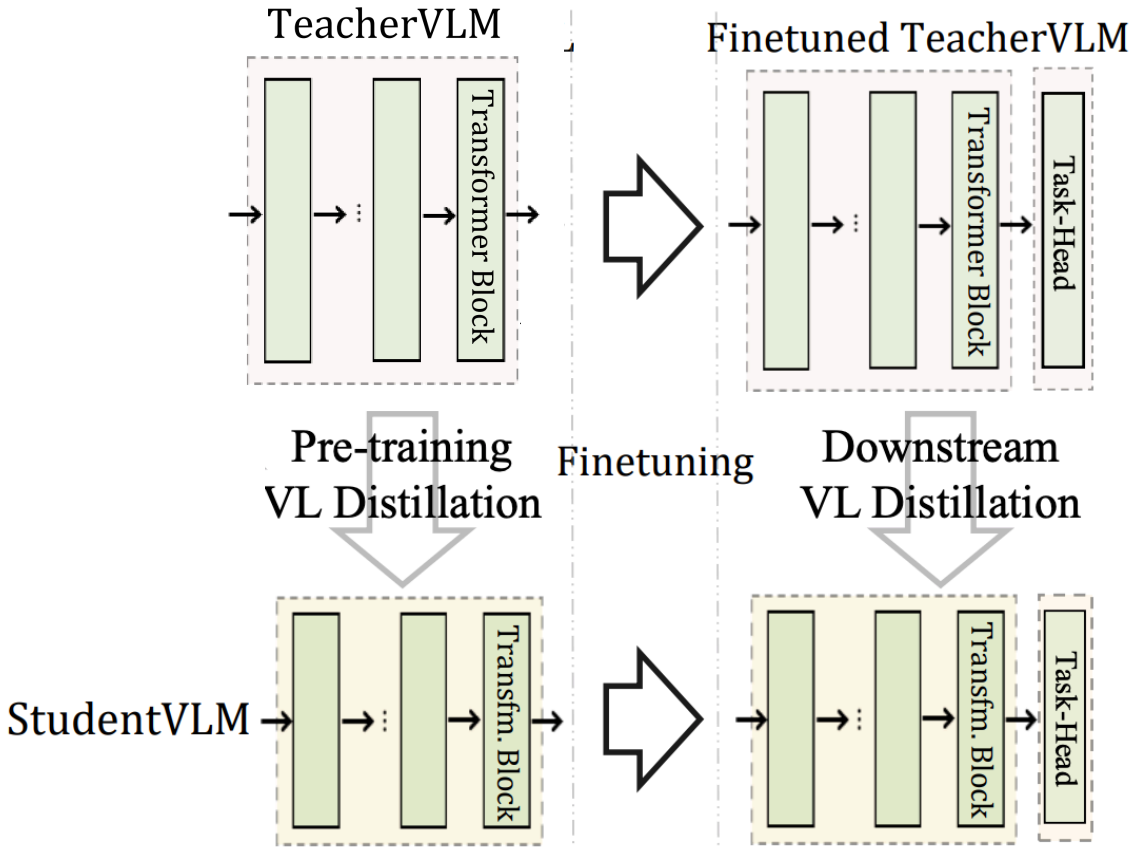
- Large VLP Model (Teacher)
  - Region feature extractor
    - R101 (BUTD, 2017)
    - X152 (VinVL, 2021)
  - Transformer
    - $\text{BERT}_{\text{BASE}}$ (12/768/3072)
    - $\text{BERT}_{\text{LARGE}}$
- Compact VLP Model (Student)
  - Region feature extractor
    - Two-stage Efficient Extractor (TEE)
  - Transformer
    - MiniLM (12/384/1536)



MiniVLM (Wang et al. 2020)

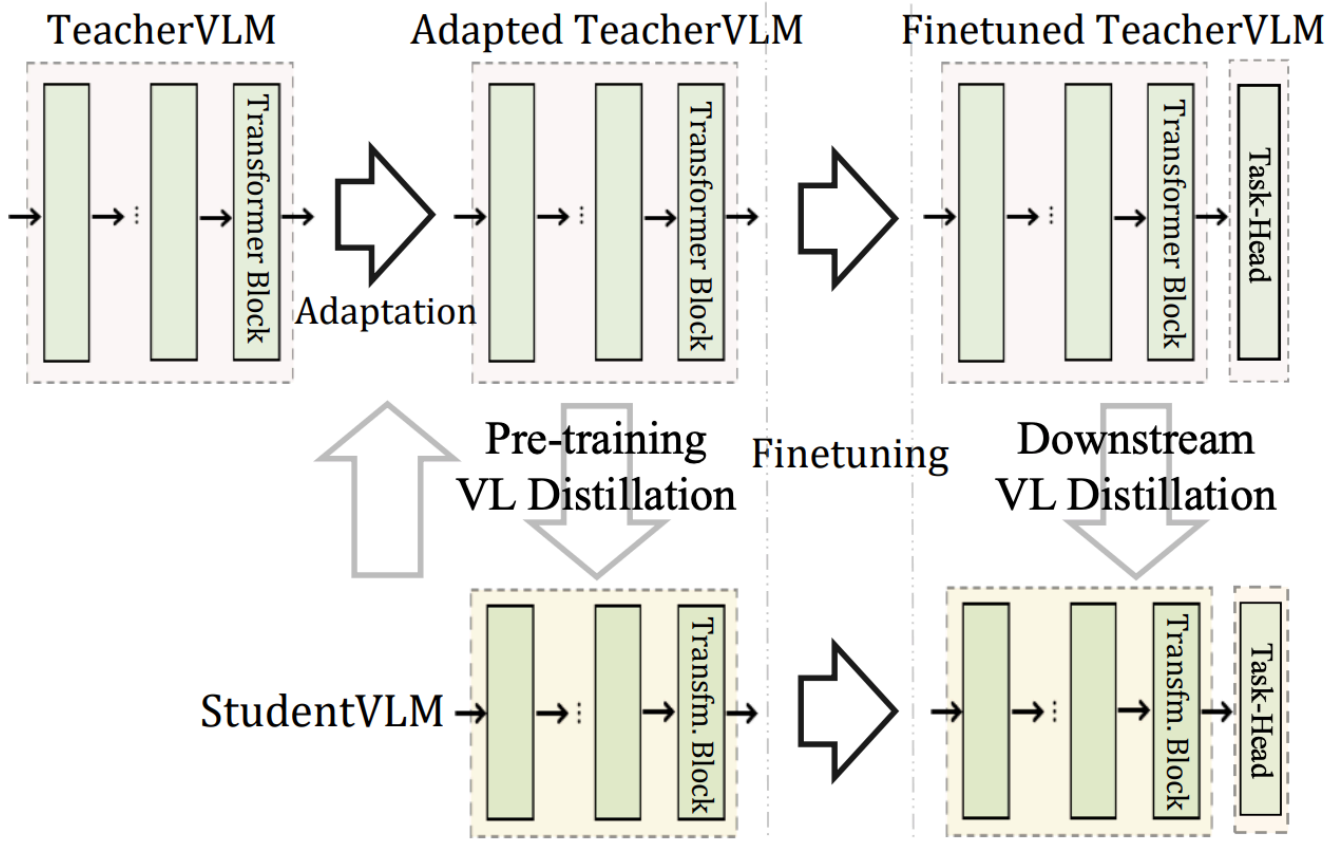
Credit to: Jianfeng wang

Compressing VLP Models via Distillation



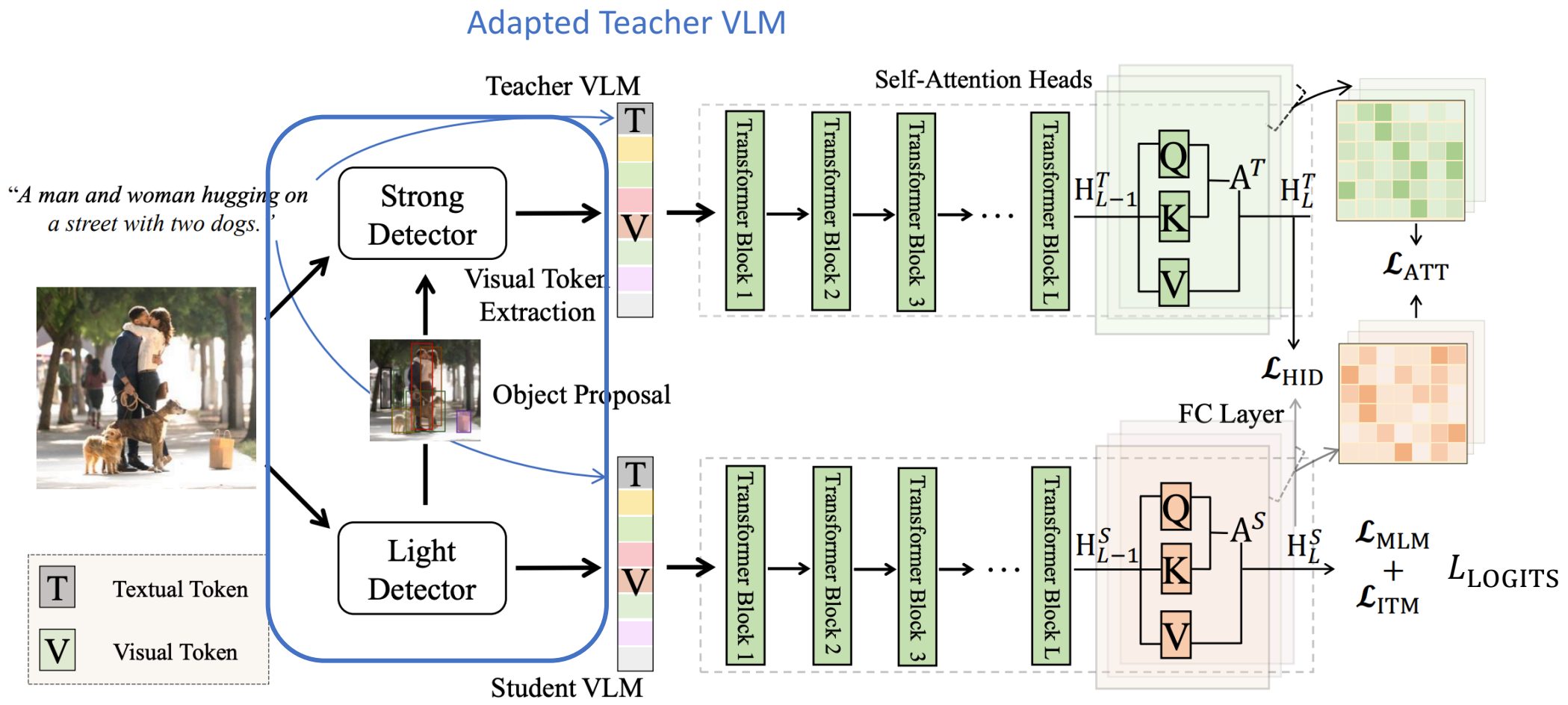
DistillVLM (Fang et al. 2021)

Compressing VLP Models via Distillation



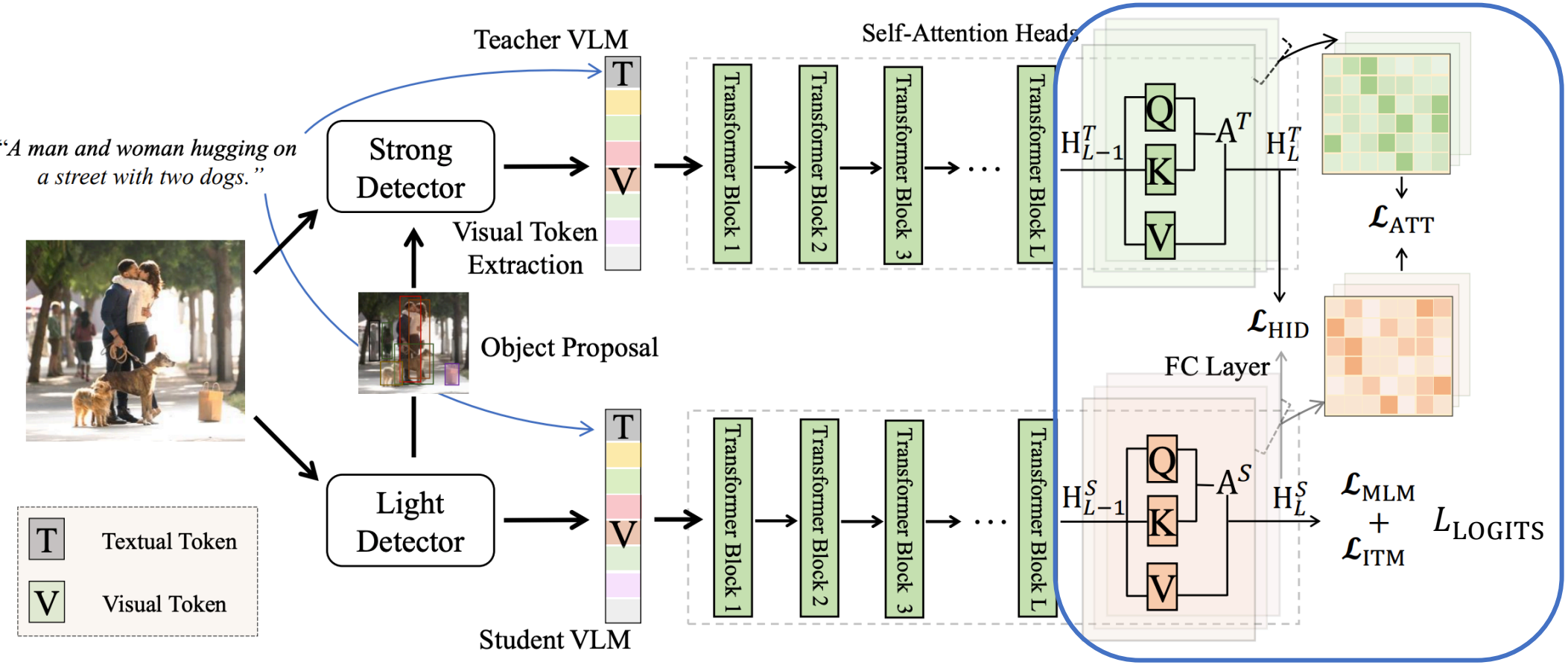
DistillVLM (Fang et al. 2021)

Compressing VLP Models via Distillation



DistillVLM (Fang et al. 2021)

Compressing VLP Models via Distillation



DistillVLM (Fang et al. 2021)

Distillation on (1) logits, (2) hidden states and (3) attention matrix

Compressing VLP Models via Distillation

Method	# Param	# *I-T* Pairs	Visual Feat.	P. D.	F. D.	COCO Captioning				VQA	
						B@4	M	C	S	test-std	test-dev
UVLP [74]	111.7M	3M	ResNeXt101	✗	✗	36.5	28.4	116.9	21.2	70.7	–
OSCAR_{B} [37]	111.7M	7M	R101-F	✗	✗	36.5	30.3	123.7	23.1	73.4	73.2
MiniVLM [65]	34.5M	7M	TEE	✗	✗	34.3	28.1	116.7	21.3	-	-
MiniVLM [65]	34.5M	14M	TEE	✗	✗	35.6	28.6	119.8	21.6	69.4	69.1
DistillVLM	34.5M	7M	TEE	✗	✗	34.0	28.0	115.7	21.1	69.0	68.8
				✗	✓	34.5	28.2	117.1	21.5	69.2	69.0
				✓	✗	35.2	28.6	120.1	21.9	69.7	69.6
				✓	✓	35.6	28.7	120.8	22.1	69.8	69.6

DistillVLM improves over MiniVLM, but still more compact and faster than large VLP models!

DistillVLM (Fang et al. 2021)

Compressing VLP Models via Pruning

- A popular direction: *lottery ticket hypothesis*

What to prune? | How to prune? | How often? | When to prune?

Unstructed | Magnitude | Iterative | ‚Before‘

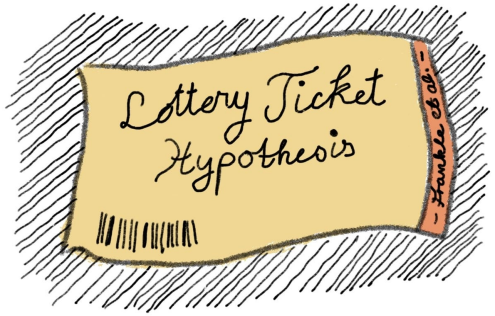
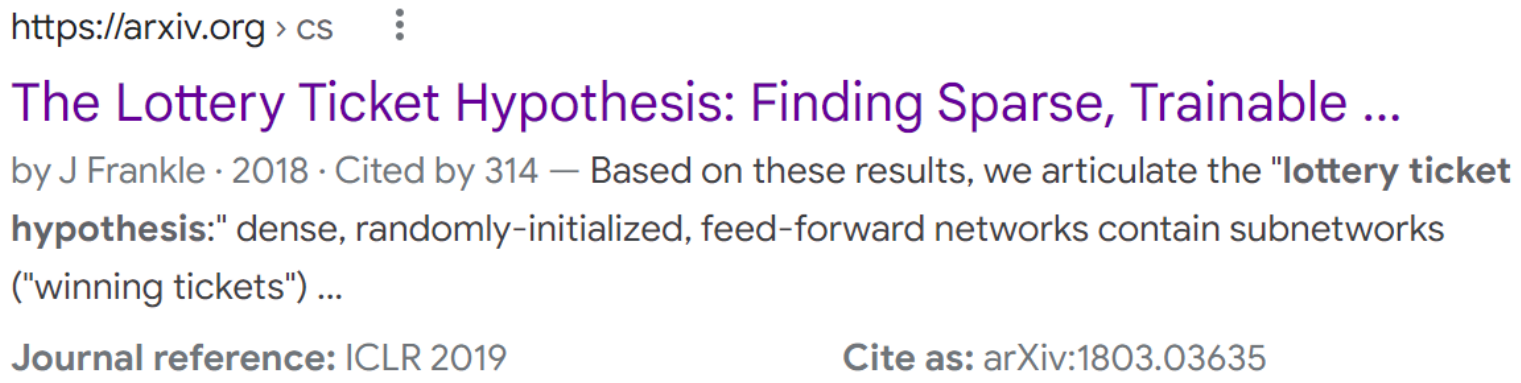


Image credit to: https://roberttlange.github.io/posts/2020/06/lottery-ticket-hypothesis/

Lottery Ticket Hypothesis

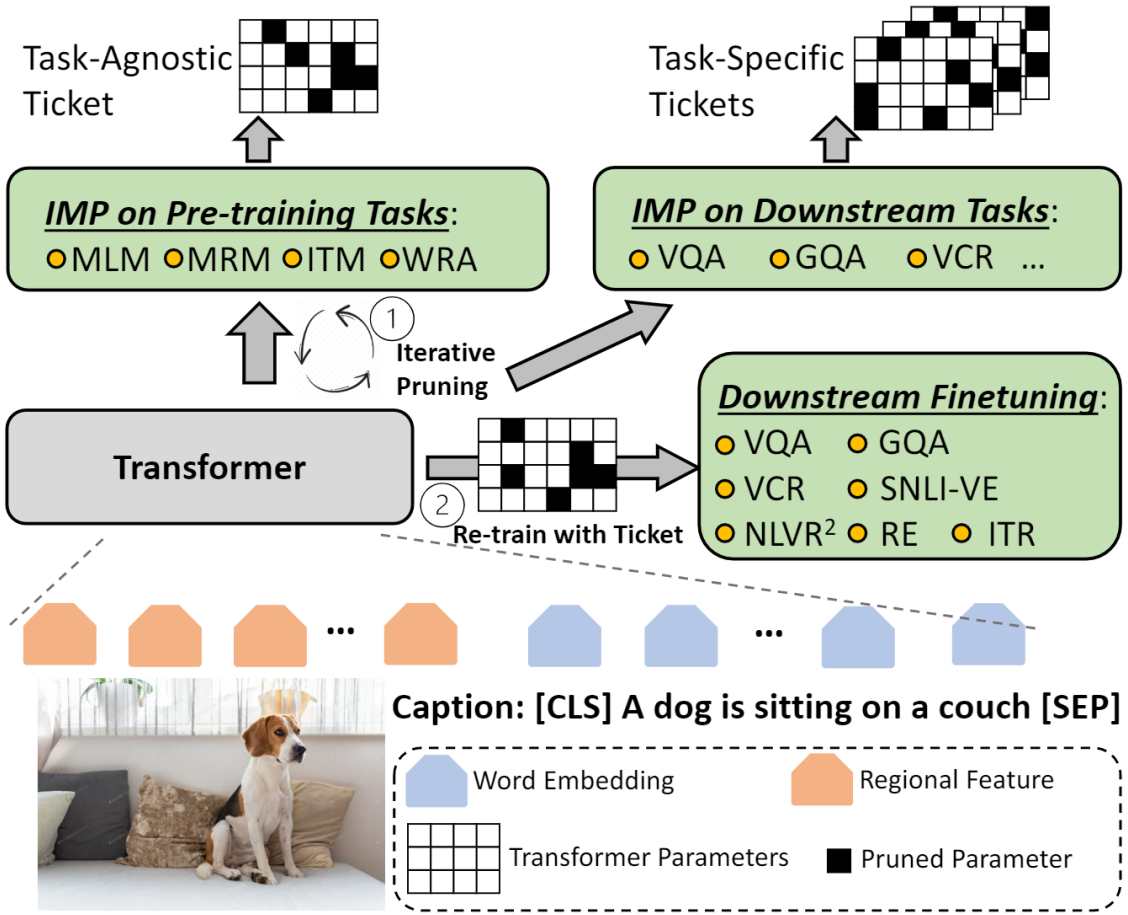
- ICLR 2019 Best Paper by MIT: The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks
- LTH: A randomly initialized, dense neural network contains a subnetwork that is initialized such that — when trained in isolation — it can match the test accuracy of the original network after training for at most the same number of iterations. - Frankle & Carbin (2019, p.2)
- An emerging sub-field in deep learning regarding sparse neural networks



Playing Lottery Tickets with Vision and Language

- *Existence*: Can we draw VLP winning tickets successfully for VL downstream tasks?
- *Transferability*: Can we find tickets that transfer universally to all downstream VL tasks?
- *Compatibility*: Can we find tickets compatible with adversarial training to enhance the performance?

Playing Lottery Tickets with Vision and Language



Algorithm 1 Iterative Magnitude Pruning for V+L Tickets.

Input Initial mask $\boldsymbol{m} = 1^{d_1}$; Pre-trained parameters $\boldsymbol{\theta}_0$ and task-specific parameters $\boldsymbol{\phi}_0$; rewinding step i (could be 0), sparsity level s, total training step t.
Train the pre-trained V+L model $f(\boldsymbol{x}; \boldsymbol{m} \odot \boldsymbol{\theta}_0, \boldsymbol{\phi}_0)$ to step i: $f(\boldsymbol{x}; \boldsymbol{m} \odot \boldsymbol{\theta}_i, \boldsymbol{\phi}_i)$.
repeat
  Train $f(\boldsymbol{x}; \boldsymbol{m} \odot \boldsymbol{\theta}_i, \boldsymbol{\phi}_i)$ to step t: $f(\boldsymbol{x}; \boldsymbol{m} \odot \boldsymbol{\theta}_t, \boldsymbol{\phi}_t)$.
  Prune 10% of non-zero weights of $\boldsymbol{m} \odot \boldsymbol{\theta}_t$ based on the magnitudes and update $\boldsymbol{m}$ accordingly.
until the sparsity of $\boldsymbol{m}$ reaches s
Return $f(\boldsymbol{x}; \boldsymbol{m} \odot \boldsymbol{\theta}_i, \cdot)$ =0

A *winning ticket* is a sub-network that matches the performance of the original full dense network

Playing lottery tickets with VL (Gan et al. 2021)

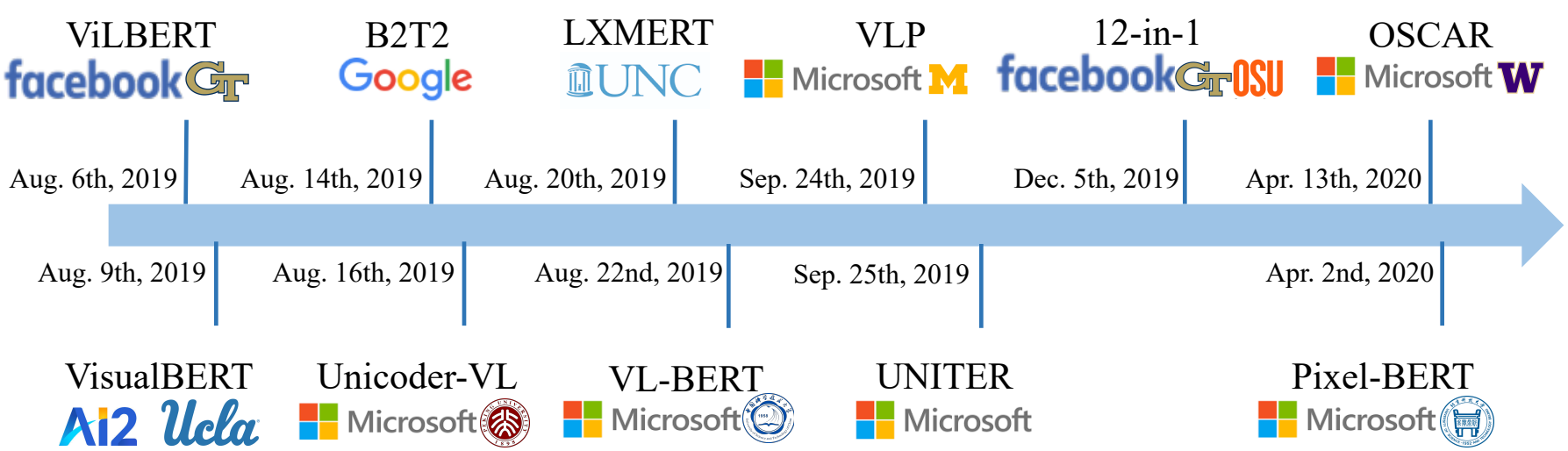
Playing Lottery Tickets with Vision and Language

- ~~*Existence*: Can we draw VLP winning tickets successfully for VL downstream tasks?~~
- *VLM can play lottery tickets too*: We confirm that "relaxed" winning tickets that match 99% of the full accuracy can be found at 50%-70% sparsity across all the tasks.
- ~~*Transferability*: Can we find tickets that transfer universally to all downstream VL tasks?~~
- *One ticket to win them all*: Matching subnetworks found via IMP on pre-training tasks transfer universally. Unexpectedly, matching subnetworks found via IMP on each downstream task also transfer to other tasks reasonably well.
- ~~*Compatibility*: Can we find tickets compatible with adversarial training to enhance the performance?~~
- *Enhancing tickets with adversarial training*: Though the found winning tickets are sparse neural networks, adversarial training can be still helpful to enhance the performance across all the tasks considered.

Agenda

- Advanced training strategies in VLP
- Diverse applications of VLP
- VL for V/L
- Compressing VLP models
- Robustness/fairness of VLP models
- Multilingual VLP

Great success of VLP models



V+L Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

How robust are these pre-trained V+L Models?

Similar Data Distribution

Training

~

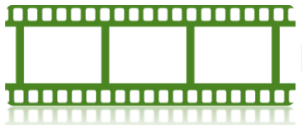
Test

Little-to-None Linguistic Variations

Original	Q: What is in the basket? A: Remote
Rephrasing	Q: What *can be seen* inside the basket? A: Remote
Logical Transformation	Q: *Is remote* in the basket? A: *Yes*

Without Visual Content Manipulations

COCO
Common Objects in Context



Movie Clips

Standard V+L Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

Robust VQA Benchmarks

- Compilation of 9 diverse VQA datasets covering 4 types of robustness
- Note that robustness here *is not* adversarial robustness

Linguistic Variation (Lingual)

- *VQA-Rephrasings*

Logical Reasoning (Reason)

- *VQA-LOL Compose*
- *VQA-LOL Supplement*
- *VQA-Introspect*
- *GQA*

Visual Content Manipulation (Visual)

- *IV-VQA*
- *CV-VQA*

Answer Distribution Shift (Answer)

- *VQA-CP v2*
- *GQA-OOD*

MANGO (Li et al. 2020)

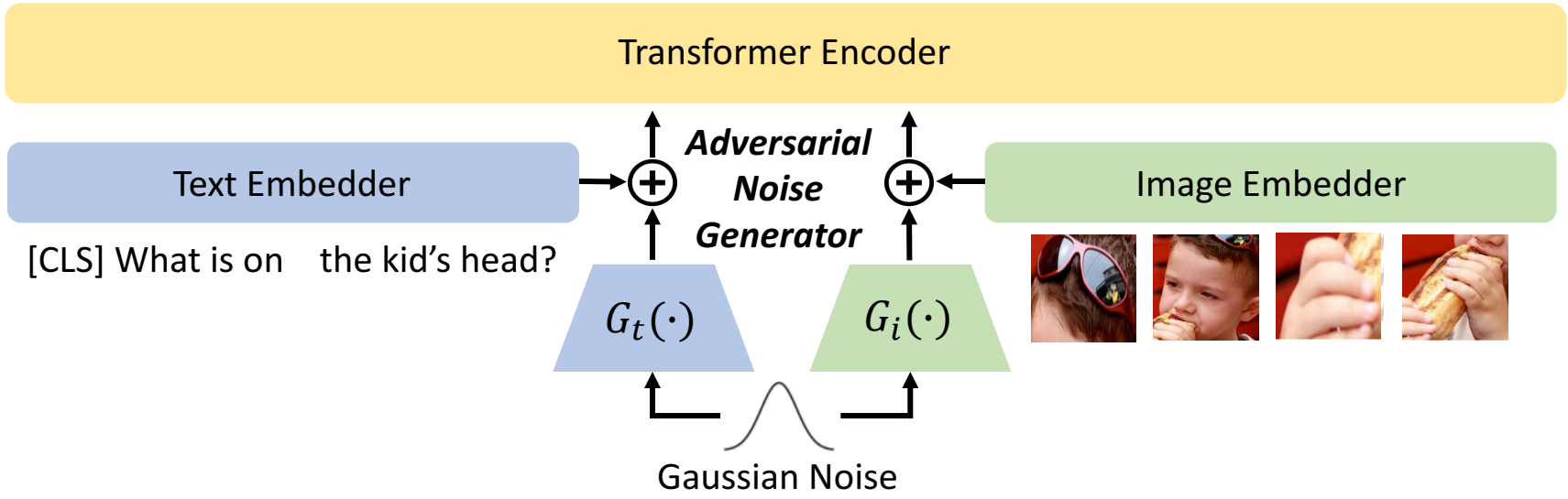
Robust VQA Benchmarks

Type	Benchmark	Metric	Q Type	Train			Val		Test	
				Source	#IQ	len(Q)	#IQ	len(Q)	#IQ	len(Q)
Lingual	VQA-Rep. [58]	Acc.	All	VQA v2 [20] train	444K	6.20	162K	7.15	-	-
Reason	VQA-LOL Comp. [18]	Acc.	Y/N	VQA v2 train	444K	6.20	43K	12.09	291K	12.12
	VQA-LOL Supp. [18]	Acc.	Y/N	VQA v2 train	444K	6.20	9K	15.15	669K	15.19
	VQA-Intro. [56]	M✓S✓	All	VQA v1 [6] train	248K	6.21	-	-	95K	6.36
	GQA [26]	Acc.	All	-	943K	8.76	132K	8.77	13K	8.51
Visual	IV-VQA [2]	#flips	All	VQA v2 train	444K	6.20	120K	5.85	-	-
	CV-VQA [2]	#flips	Num.	VQA v2 train	444K	6.20	4K	5.83	-	-
Answer	VQA-CP v2 [3]	Acc.	All	-	438K	6.14	-	-	220K	6.31
	GQA-OOD [32]	Acc.	All	GQA train	943K	8.76	51K	8.09	3K	7.70

Table 1: Detailed descriptions of each downstream benchmark, including robustness type, evaluation metric, question type, training data source and statistics on train, val, test data in terms of number of Image-Question pairs (#IQ) and average question length (len(Q)). We use the training data provided with the benchmark unless specified otherwise. Results on val split are reported when test split is not available. Acc. is short for Accuracy. M✓S✓ is a consistency measure between main questions and sub-questions in VQA-Introspect. #flips is the number of predictions mismatched before and after visual content manipulation.

MANGO Framework

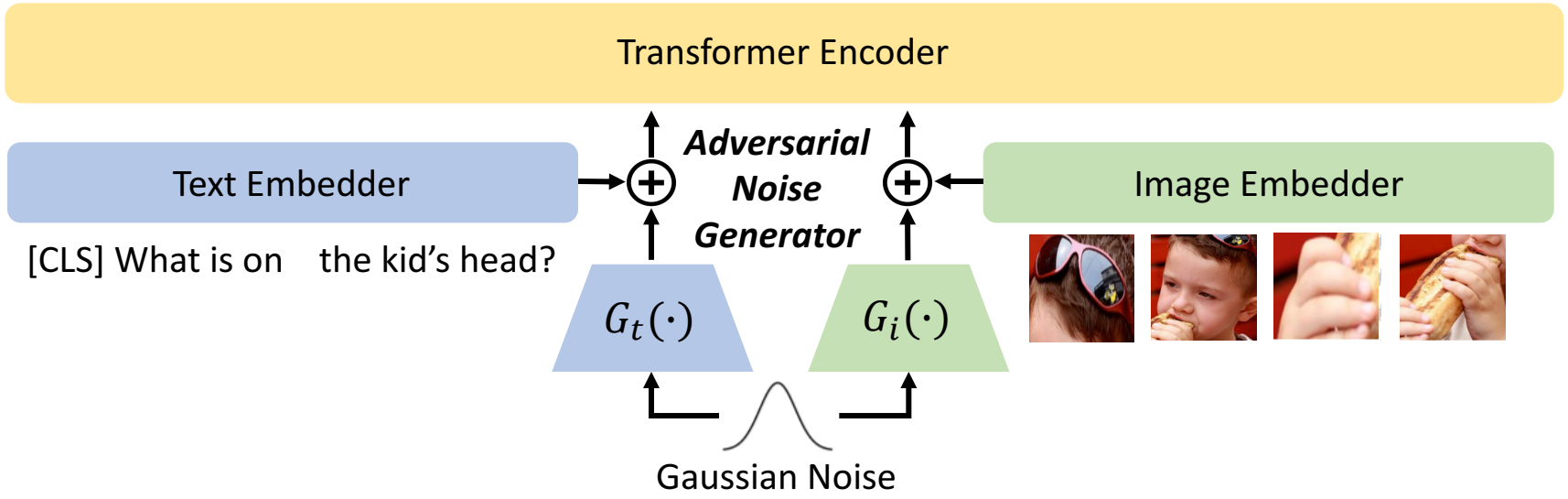
- Adversarial Noise Generator



MANGO (Li et al. 2020)

MANGO Framework

- Adversarial Noise Generator



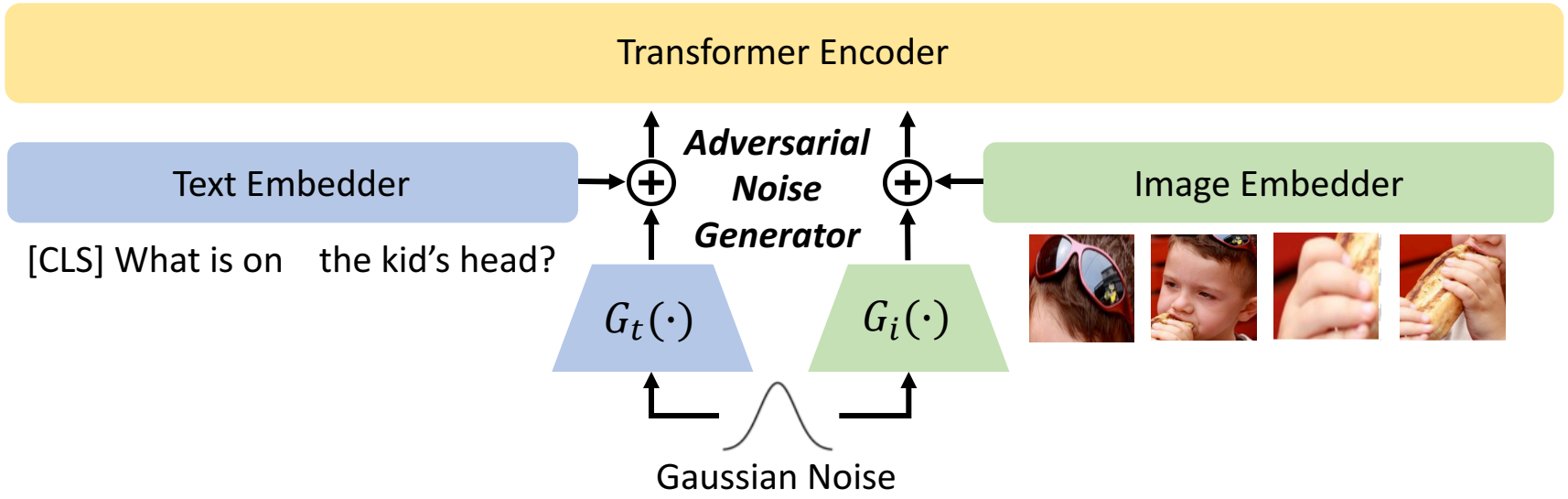
Minimize for V+L models

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}_v} \mathbb{E}_{(\boldsymbol{v},\boldsymbol{w},\boldsymbol{y})\sim\mathcal{D}} \mathbb{E}_{\boldsymbol{\alpha}\in\mathcal{N}(\mathbf{0},\mathbf{1})} [\mathcal{L}_{std}(\boldsymbol{\theta},\boldsymbol{\phi}_v) + \beta\mathcal{R}_{at}(\boldsymbol{\theta},\boldsymbol{\phi}_v)]$$

Maximize for Adv. Noise Generator

MANGO Framework

- Adversarial Noise Generator



$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}_v} \mathop{\mathbb{E}}_{(\boldsymbol{v},\boldsymbol{w},\boldsymbol{y})\sim\mathcal{D}} \mathop{\mathbb{E}}_{\boldsymbol{\alpha}\in\mathcal{N}(\mathbf{0},\mathbf{1})} [\mathcal{L}_{std}(\boldsymbol{\theta},\boldsymbol{\phi}_v) + \beta\mathcal{R}_{at}(\boldsymbol{\theta},\boldsymbol{\phi}_v)]$$

$$\mathcal{L}_{std}(\boldsymbol{\theta},\boldsymbol{\phi}_v) = \mathcal{L}_{\text{BCE}}(f_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{w}),\boldsymbol{y})$$

VQA task loss on clean inputs

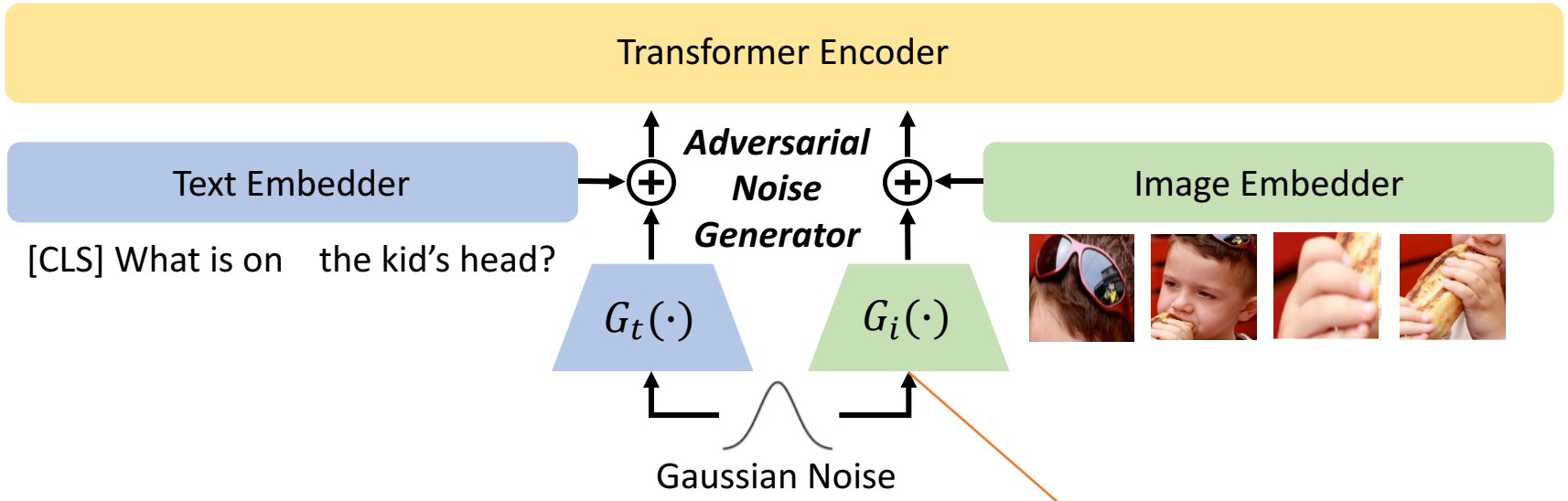
$$\mathcal{R}_{at}(\boldsymbol{\theta},\boldsymbol{\phi}_v) = \mathcal{L}_{\text{BCE}}(f_{\boldsymbol{\theta}}(\boldsymbol{v}+g_{\boldsymbol{\phi}_v}(\boldsymbol{\alpha}),\boldsymbol{w}),\boldsymbol{y}) + \mathcal{L}_{kl}(f_{\boldsymbol{\theta}}(\boldsymbol{v}+g_{\boldsymbol{\phi}_v}(\boldsymbol{\alpha}),\boldsymbol{w}),f_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{w}))$$

VQA task loss on perturbed inputs

KL Divergence between clean inputs and perturbed inputs

MANGO Framework

- Adversarial Noise Generator



$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}_v} \mathop{\mathbb{E}}_{(\boldsymbol{v},\boldsymbol{w},\boldsymbol{y})\sim\mathcal{D}} \mathop{\mathbb{E}}_{\boldsymbol{\alpha}\in\mathcal{N}(\mathbf{0},\mathbf{1})} [\mathcal{L}_{std}(\boldsymbol{\theta},\boldsymbol{\phi}_v) + \beta\mathcal{R}_{at}(\boldsymbol{\theta},\boldsymbol{\phi}_v)]$$

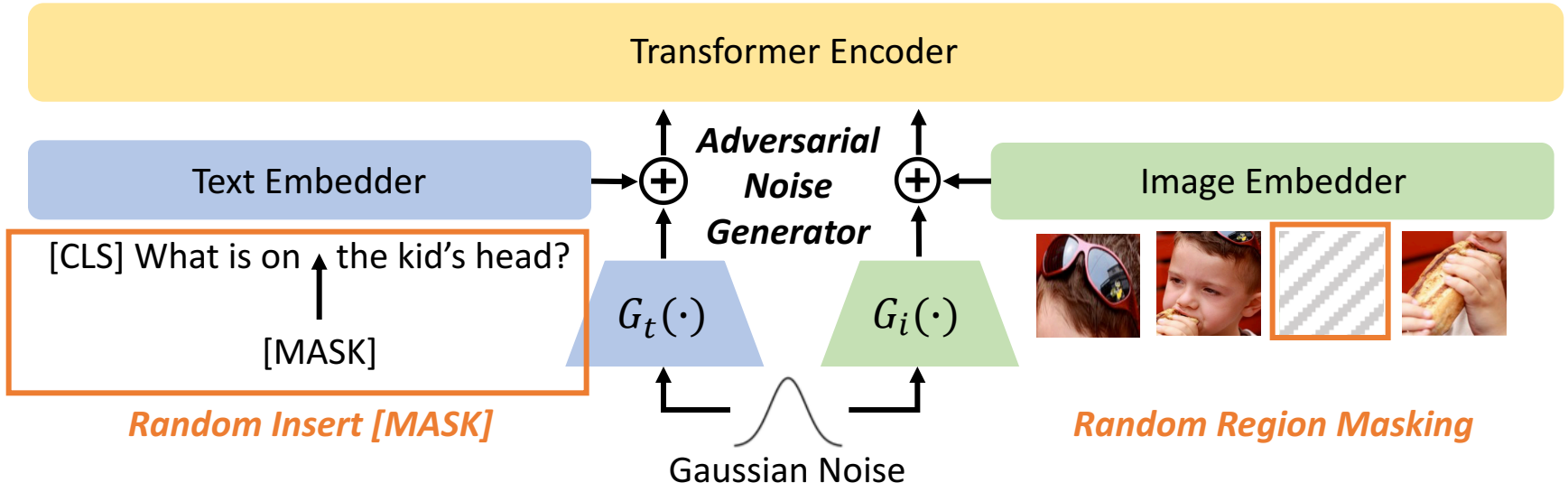
$$\mathcal{L}_{std}(\boldsymbol{\theta},\boldsymbol{\phi}_v) = \mathcal{L}_{\text{BCE}}(f_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{w}),\boldsymbol{y})$$

$$\mathcal{R}_{at}(\boldsymbol{\theta},\boldsymbol{\phi}_v) = \mathcal{L}_{\text{BCE}}(f_{\boldsymbol{\theta}}(\boldsymbol{v} + g_{\boldsymbol{\phi}_v}(\boldsymbol{\alpha}), \boldsymbol{w}),\boldsymbol{y}) + \mathcal{L}_{kl}(f_{\boldsymbol{\theta}}(\boldsymbol{v} + g_{\boldsymbol{\phi}_v}(\boldsymbol{\alpha}), \boldsymbol{w}), f_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{w}))$$

Perturbations generated via a small neural network

MANGO Framework

- Random Masking



Motivation: significant mismatch in the distribution of question lengths and image regions between training and test splits of robustness benchmarks

MANGO (Li et al. 2020)

Experimental Results

	Model		Lingual	Reason				Visual		Answer		
			VQA-Rep.	VQA-LOL Comp.	VQA-LOL Supp.	VQA-Intro.	GQA	IV-VQA	CV-VQA	VQA-CP v2	GQA-OOD	VQA v2
		Meta-Ave. ↑	Acc. ↑	Acc. ↑	Acc. ↑	M✓ S✓ ↑	Acc. ↑	#flips ↓	#flips ↓	Acc. ↑	Acc. ↑	Acc. ↑
1	SOTA	N/A	56.59	48.99	50.54	50.05	**63.17**	7.53	78.44	**69.52**	52.70	**74.69**
2	UNITER$_B$	40.98	64.56	54.54	50.00	56.80	59.99	8.47	40.67	46.93	53.43	72.70
3	MANGO$_B$	42.80	65.80	56.22	56.49	58.33	60.65	7.32	38.11	47.52	55.15	73.24
4	VILLA$_B$	42.37	*+11.74*	*+12.50*	*+9.96*	*+12.55*	60.26	*+0.84*	*+42.92*	45.39	*+3.70*	73.59
5	MANGO$_{VB}$	43.08	65.91	55.44	57.58	58.94	60.73	7.43	38.25	48.63	55.79	73.45
6	UNITER$_L$	43.37	67.64	58.60	55.95	57.64	60.30	8.20	36.66	50.98	53.65	73.82
7	MANGO$_L$	45.27	**68.33**	59.45	**60.50**	62.14	61.10	**6.69**	**35.52**	52.76	**56.40**	74.26
8	VILLA$_L$	44.33	68.16	58.66	58.29	62.00	61.38	6.70	37.55	49.10	55.26	**74.69**
9	MANGO$_{VL}$	**45.31**	68.27	**61.49**	58.83	**62.60**	61.41	6.73	35.64	52.55	56.08	74.20

- Comparison with SOTA, MANGO pushes state-of-the-art performance by a large margin on 7 out of 9 benchmarks
- On VQA-CP v2 and GQA, the SOTA methods exploit additional task-specific information (for example, scene graphs)

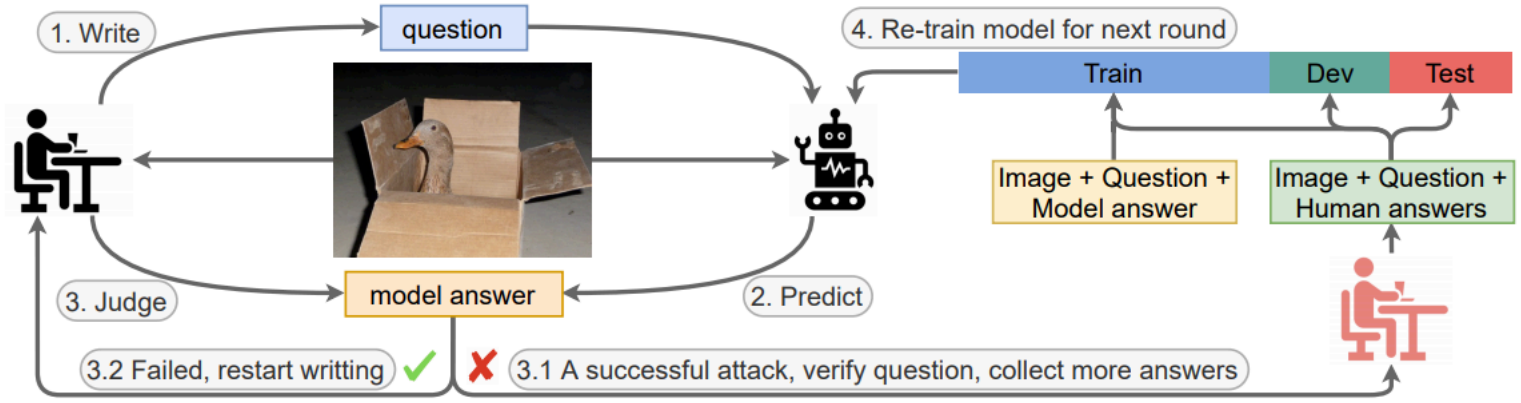
MANGO (Li et al. 2020)

What about Adversarial Robustness?

- MANGO => Adversarial VQA

Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models

Linjie Li[1], Jie Lei[2], Zhe Gan[1], Jingjing Liu[1]
[1]Microsoft [2]UNC Chapel Hill
{lindesy.li, zhe.gan, jingjl}@microsoft.com
jielei@cs.unc.edu



Adversarial VQA

Q: What surrounds the sign?

Pred: Grass

Q: What about the sign?

Pred: Nothing

Sears (Ribeiro et al. 2018)

Q: What kind of cars are featured in the picture?

Pred: Trucks

Q: What kind of automobiles are featured in the picture?

Pred: Motocycles

TextFooler (Jin et al. 2020)

Q: How many windows are on the right side of the train?

Pred: 3

Q: How many skylights are on the right side of the train?

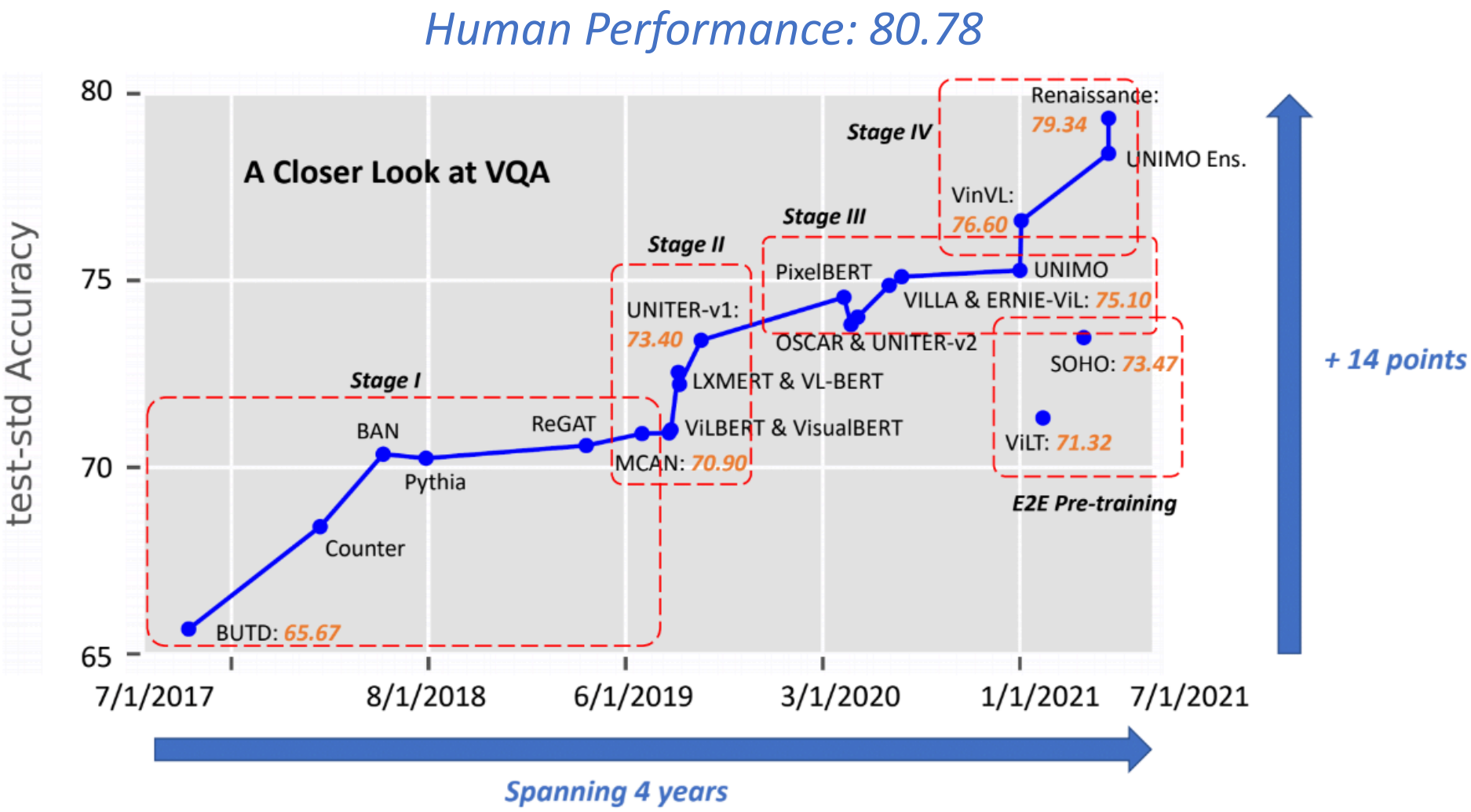
Pred: 0

Semem+PSO (Zang et al. 2020)

Automatically generated adversarial questions are often incorrect.

AVQA (Li et al. 2021)

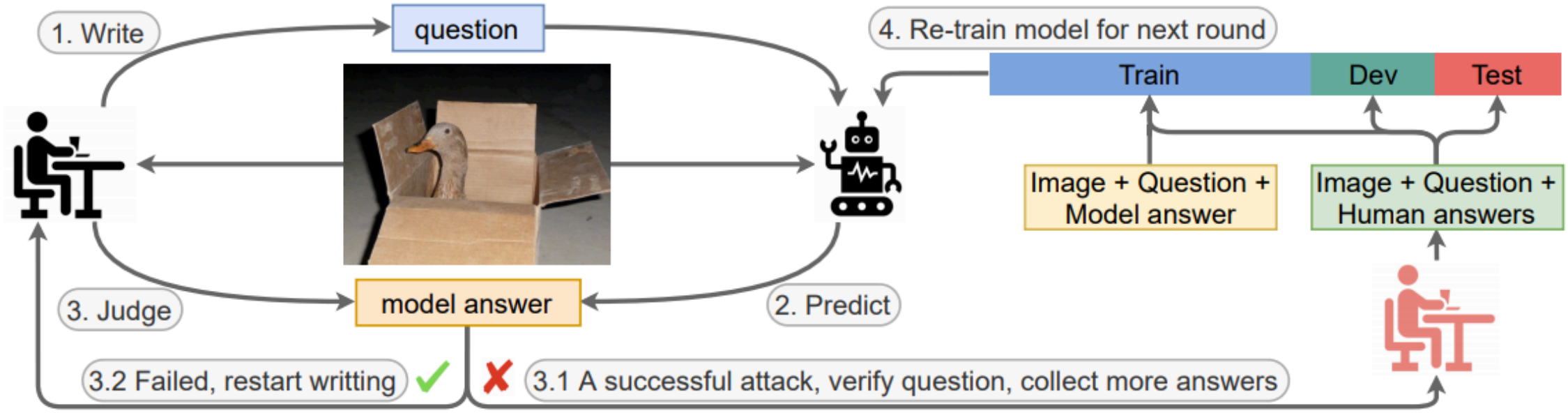
Adversarial VQA



AVQA (Li et al. 2021)

Adversarial VQA

- A *dynamically evolving adversarial* VQA benchmark with *human-and-model-in-the-loop*



Adversarial VQA

Dataset	Image Source	#Img	IsCollected	#IQ	Model error rate	#Tries	Time (sec.)	Data Split
				Total/Verified	Total/Verified	Mean/Median	per verified ex.	Train/Dev/Test
R1	CC	13.7K	✓	93.1K/45.6K	48.9%/35.2%	1.6/1	71.0	53.6K/3.3K/10.0K
R2	CC	13.1K	✓	70.4K/37.8K	56.1%/49.0%	1.5/1	54.2	42.8K/2.7K/8.3K
R3	Various	11.1K	✓	60.4K/30.0K	49.8%/36.3%	1.6/1	57.3	35.3K/2.2K/6.6K
AVQA	Various	37.9K	✓	223.9K/113.4K	50.7%/40.0%	1.6/1	61.6	131.7K/8.2K/24.9K

SOTA VQA models fail within 2 tries on average

Adversarial VQA

Model	Training Data	R1	R2	R3	AVQA	VQA v2	Δ(v2, AVQA)
		dev/test	dev/test	dev/test	dev/test	test-dev	test-dev, test
BUTD	VQA v2 +VGQA	20.80/19.28	18.77/18.85	21.17/21.31	20.18/19.60	67.60	48.00
	ALL	24.96/22.11	22.62/22.78	**24.03/23.70**	23.94/22.71	67.52	**44.81**
UNITER-B	VQA v2 +VGQA	20.60/17.91	17.86/18.55	19.45/20.20	19.39/18.66	72.70	54.04
	+R1	26.03/22.94	17.30/17.36	19.41/20.23	21.48/20.37	72.98	52.61
	+R1+R2	26.60/24.76	23.21/23.86	18.60/19.09	23.58/23.14	72.75	49.61
	ALL	26.85/24.93	23.38/23.92	23.76/23.02	24.94/24.14	72.66	48.52
UNITER-L	VQA v2 +VGQA	25.04/23.72	17.82/17.49	18.86/19.34	21.11/20.54	73.82	53.28
	+R1	29.31/26.63	19.34/18.66	19.53/18.30	23.60/21.93	73.89	51.96
	+R1+R2	30.13/28.15	23.11/23.54	16.09/16.47	24.46/23.84	73.77	49.93
	ALL	**30.80/28.45**	22.95/23.11	23.75/21.88	**26.11/25.07**	**74.15**	49.08
LXMERT	VQA v2 +VGQA	19.76/18.15	18.98/18.79	20.75/20.26	19.72/18.86	72.31	53.45
	+R1	23.89/22.65	19.01/17.91	21.47/20.85	21.64/20.58	72.51	51.93
	+R1+R2	26.76/24.86	23.28/24.11	19.16/18.93	23.80/23.23	72.61	49.38
	ALL	26.35/24.55	**23.84/24.02**	24.00/22.90	24.94/23.98	72.42	48.44

Adversarial VQA

Round	Count	OCR	Reasoning				Visual Concept Recognition				
			Position	Relation	Common-sense	Other	Low-level	Action	Small Object	Occlusion	Abstract
R1	23.3%	10.7%	14.7%	8.3%	17.3%	0.7%	9.7%	4.3%	13.3%	14.7%	6.3%
R2	30.0%	22.7%	12.0%	27.7%	20.0%	4.3%	12.7%	9.3%	22.7%	10.0%	15.3%
R3	35.3%	13.0%	13.0%	28.3%	25.0%	6.3%	11.7%	4.3%	20.0%	20.0%	6.0%
Ave.	29.6%	15.4%	13.2%	21.4%	20.8%	3.8%	11.3%	6.0%	18.7%	14.9%	9.2%

AVQA includes diverse question types.

Adversarial VQA

Round	Count	OCR	Reasoning				Visual Concept Recognition				
			Position	Relation	Common-sense	Other	Low-level	Action	Small Object	Occlusion	Abstract
R1	23.3%	10.7%	14.7%	8.3%	17.3%	0.7%	9.7%	4.3%	13.3%	14.7%	6.3%
R2	30.0%	22.7%	12.0%	27.7%	20.0%	4.3%	12.7%	9.3%	22.7%	10.0%	15.3%
R3	35.3%	13.0%	13.0%	28.3%	25.0%	6.3%	11.7%	4.3%	20.0%	20.0%	6.0%
Ave.	29.6%	15.4%	13.2%	21.4%	20.8%	3.8%	11.3%	6.0%	18.7%	14.9%	9.2%

Q1: What creature is in the box?
A1: Bird, Conf: 53.2%

Q2: What is the box made of?
A2: Wood, Conf: 99.8%
A2: Carboard (0.9), Paper (0.3)

R1

Q1: How many football players can you see off the ground?
A1: 0, Conf: 92.2%
A1: 1 (1.0)

R2

Q1: Is the person next to the dog standing or squatting down?
A1: walking, Conf: 15.4%
A1: squatting down (1.0)

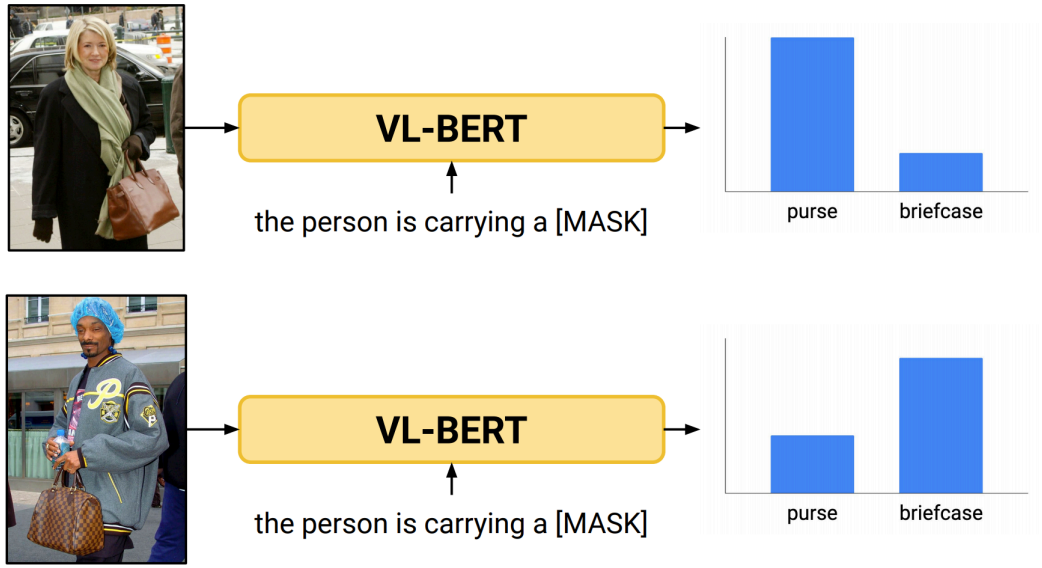
R3

What about Fairness?

Worst of Both Worlds:
Biases Compound in Pre-trained Vision-and-Language Models

Tejas Srinivasan
University of Southern California
tejas.srinivasan@usc.edu

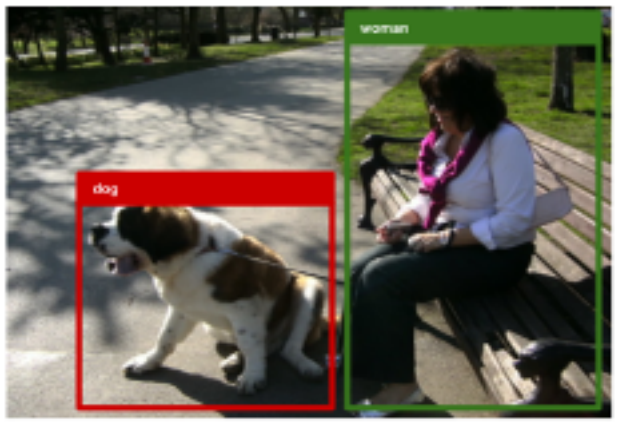
Yonatan Bisk
Carnegie Mellon University
ybisk@cs.cmu.edu



Agenda

- Advanced training strategies in VLP
- Diverse applications of VLP
- VL for V/L
- Model Compression
- Robustness/fairness of VLP models
- Multilingual VLP

Multilingual VLP



Visual Question Answering

What is the animal in the picture? **dog**

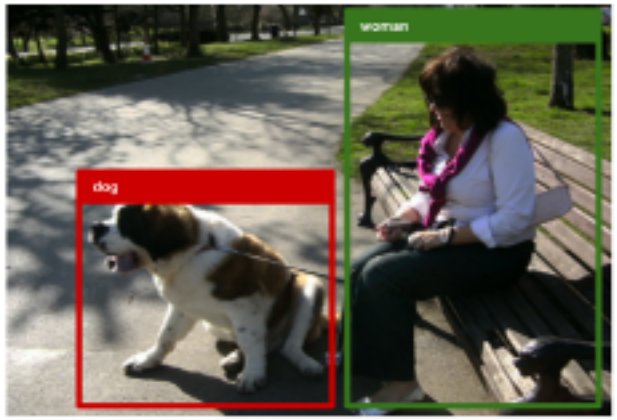
Multi-modal Verification

The woman is sitting next to a cat. **false**

Caption-based Image Retrieval

A woman sitting on a bench next to a dog.

Multilingual VLP



Visual Question Answering

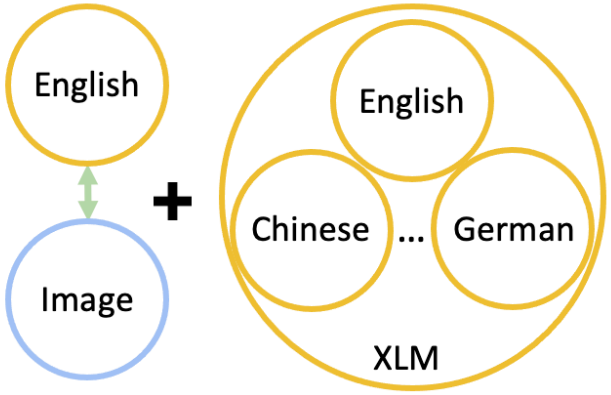
What is the animal in the picture? **dog**

Multi-modal Verification

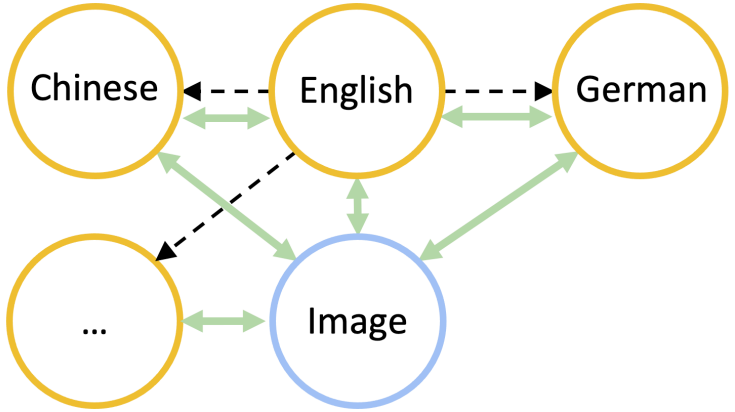
The woman is sitting next to a cat. **false**

Caption-based Image Retrieval

A woman sitting on a bench next to a dog.



M3P (Ni et al. 2021)



UC2 (Zhou et al. 2021)

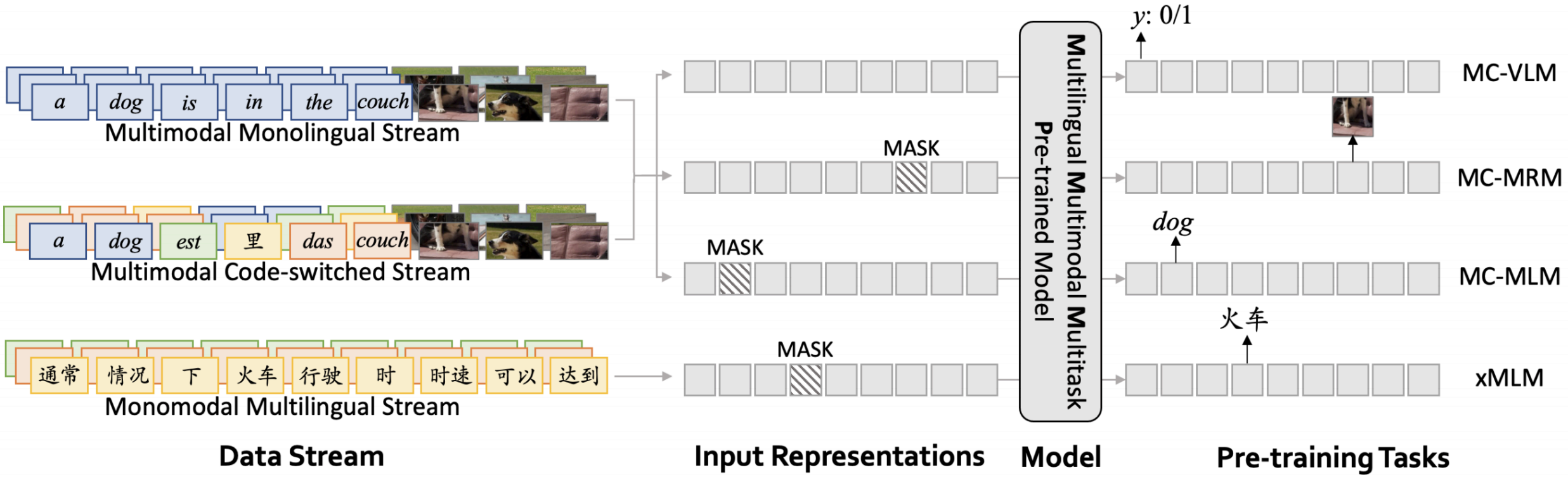
Cross-modal Encoder

Translation

Language Encoder

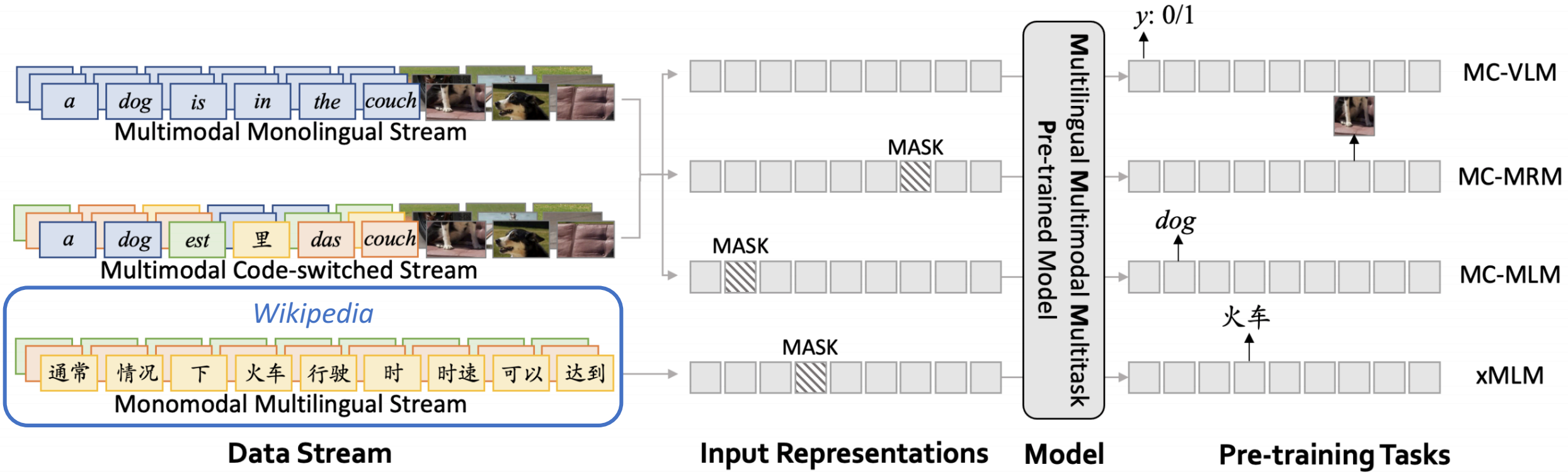
Image Encoder

M3P: Multitask Multilingual Multimodal Pre-training



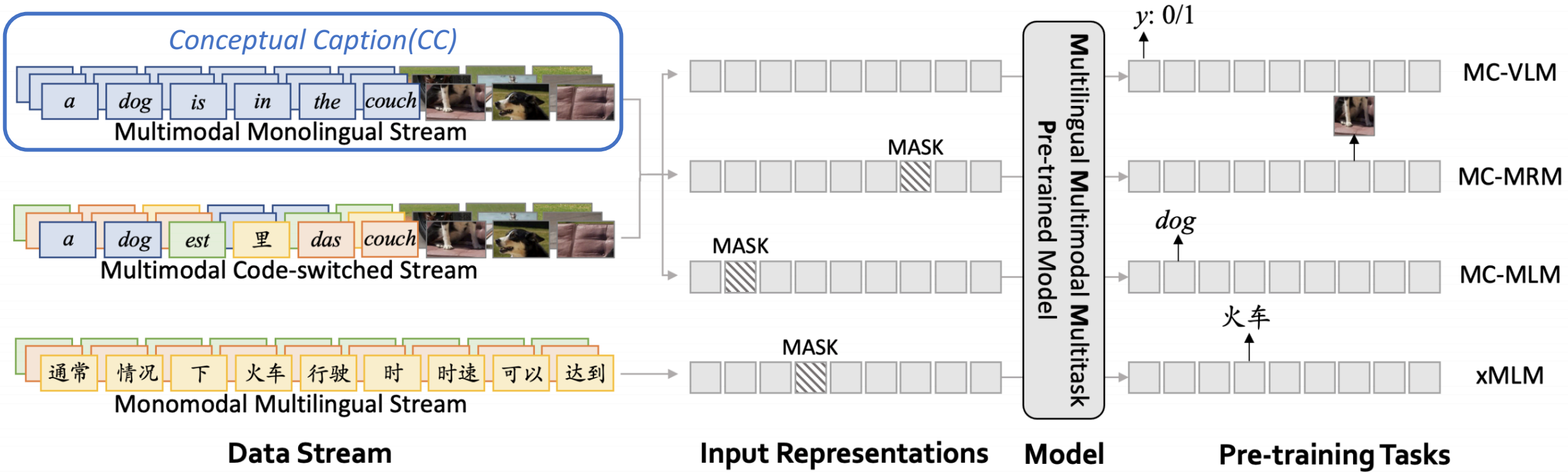
M3P (Ni et al. 2021)

M3P: Multitask Multilingual Multimodal Pre-training



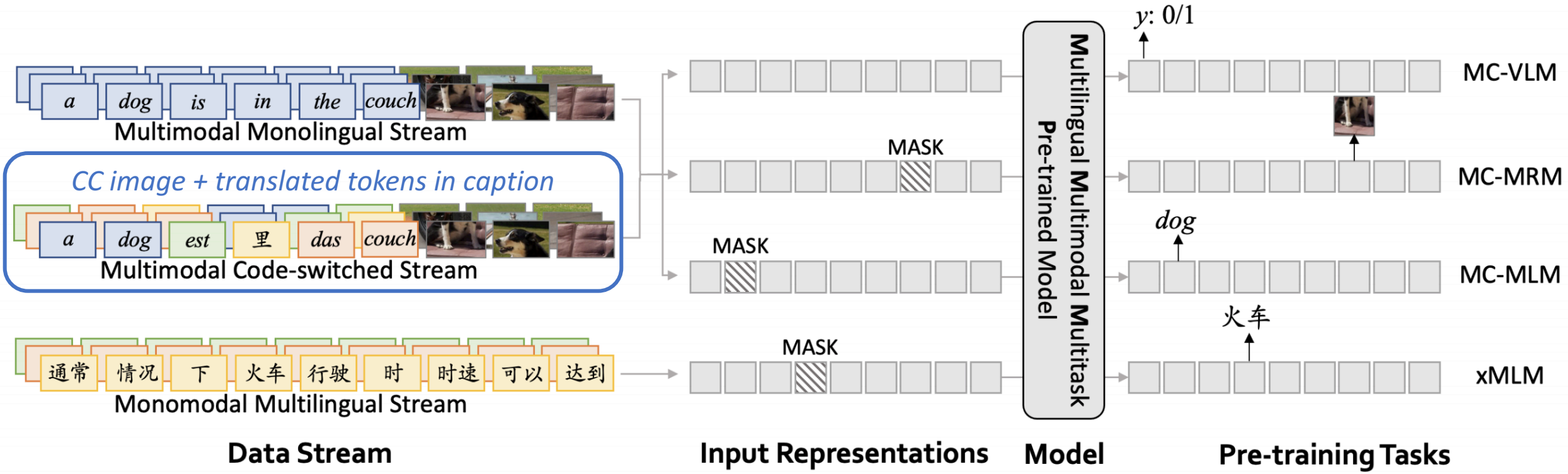
M3P (Ni et al. 2021)

M3P: Multitask Multilingual Multimodal Pre-training



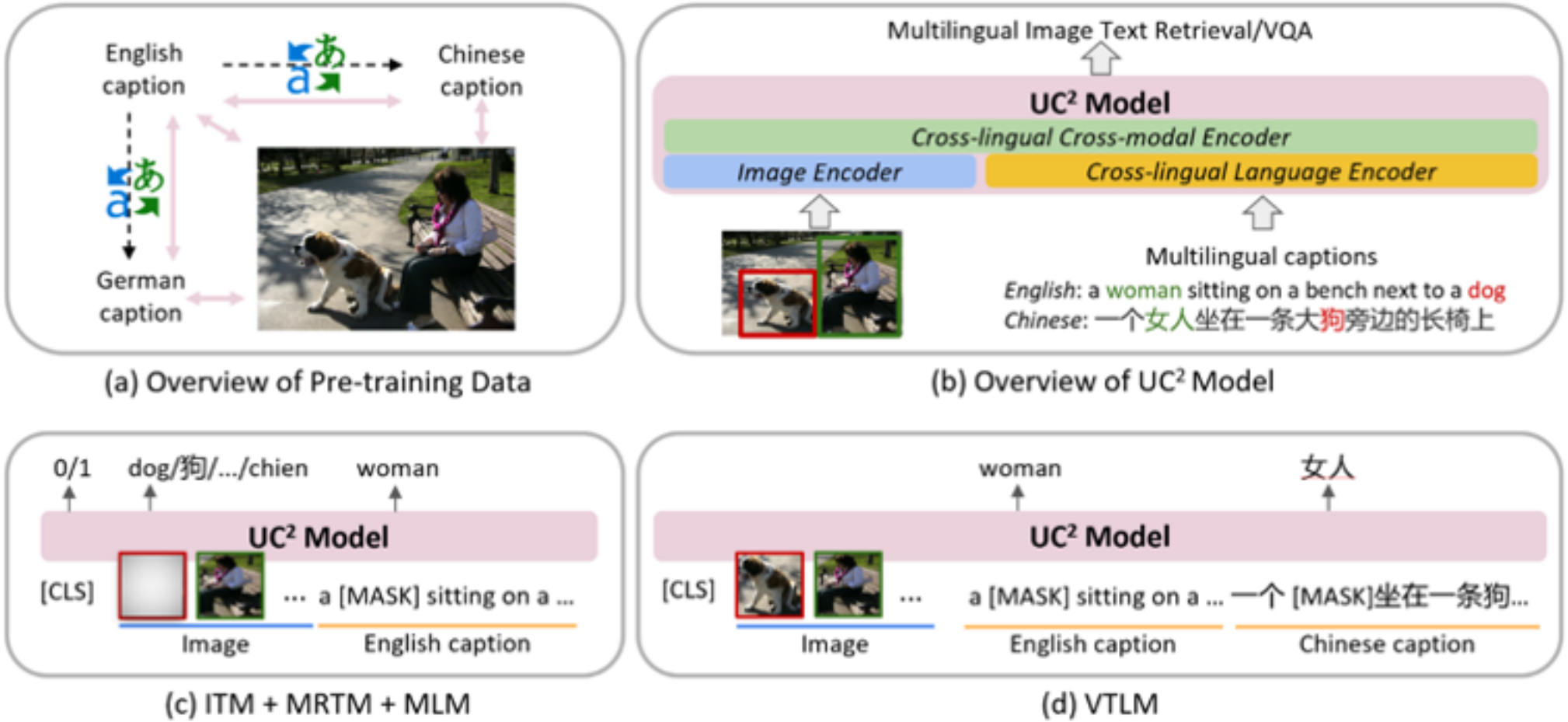
M3P (Ni et al. 2021)

M3P: Multitask Multilingual Multimodal Pre-training



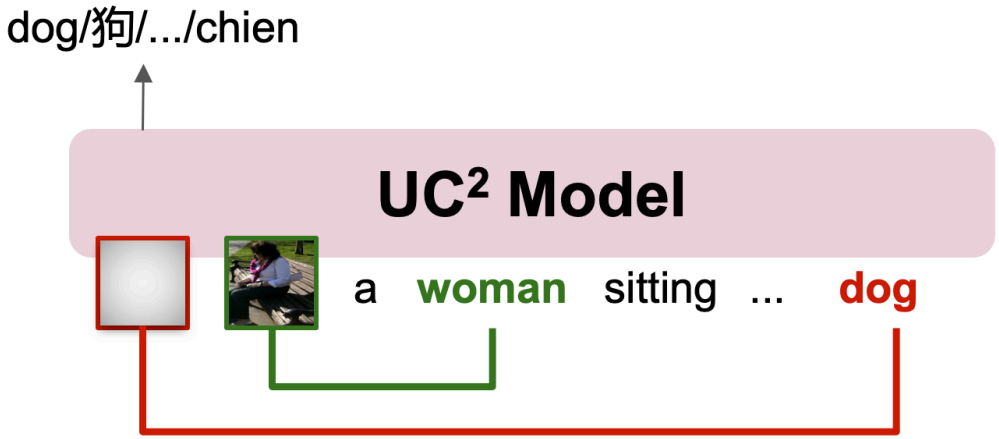
M3P (Ni et al. 2021)

UC2: Universal Cross-lingual Cross-modal Pre-training



UC2 (Zhou et al. 2021)

UC2: Masked Region-to-Token Modeling



English: a **woman** sitting on a bench next to a **dog**

Chinese: 一个**女人**坐在一条大**狗**旁边的长椅上

Japanese: **犬**の隣のベンチに座っている**女性**

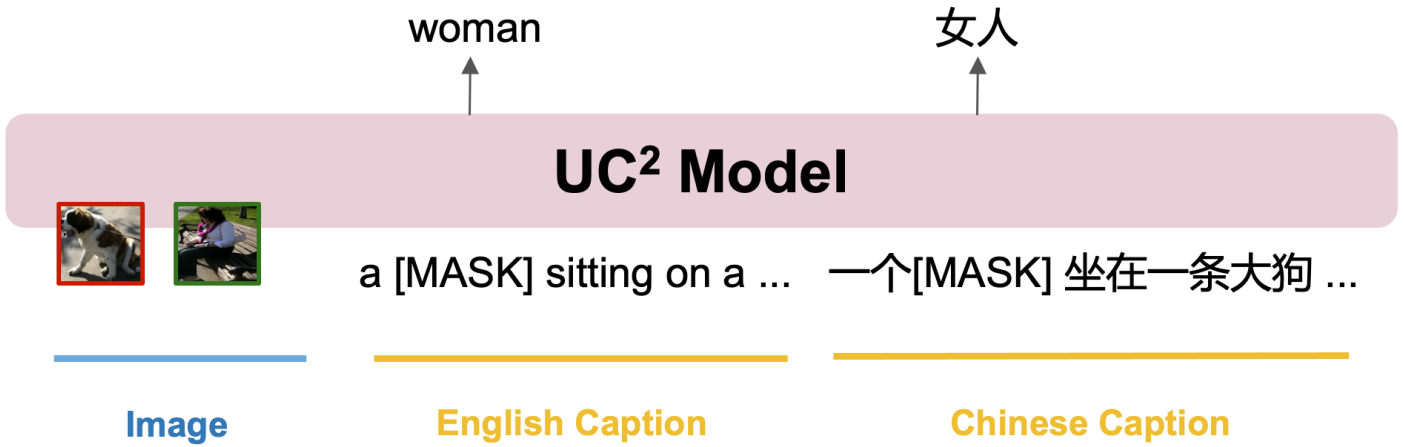
French: une **femme** assise sur un banc à côté d'un gros **chien**

...

UC2 (Zhou et al. 2021)

Credit to: Mingyang Zhou

UC2: Visual Translation Language Modeling



English: a **woman** sitting on a bench next to a **dog**

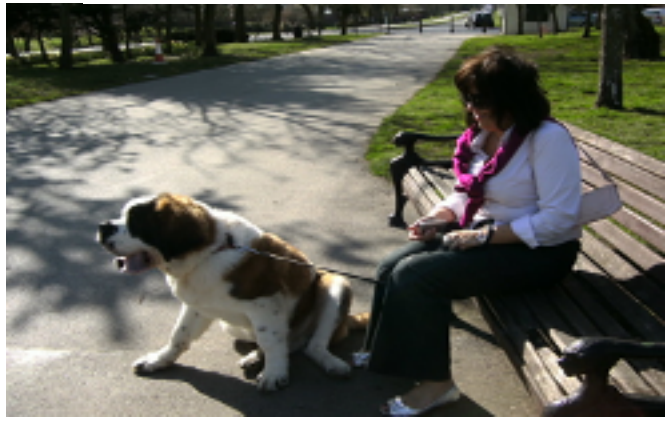
Chinese: 一个 **女人** 坐在 一条 大**狗** 旁边的 长椅 上

UC2 (Zhou et al. 2021)

Credit to: Mingyang Zhou

UC2: Experimental Results

	Flickr30K				MSCOCO			
Method	**EN**	**DE**	**FR**	**CS**	**EN**	**ZH**	**JA**	**Meta-Ave**
SMALR	74.5	69.8	65.9	64.8	81.5	77.5	76.7	73.0
M^{3}P	87.7	82.7	73.9	72.2	**88.7**	86.2	**87.9**	82.8
UNITER	87.7	81.2	81.9	80.2	88.4	87.3	82.2	84.1
UC^2	**88.2**	**84.5**	**83.9**	**81.2**	88.1	**89.8**	87.5	**86.2**



English: a woman sitting on a bench next to a dog

Chinese: 一个女人坐在一条大狗旁边的长椅上

Japanese: 犬の隣のベンチに座っている女性

French: une femme assise sur un banc à côté d'un gros chien

UC2 (Zhou et al. 2021)

Credit to: Mingyang Zhou

Conclusion

- Advanced training strategies in VLP
- Diverse applications of VLP
- VL for V/L
- Compressing VLP models
- Robustness/fairness of VLP models
- Multilingual VLP

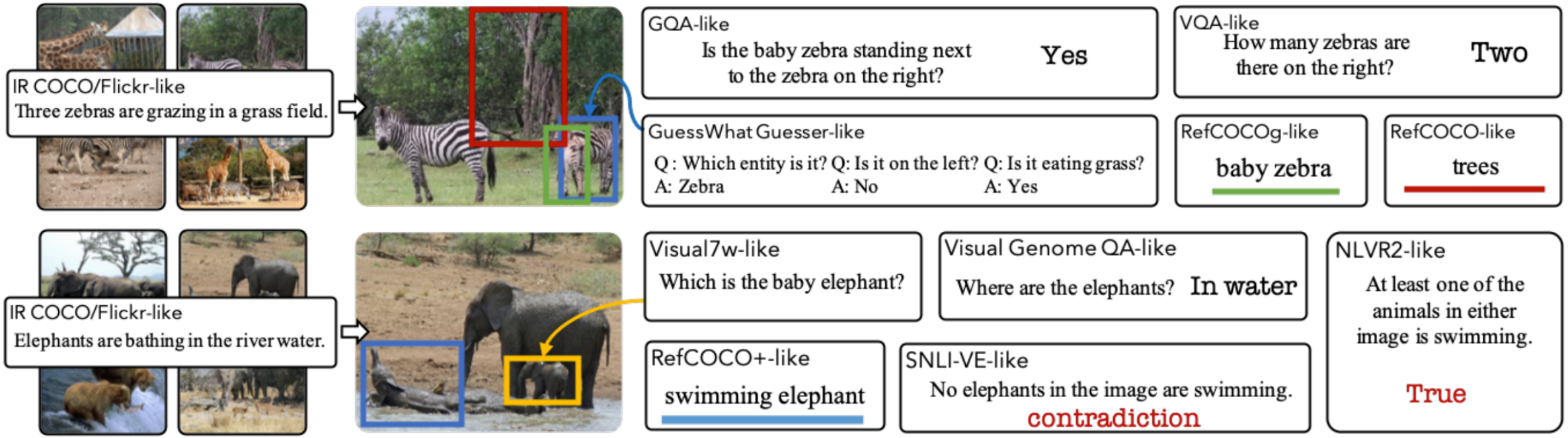
Challenges and Future Directions

- *Fairness*:
  - We observe that there are severe biases in VLP models. How can we improve the fairness of VLP models?
- *Adversarial* robustness:
  - We observe that VLP models can be easily attacked. How can we enhance the adversarial robustness of VLP models?
- *Training* efficiency:
  - How can we obtain training efficiency rather than inference/parameter efficiency? This could be especially useful for pre-training.

Thank you!

CVPR 2021 Tutorial

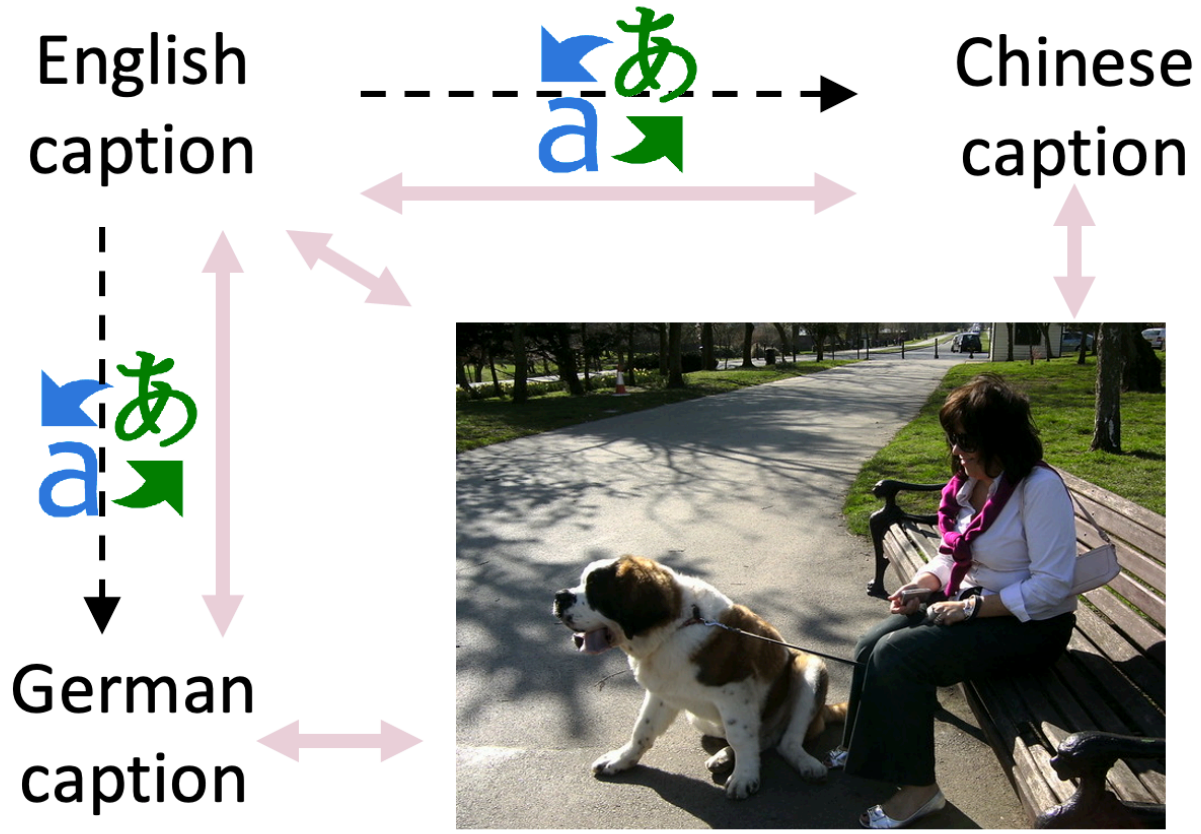
Multi-task training



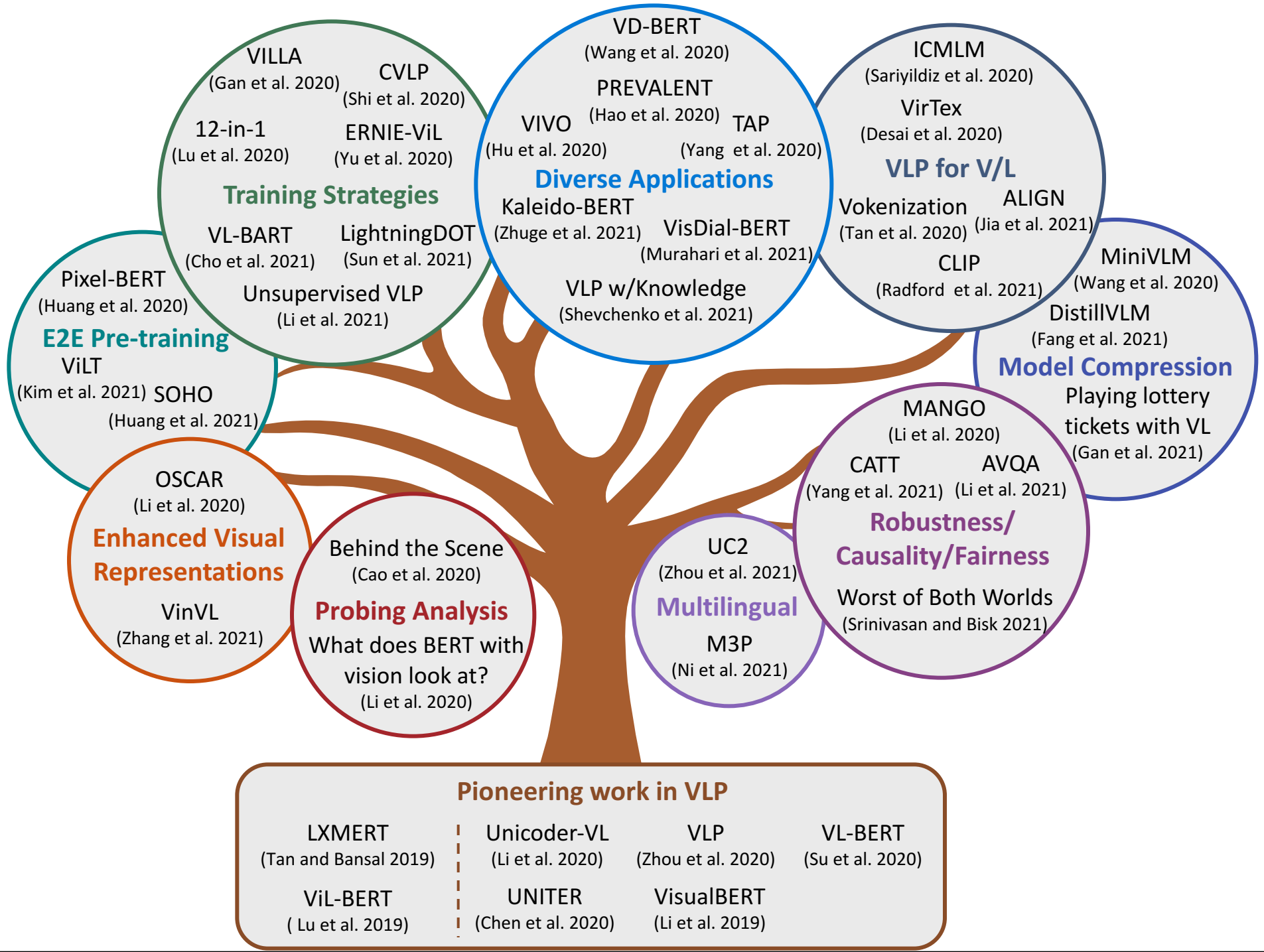
Model	# models	# parameters	Average Performance
Independently train for each task	12	12X270M = 3B	67.24
Train all tasks together (12-in-1)	**1**	**270M**	69.08
Further fineune on each task	12	12X270M = 3B	**70.24** (+3)

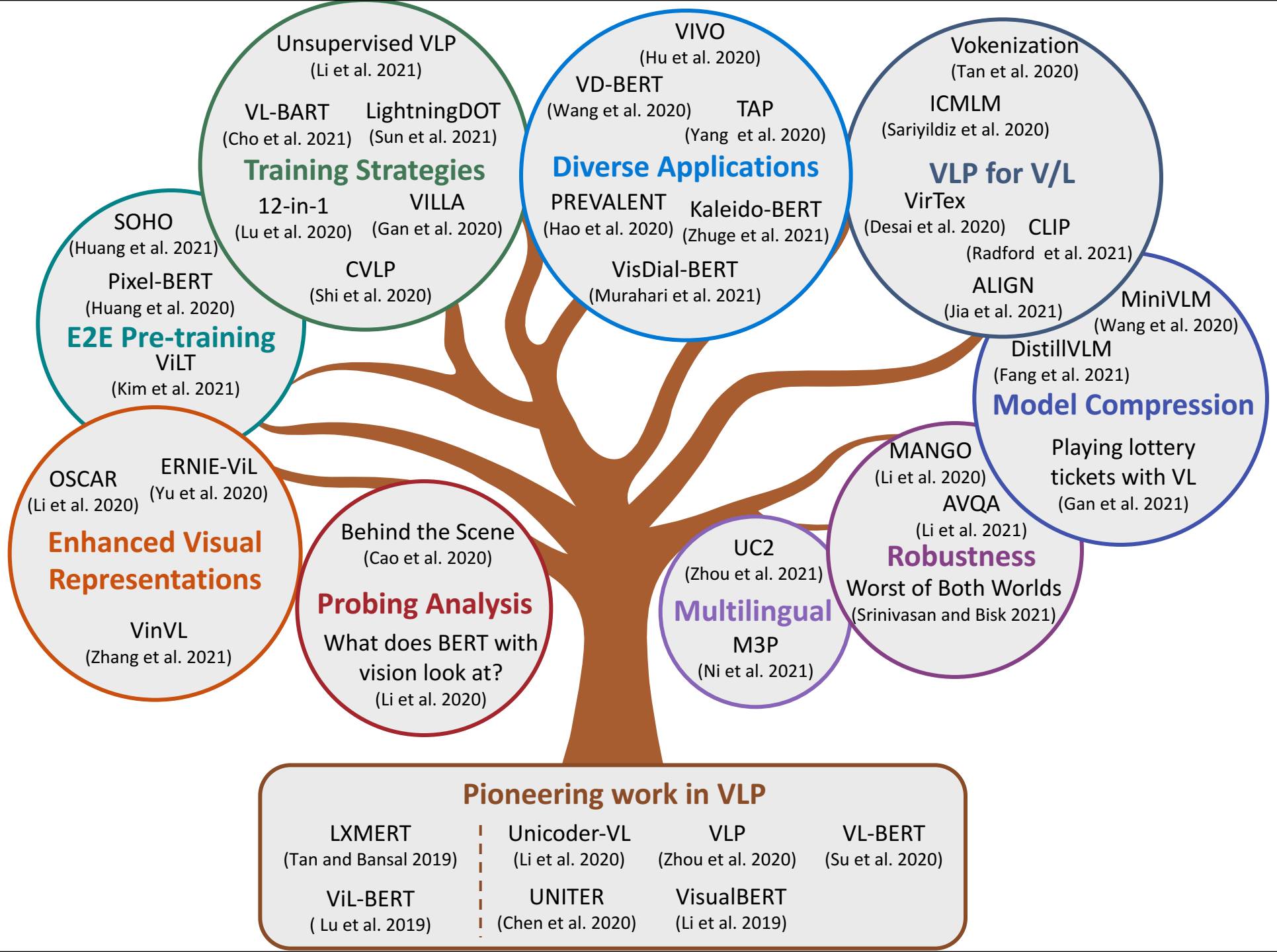
12-in-1 (Lu et al. 2020)

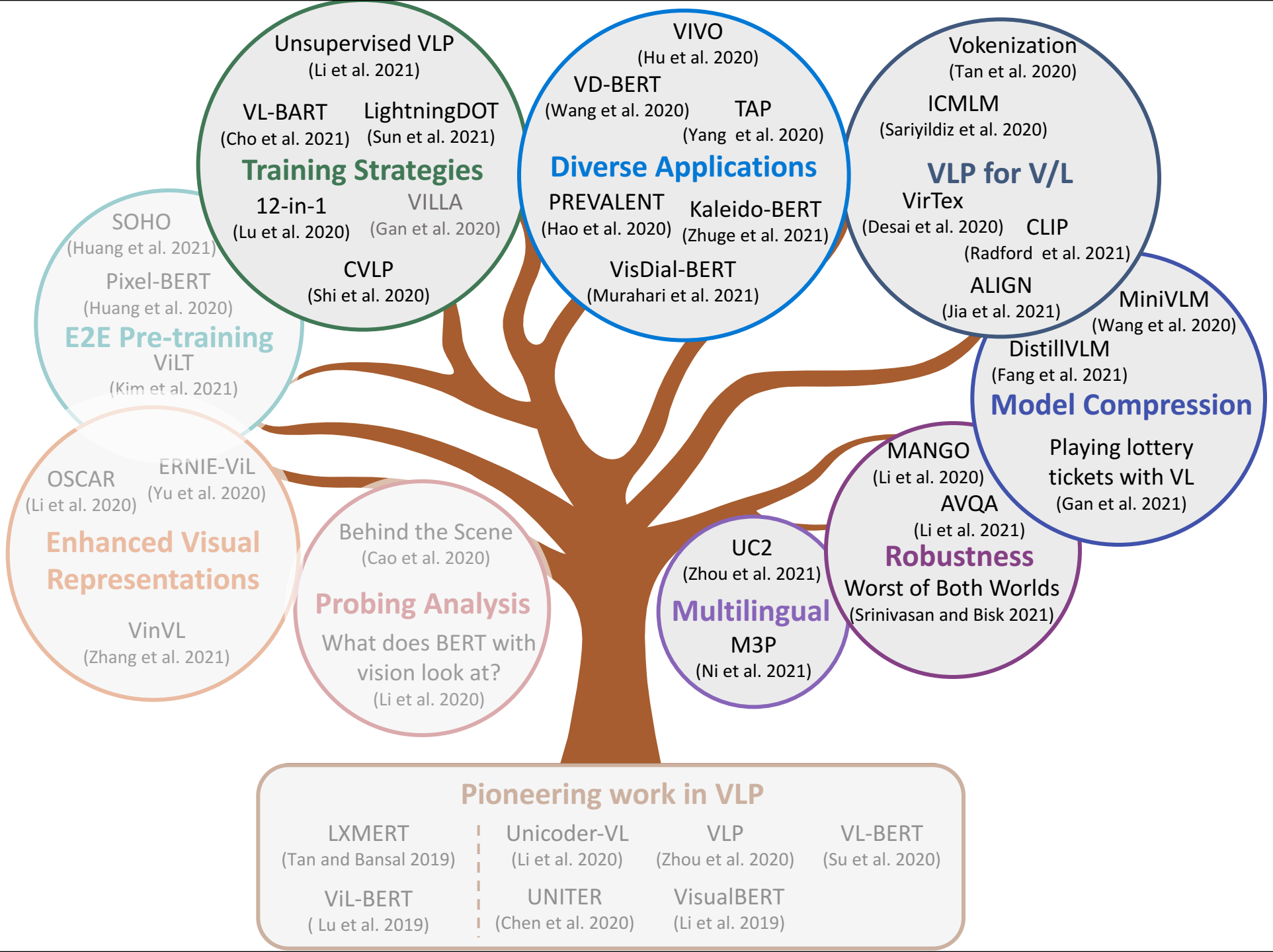
UC2: Data Augmentation via Machine Translation



UC2 (Zhou et al. 2021)


VILLA
(Gan et al. 2020)
CVLP
(Shi et al. 2020)
12-in-1
(Lu et al. 2020)
ERNIE-ViL
(Yu et al. 2020)
Training Strategies
VL-BART
(Cho et al. 2021)
LightningDOT
(Sun et al. 2021)
Unsupervised VLP
(Li et al. 2021)
VD-BERT
(Wang et al. 2020)
PREVALENT
(Hao et al. 2020)
VIVO
(Hu et al. 2020)
TAP
(Yang et al. 2020)
Diverse Applications
Kaleido-BERT
(Zhuge et al. 2021)
VisDial-BERT
(Murahari et al. 2021)
VLP w/Knowledge
(Shevchenko et al. 2021)
ICMLM
(Sariyildiz et al. 2020)
VirTex
(Desai et al. 2020)
VLP for V/L
Vokenization
(Tan et al. 2020)
ALIGN
(Jia et al. 2021)
CLIP
(Radford et al. 2021)
Pixel-BERT
(Huang et al. 2020)
E2E Pre-training
ViLT
(Kim et al. 2021)
SOHO
(Huang et al. 2021)
MiniVLM
(Wang et al. 2020)
DistillVLM
(Fang et al. 2021)
Model Compression
Playing lottery tickets with VL
(Gan et al. 2021)
OSCAR
(Li et al. 2020)
Enhanced Visual Representations
VinVL
(Zhang et al. 2021)
Behind the Scene
(Cao et al. 2020)
Probing Analysis
What does BERT with vision look at?
(Li et al. 2020)
UC2
(Zhou et al. 2021)
Multilingual
M3P
(Ni et al. 2021)
MANGO
(Li et al. 2020)
CATT
(Yang et al. 2021)
AVQA
(Li et al. 2021)
Robustness/ Causality/Fairness
Worst of Both Worlds
(Srinivasan and Bisk 2021)
Pioneering work in VLP
LXMERT
(Tan and Bansal 2019)
Unicoder-VL
(Li et al. 2020)
VLP
(Zhou et al. 2020)
VL-BERT
(Su et al. 2020)
ViL-BERT
(Lu et al. 2019)
UNITER
(Chen et al. 2020)
VisualBERT
(Li et al. 2019)


Unsupervised VLP
(Li et al. 2021)
VL-BART
(Cho et al. 2021)
LightningDOT
(Sun et al. 2021)
Training Strategies
12-in-1
(Lu et al. 2020)
VILLA
(Gan et al. 2020)
CVLP
(Shi et al. 2020)
VIVO
(Hu et al. 2020)
VD-BERT
(Wang et al. 2020)
TAP
(Yang et al. 2020)
Diverse Applications
PREVALENT
(Hao et al. 2020)
Kaleido-BERT
(Zhuge et al. 2021)
VisDial-BERT
(Murahari et al. 2021)
Vokenization
(Tan et al. 2020)
ICMLM
(Sariyildiz et al. 2020)
VLP for V/L
VirTex
(Desai et al. 2020)
CLIP
(Radford et al. 2021)
ALIGN
(Jia et al. 2021)
SOHO
(Huang et al. 2021)
Pixel-BERT
(Huang et al. 2020)
E2E Pre-training
ViLT
(Kim et al. 2021)
MiniVLM
(Wang et al. 2020)
DistillVLM
(Fang et al. 2021)
Model Compression
Playing lottery tickets with VL
(Gan et al. 2021)
OSCAR
(Li et al. 2020)
ERNIE-ViL
(Yu et al. 2020)
Enhanced Visual Representations
VinVL
(Zhang et al. 2021)
Behind the Scene
(Cao et al. 2020)
Probing Analysis
What does BERT with vision look at?
(Li et al. 2020)
UC2
(Zhou et al. 2021)
Multilingual
M3P
(Ni et al. 2021)
MANGO
(Li et al. 2020)
AVQA
(Li et al. 2021)
Robustness
Worst of Both Worlds
(Srinivasan and Bisk 2021)
Pioneering work in VLP
LXMERT
(Tan and Bansal 2019)
Unicoder-VL
(Li et al. 2020)
VLP
(Zhou et al. 2020)
VL-BERT
(Su et al. 2020)
ViL-BERT
(Lu et al. 2019)
UNITER
(Chen et al. 2020)
VisualBERT
(Li et al. 2019)


Unsupervised VLP
(Li et al. 2021)
VL-BART
(Cho et al. 2021)
LightningDOT
(Sun et al. 2021)
Training Strategies
12-in-1
(Lu et al. 2020)
VILLA
(Gan et al. 2020)
CVLP
(Shi et al. 2020)
VIVO
(Hu et al. 2020)
VD-BERT
(Wang et al. 2020)
TAP
(Yang et al. 2020)
Diverse Applications
PREVALENT
(Hao et al. 2020)
Kaleido-BERT
(Zhuge et al. 2021)
VisDial-BERT
(Murahari et al. 2021)
Vokenization
(Tan et al. 2020)
ICMLM
(Sariyildiz et al. 2020)
VLP for V/L
VirTex
(Desai et al. 2020)
CLIP
(Radford et al. 2021)
ALIGN
(Jia et al. 2021)
SOHO
(Huang et al. 2021)
Pixel-BERT
(Huang et al. 2020)
E2E Pre-training
ViLT
(Kim et al. 2021)
MiniVLM
(Wang et al. 2020)
DistillVLM
(Fang et al. 2021)
Model Compression
Playing lottery tickets with VL
(Gan et al. 2021)
OSCAR
(Li et al. 2020)
ERNIE-ViL
(Yu et al. 2020)
Enhanced Visual Representations
VinVL
(Zhang et al. 2021)
Behind the Scene
(Cao et al. 2020)
Probing Analysis
What does BERT with vision look at?
(Li et al. 2020)
UC2
(Zhou et al. 2021)
Multilingual
M3P
(Ni et al. 2021)
MANGO
(Li et al. 2020)
AVQA
(Li et al. 2021)
Robustness
Worst of Both Worlds
(Srinivasan and Bisk 2021)
Pioneering work in VLP
LXMERT
(Tan and Bansal 2019)
Unicoder-VL
(Li et al. 2020)
VLP
(Zhou et al. 2020)
VL-BERT
(Su et al. 2020)
ViL-BERT
(Lu et al. 2019)
UNITER
(Chen et al. 2020)
VisualBERT
(Li et al. 2019)