Multimodal AI

JJ Liu

Senior Principal Research Manager
Microsoft

Collaborators

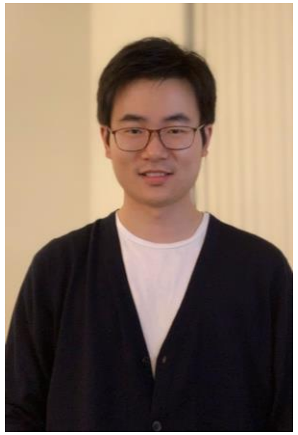

Lindsey Li


Yen-Chun Chen


Zhe Gan


Yu Cheng


Jize Cao


Licheng Yu


Jingzhou Liu

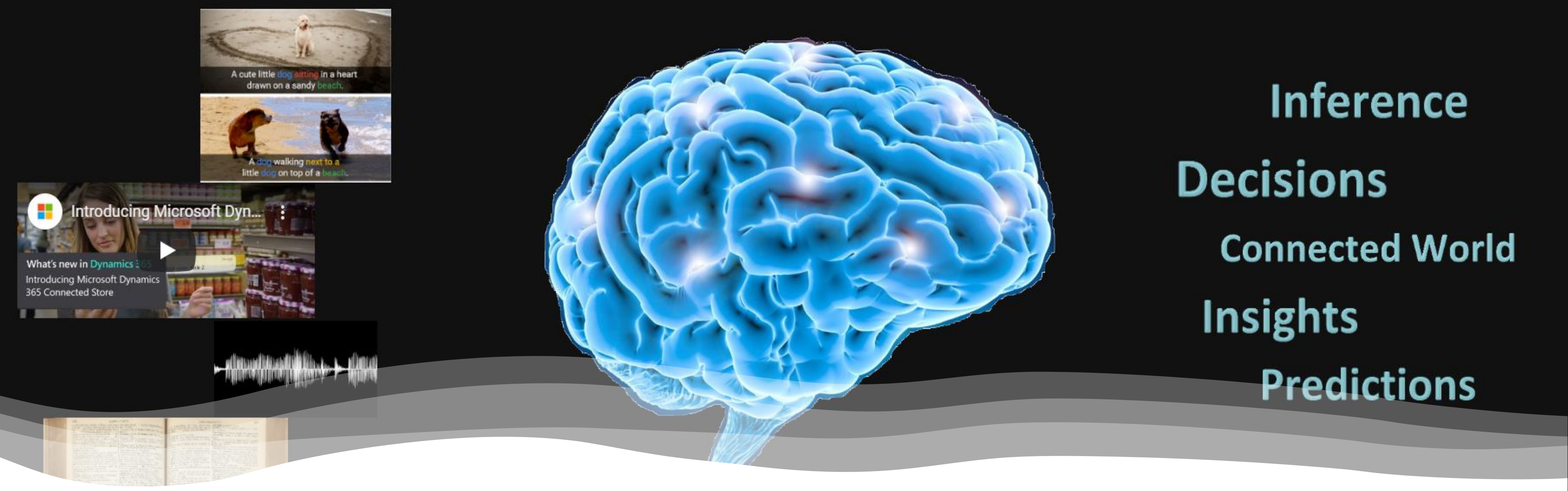

Wenhu Chen


Yandong Li


Chen Zhu


Ahmed El Kholy


Faisal Ahmed



Self-supervised Learning for Multimodal Pre-training

Large-scale Adversarial Training for Vision+Language

AI Explainability and Interpretability

Vision-and-Language Inference

High-Resolution Image Synthesis

Self-supervised Learning for Multimodal Pre-training

UNITER: Universial Image-Text Representation

Self-Supervised Learning for Computer Vision

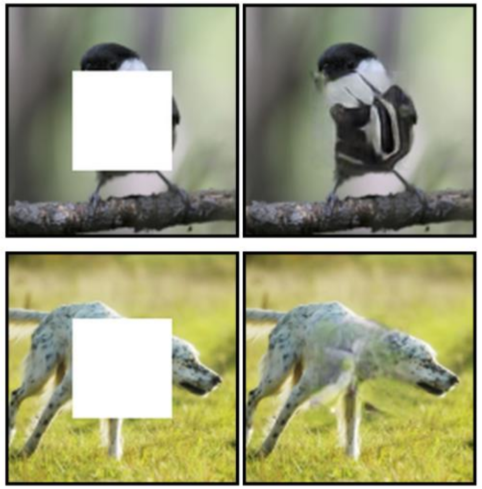
Image Colorization

[Zhang et al. ECCV 2016]

Jigsaw puzzles

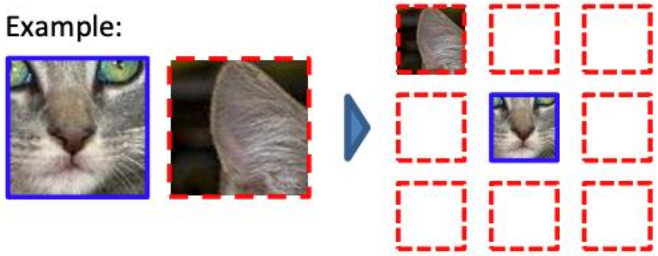
[Noroozi et al. ECCV 2016]

Image Inpainting



[Pathak et al. CVPR 2016]

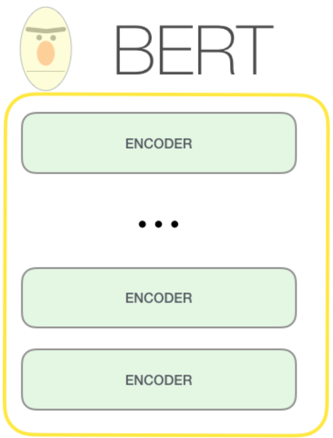
Relative Location Prediction



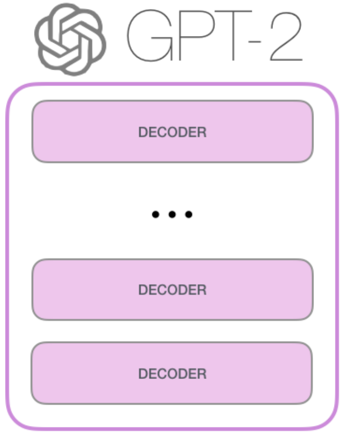
[Doersch et al. ICCV 2015]

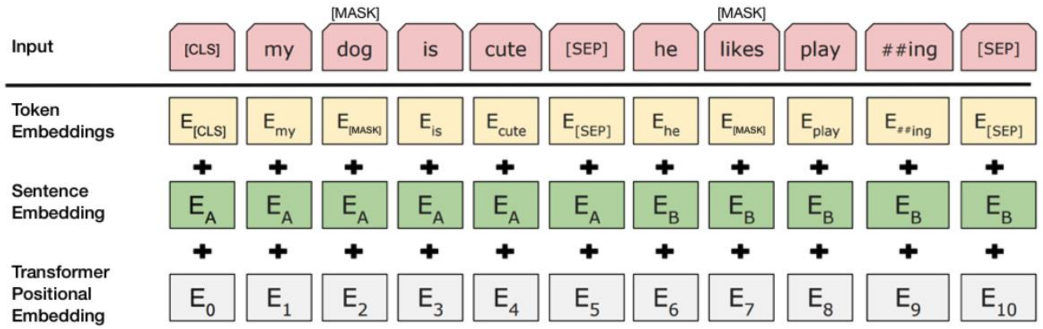
Self-Supervised Learning for NLP

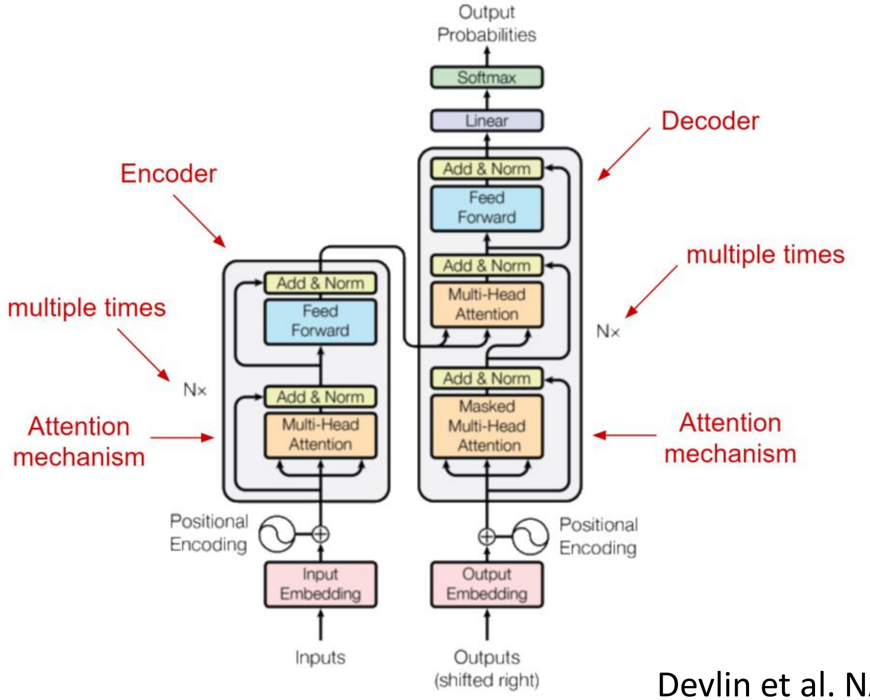
Language Understanding



Language Generation

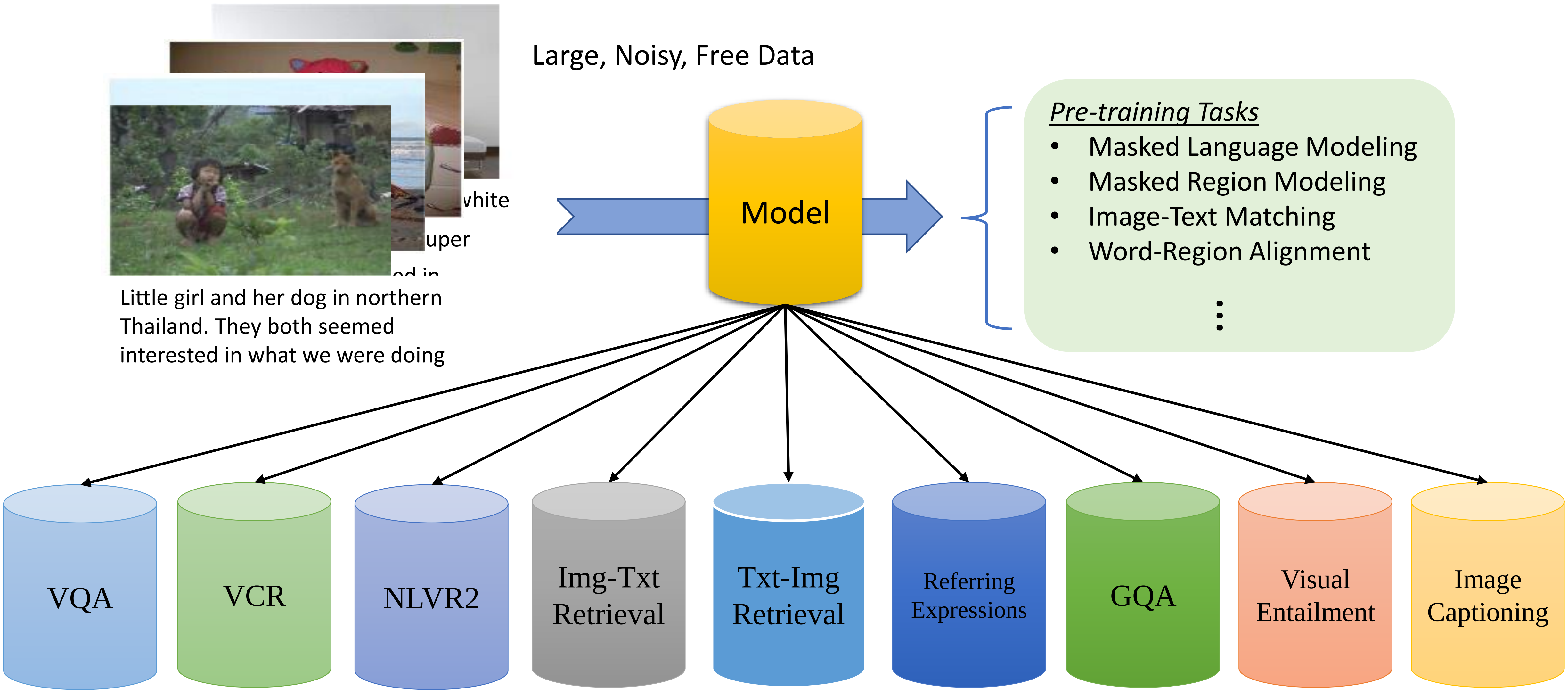






Devlin et al. NAACL 2019; Radford et al. 2019

Self-Supervised Learning for Vision+Language



Landscape

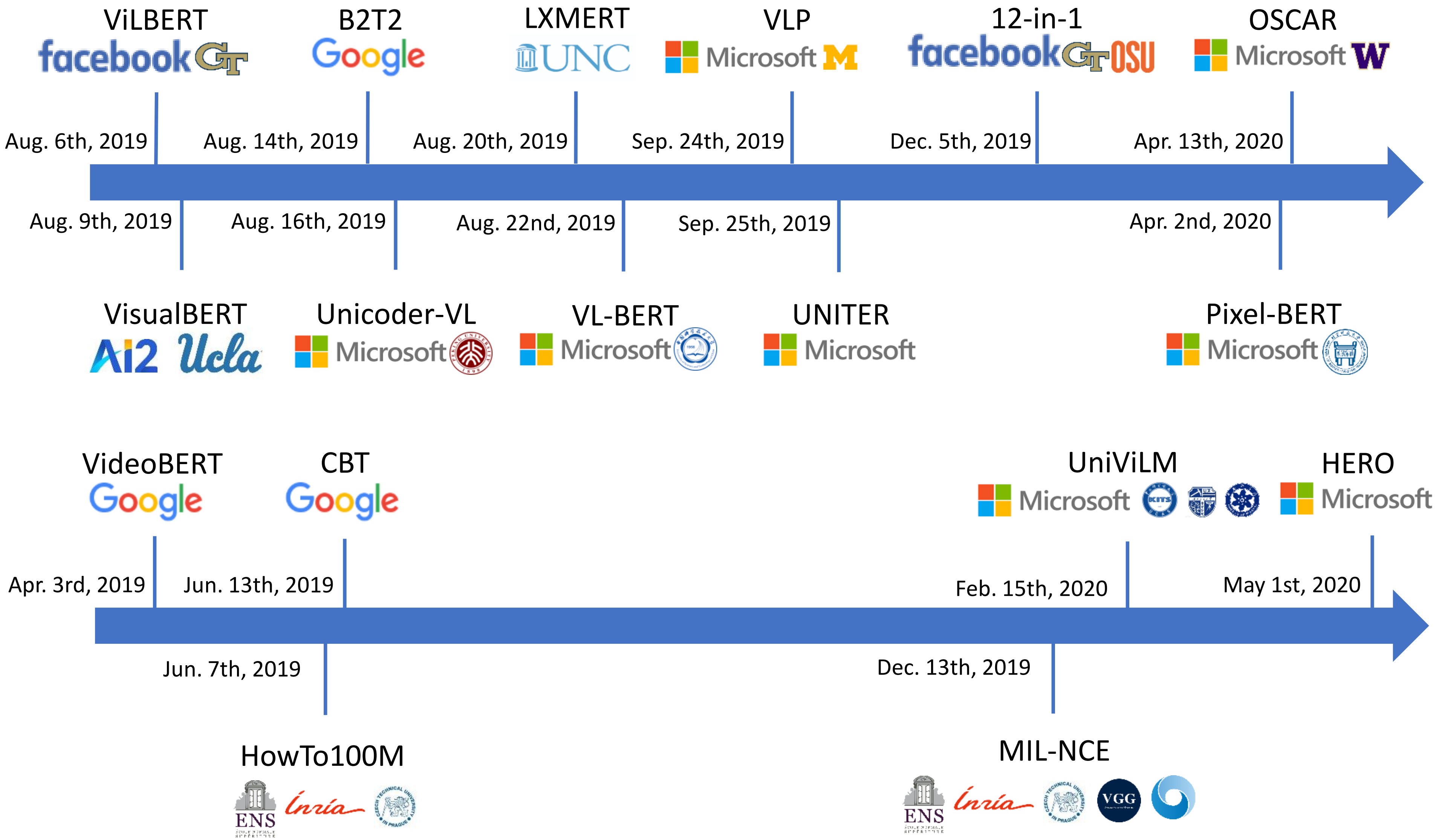


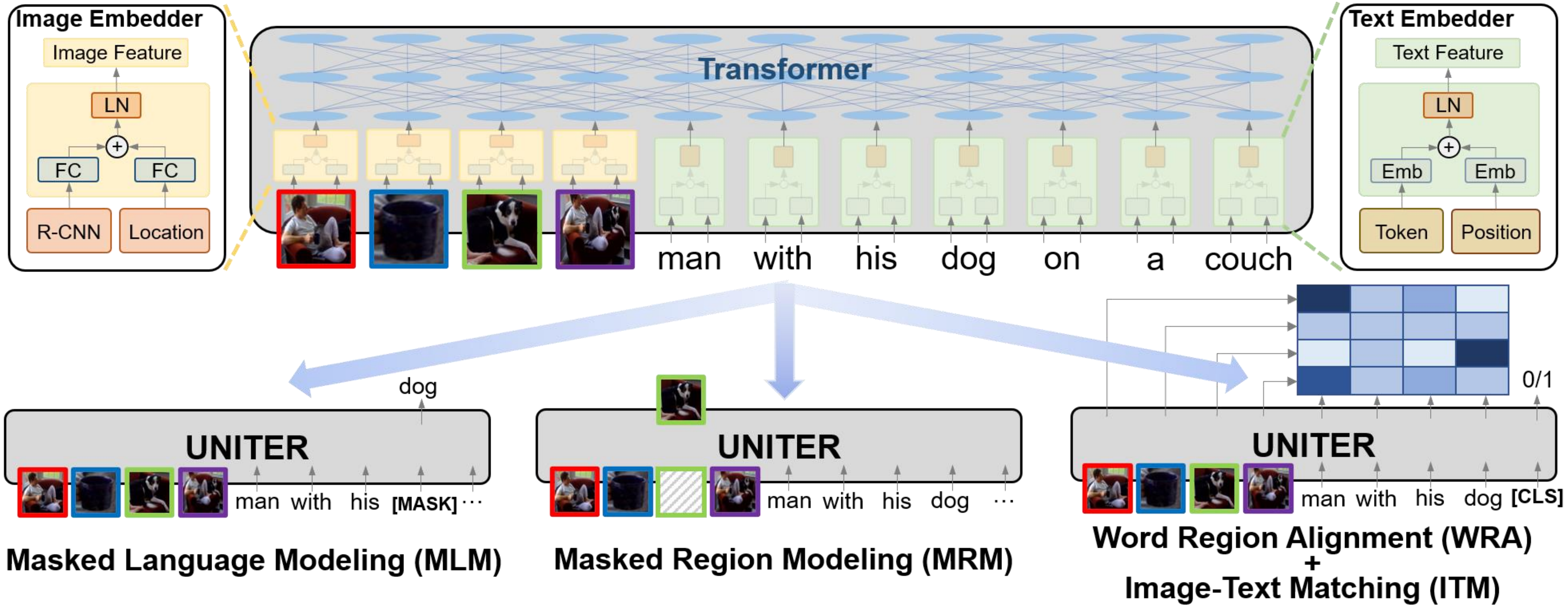
__Image Downstream Tasks__

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

__Video Downstream Tasks__

- Video QA
- Video-and-Language Inference
- Video Captioning
- Video Moment Retrieval

UNITER: Universial Image-Text Representations



Chen et al., UNITER: Universal Image-Text Representation Learning, ECCV 2020

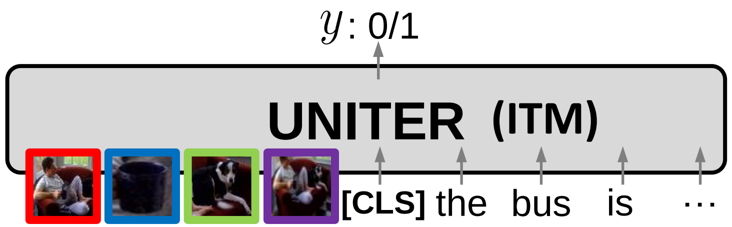
Pre-training Tasks: MLM, ITM & WRA

Masked Language Modeling (MLM)

$$\mathcal{L}_{\mathrm{MLM}}(\theta) = -E_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta(\mathbf{w_m}|\mathbf{w}_{\backslash \mathbf{m}}, \mathbf{v})$$

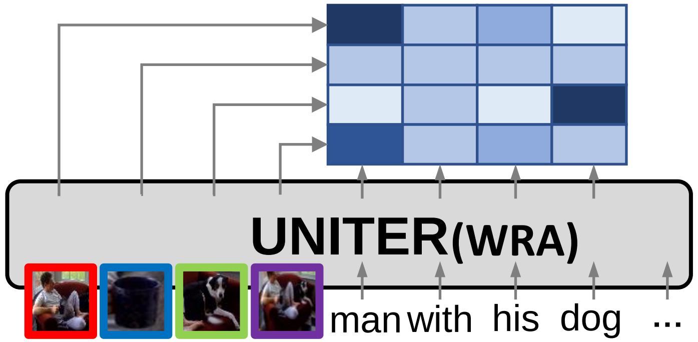
Image-Text Matching (ITM)

$$\mathcal{L}_{\mathrm{ITM}}(\theta) = -E_{(\mathbf{w},\mathbf{v})\sim D}[y \log s_\theta(\mathbf{w},\mathbf{v}) + (1-y)\log(1-s_\theta(\mathbf{w},\mathbf{v}))])$$

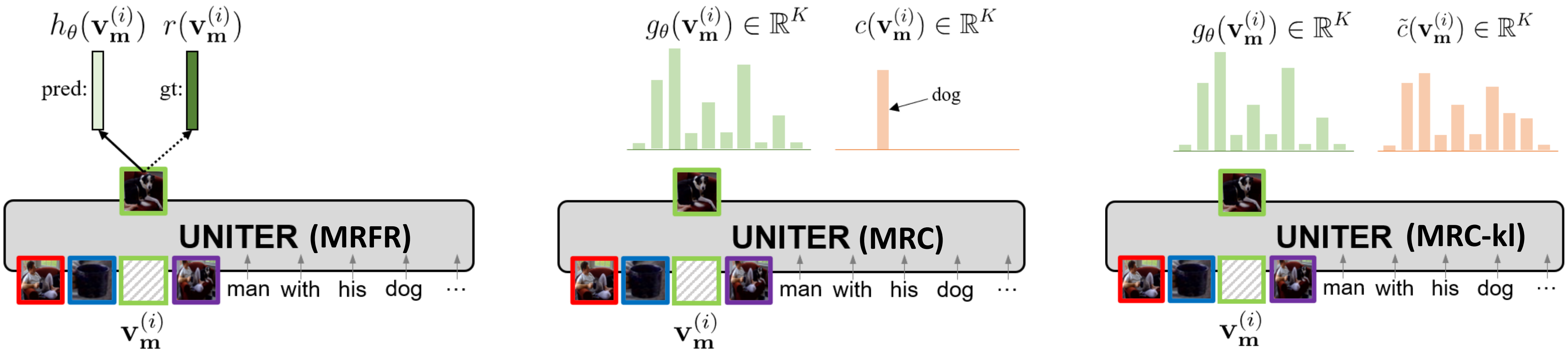


Word Region Alignment (WRA)

$$\mathcal{L}_{\mathrm{WRA}}(\theta) = \mathcal{D}_{ot}(\boldsymbol{\mu},\boldsymbol{\nu}) = \min_{\mathbf{T}\in \Pi(\mathbf{a},\mathbf{b})} \sum_{i=1}^{T}\sum_{j=1}^{K} \mathbf{T}_{ij}\cdot c(\mathbf{w}_i,\mathbf{v}_j)$$



Pre-training Tasks: MRM



Loss Function of Masked Region Modeling (MRM)

$$\mathcal{L}_{\mathrm{MRM}}(\theta) = E_{(\mathbf{w},\mathbf{v})\sim D} f_\theta(\mathbf{v}_\mathbf{m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w})$$

1) **Masked Region Feature Regression (MRFR)**

$$f_\theta(\mathbf{v}_\mathbf{m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} \|h_\theta(\mathbf{v}_\mathbf{m}^{(i)}) - r(\mathbf{v}_\mathbf{m}^{(i)})\|_2^2$$

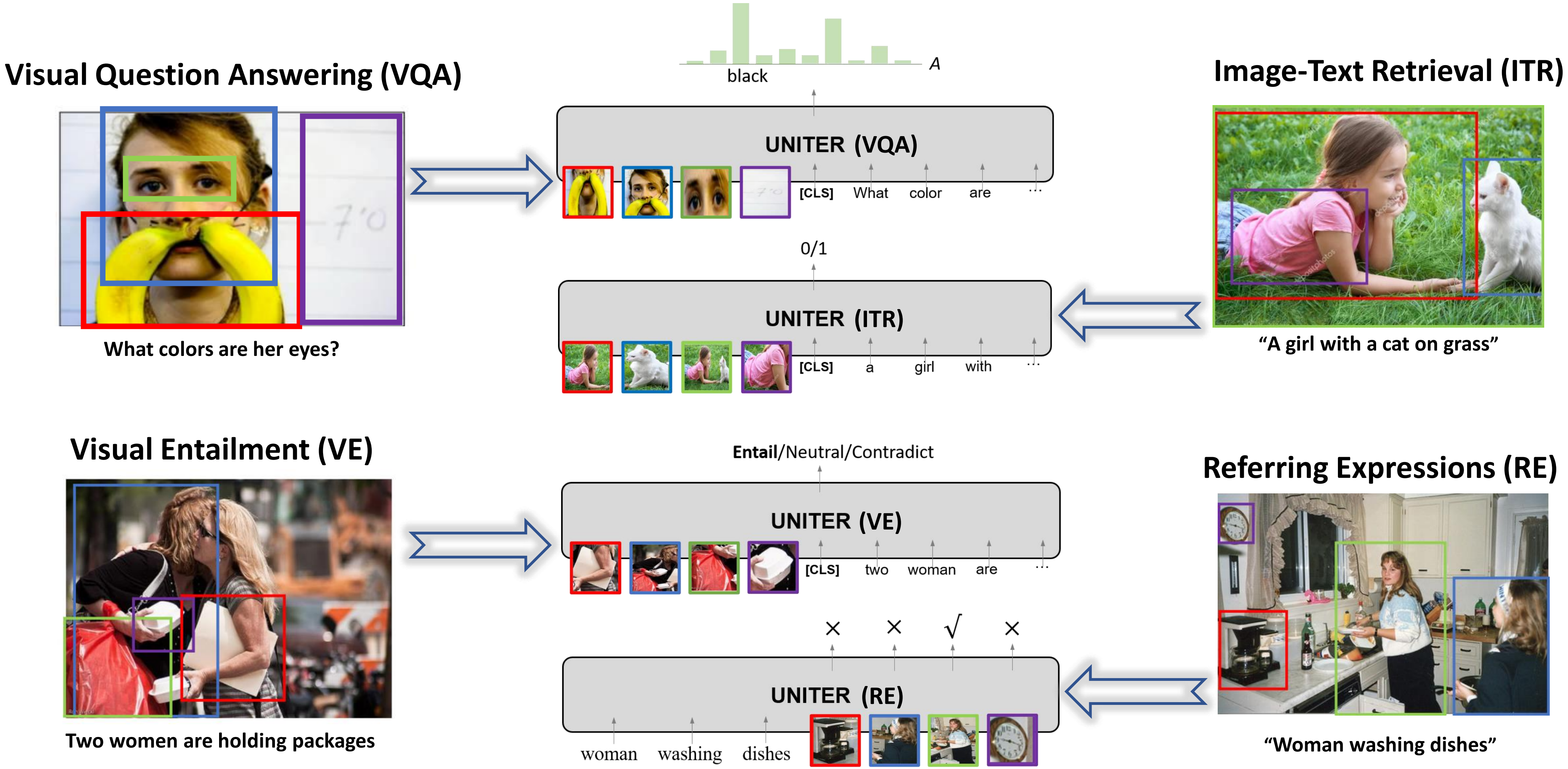
2) **Masked Region Classification (MRC)**

$$f_\theta(\mathbf{v}_\mathbf{m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} \mathrm{CE}(c(\mathbf{v}_\mathbf{m}^{(i)}), g_\theta(\mathbf{v}_\mathbf{m}^{(i)}))$$

3) **Masked Region Classification – KL Divergence (MRC-kl)**

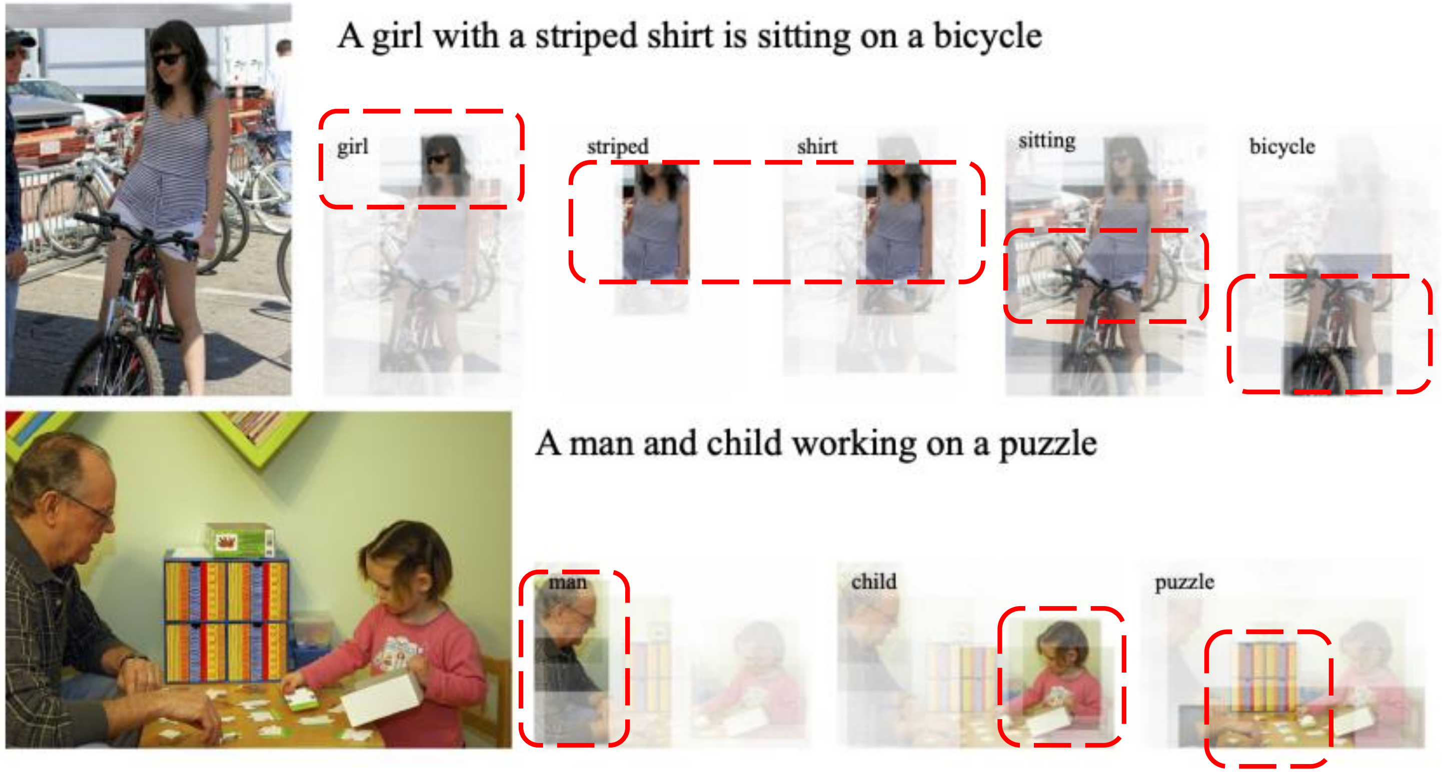
$$f_\theta(\mathbf{v}_\mathbf{m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} D_{KL}(\tilde{c}(\mathbf{v}_\mathbf{m}^{(i)})||g_\theta(\mathbf{v}_\mathbf{m}^{(i)}))$$

Downstream Tasks: VQA, VE, ITR, RE



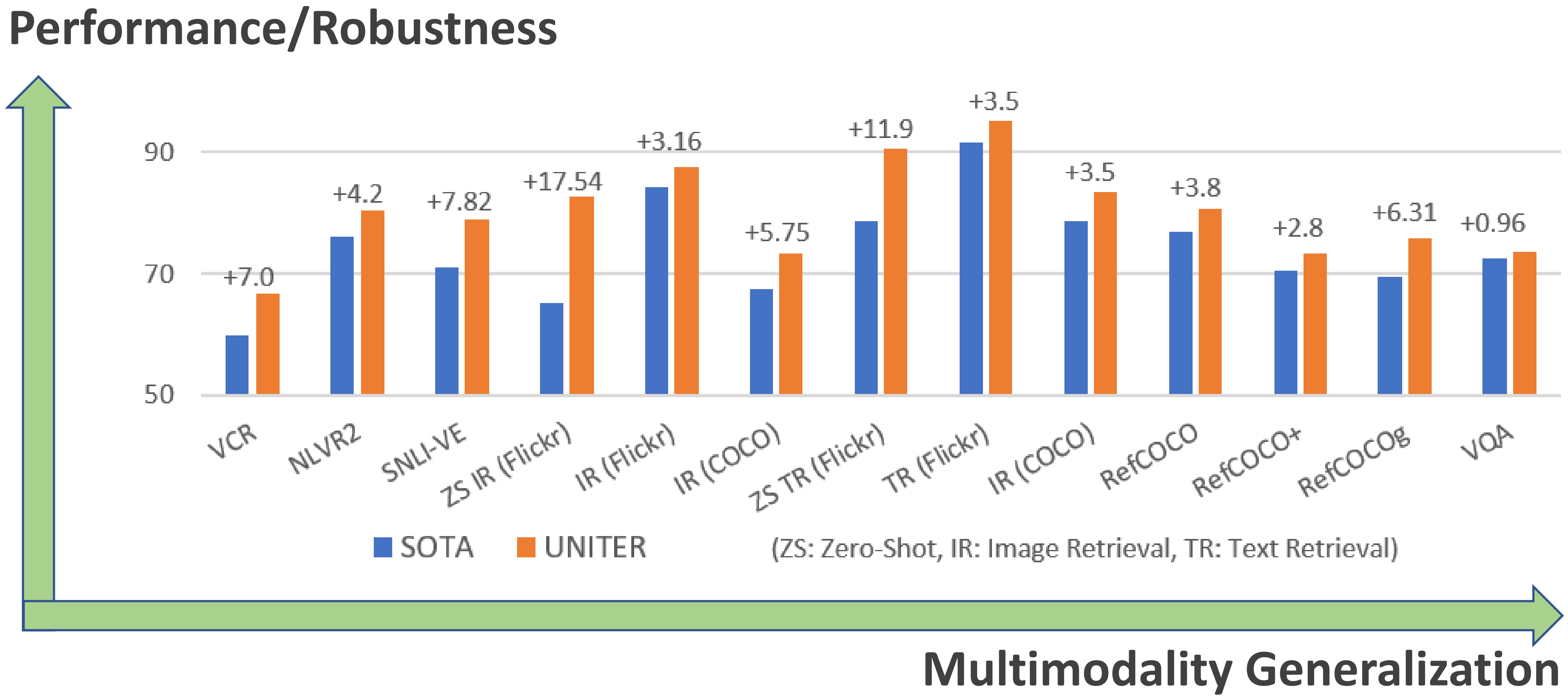
Visualization (Text-to-Image Attention)

- UNITER learns local cross-modality alignment between regions and tokens



State-of-the-Art Results

- UNITER outperformed both task-specific and pre-trained SOTA models over nine V+L tasks (as of Sep 2019 until early 2020)



Large-scale Adversarial Training for Vision+Language

VILLA: Vision-and-Language Large-scale Adversarial Training

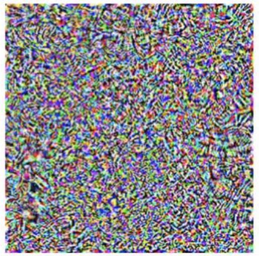
What’s Adversarial Training?

- Neural Networks are prone to label-preserving adversarial examples

Computer Vision:

“pig”

+ 0.005 x



=

“airliner”

Natural Language Processing:

Original: What is the oncorhynchus also called? **A:** chum salmon	**Original:** How long is the Rhine? **A:** 1,230 km
Changed: What’s the oncorhynchus also called? **A:** keta	**Changed:** How long is the Rhine?? **A:** more than 1,050,000

- *What doesn't kill you makes you stronger!*
  - Find adversarial examples that maximize the empirical risk
  - Train the model to predict correctly on adversarial examples

Goodfellow et al., Explaining and harnessing adversarial examples, ICLR 2015

Adversarial Training for Vision+Language

- Aggressive finetuning often falls into the overfitting trap in existing multimodal pre-training methods
- Adversarial training (e.g., FreeLB) has shown great success in improving large-scale NLP models via finetuning
- *1+1>2?*

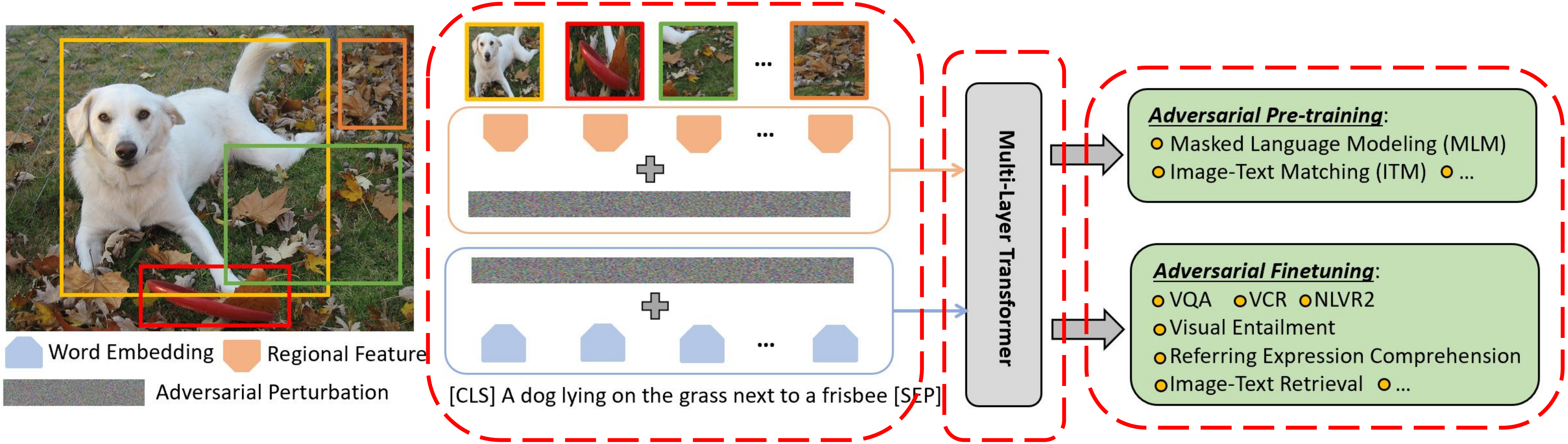
Multimodal Pre-training  Adversarial Training

  - How to enable adversarial training in pre-training stage?
  - How to add perturbations to multiple modalities?
  - How to design advanced adversarial algorithm for V+L?

Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR 2020

Recipe in VILLA

- Ingredient #1: Perturbations in the embedding space
- Ingredient #2: Enhanced adversarial training algorithm
- Ingredient #3: Adversarial pre-training + finetuning



Perturbations in the Embedding Space

- Adversarial label-preserving examples should *preserve semantics*

Original: He has a natural gift for writing scripts.

Adversarial: He has a natural talent for writing scripts. ✓

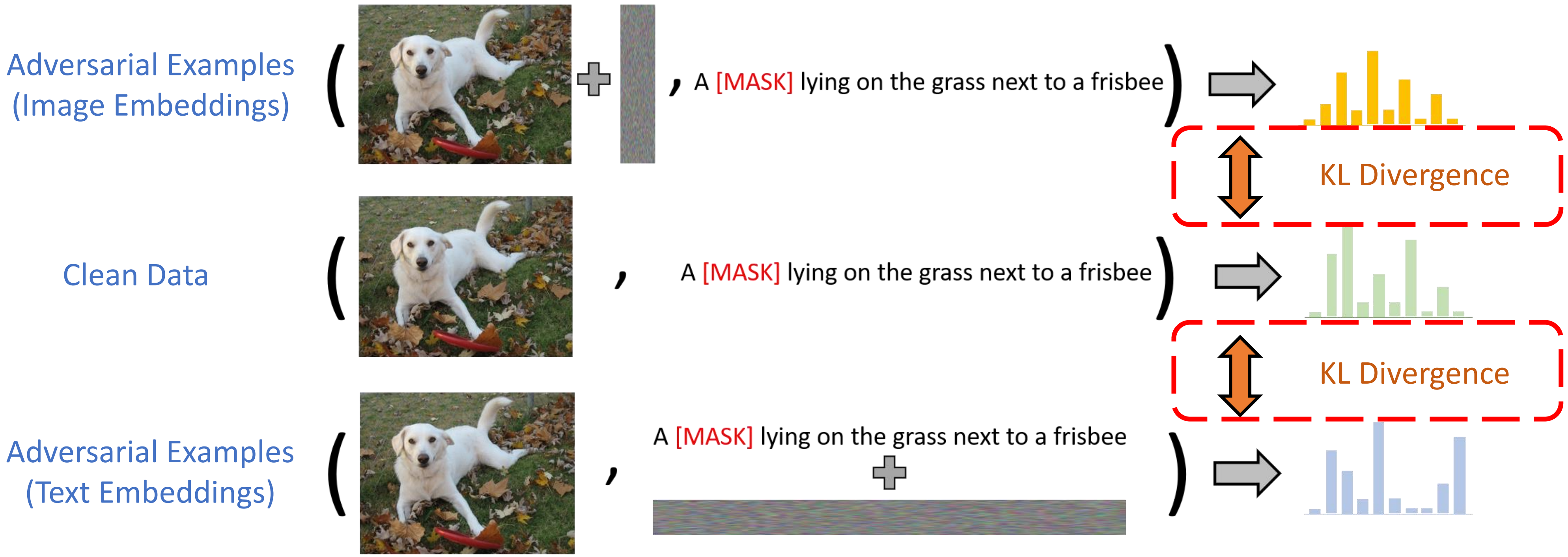
Adversarial: He has a natural present for writing scripts. ✗

- Possible solutions
  - Use back-translation scores to filter out invalid adversaries: *Expensive*
  - Searching for semantically equivalent adversarial rules: *Heuristic*
- Our proposal: add perturbations to the embedding space directly, as the goal is *end results* of adversarial training

Ribeiro et al., Semantically Equivalent Adversarial Rules for Debugging NLP Models, ACL 2018
Cheng et al., Robust Neural Machine Translation with Doubly Adversarial Inputs, ACL 2019

Adversarial Training Algorithm

- Training objective: $\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}, \boldsymbol{y}) \sim \mathcal{D}} \left[\boxed{\mathcal{L}_{std}(\boldsymbol{\theta})} + \boxed{\mathcal{R}_{at}(\boldsymbol{\theta})} + \alpha \cdot \boxed{\mathcal{R}_{kl}(\boldsymbol{\theta})} \right]$
  - $\mathcal{L}_{std}(\boldsymbol{\theta})$: Cross-entropy loss on clean data
  - $\mathcal{R}_{at}(\boldsymbol{\theta})$: Cross-entropy loss on adversarial embeddings
  - $\mathcal{R}_{kl}(\boldsymbol{\theta})$: KL-divergence loss for fine-grained adversarial regularization



Results (VQA, VCR, NLVR2, SNLI-VE)

- Established new state of the art on all the tasks considered
- Gain: +0.85 on VQA, +2.9 on VCR, +1.49 on NLVR2, +0.64 on SNLI-VE

Method	VQA		VCR			$NLVR^2$		SNLI-VE	
	test-dev	test-std	Q→A	QA→R	Q→AR	dev	test-P	val	test
ViLBERT	70.55	70.92	72.42 (73.3)	74.47 (74.6)	54.04 (54.8)	-	-	-	-
VisualBERT	70.80	71.00	70.8 (71.6)	73.2 (73.2)	52.2 (52.4)	67.4	67.0	-	-
LXMERT	72.42	72.54	-	-	-	74.90	74.50	-	-
Unicoder-VL	-	-	72.6 (73.4)	74.5 (74.4)	54.4 (54.9)	-	-	-	-
12-in-1	73.15	-	-	-	-	-	78.87	-	76.95
$\text{VL-BERT}_{\text{BASE}}$	71.16	-	73.8 (-)	74.4 (-)	55.2 (-)	-	-	-	-
$\text{Oscar}_{\text{BASE}}$	73.16	73.44	-	-	-	78.07	78.36	-	-
$\text{UNITER}_{\text{BASE}}$	72.70	72.91	74.56 (75.0)	77.03 (77.2)	57.76 (58.2)	77.18	77.85	78.59	78.28
$\text{VILLA}_{\text{BASE}}$	**73.59**	**73.67**	**75.54 (76.4)**	**78.78 (79.1)**	**59.75 (60.6)**	**78.39**	**79.30**	**79.47**	**79.03**
$\text{VL-BERT}_{\text{LARGE}}$	71.79	72.22	75.5 (75.8)	77.9 (78.4)	58.9 (59.7)	-	-	-	-
$\text{Oscar}_{\text{LARGE}}$	73.61	73.82	-	-	-	79.12	80.37	-	-
$\text{UNITER}_{\text{LARGE}}$	73.82	74.02	77.22 (77.3)	80.49 (80.8)	62.59 (62.8)	79.12	79.98	79.39	79.38
$\text{VILLA}_{\text{LARGE}}$	**74.69**	**74.87**	**78.45 (78.9)**	**82.57 (82.8)**	**65.18 (65.7)**	**79.76**	**81.47**	**80.18**	**80.02**

(a) Results on VQA, VCR, $NLVR^2$, and SNLI-VE.

Results (ITR, RE)

- Gain: +1.52/+0.60 on Flickr30k IR & TR (R@1), and +0.99 on 3 RE datasets

Method	RefCOCO+						RefCOCO					
	val	testA	testB	val^d	testA^d	testB^d	val	testA	testB	val^d	testA^d	testB^d
ViLBERT	-	-	-	72.34	78.52	62.61	-	-	-	-	-	-
$\text{VL-BERT}_{\text{BASE}}$	79.88	82.40	75.01	71.60	77.72	60.99	-	-	-	-	-	-
$\text{UNITER}_{\text{BASE}}$	83.66	86.19	78.89	75.31	81.30	65.58	91.64	92.26	90.46	81.24	86.48	73.94
$\text{VILLA}_{\text{BASE}}$	**84.26**	**86.95**	**79.22**	**76.05**	**81.65**	**65.70**	**91.93**	**92.79**	**91.38**	**81.65**	**87.40**	**74.48**
$\text{VL-BERT}_{\text{LARGE}}$	80.31	83.62	75.45	72.59	78.57	62.30	-	-	-	-	-	-
$\text{UNITER}_{\text{LARGE}}$	84.25	**86.34**	79.75	75.90	81.45	66.70	91.84	92.65	91.19	81.41	87.04	74.17
$\text{VILLA}_{\text{LARGE}}$	**84.40**	86.22	**80.00**	**76.17**	**81.54**	**66.84**	**92.58**	**92.96**	**91.62**	**82.39**	**87.48**	**74.84**

(b) Results on RefCOCO+ and RefCOCO. The superscript d denotes evaluation using detected proposals.

Method	RefCOCOg				Flickr30k IR			Flickr30k TR		
	val	test	val^d	test^d	R@1	R@5	R@10	R@1	R@5	R@10
ViLBERT	-	-	-	-	58.20	84.90	91.52	-	-	-
Unicoder-VL	-	-	-	-	71.50	90.90	94.90	86.20	96.30	99.00
$\text{UNITER}_{\text{BASE}}$	86.52	86.52	74.31	74.51	72.52	92.36	**96.08**	85.90	97.10	98.80
$\text{VILLA}_{\text{BASE}}$	**88.13**	**88.03**	**75.90**	**75.93**	**74.74**	**92.86**	95.82	**86.60**	**97.90**	**99.20**
$\text{UNITER}_{\text{LARGE}}$	87.85	87.73	74.86	75.77	75.56	94.08	96.76	87.30	**98.00**	**99.20**
$\text{VILLA}_{\text{LARGE}}$	**88.42**	**88.97**	**76.18**	**76.71**	**76.26**	**94.24**	**96.84**	**87.90**	97.50	98.80

(c) Results on RefCOCOg and Flickr30k Image Retrieval (IR) and Text Retrieval (TR).

Ablation Study and Generalization

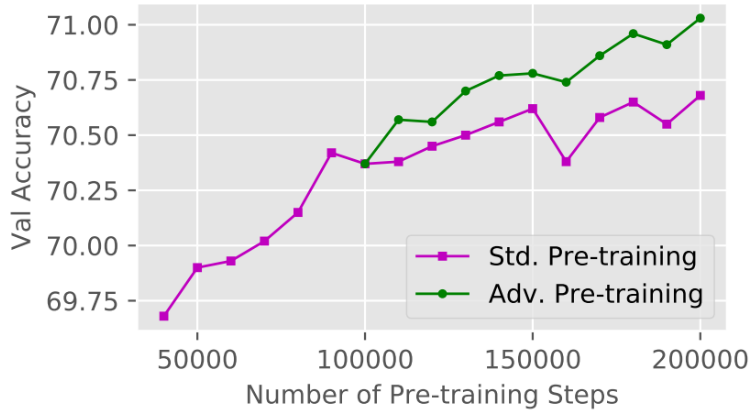
- Both adversarial pre-training and finetuning contribute to performance boost

Method	VQA	VCR (val)			NLVR^2	VE	Flickr30k IR			RefCOCO		Ave.
	test-dev	Q→A	QA→R	Q→AR	test-P	test	R@1	R@5	R@10	testA^d	testB^d	
UNITER (reimp.)	72.70	74.24	76.93	57.31	77.85	78.28	72.52	92.36	96.08	86.48	73.94	78.06
VILLA-pre	73.03	74.76	77.04	57.82	78.44	78.43	73.76	93.02	96.28	87.34	74.35	78.57
VILLA-fine	73.29	75.18	78.29	59.08	78.84	78.86	73.46	92.98	96.26	87.17	74.31	78.88
VILLA	73.59	75.54	78.78	59.75	79.30	79.03	74.74	92.86	95.82	87.40	74.48	**79.21**

+0.51 Pre-train
+0.82 Finetune
+1.15 Both

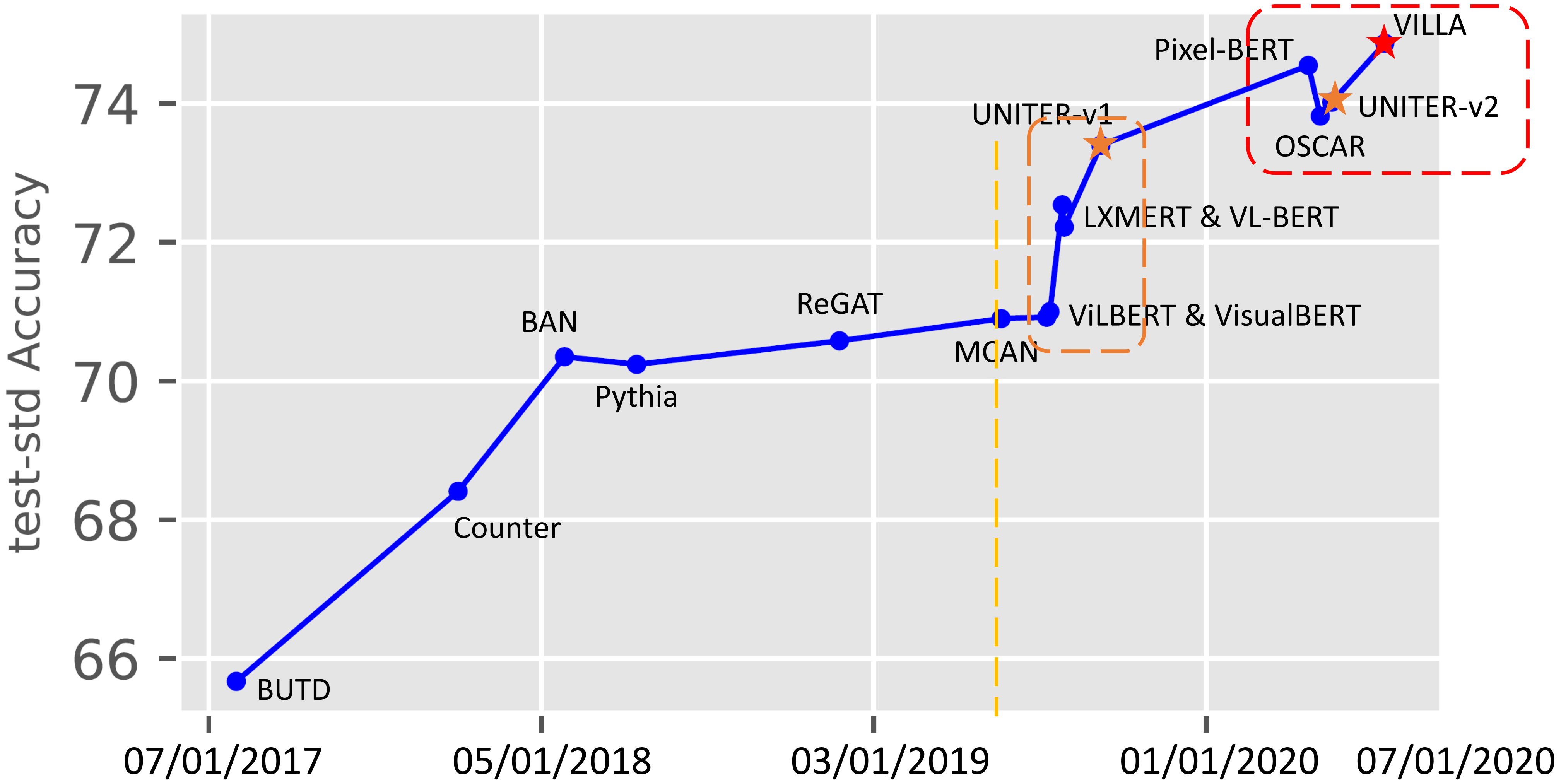
- VILLA can be applied to any pre-trained V+L models

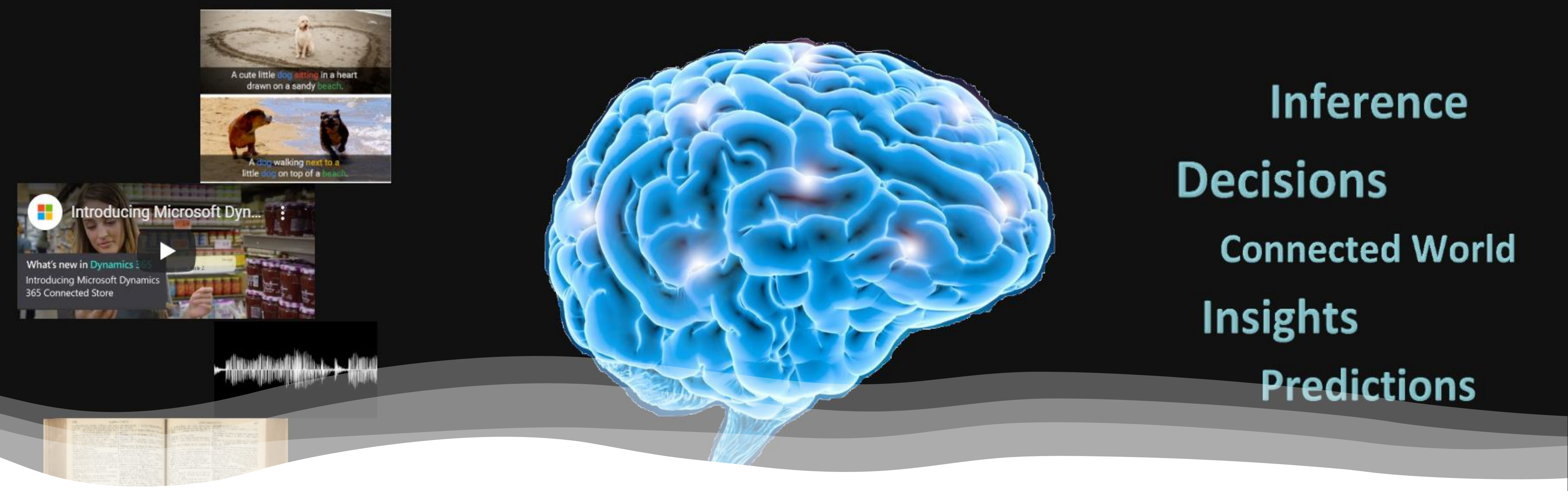
Method	VQA		GQA		NLVR^2		Meta-Ave.
	test-dev	test-std	test-dev	test-std	dev	test-P	
LXMERT	72.42	72.54	60.00	60.33	74.95	74.45	69.12
LXMERT (reimp.)	72.50	72.52	59.92	60.28	74.72	74.75	69.12
VILLA-fine	**73.02**	**73.18**	**60.98**	**61.12**	**75.98**	**75.73**	**70.00**



Gan et al., Large-Scale Adversarial Training for Vision-and-Language Representation Learning, 2020

A Closer Look at VQA



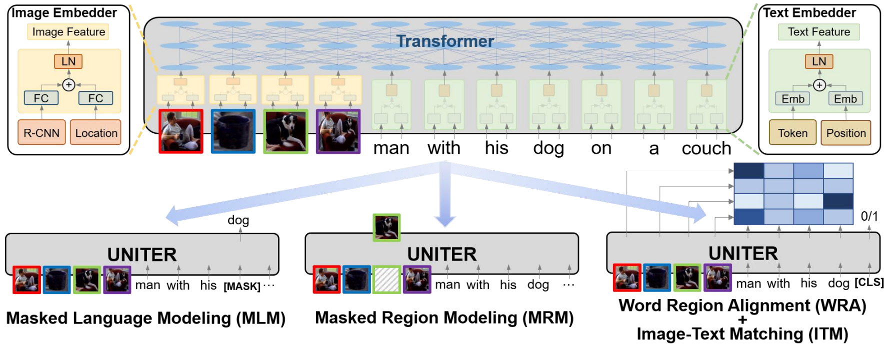


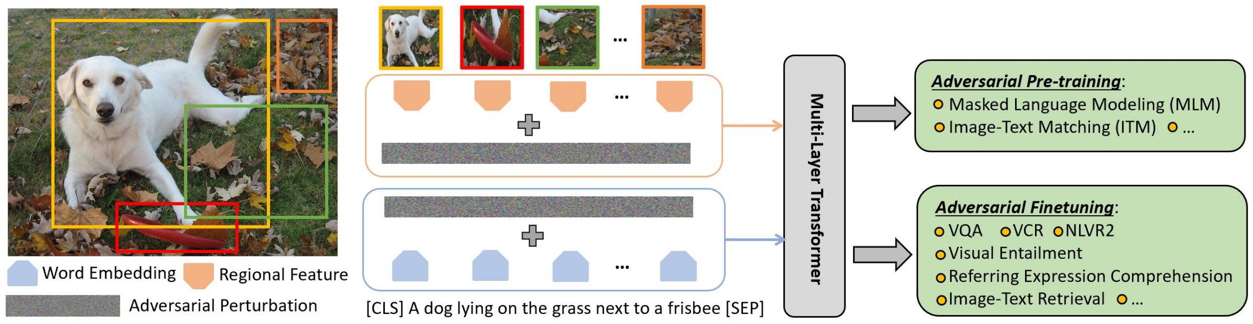
AI Explainability and Interpretability

VALUE: Vision-And-Language Understanding Evaluation

What Have Pretrained Models Learned?

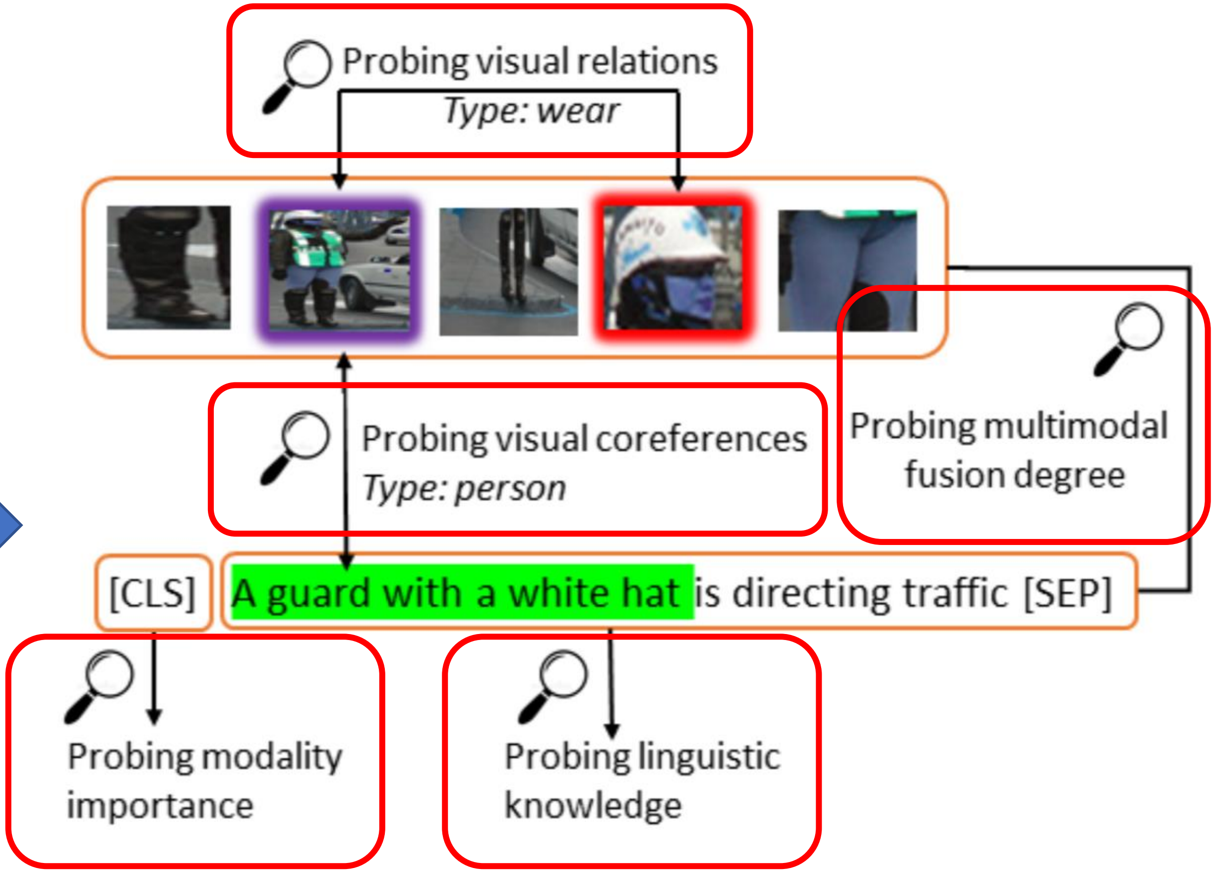
- What is the *correlation* between multimodal fusion and network layers?
- Which *modality* plays a more important role?
- What *cross-modal knowledge* is encoded in pre-trained models?
- What *intra-modal knowledge* has been learned?
- What *linguistic knowledge* do pre-trained V+L models encode?





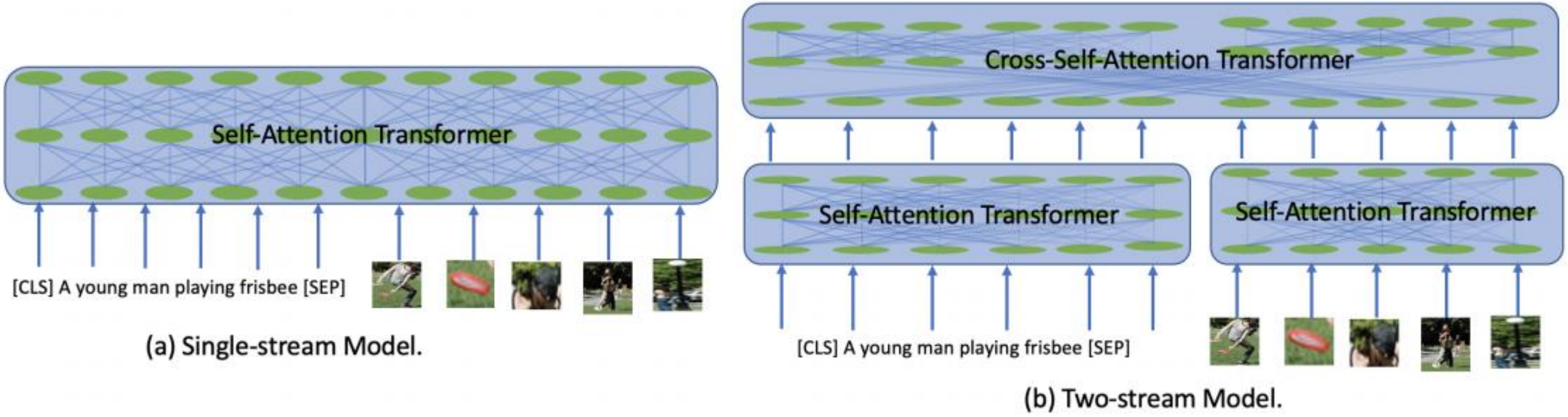
VALUE: Vision-And-Language Understanding Evaluation

- Visual probing
- Linguistic probing
- Cross-modality probing



Cao et al., Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models, ECCV 2020

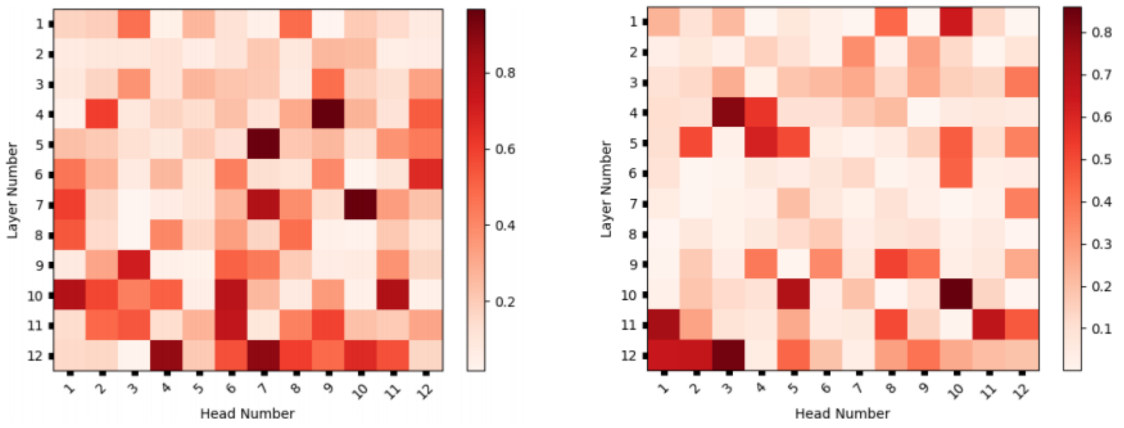
Single-Stream vs. Two-Stream Architecture



- Models: UNITER (single-stream) vs. LXMERT (two-stream)
- Probing targets: 144 attention weight matrices (12 layers x 12 heads)
- Datasets: Visual Genome (for visual relations), Flickr30k (for visual coreference)
- Toolkit: SentEval (for linguistic probing)

Take-home Message

- *Deep to Profound*: Deeper layers lead to more intertwined multimodal fusion
- *Who Pulls More Strings*: Textual modality is more dominant than image
- *Winner Takes All*: A subset of heads is specialized for cross-modal interaction
- *Secret Liaison Revealed*: Cross-modality fusion registers visual relations
- *No Lost in Translation*: Pre-trained V+L models encode rich linguistic knowledge



(a) Textual modality importance (b) Visual modality importance

High-Resolution Image Synthesis: BachGAN

Vision-and-Language Inference: VIOLIN

BachGAN: Background Hallucination GAN

Task: Image Synthesis from Object Layout

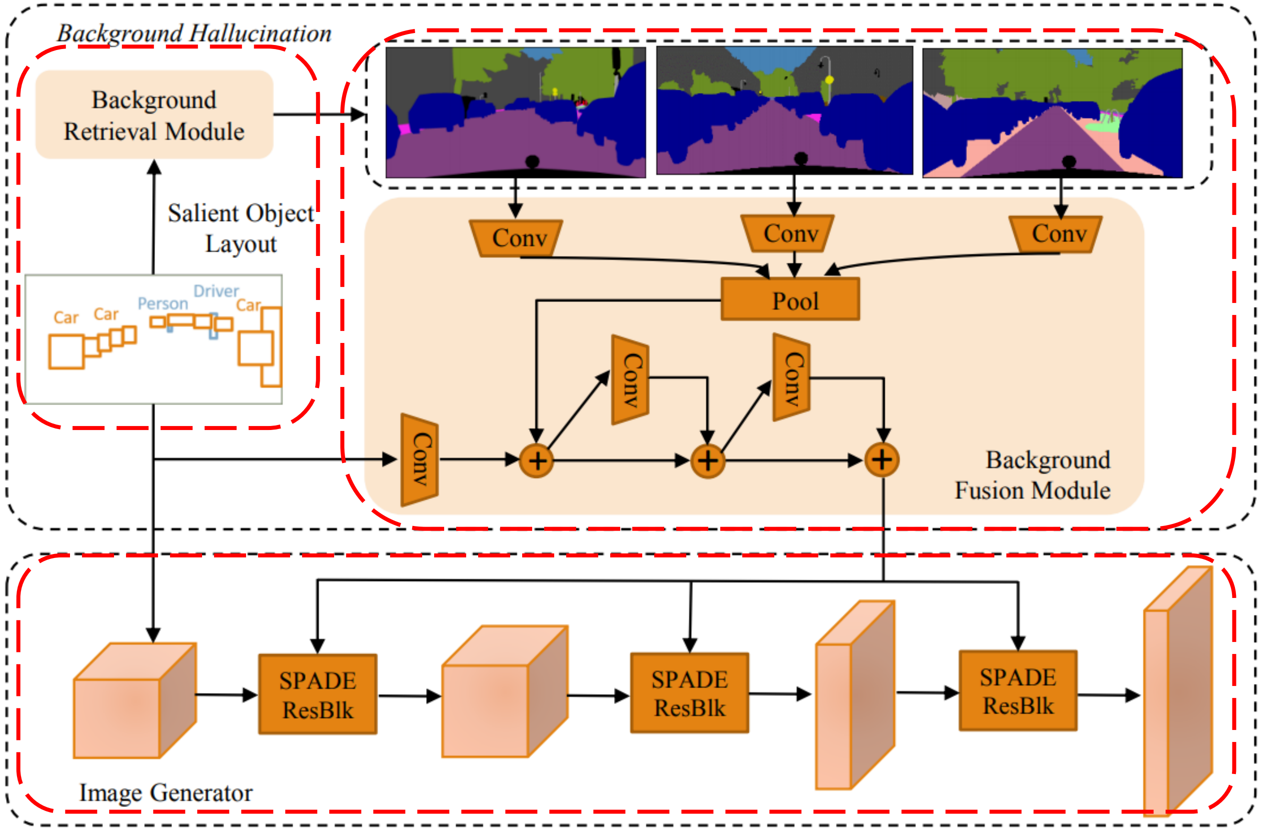
Segmentation Map Input *(Prior work)*

Bounding Box Input *(BachGAN)*

Synthesized Results *(BachGAN vs. Baselines)*

BachGAN: Background Hallucination GAN

- BachGAN outperforms baseline models in both quantitative and human evaluations



Model	Cityscapes Acc	Cityscapes FID	ADE20K Acc	ADE20K FID
Layout2im [40]	-	99.1	-	-
SPADE	57.6	86.7	55.3	59.4
SPADE-SEG	60.2	81.2	60.9	57.2
BachGAN-r	67.3	74.4	64.5	53.2
BachGAN	**70.4**	**73.3**	**66.8**	**49.8**

Results on automatic metrics

Dataset	BachGAN vs. SPADE			BachGAN vs. SPADE-Seg			BachGAN vs. BachGAN-r		
	win	loss	tie	win	loss	tie	win	loss	tie
Cityscapes	85.5	3.4	11.1	71.7	12.4	15.9	61.6	24.1	14.3
ADE20K	75.9	12.8	11.3	66.8	17.4	15.8	57.2	18.7	24.1

Results from human study

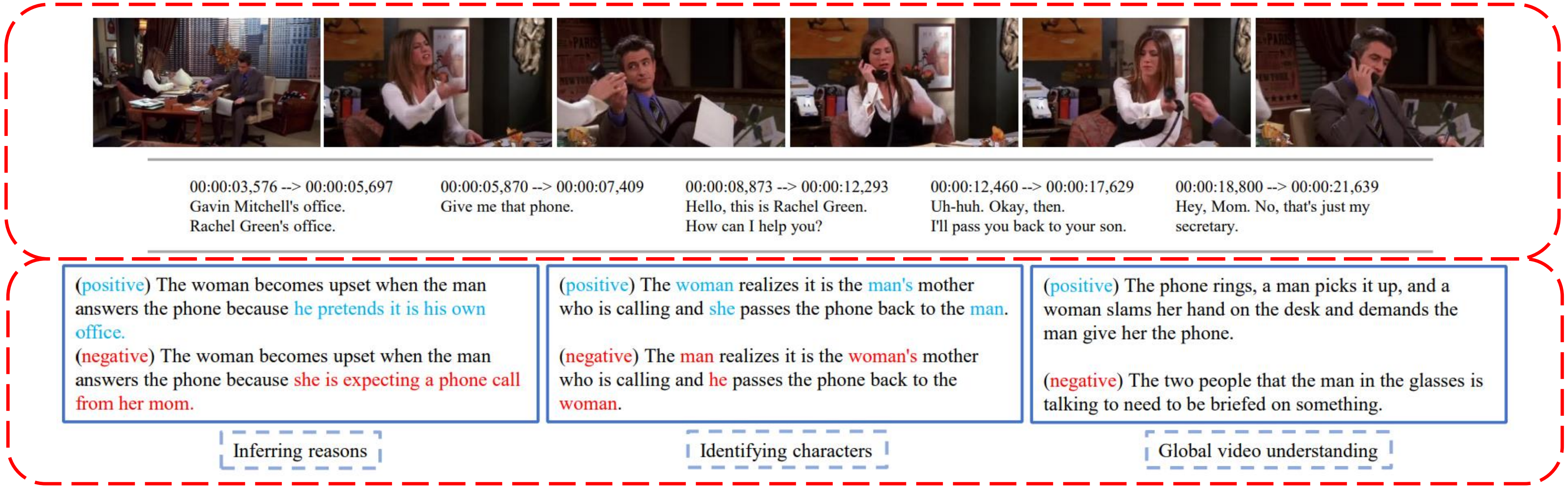
Li et al., BachGAN: High-Resolution Image Synthesis from Salient Object Layout, CVPR 2020

VIOLIN: Video-and-Language Inference

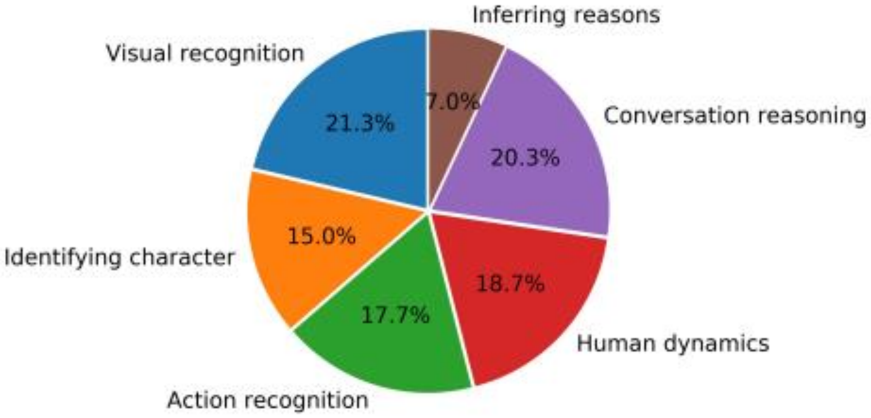
- 95K video+statement pairs collected from 16K video clips (TV shows & movies clips)
- Each video is 35-second long on average, paired with 6 statements
- Each statement is either 'Entailment' or 'Contradiction' to the video

Dataset	Visual Domain	Source	Subtitles	Inference	Task	# images/videos	# samples
Movie-QA [54]	video	movie	✓	✗	QA	6.8K	6.5K
MovieFIB [44]	video	movie	✗	✗	QA	118.5K	349K
TVQA [35]	video	TV show	✓	✗	QA	21.8K	152.5K
VCR [72]	image	movie	✗	✓	QA	110K	290K
GQA [25]	image	indoor	✗	✓	QA	113K	22M
SNLI-VE [61]	image	natural	✗	✓	Entailment	31.8K	565.3K
$NLVR^2$ [52]	image	natural	✗	✓	Entailment	127.5K	107.3K
VIOLIN (ours)	video	TV show/movie	✓	✓	Entailment	15.9K	95.3K

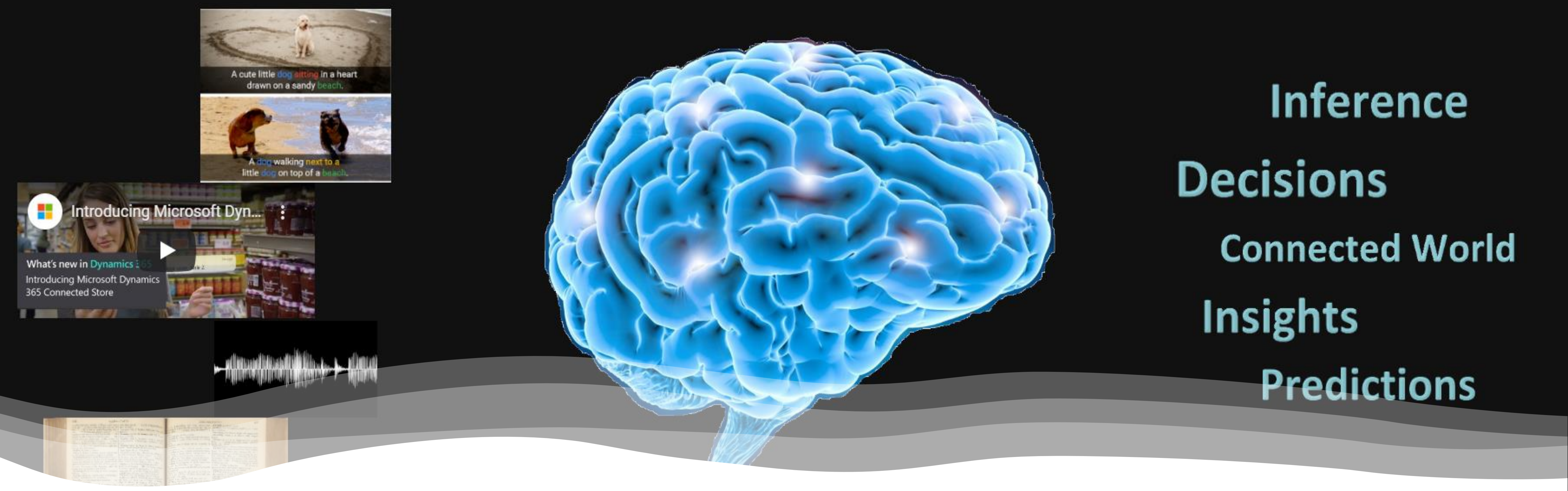
VIOLIN: Video-and-Language Inference



- Explicit Visual Understanding (54%): Visual recognition, Identifying character, Action Recognition
- Deeper Inference (46%): Inferring reasons/causal relations, Conversation reasoning, Social dynamics



Liu et al., VIOLIN: A Large-Scale Dataset for Video-and-Language Inference, CVPR 2020



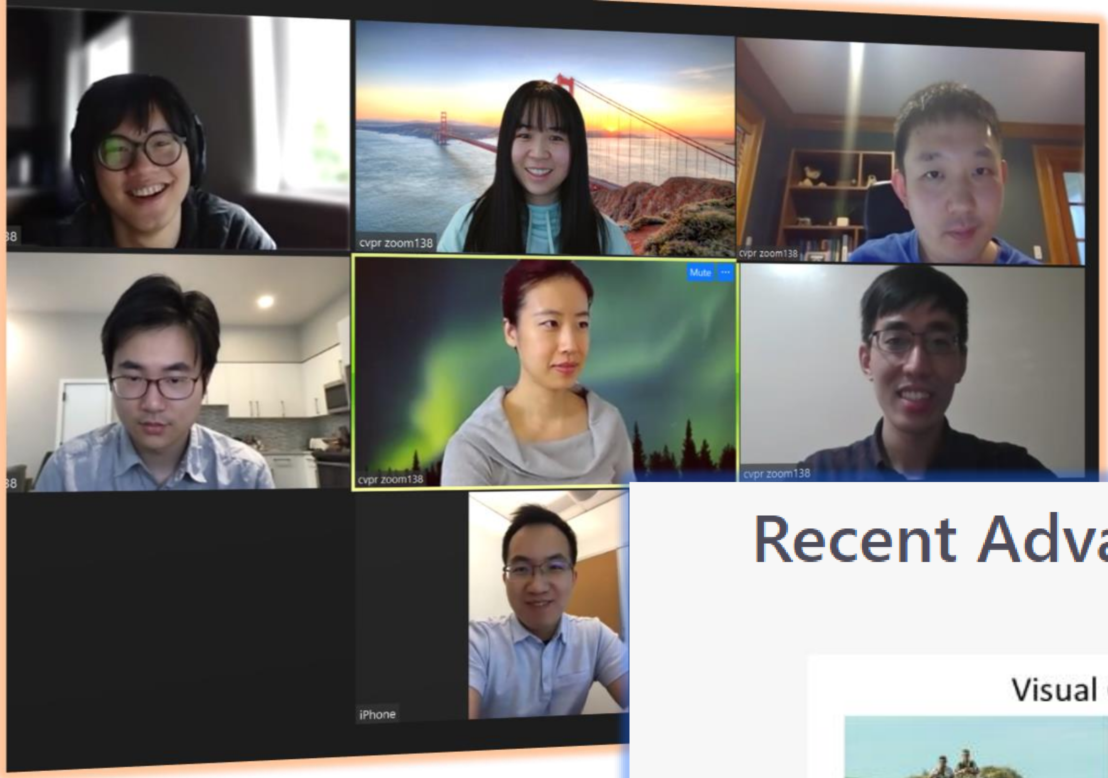
Self-supervised Learning for Multimodal Pre-training: **UNITER**

Large-scale Adversarial Training for Vision+Language: **VILLA**

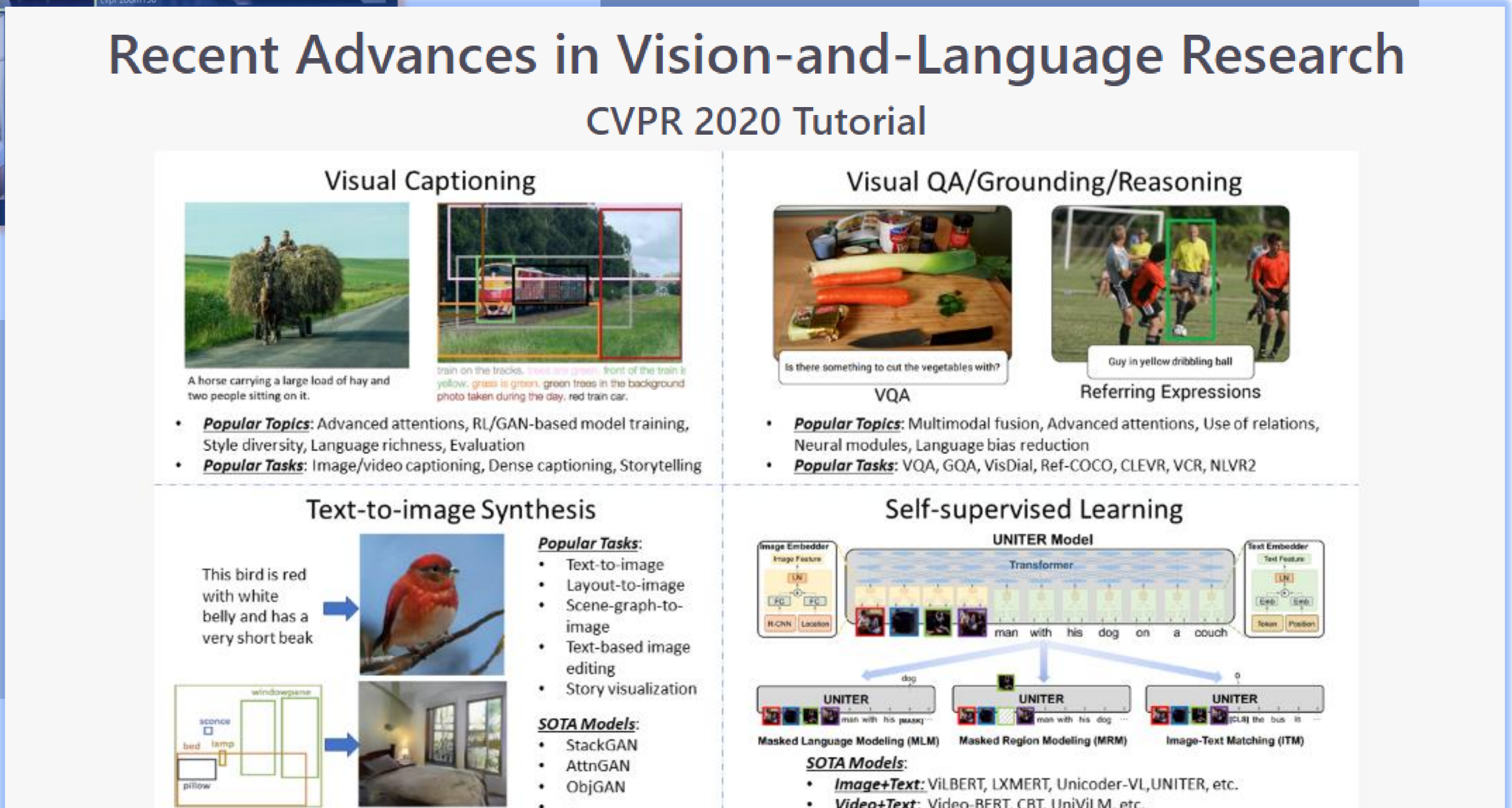
AI Explainability: **VALUE**

Image Synthesis: **BachGAN**

Vision-and-Language Inference: **VIOLIN**


cvpr zoom138
cvpr zoom138
cvpr zoom138
cvpr zoom138
iPhone


CVPR
VIRTUAL JUNE 14-19


Recent Advances in Vision-and-Language Research
CVPR 2020 Tutorial
Visual Captioning
A horse carrying a large load of hay and two people sitting on it.
photo taken during the day. red train car.
• Popular Topics: Advanced attentions, RL/GAN-based model training, Style diversity, Language richness, Evaluation
• Popular Tasks: Image/video captioning, Dense captioning, Storytelling
Visual QA/Grounding/Reasoning
Is there something to cut the vegetables with?
VQA
Guy in yellow dribbling ball
Referring Expressions
• Popular Topics: Multimodal fusion, Advanced attentions, Use of relations, Neural modules, Language bias reduction
• Popular Tasks: VQA, GQA, VisDial, Ref-COCO, CLEVR, VCR, NLVR2
Text-to-image Synthesis
This bird is red with white belly and has a very short beak
Popular Tasks:
• Text-to-image
• Layout-to-image
• Scene-graph-to-image
• Text-based image editing
• Story visualization
SOTA Models:
• StackGAN
• AttnGAN
• ObjGAN
Self-supervised Learning
UNITER Model
Transformer
man with his dog on a couch
UNITER
UNITER
UNITER
Masked Language Modeling (MLM)
Masked Region Modeling (MRM)
Image-Text Matching (ITM)
SOTA Models:
• Image+Text: ViLBERT, LXMERT, Unicoder-VL, UNITER, etc.
• Video+Text: Video-BERT, CBT, UniViLM, etc.

Thank You

Microsoft Multimodal AI Group: http://aka.ms/mmai