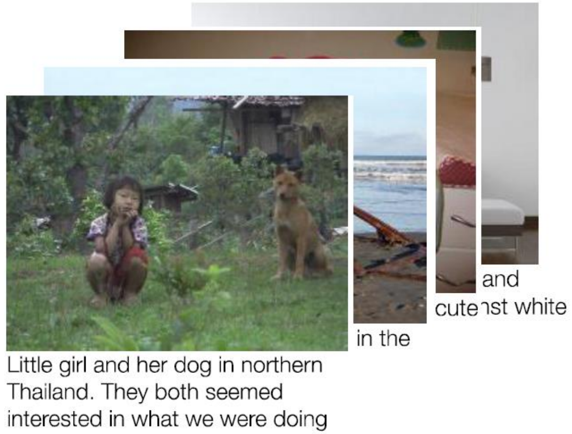
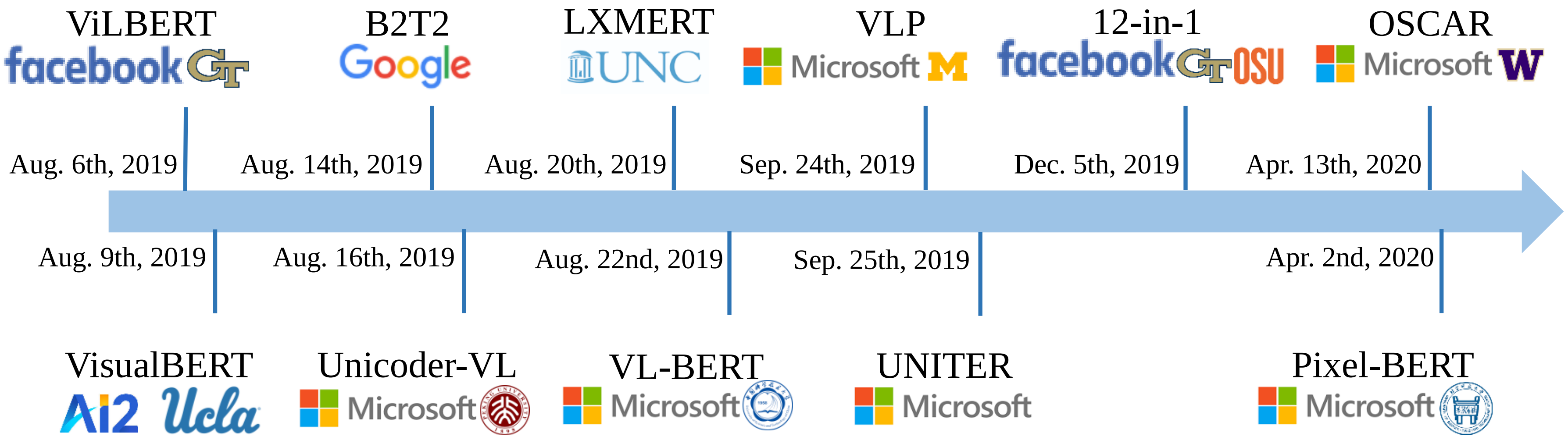
A Closer Look at the Robustness of Vision-and-Language Pre-trained Models

Linjie Li, Zhe Gan, Jingjing Liu

Microsoft

Image: Single image
Language: Textual Descriptions

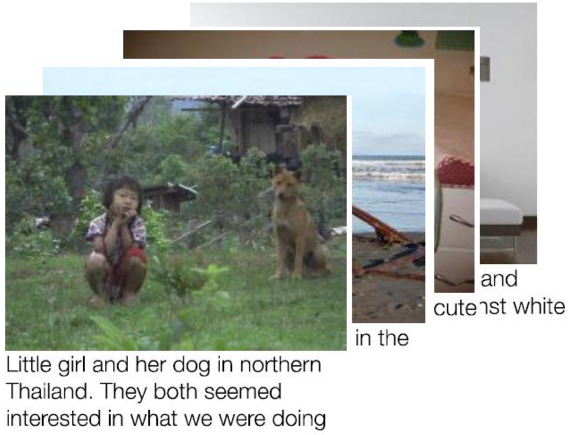


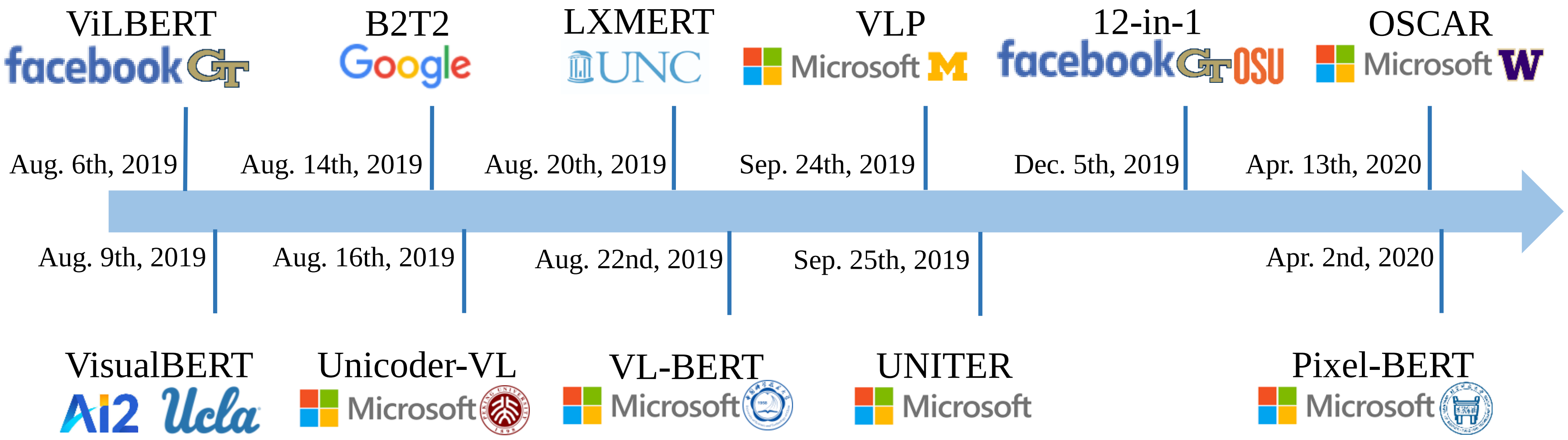


V+L Tasks
- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

- Pre-trained multimodal Transformers have achieved SOTA performance across a wide range of V+L tasks

Image: Single image
Language: Textual Descriptions



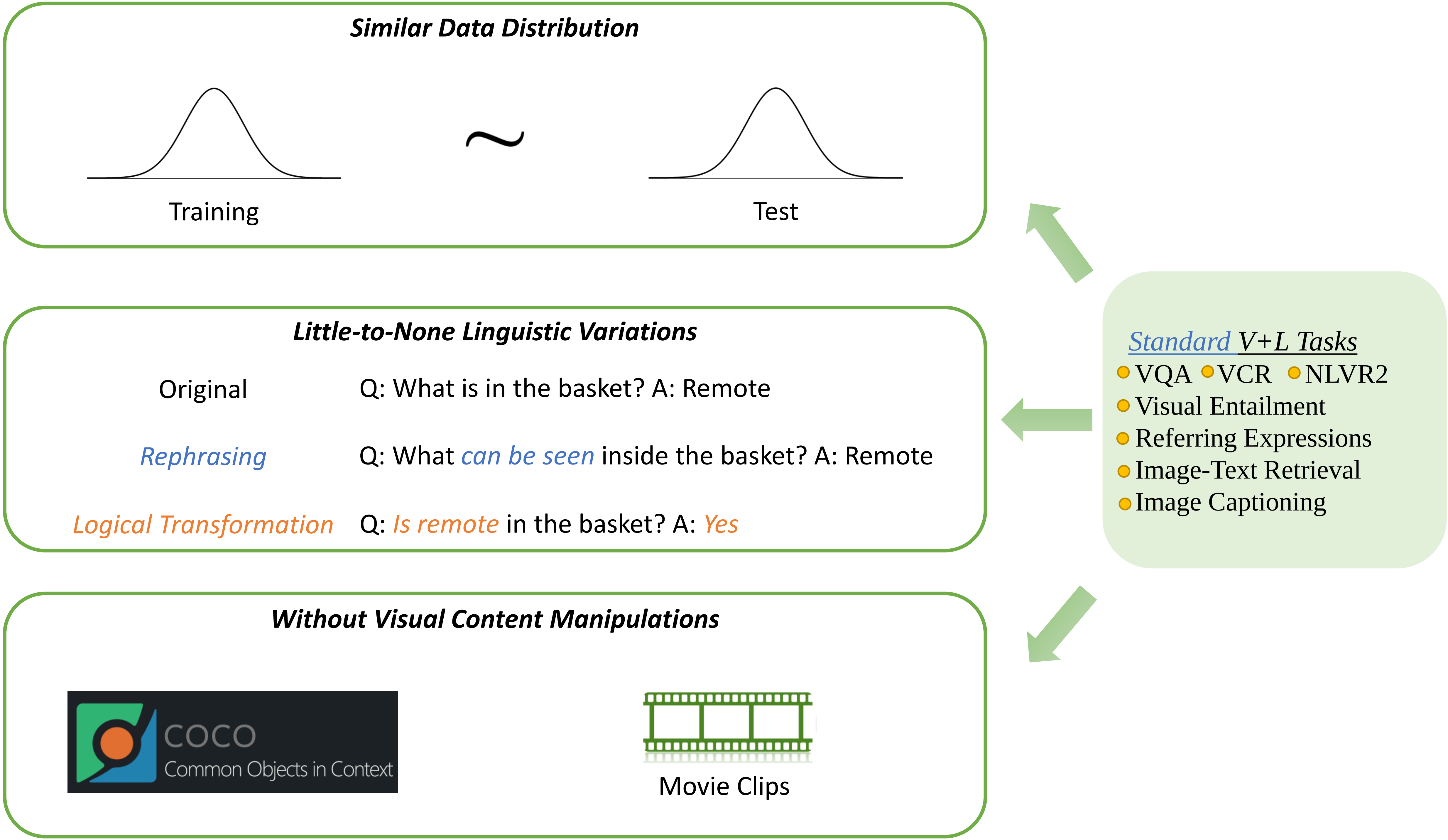


V+L Tasks

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

- Pre-trained multimodal Transformers have achieved SOTA performance across a wide range of V+L tasks

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

A Closer Look At the Robustness of Vision-and-Language Pre-trained Models

- The first systematical examination of pre-trained V+L model robustness over 4 generic robustness types
  - Linguistic Variation
  - Logical Reasoning
  - Visual Content Manipulation
  - Answer Distribution Shift
- We present MANGO (Multimodal Adversarial Noise GeneratOr)
  - A generic and efficient adversarial training approach
  - Surpasses SOTA on 7 out of 9 robustness benchmarks by a large margin

Preliminary: Robust Evaluation of ~~V+L~~ VQA Models

- *VQA-CP* [Agrawal et al. CVPR 2018]
  - The first VQA robustness benchmark on *answer distribution shift*
  - Constructed by reshuffling original VQA train and test splits
  - Previous methods focusing on VQA-CP
    - Auxiliary model as regularizer [NeurIPS 19, EMNLP 19]
    - Additional supervision with human-generated attention maps [ICCV 19, EMNLP 20]
    - Synthesize counterfactual examples as data augmentation [CVPR 2020]
- Recent work: *GQA-OOD* [Kervadec et al. arXiv Preprint 2020]
  - Designed based on a fine-grained reorganization of GQA dataset

Preliminary: Robust Evaluation of ~~V+L~~ VQA Models

- *Linguistic Variation*
  - *VQA-Rephrasings* [Shah et al. CVPR 2019] collects human-generated rephrasings of original VQA question
- *Logical Reasoning*
  - *VQA-LOL* [Gokhale et al. ECCV 2020] with logical combination of Y/N questions
  - *VQA-Introspect* [Selvaraju et al. CVPR 2020] with high-level reasoning questions and low-level perceptual questions
  - *GQA* [Hudson and Manning CVPR 2019] with rule-based questions to analyze reasoning skills of VQA model
- *Visual Content Manipulation*
  - *IV-VQA* and *CV-VQA* [Agarwal et al. CVPR 2020] introduces manipulated images with irrelevant objects removed

Robust VQA Benchmarks

- Compilation of 9 diverse VQA datasets covering 4 types of 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*

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.

Robust VQA Benchmarks

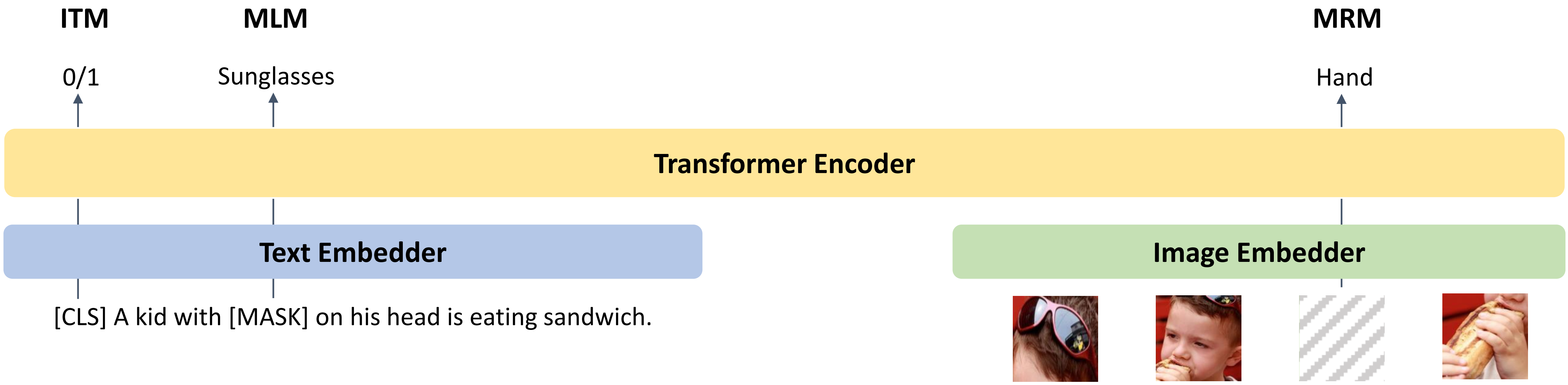
Type	Benchmark	Metric	Q Type	Train			Val		Test	
				Source	#IQ	len(Q)	#IQ	len(Q)	#IQ	len(Q)
Lingual	VQA-Rep. [58]	Acc.	All	VQA v2 [20] train	444K	6.20	162K	7.15	-	-
Reason	VQA-LOL Comp. [18]	Acc.	Y/N	VQA v2 train	444K	6.20	43K	12.09	291K	12.12
	VQA-LOL Supp. [18]	Acc.	Y/N	VQA v2 train	444K	6.20	9K	15.15	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

MANGO Framework

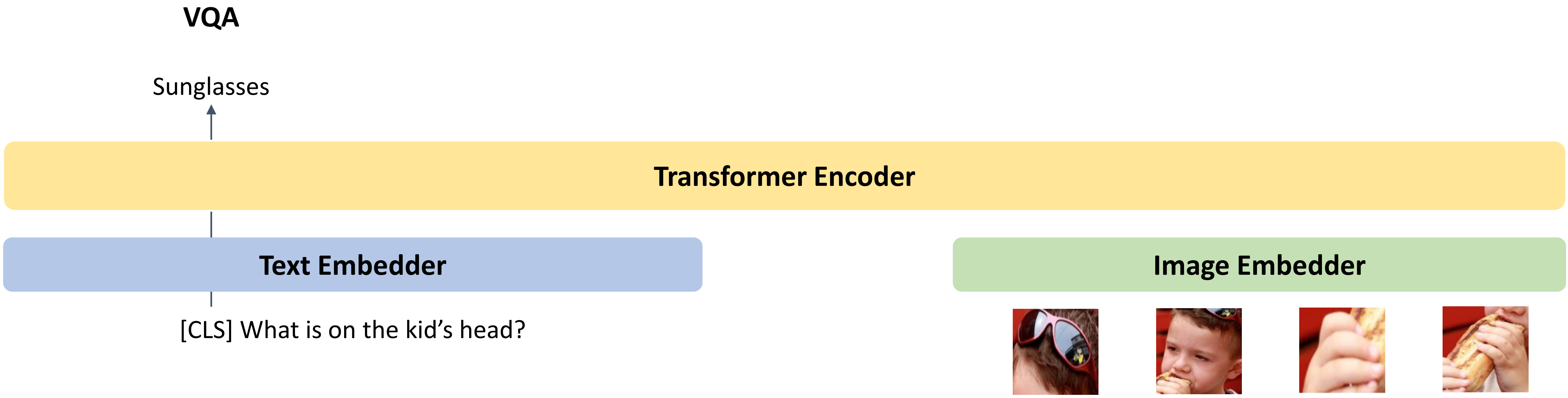
- Preliminary: Pre-trained V+L Models



V+L Pre-training

MANGO Framework

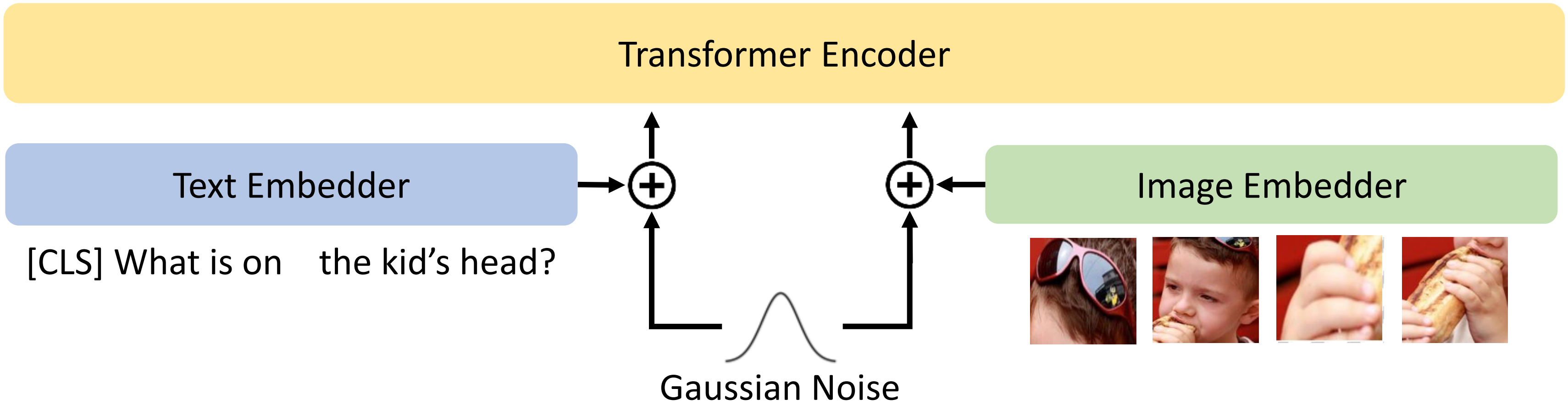
- Preliminary: Pre-trained V+L Models



VQA Finetuning

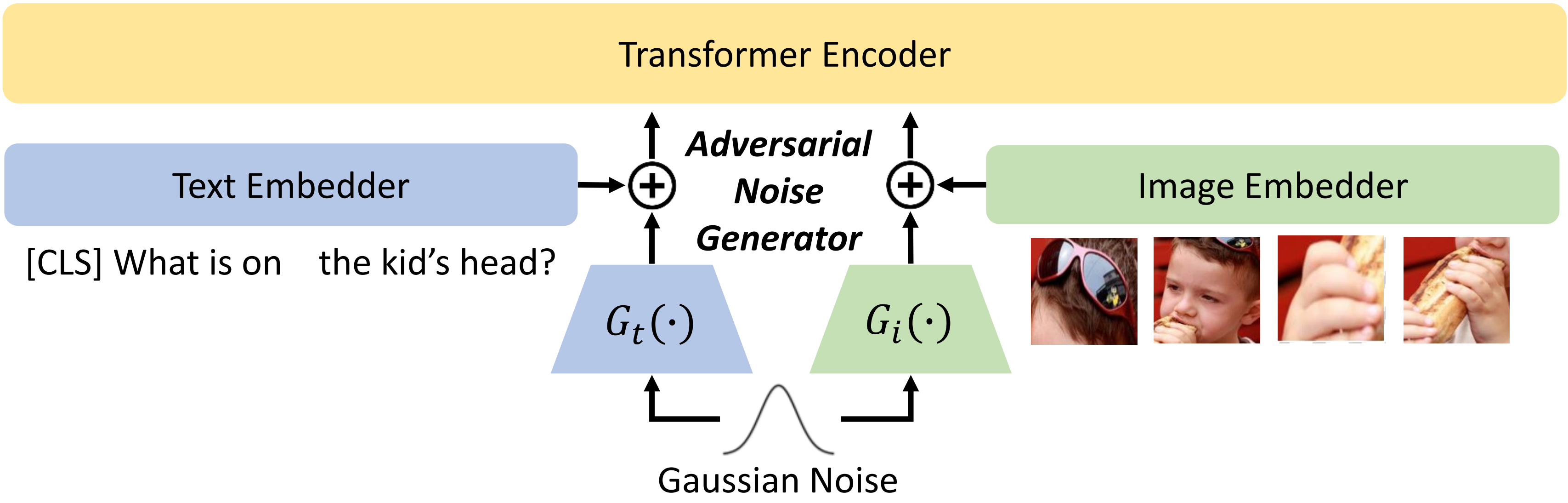
MANGO Framework

- Baseline: Gaussian Noise Augmentation



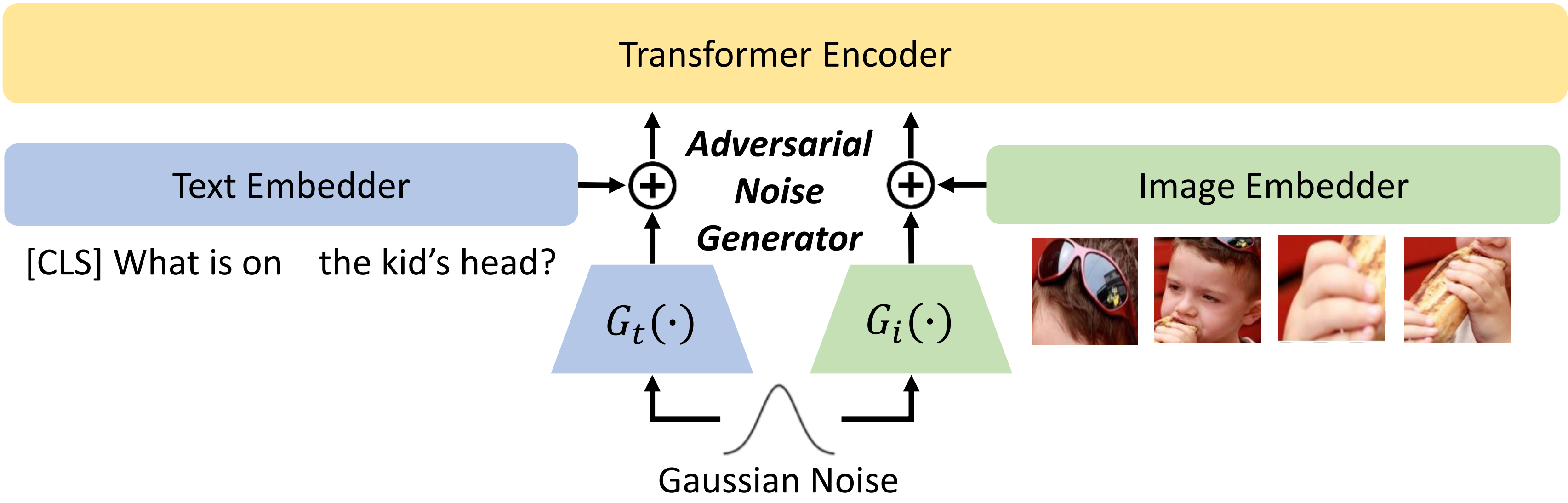
MANGO Framework

- Adversarial Noise Generator



MANGO Framework

- Adversarial Noise Generator



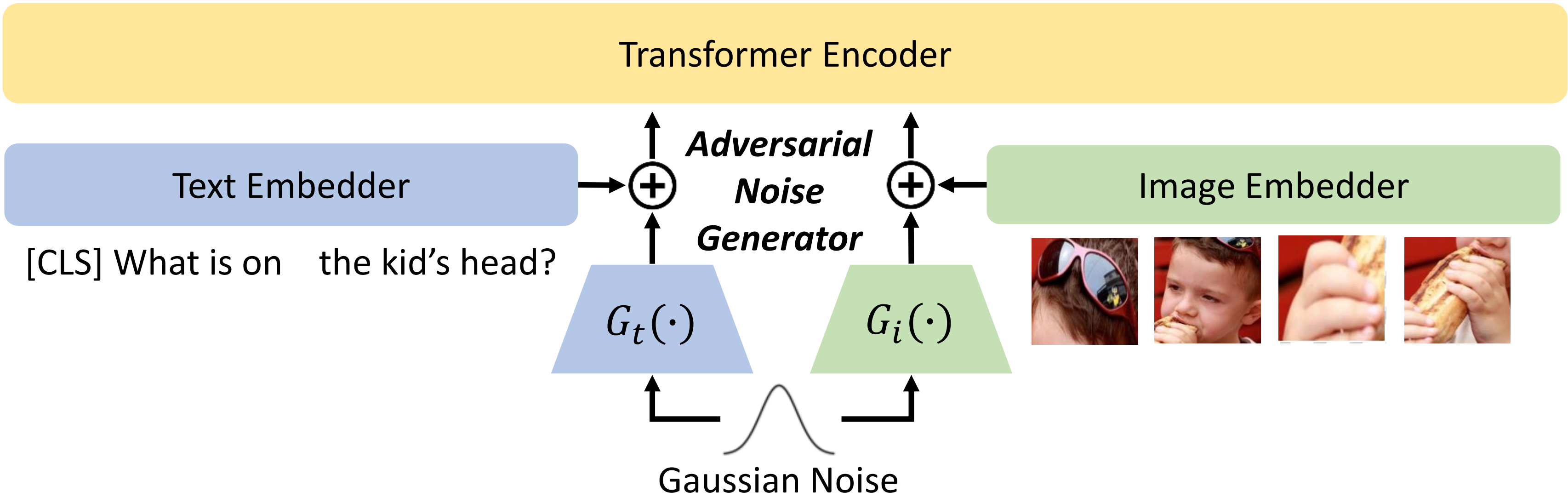
Minimize for
V+L models

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

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

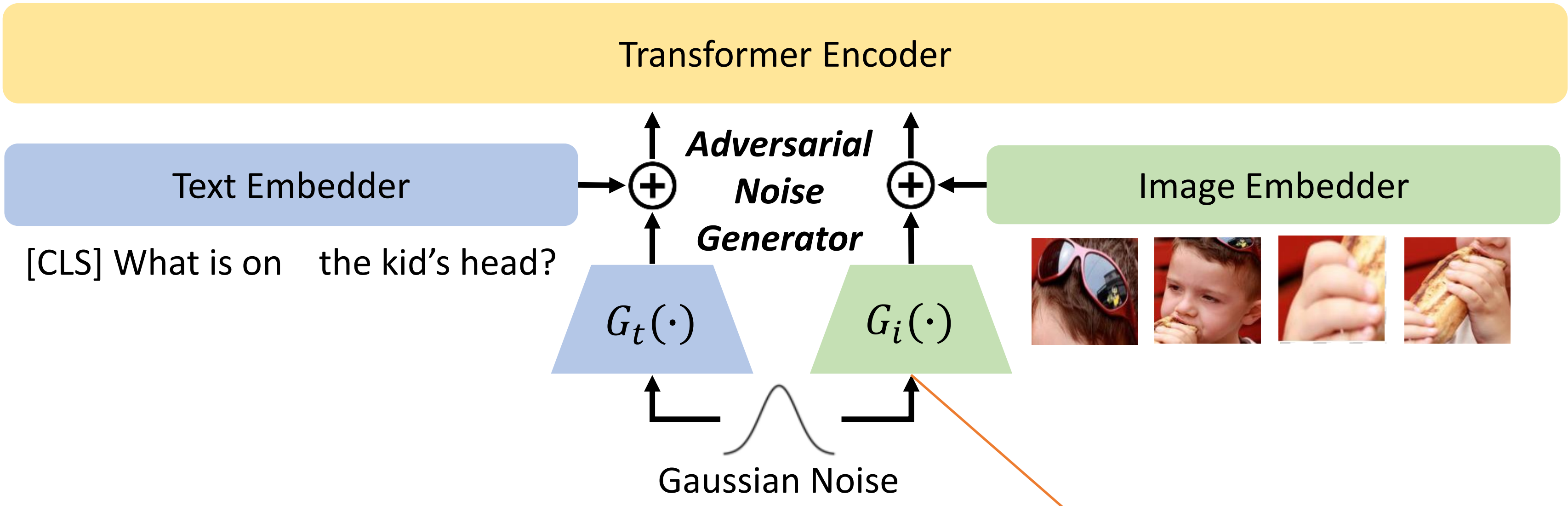
$$\mathcal{R}_{at}(\boldsymbol{\theta},\boldsymbol{\phi}_v) = \mathcal{L}_{\text{BCE}}(f_{\boldsymbol{\theta}}(\boldsymbol{v}+g_{\boldsymbol{\phi}_v}(\boldsymbol{\alpha}),\boldsymbol{w}),\boldsymbol{y}) + \mathcal{L}_{kl}(f_{\boldsymbol{\theta}}(\boldsymbol{v}+g_{\boldsymbol{\phi}_v}(\boldsymbol{\alpha}),\boldsymbol{w}),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

VQA task loss on perturbed inputs

KL Divergence between clean inputs and perturbed inputs

MANGO Framework

- Adversarial Noise Generator



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

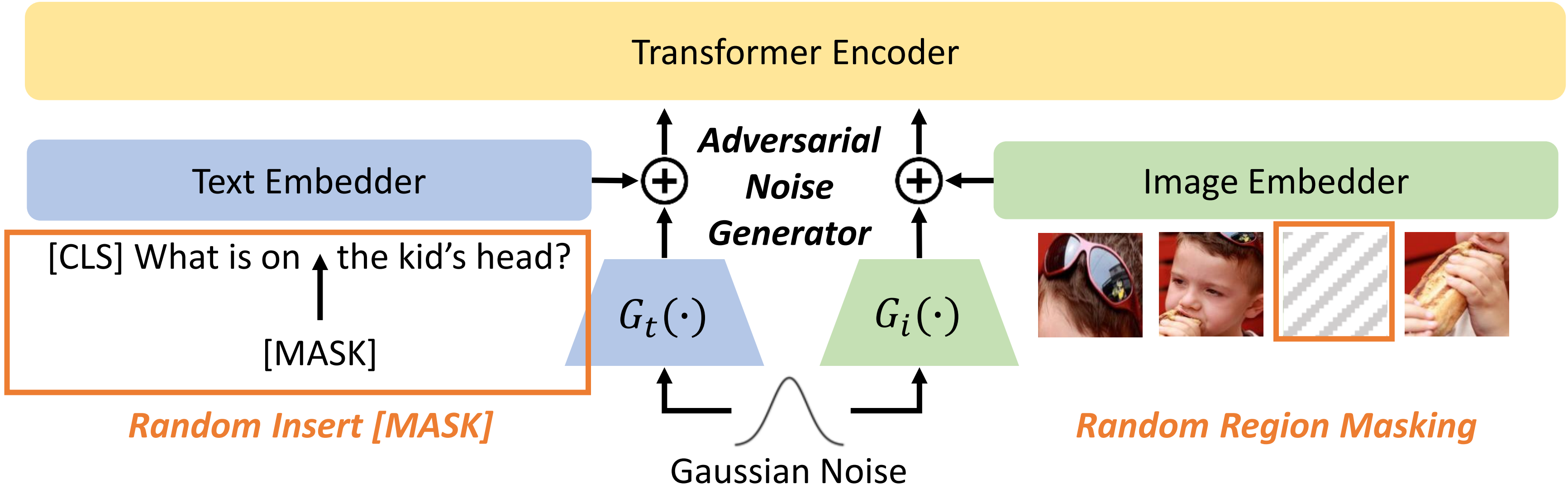
$$\mathcal{L}_{std}(\boldsymbol{\theta}, \boldsymbol{\phi}_v) = \mathcal{L}_{\text{BCE}}(f_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{w}), \boldsymbol{y})$$

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

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

Experimental Results

- Benchmarks: 9 Robust VQA benchmarks + standard VQA v2
- Methods for comparison:
  - SOTA – task-specific state of the art
  - UNITER-B and UNITER-L
  - VILLA-B and VILLA-L
  - MANGO-B and MANGO-L – applying MANGO to UNITER-pretrained models
  - MANGO-VB and MANGO-VL – applying MANGO to VILLA-pretrained models

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	65.35	54.90	56.17	58.29	60.26	7.07	38.28	46.39	54.11	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

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	65.35	54.90	56.17	58.29	60.26	**7.07**	38.28	46.39	54.11	**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

- UNITER-B establishes a strong baseline
- MANGO-B achieves across-the-board performance lift on all benchmarks over UNTIER-B, including VQA v2
- MANGO-VB outperforms VILLA-B on 7 out of 9 robustness benchmarks, but is 25% faster

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	65.35	54.90	56.17	58.29	60.26	7.07	38.28	46.39	54.11	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

- Scaling up to large model size, we observe consistent performance improvement as in other V+L pretraining works
- MANGO further pushes the margins of performance gain across all benchmarks

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
			+11.74	+12.50	+9.96	+12.55		+0.84	+42.92		+3.70	
4	VILLA$_B$	42.37	[illegible]	[illegible]	[illegible]	[illegible]	60.26	[illegible]	[illegible]	[illegible]5.39	[illegible]	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)

A Closer Look at Robustness: Lingual

	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	**65.35**	54.90	56.17	58.29	60.26	7.07	38.28	46.39	54.11	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

- Excessive variations of textual inputs seen during pre-training may help UNITER defending model robustness
- Random masking introduced from the text modality enables more diverse adversarial examples for MANGO, compared to VILLA

A Closer Look at Robustness: Reason

	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	65.35	54.90	56.17	58.29	60.26	7.07	38.28	46.39	54.11	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

- UNITER suffers on VQA-LOL, especially VQA-LOL Supplement
- VILLA brings performance lift on all 4 reasoning benchmarks
- MANGO-VB outperforms VILLA-B, especially on VQA-LOL, whose question length is much longer than that in VQA v2

A Closer Look at Robustness: Visual

	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	65.35	54.90	56.17	58.29	60.26	**7.07**	38.28	46.39	54.11	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

- Diverse images seen during pre-training may help UNITER defending model robustness
- MANGO-B, VILLA-B, MANGO-VB performs on par to each other

A Closer Look at Robustness: Answer

	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	65.35	54.90	56.17	58.29	60.26	7.07	38.28	46.39	54.11	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

- VILLA suffers on VQA-CP v2, with performance degradation comparing to UNITER
- MANGO outperforms VILLA on both benchmarks, with better generalizability to challenging OOD datasets

Ablation Study

Modality		Method	VQA-Rep.	VQA-LOL Comp.	VQA-LOL Supp.	IV-VQA	VQA-CP v2
			Acc. ↑	Acc. ↑	Acc. ↑	#flips ↓	Acc. ↑
None	1	None	64.56	54.54	50.00	8.47	47.29
Image	2	GN	65.17	54.46	50.68	8.45	47.29
	3	AN	65.42	54.59	52.54	7.52	47.38
	4	MANGO	65.51	**56.67**	55.20	7.39	47.51
Text	5	GN	64.73	53.66	54.59	8.46	46.59
	6	AN	65.36	54.12	52.95	7.99	47.09
	7	MANGO	65.63	55.79	**56.54**	7.53	47.45
Both	8	MANGO	**65.80**	56.22	56.49	**7.32**	**47.52**

Ablation Study

Modality		Method	VQA-Rep. Acc. ↑	VQA-LOL Comp. Acc. ↑	VQA-LOL Supp. Acc. ↑	IV-VQA #flips ↓	VQA-CP v2 Acc. ↑
None	1	None	64.56	54.54	50.00	8.47	47.29
Image	2	GN	65.17	54.46	50.68	8.45	47.29
	3	AN	65.42	54.59	52.54	7.52	47.38
	4	MANGO	65.51	**56.67**	55.20	7.39	47.51
Text	5	GN	64.73	53.66	54.59	8.46	46.59
	6	AN	65.36	54.12	52.95	7.99	47.09
	7	MANGO	65.63	55.79	**56.54**	7.53	47.45
Both	8	MANGO	**65.80**	56.22	56.49	**7.32**	**47.52**

- Adding Gaussian noise (GN) to multimodal embeddings is not always helpful

Ablation Study

Modality		Method	VQA-Rep. Acc. ↑	VQA-LOL Comp. Acc. ↑	VQA-LOL Supp. Acc. ↑	IV-VQA #flips ↓	VQA-CP v2 Acc. ↑
None	1	None	64.56	54.54	50.00	8.47	47.29
Image	2	GN	65.17	54.46	50.68	8.45	47.29
	3	AN	65.42	54.59	52.54	7.52	47.38
	4	MANGO	65.51	**56.67**	55.20	7.39	47.51
Text	5	GN	64.73	53.66	54.59	8.46	46.59
	6	AN	65.36	54.12	52.95	7.99	47.09
	7	MANGO	65.63	55.79	**56.54**	7.53	47.45
Both	8	MANGO	**65.80**	56.22	56.49	**7.32**	**47.52**

- Adding Gaussian noise (GN) to multimodal embeddings is not always helpful
- Adversarial Noise (AN) brings universal performance improvements over GN

Ablation Study

Modality		Method	VQA-Rep. Acc. ↑	VQA-LOL Comp. Acc. ↑	VQA-LOL Supp. Acc. ↑	IV-VQA #flips ↓	VQA-CP v2 Acc. ↑
None	1	None	64.56	54.54	50.00	8.47	47.29
Image	2	GN	65.17	54.46	50.68	8.45	47.29
	3	AN	65.42	54.59	52.54	7.52	47.38
	4	MANGO	65.51	**56.67**	55.20	7.39	47.51
Text	5	GN	64.73	53.66	54.59	8.46	46.59
	6	AN	65.36	54.12	52.95	7.99	47.09
	7	MANGO	65.63	55.79	**56.54**	7.53	47.45
Both	8	MANGO	**65.80**	56.22	56.49	**7.32**	**47.52**

- Adding Gaussian noise (GN) to multimodal embeddings is not always helpful
- Adversarial Noise (AN) brings universal performance improvements over GN
- Through random masking, MANGO is better than using AN alone

Ablation Study

Modality		Method	VQA-Rep.	VQA-LOL Comp.	VQA-LOL Supp.	IV-VQA	VQA-CP v2
			Acc. ↑	Acc. ↑	Acc. ↑	#flips ↓	Acc. ↑
None	1	None	64.56	54.54	50.00	8.47	47.29
Image	2	GN	65.17	54.46	50.68	8.45	47.29
	3	AN	65.42	54.59	52.54	7.52	47.38
	4	MANGO	65.51	**56.67**	55.20	7.39	47.51
Text	5	GN	64.73	53.66	54.59	8.46	46.59
	6	AN	65.36	54.12	52.95	7.99	47.09
	7	MANGO	65.63	55.79	**56.54**	7.53	47.45
Both	8	MANGO	**65.80**	56.22	56.49	**7.32**	**47.52**

- Adding Gaussian noise (GN) to multimodal embeddings is not always helpful
- Adversarial Noise (AN) brings universal performance improvements over GN
- Through random masking, MANGO is better than using AN alone
- Empirically, MANGO on both modalities performs on par with on single modality

Ablation Study

Task	LXMERT	Ours
VQA-Rep.	67.20	**68.61**
VQA-LOL Comp.	49.34	**53.83**
VQA-LOL Supp.	47.33	**53.54**
GQA	59.78	**60.06**
GQA-OOD	53.86	**54.94**
VQA v2	72.31	**72.70**

- MANGO is also generalizable to two-stream backbone LXMERT, with universal performance lift

Model	$NLVR^2$	RefCOCO	RefCOCOg	VE
UNITER_B	77.52	80.55	74.41	78.44
MANGO_B	**78.36**	**80.95**	**75.37**	**78.87**

- MANGO can be also applied to other standard V+L tasks, with performance improvements over baseline

Conclusion

- First known systematic study on robustness of pre-trained V+L models
- A simple yet efficient adversarial training method to enhance model robustness: MANGO
- MANGO advances SOTA on 7 out of 9 robustness benchmarks by a large margin