Visual Question Answering and Beyond

Aishwarya Agrawal

Ph.D. Candidate

School of Interactive Computing

Georgia Institute of Technology

VQA Task

VQA Task

What is the mustache made of?

VQA Task

What is the mustache made of?

AI System

VQA Task

What is the mustache made of?

AI System

bananas

Applications of VQA

- An aid to visually-impaired

  Is it safe to cross the street now?

Applications of VQA

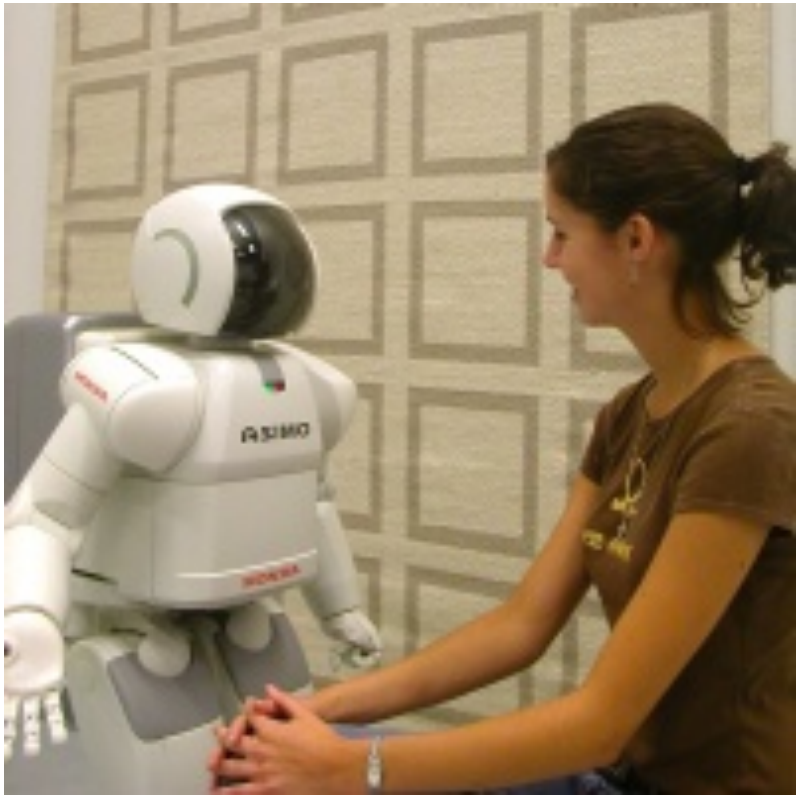
- Surveillance

  What kind of car did the man in red shirt leave in?

Applications of VQA

- Interacting with personal assistants

  Is my laptop in my bedroom upstairs?

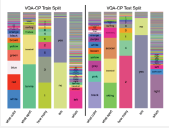


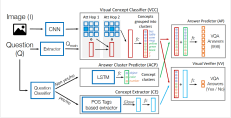
Outline

Overview of VQA
[ICCV'15, IJCV'16, AI Mag'16]

Problem with existing setup + models
[EMNLP'16]

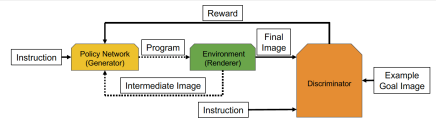
Overcoming priors
- A new evaluation protocol [CVPR'18]
- A novel architecture [CVPR'18]
- A novel objective function [NIPS'18]





$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_Q(f_Q, g) - \lambda_H \mathcal{L}_H(f,g,h,f_Q)$$

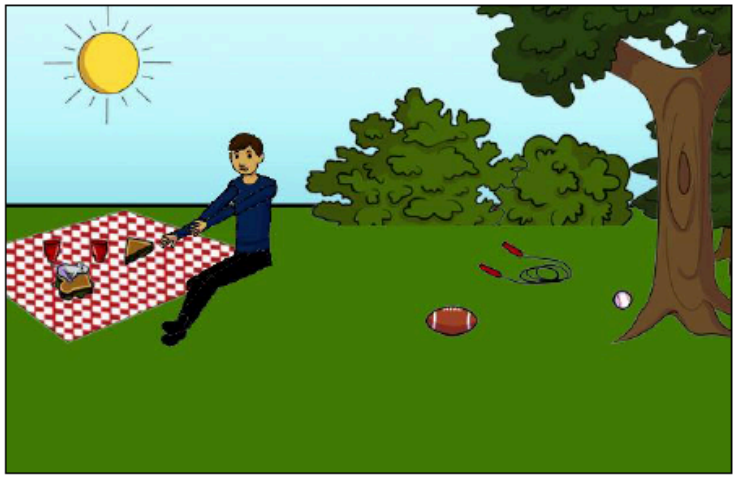
Beyond VQA
[Work in progress]



VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

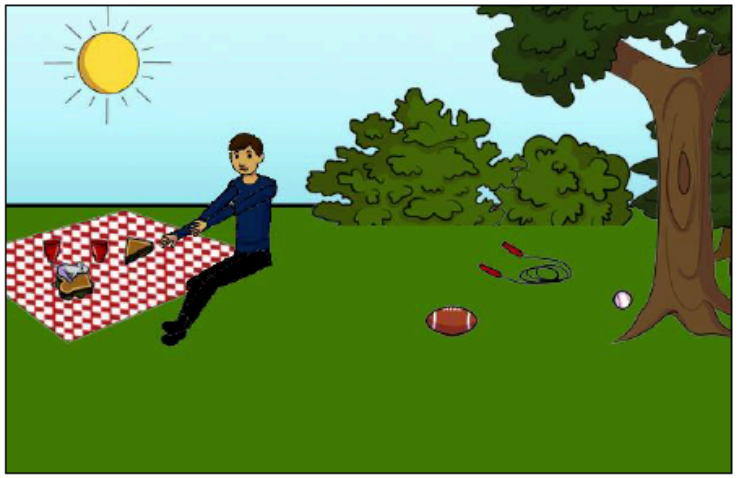
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA Dataset

What color are her eyes?
What is the mustache made of?

About objects

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

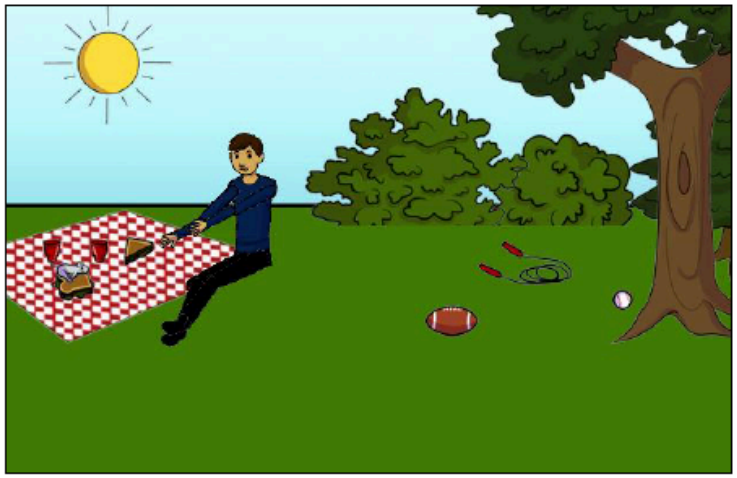
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Counting



Is this person expecting company?
What is just under the tree?

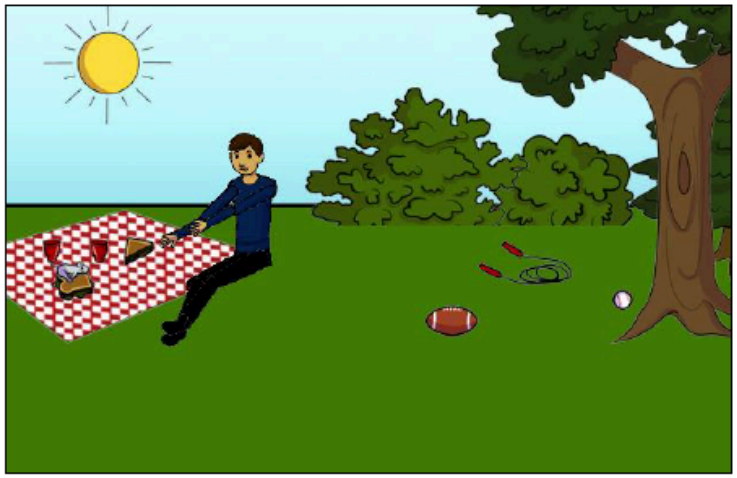
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Fine-grained recognition



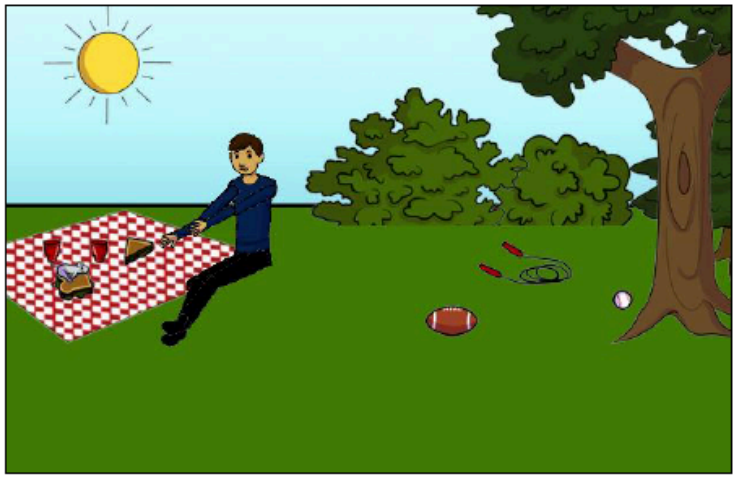
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Common sense

VQA

- Multimodal inputs – Image and Question
- Details of the image
- Common sense + knowledge base
- Task-driven
- Holy-grail of automatic image understanding

VQA Dataset Stats

>0.25 million images

>0.76 million questions

~10 million answers

Please visit www.visualqa.org for more details.

VQA Visual Question Answering

VirginiaTech Invent the Future®

Microsoft Research

Home | People | Download | Evaluation | Challenge | Browse | Visualize | Terms

Full release is out! 254,721 images, 764,163 questions, 9,934,119 answers.
VQA challenge announced! The VQA evaluation server is now up.

What is VQA?

VQA is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.

- Over 250K images (MSCOCO and abstract scenes)
- 3 questions per image
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question
- Open-ended and multiple-choice answering tasks
- Automatic evaluation metric

Subscribe to our group for updates!

Dataset

Details on downloading the latest dataset may be found on the download webpage.

October 2015: Full release (v1.0)

Real Images
- 204,721 MSCOCO images (all of current train/val/test)
- 614,163 questions
- 6,141,630 ground truth answers
- 1,842,489 plausible answers

Abstract Scenes
- 50,000 abstract scenes
- 150,000 questions
- 1,500,000 ground truth answers
- 450,000 plausible answers
- 250,000 captions

⊕ **July 2015: Beta v0.9 release**

⊕ **June 2015: Beta v0.1 release**

Paper

Download the paper

BibTeX

What color are her eyes?
What is the mustache made of?

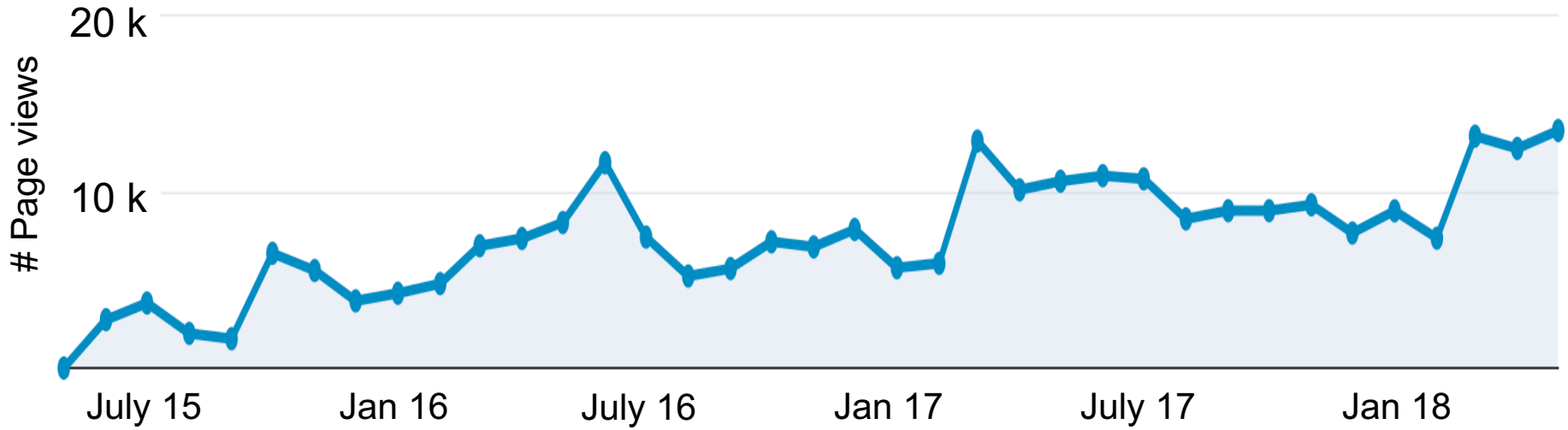
How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Contact: visualqa@gmail.com

Interest in VQA (http://www.visualqa.org/)



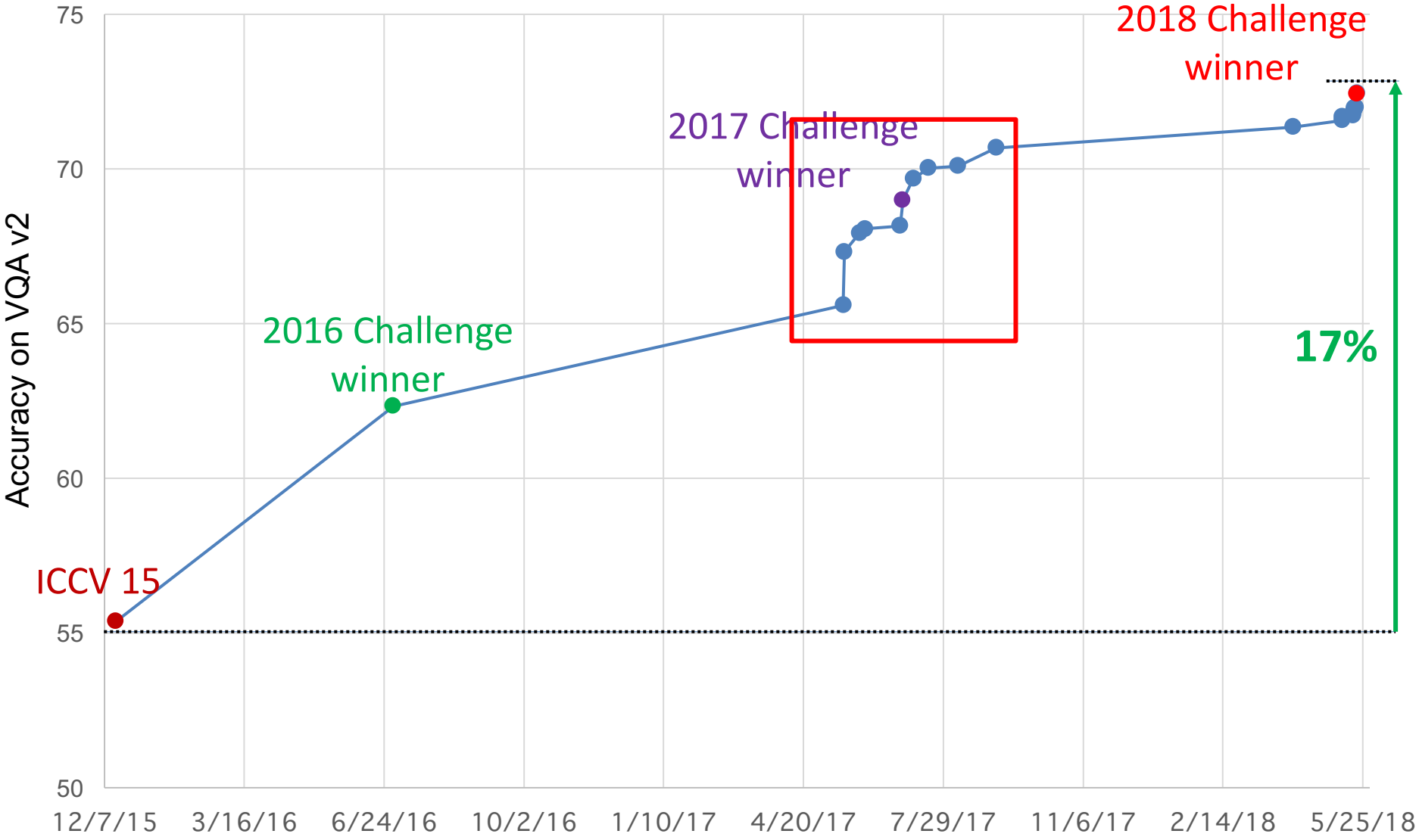
13k page views/month during VQA Challenge 2018

Slide Credit: Yash Goyal

Other VQA Datasets

- Visual Turing Test [Geman et al., PNAS 2014]
- DAQUAR [Malinowski & Fritz, NIPS 2014]
- COCO-QA [Ren et al., NIPS 2015]
- FM-IQA [Gao et al., NIPS 2015]
- Visual7W [Zhu et al., CVPR 2016]
- Visual Genome [Krishna et al., IJCV 2016]
- CLEVR [Johnson et al., CVPR 2017]
- VQA v2.0 [Goyal et al., CVPR 2017]

SOTA in VQA over the years



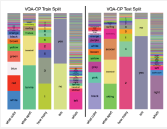
Slide Credit: Yash Goyal

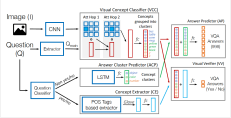
Outline

Overview of VQA
[ICCV'15, IJCV'16, AI Mag'16]

Problem with existing setup + models
[EMNLP'16]

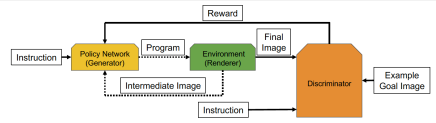
Overcoming priors
- A new evaluation protocol [CVPR'18]
- A novel architecture [CVPR'18]
- A novel objective function [NIPS'18]





$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_Q(f_Q, g) - \lambda_H \mathcal{L}_H(f,g,h,f_Q)$$

Beyond VQA
[Work in progress]



VQA models lack compositionality

Compositionality

Training

Q: What color is the plate?

A: Green

Q: What color are stop lights?

A: Red

Testing

Q: What color is the stop light?

A: Green

Q: What is the color of the plate?

A: Red

Test Sample

Q: What color are the safety cones?

Test Sample

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

Nearest Neighbor Training Samples

Q: What color are the cones?

GT Ans: orange

Q: What color is the cone?

GT Ans: orange

Q: What color are the cones?

GT Ans: orange

VQA models lack compositionality

VQA models are driven by language priors in training data

Q: Are **A:** military

Q: Are they **A:** yes

Q: Are they playing **A:** yes

Q: Are they playing a **A:** yes

Q: Are they playing a game? **A:** yes

GT Ans: yes

VQA models lack compositionality

VQA models are driven by language priors in training data

VQA models lack sufficient image grounding

Looking at the Image

Q: What does the red sign say?

Predicted Ans: stop

Correct Response

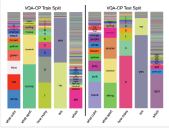
Outline

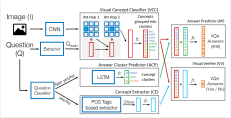
Overview of VQA
[ICCV'15, IJCV'16, AI Mag'16]

Problem with existing setup + models
[EMNLP'16]

Overcoming priors

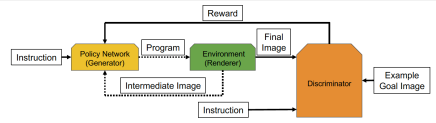
- A new evaluation protocol [CVPR'18]
- A novel architecture [CVPR'18]
- A novel objective function [NIPS'18]



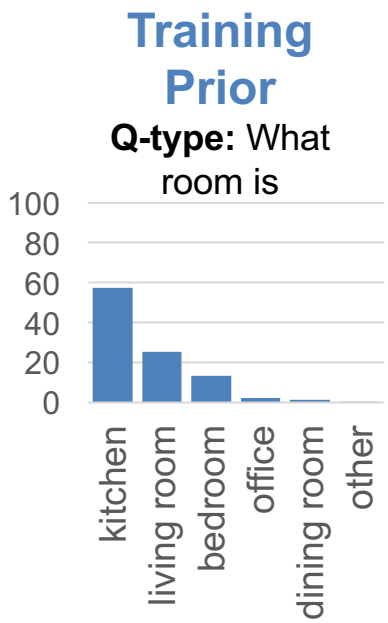


$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_Q(f_Q, g) - \lambda_H \mathcal{L}_H(f,g,h,f_Q)$$

Beyond VQA
[Work in progress]



Problem with existing setup + models



Problem with existing setup + models

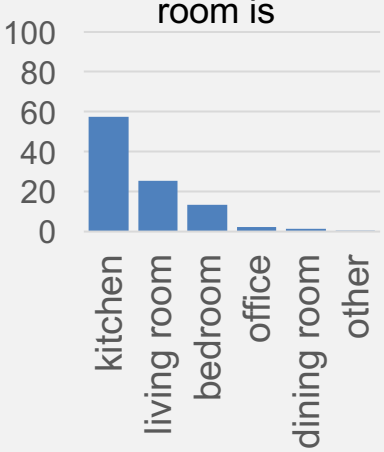
Train

Q: What room is this?

A: Kitchen

Training Prior

Q-type: What room is



Problem with existing setup + models

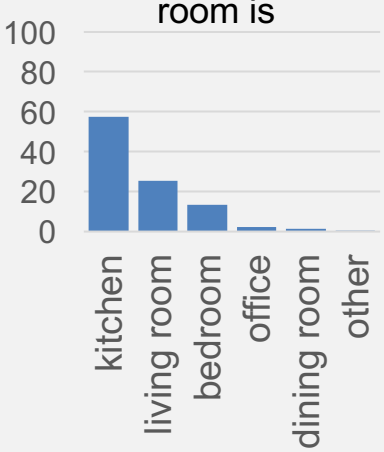
Train

Q: What room is this?

A: Kitchen

Training Prior

Q-type: What room is



Test

Q: What room is this?

A: Bathroom

Problem with existing setup + models

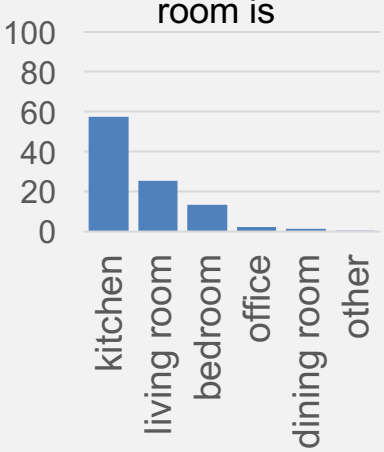
Train

Q: What room is this?

A: Kitchen

Training Prior

Q-type: What room is



Test

Q: What room is this?

A: Bathroom

Prediction

Kitchen

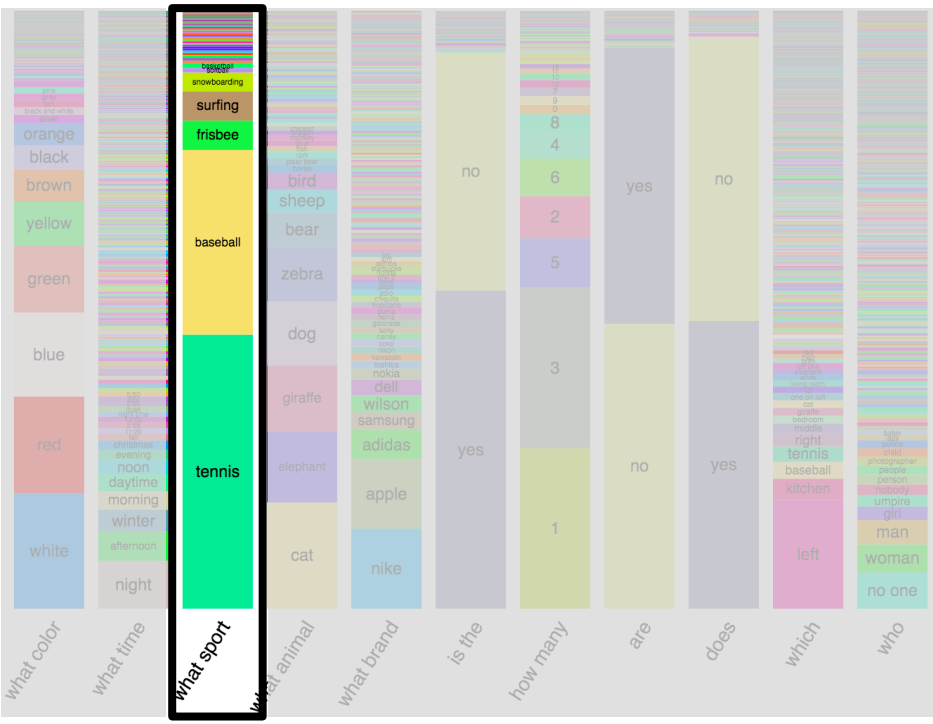
Problem with existing setup + models

- IID splits → similar priors in train and test
- Memorization of priors does not hurt as much
- Problematic for benchmarking progress

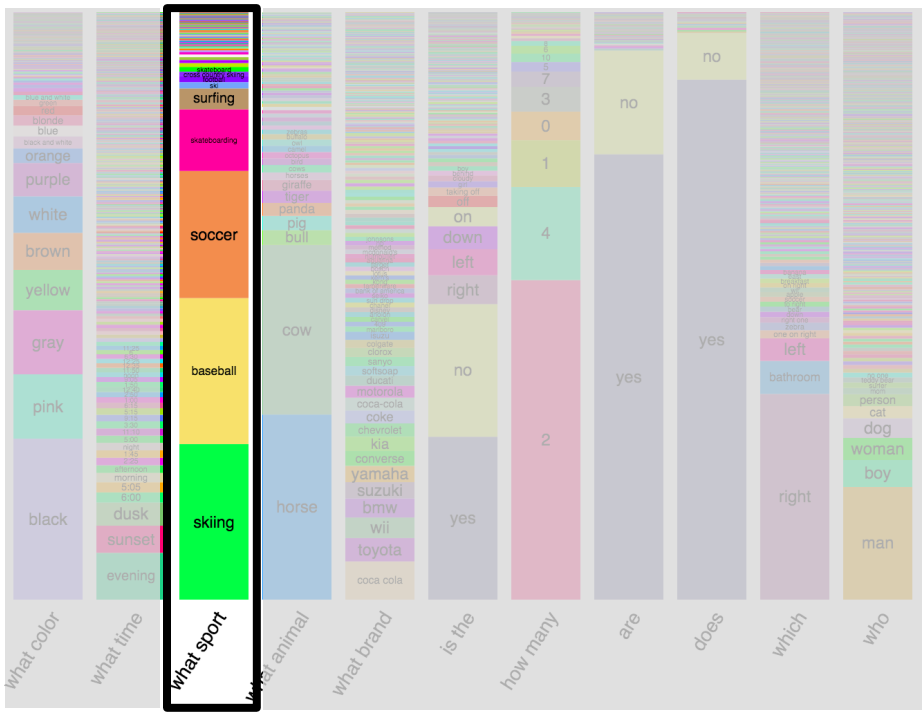
Meet VQA-CP!

- New splits of the VQA v1 and VQA v2 datasets
- Visual Question Answering under Changing Priors (VQA-CP v1/v2)

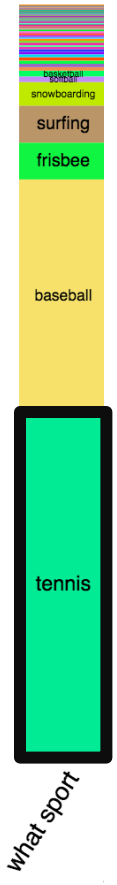
VQA-CP Train Split



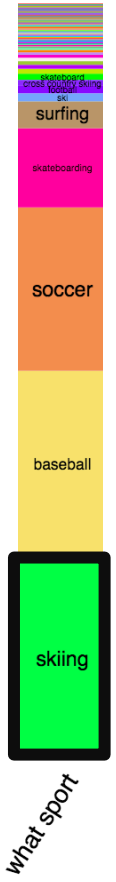
VQA-CP Test Split



VQA-CP Train Split



VQA-CP Test Split



Performance of VQA models on VQA-CP

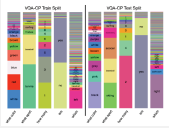
Model	Dataset	Overall	
d-LSTM Q + norm I (Antol et al. ICCV15)	VQA v1	54.40	-31%
	VQA-CP v1	23.51	
NMN (Andreas et al. CVPR16)	VQA v1	54.83	-25%
	VQA-CP v1	29.64	
SAN (Yang et al. CVPR16)	VQA v1	55.86	-29%
	VQA-CP v1	26.88	
MCB (Fukui et al. EMNLP16)	VQA v1	60.97	-27%
	VQA-CP v1	34.39	

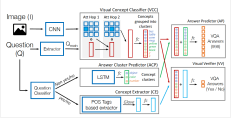
Outline

Overview of VQA
[ICCV'15, IJCV'16, AI Mag'16]

Problem with existing setup + models
[EMNLP'16]

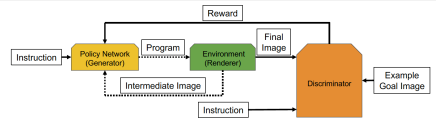
Overcoming priors
- A new evaluation protocol [CVPR'18]
- A novel architecture [CVPR'18]
- A novel objective function [NIPS'18]





$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_Q(f_Q,g) - \lambda_H \mathcal{L}_H(f,g,h,f_Q)$$

Beyond VQA
[Work in progress]



Grounded Visual Question Answering (GVQA) Model

- Inductive biases in model architecture to prevent relying on priors

- Designed to disentangle:
  - What can be said?

Q: What room is this?

Grounded Visual Question Answering (GVQA) Model

- Inductive biases in model architecture to prevent relying on priors

- Designed to disentangle:
  - What can be said?

Q: What room is this?

Grounded Visual Question Answering (GVQA) Model

- Inductive biases in model architecture to prevent relying on priors

- Designed to disentangle:
  - What can be said?
  - What should be recognized?

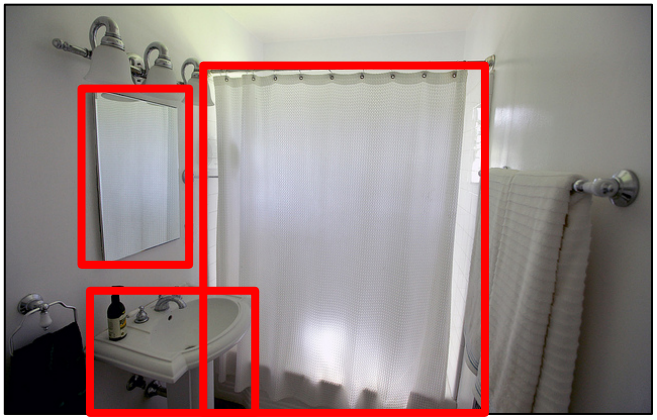
Q: What room is this?

Grounded Visual Question Answering (GVQA) Model

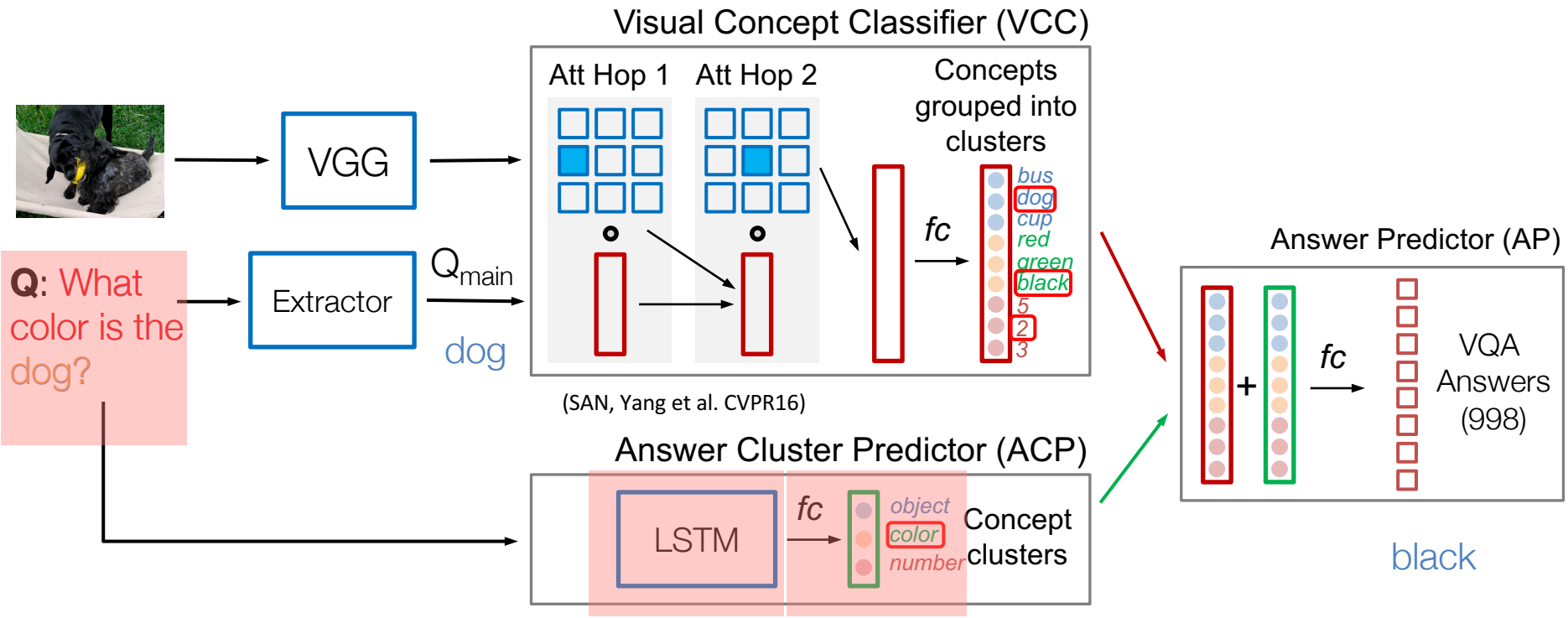
- Inductive biases in model architecture to prevent relying on priors

- Designed to disentangle:
  - What can be said?
  - What should be recognized?

Q: What room is this?



GVQA


Visual Concept Classifier (VCC)
Att Hop 1
Att Hop 2
Concepts grouped into clusters
VGG
Extractor
Q: What color is the dog?
Q_main
dog
fc
bus
dog
cup
red
green
black
5
2
3
(SAN, Yang et al. CVPR16)
Answer Cluster Predictor (ACP)
LSTM
fc
object
color
number
Concept clusters
Answer Predictor (AP)
fc
VQA Answers (998)
black

GVQA

- Disentangles visual recognition from answer-type prediction
- Explicitly enforces visual grounding
- No direct pathway from question to final answer

Results

Dataset	Model	Overall
VQA-CP v1	GVQA (Ours)	**39.23**
	SAN (Yang et al. CVPR16)	26.88
VQA-CP v2	GVQA (Ours)	**31.30**
	SAN (Yang et al. CVPR16)	24.96

+12%

+6%

Problem with existing setup + models

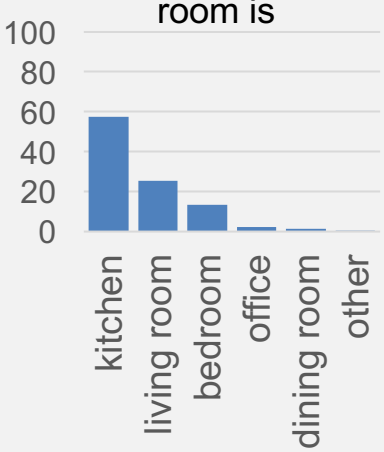
Train

Q: What room is this?

A: Kitchen

Training Prior

Q-type: What room is



Test

Q: What room is this?

A: Bathroom

Prediction

Bathroom

GVQA's output

Q: What color are the bananas?

Q-classifier	ACP	VCC	Answer
non yes/no	color	bananas	
		green	green
		many	✓
		food	

GVQA's output

Q: What is the most prominent ingredient?

Correct Ans: pasta

Q-classifier	ACP	VCC	Answer
non yes/no	vegetable	carrots	
		pasta	carrots
		green	✗
		plate	

Outline

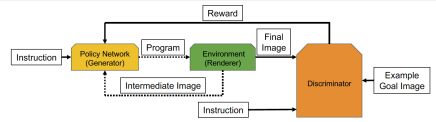
Overview of VQA
[ICCV'15, IJCV'16, AI Mag'16]

Problem with existing setup + models
[EMNLP'16]

Overcoming priors
- A new evaluation protocol [CVPR'18]
- A novel architecture [CVPR'18]
- A novel objective function [NIPS'18]

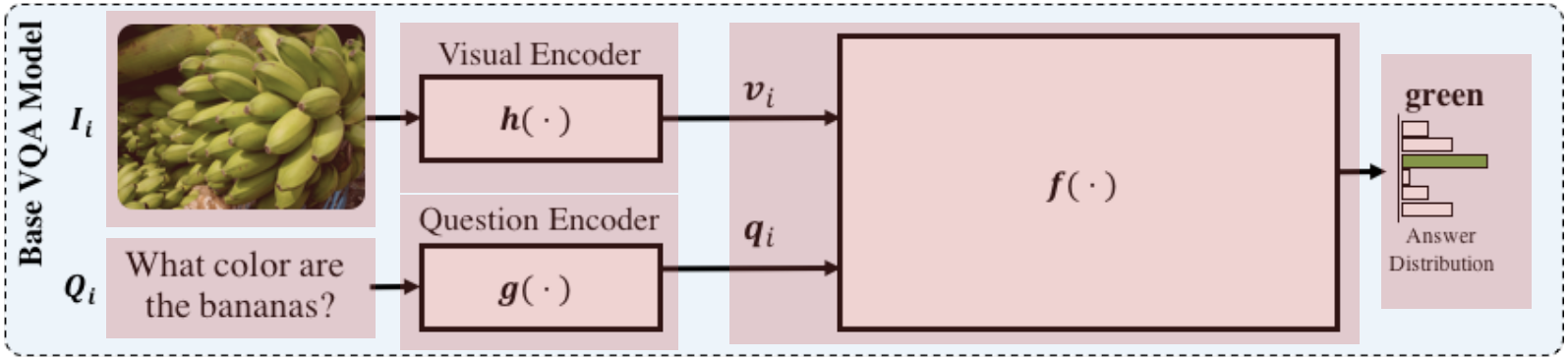
Sainandan Ramakrishnan

Beyond VQA
[Work in progress]

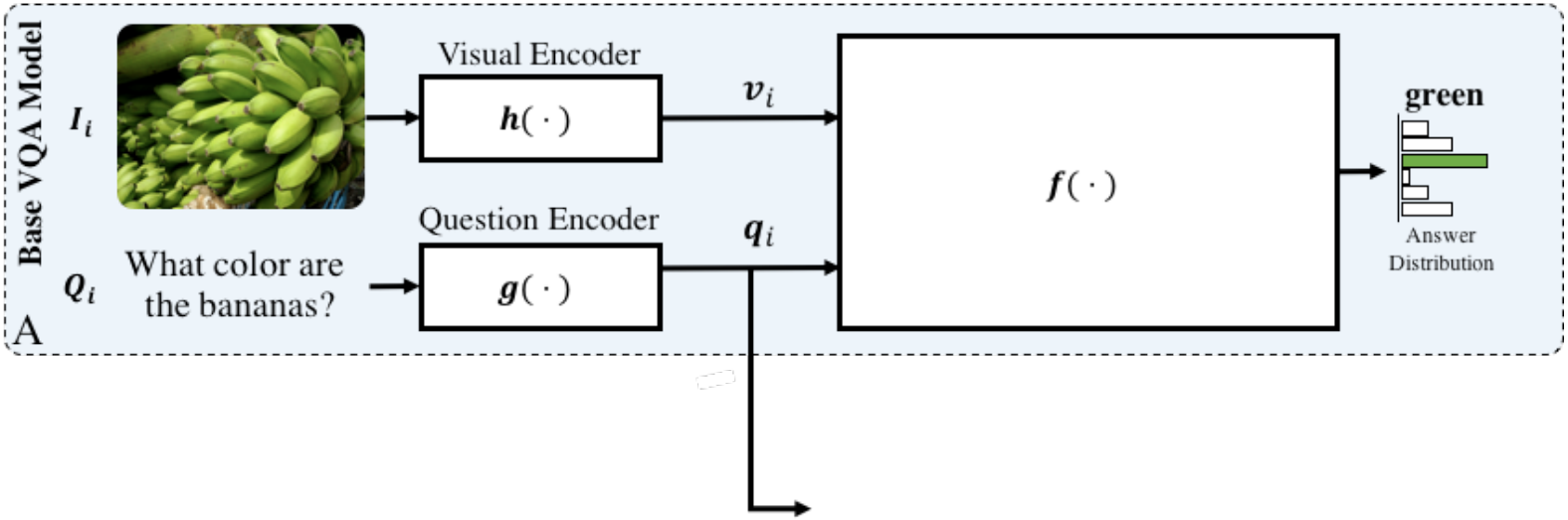


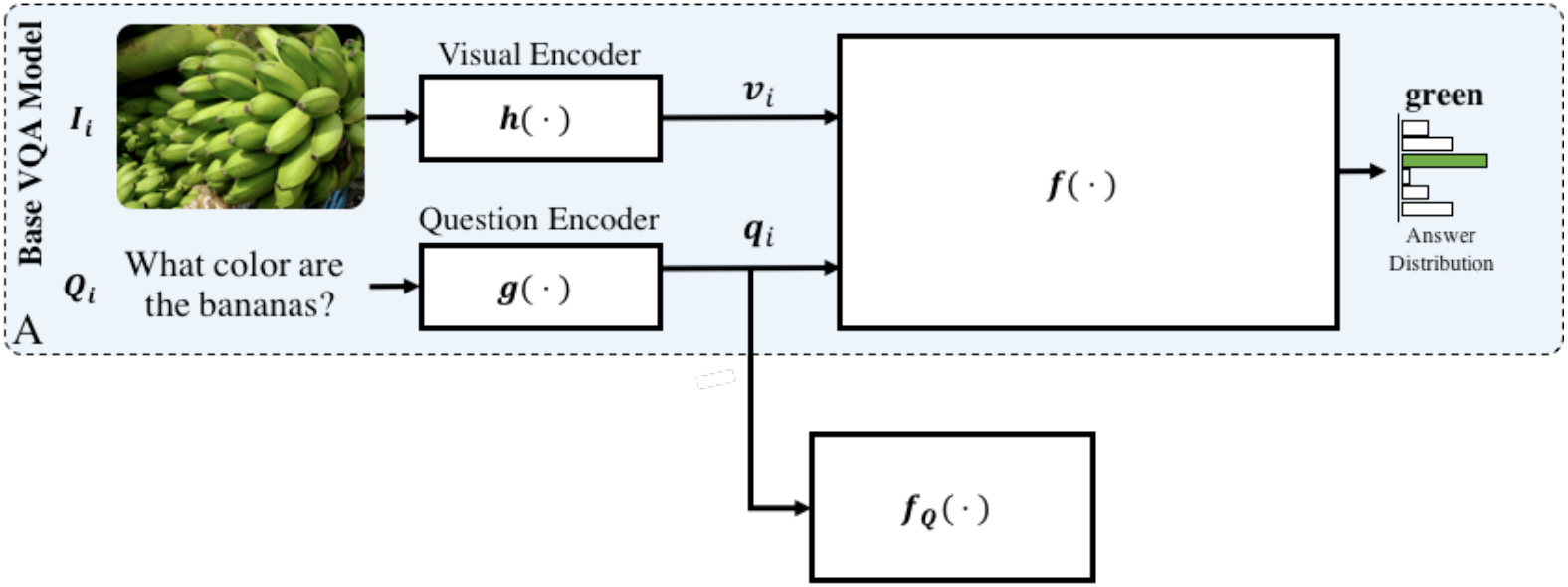
Overcoming Priors with Adversarial Regularization

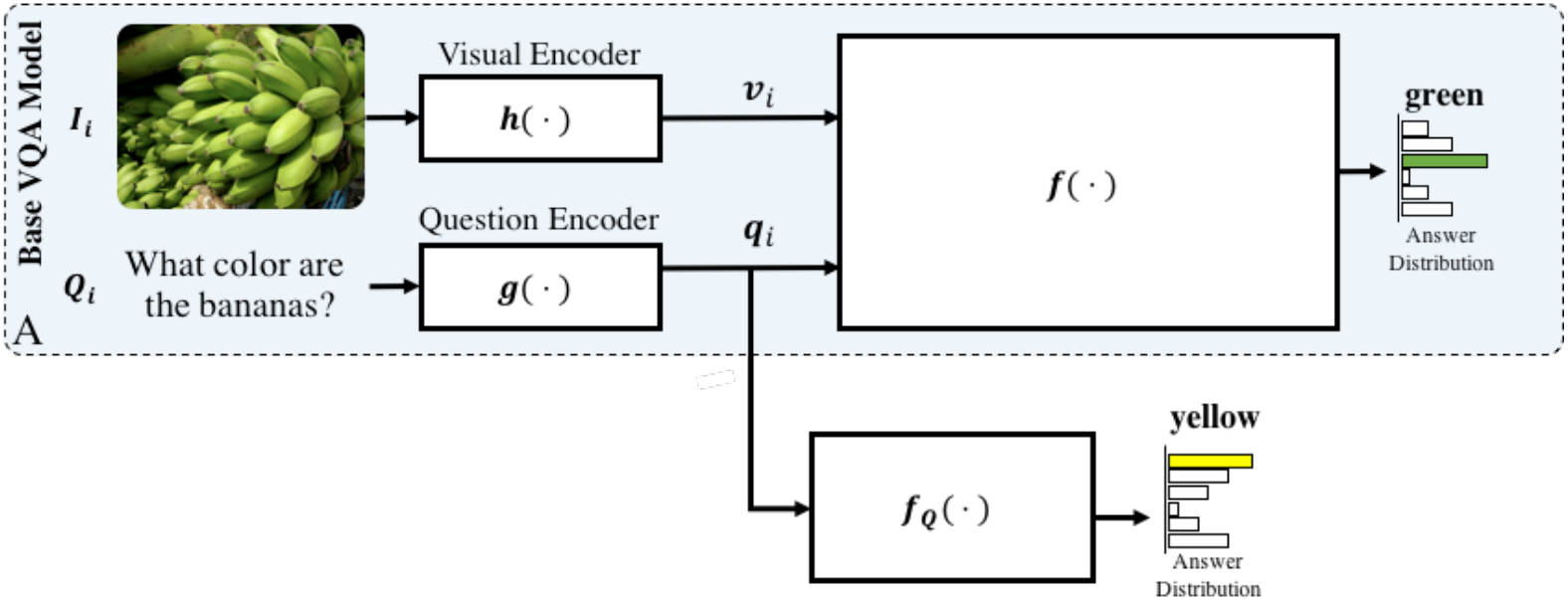
- A simple drop-in regularizer
- Question embeddings should not encode the information about the exact answer

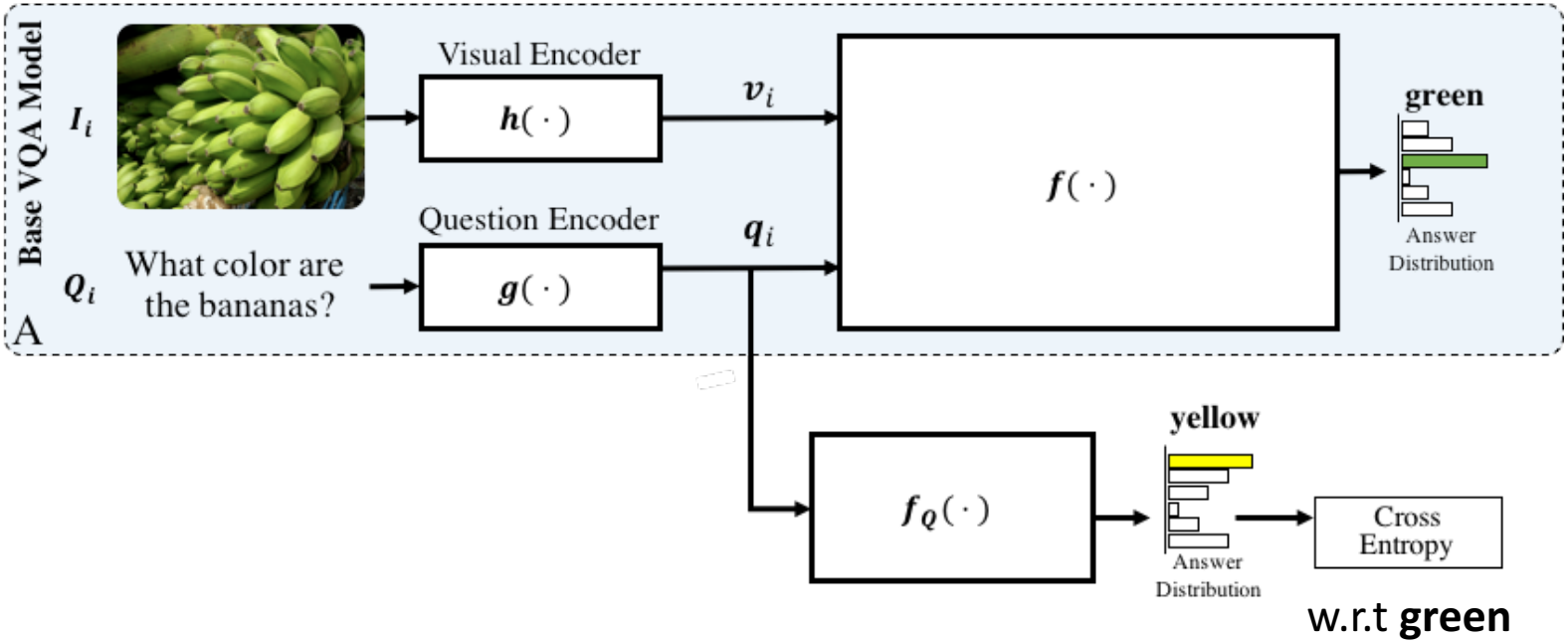


$$\mathcal{L}_{VQA}(f, g, h) \approx -\frac{1}{N}\sum_{i=1}^{N} \log f(\mathbf{v_i}, \mathbf{q}_i)[a_i]$$


Base VQA Model
I_i
Visual Encoder
$h(\cdot)$
v_i
Q_i
What color are the bananas?
Question Encoder
$g(\cdot)$
q_i
$f(\cdot)$
green
Answer Distribution
A

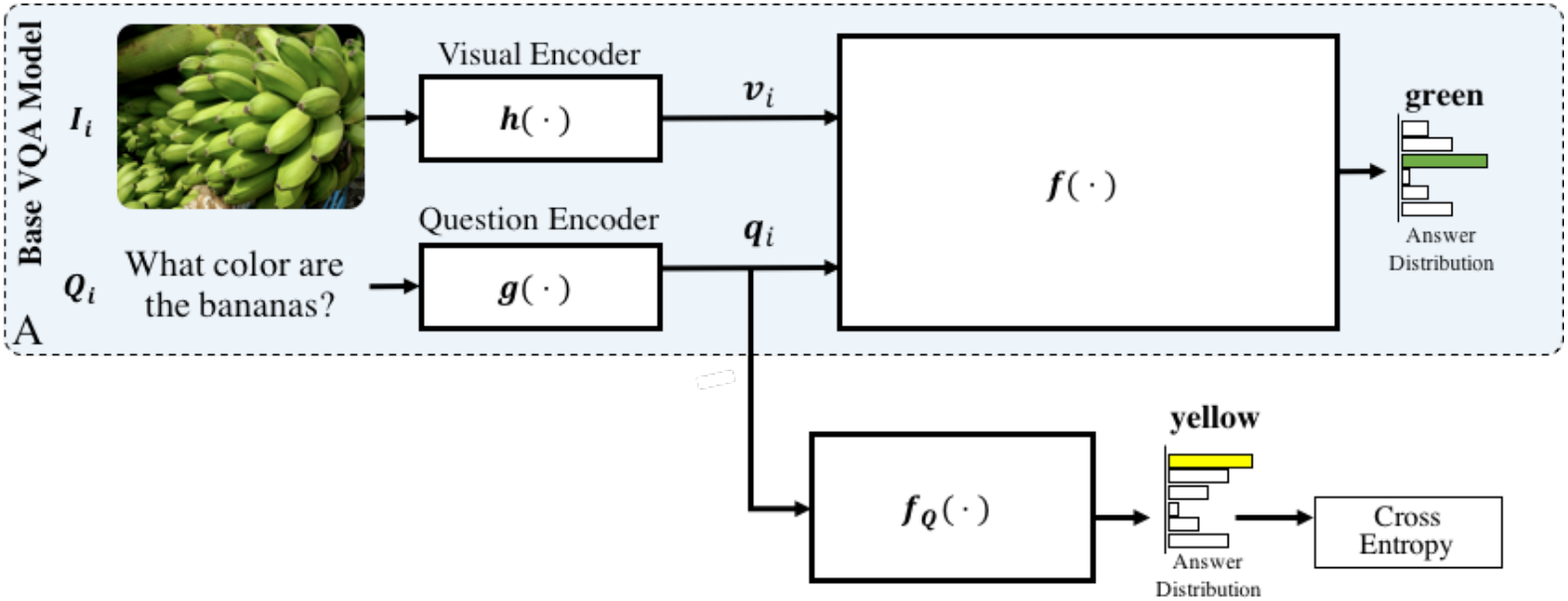

Base VQA Model
I_i
Visual Encoder
$h(\cdot)$
v_i
$f(\cdot)$
green
Answer Distribution
Question Encoder
q_i
Q_i
What color are the bananas?
$g(\cdot)$
A
$f_Q(\cdot)$


Base VQA Model
I_i
Visual Encoder
$h(\cdot)$
v_i
$f(\cdot)$
green
Answer Distribution
Q_i
What color are the bananas?
Question Encoder
$g(\cdot)$
q_i
A
$f_Q(\cdot)$
yellow
Answer Distribution



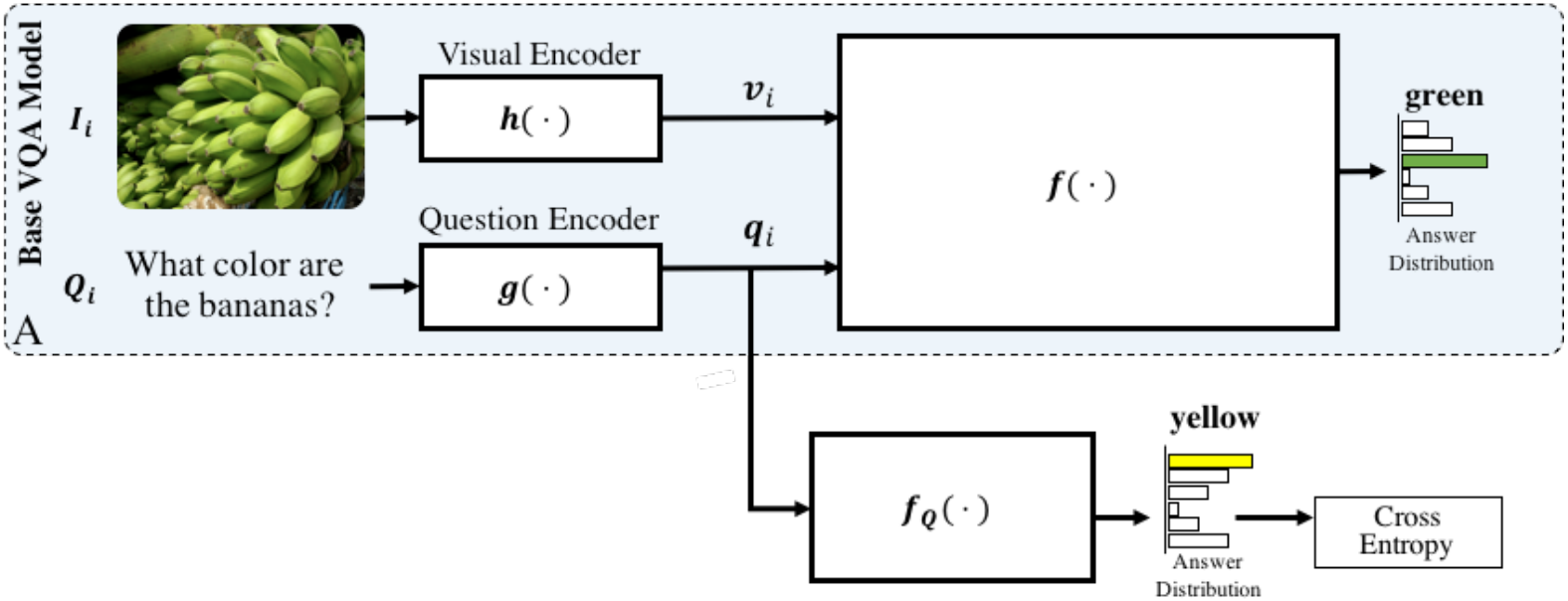
w.r.t **green**

$$\mathcal{L}_{QA}(f_Q, g) \approx -\frac{1}{N}\sum_{i=1}^{N} \log f_Q(\mathbf{q}_i)[a_i]$$

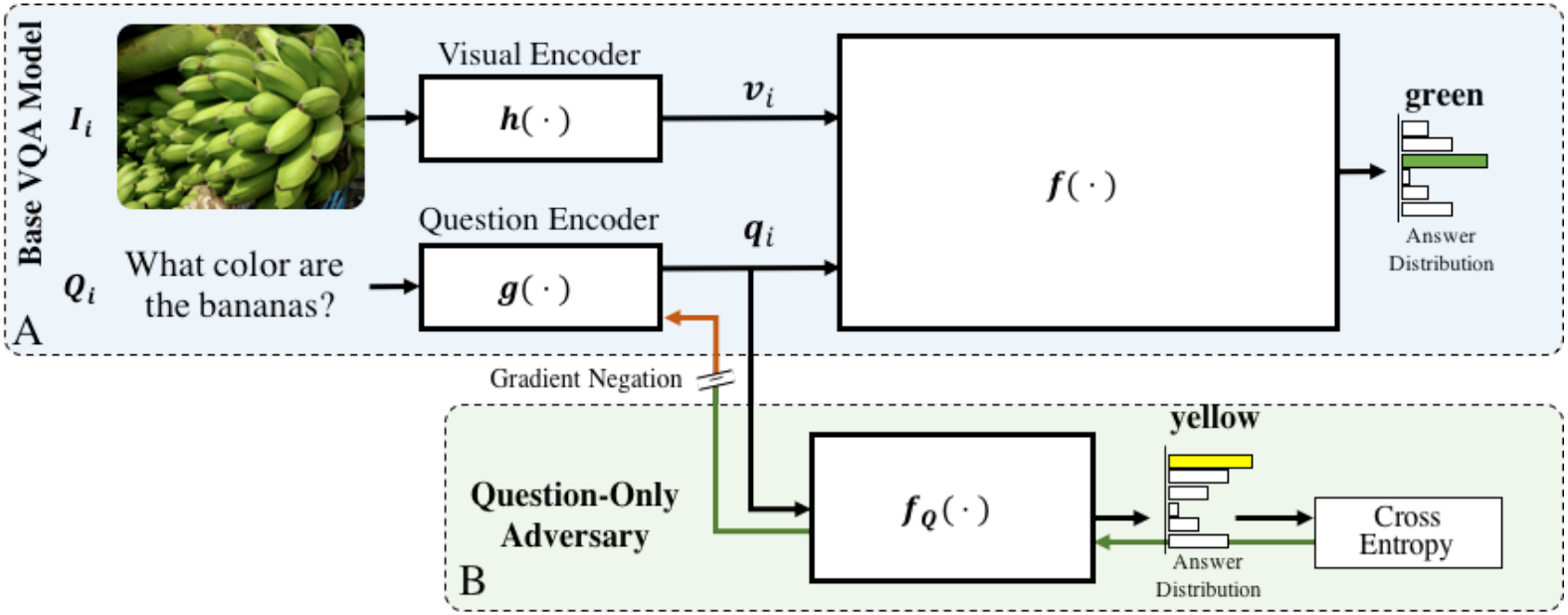


$$\mathcal{L}_{VQA}(f, g, h)$$

$$\mathcal{L}_{QA}(f_Q, g)$$



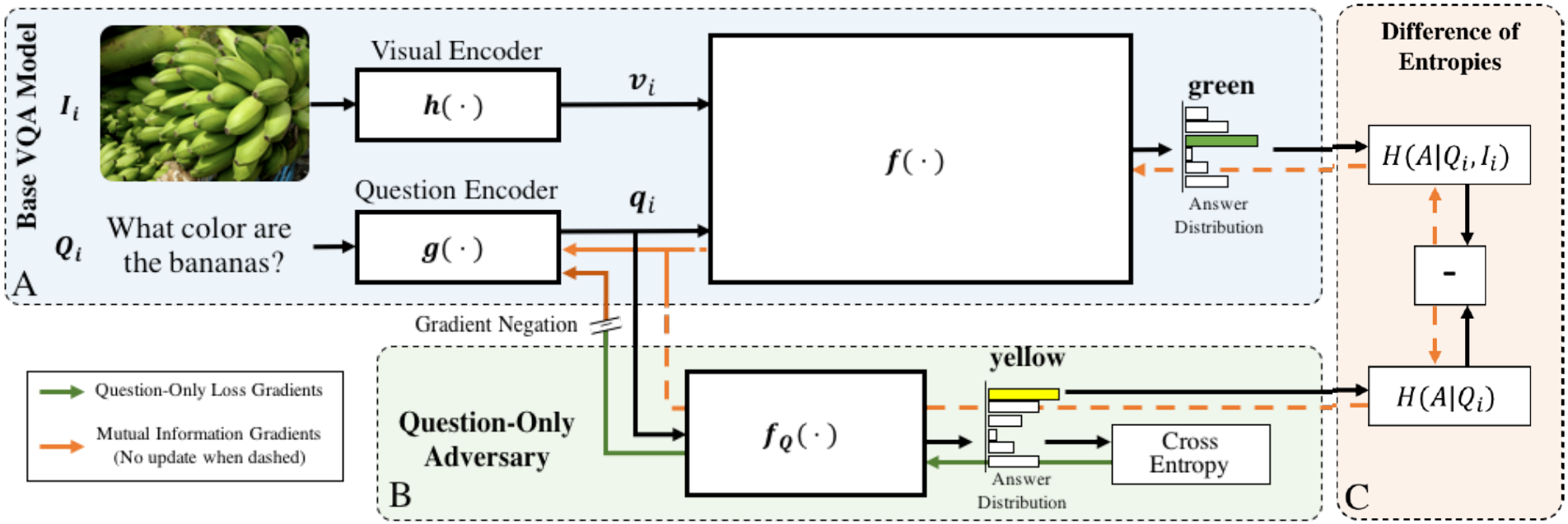
$$\min_{f,g,h} \max_{f_Q} \mathcal{L}_{VQA}(f, g, h) - \lambda_Q \mathcal{L}_{QA}(f_Q, g)$$



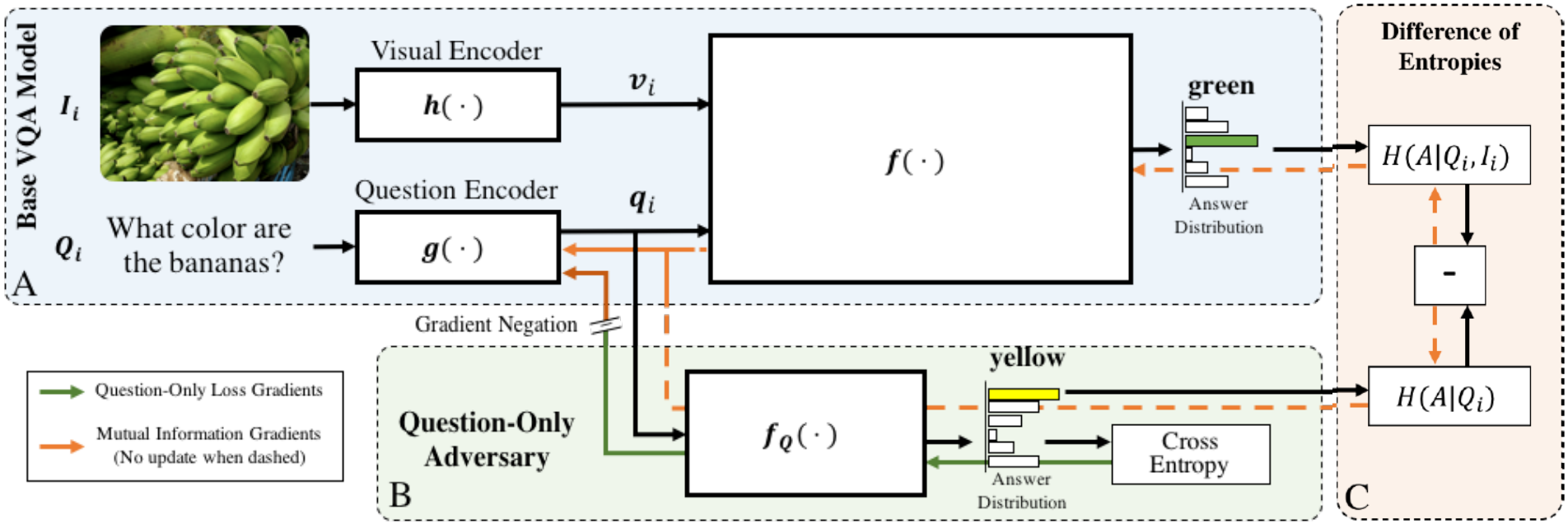
$\mathcal{L}_{VQA}(f, g, h)$

$\mathcal{L}_{QA}(f_Q, g)$

$$\min_{f,g,h} \max_{f_Q} \mathcal{L}_{VQA}(f, g, h) - \lambda_Q \mathcal{L}_{QA}(f_Q, g)$$



$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_{QA}(f_Q, g) - \lambda_H \mathcal{L}_H(f,g,h,f_Q)$$



$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_{QA}(f_Q, g) - \lambda_H \mathcal{L}_H(f,g,h,f_Q)$$

Model	VQA-CP v2 test			
	Overall	Yes/No	Number	Other
SAN (Yang et al. CVPR16)	24.96	38.35	11.14	21.74

	Model	VQA-CP v2 test			
		Overall	Yes/No	Number	Other
	SAN (Yang et al. CVPR16)	24.96	38.35	11.14	21.74
Ours	SAN + Q-Adv	27.24	54.50	14.91	16.33

+2%

	Model	VQA-CP v2 test			
		Overall	Yes/No	Number	Other
	SAN (Yang et al. CVPR16)	24.96	38.35	11.14	21.74
Ours	SAN + DoE	25.75	42.21	12.08	20.87

+1%

	Model	VQA-CP v2 test			
		Overall	Yes/No	Number	Other
	SAN (Yang et al. CVPR16)	24.96	38.35	11.14	21.74
Ours	SAN + Q-Adv	27.24	54.50	14.91	16.33
Ours	SAN + DoE	25.75	42.21	12.08	20.87
Ours	SAN + Q-Adv + DoE	**33.29**	**56.65**	**15.22**	**26.02**

+8%

	Model	VQA-CP v2 test			
		Overall	Yes/No	Number	Other
	GVQA	31.30	57.99	13.68	22.14
	SAN (Yang et al. CVPR16)	24.96	38.35	11.14	21.74
Ours	SAN + Q-Adv	27.24	54.50	14.91	16.33
Ours	SAN + DoE	25.75	42.21	12.08	20.87
Ours	SAN + Q-Adv + DoE	**33.29**	**56.65**	**15.22**	**26.02**

+2%

	Model	VQA-CP v2 test			
		Overall	Yes/No	Number	Other
	GVQA	31.30	57.99	13.68	22.14
	SAN (Yang et al. CVPR16)	24.96	38.35	11.14	21.74
Ours	SAN + Q-Adv	27.24	54.50	14.91	16.33
Ours	SAN + DoE	25.75	42.21	12.08	20.87
Ours	SAN + Q-Adv + DoE	**33.29**	**56.65**	**15.22**	**26.02**
	UpDn (Anderson et al. CVPR18)	39.74	42.27	11.93	46.05
Ours	UpDn + Q-Adv	40.08	42.34	13.02	46.33
Ours	UpDn + DoE	40.43	42.62	12.19	**47.03**
Ours	UpDn + Q-Adv + DoE	**41.17**	**65.49**	**15.48**	35.48

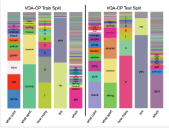
+1.5%

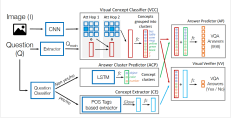
Outline

Overview of VQA
[ICCV'15, IJCV'16, AI Mag'16]

Problem with existing setup + models
[EMNLP'16]

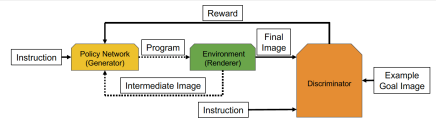
Overcoming priors
- A new evaluation protocol [CVPR'18]
- A novel architecture [CVPR'18]
- A novel objective function [NIPS'18]



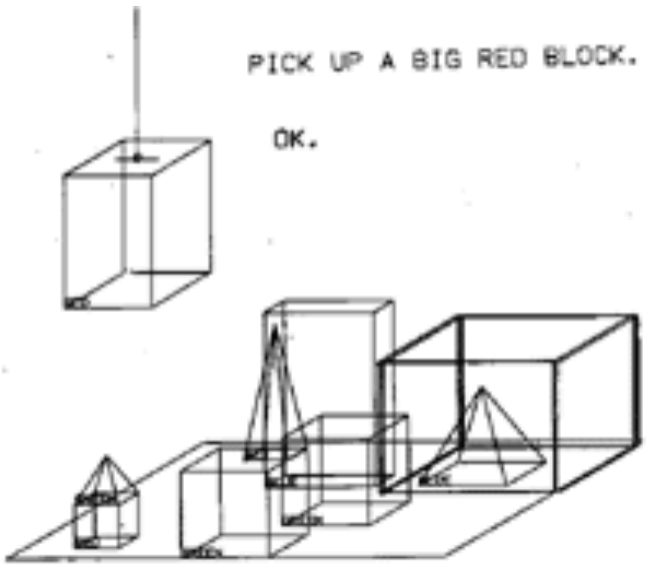


$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_Q(f_Q, g) - \lambda_H \mathcal{L}_H(f,g,h,f_Q)$$

Beyond VQA
[Work in progress]

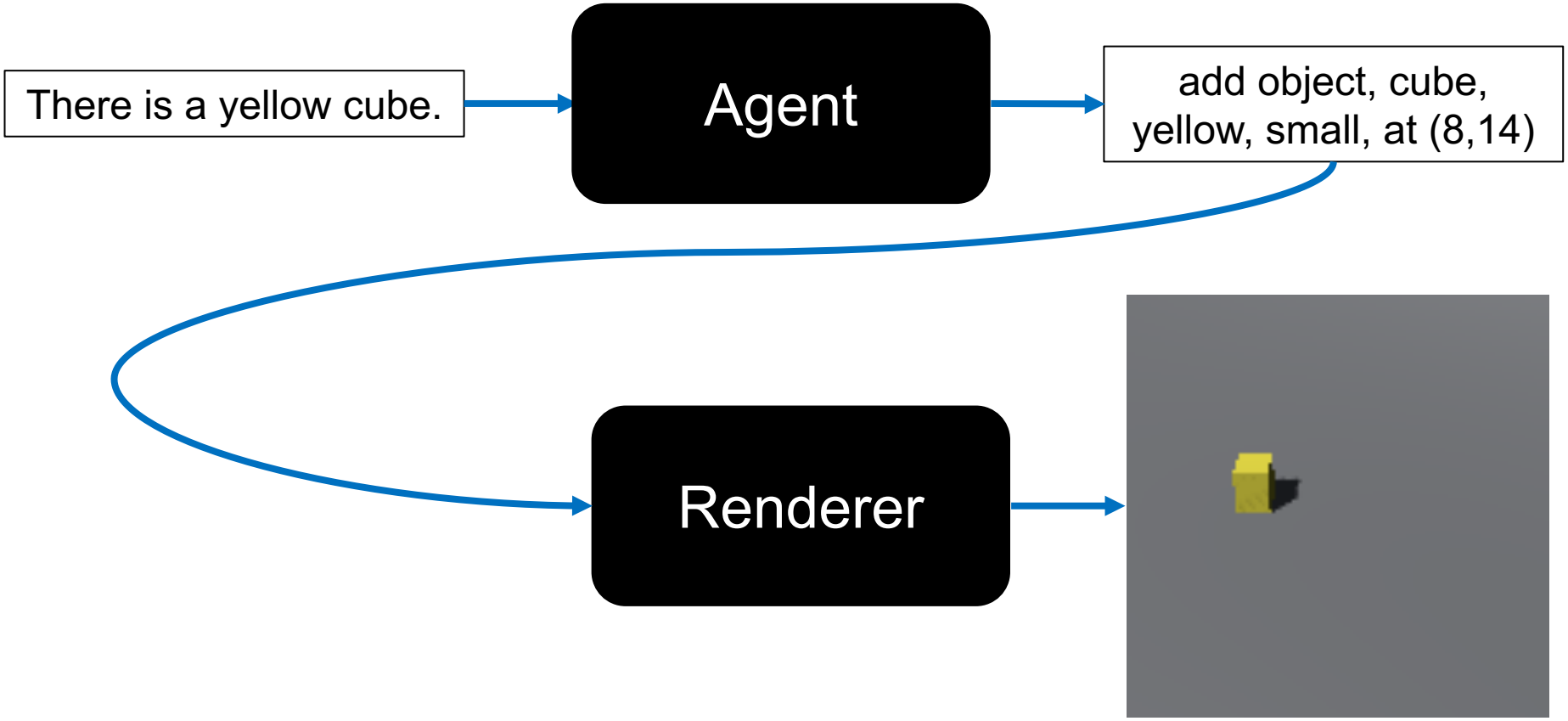


SHRDLU



Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.

Task


There is a yellow cube.
Agent
add object, cube, yellow, small, at (8,14)
Renderer

Task

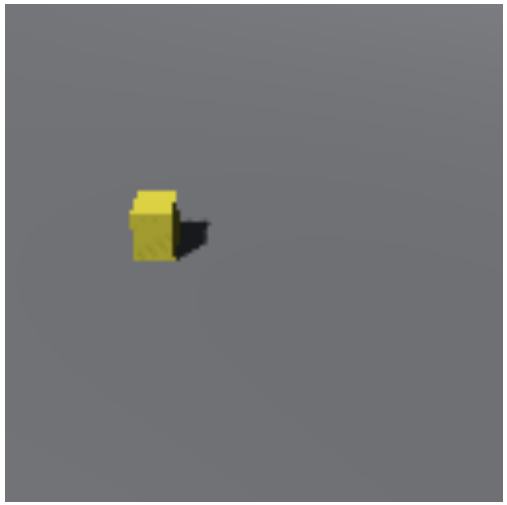
There is a yellow cube.

Agent

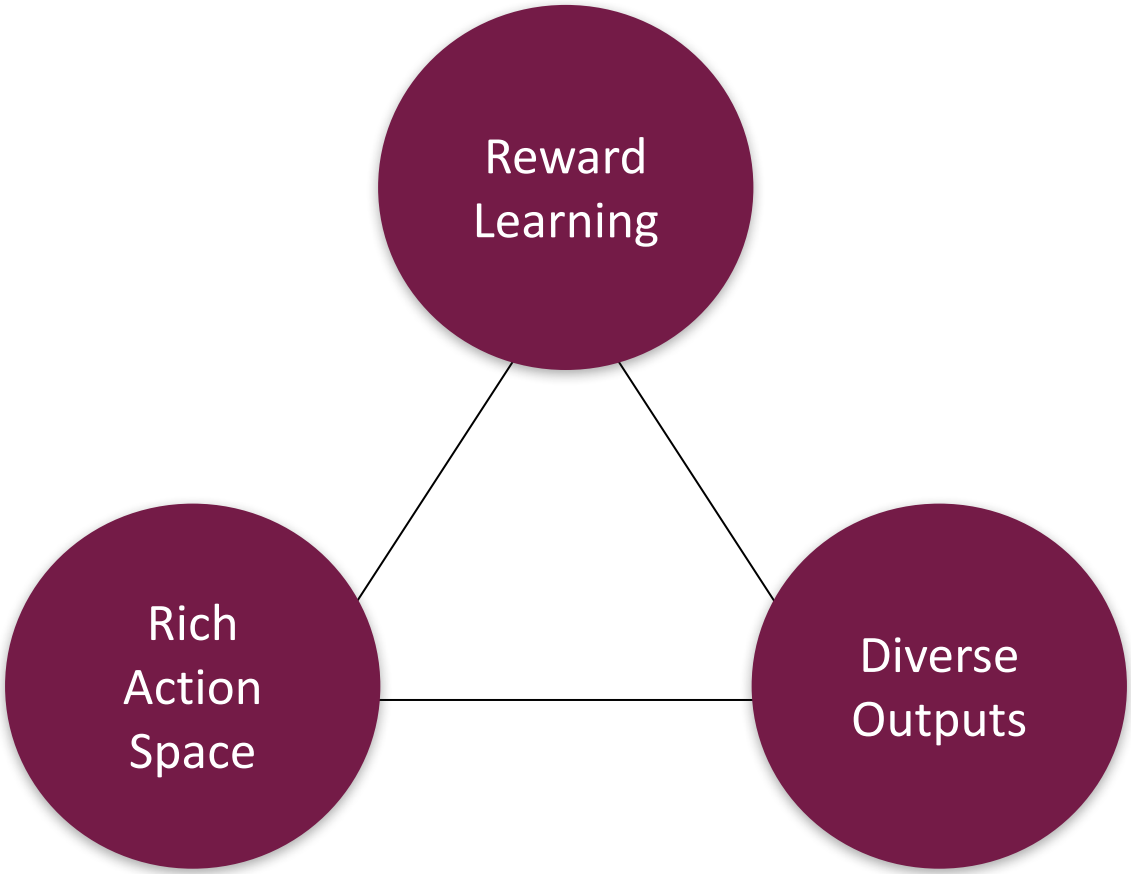
add object, cube, yellow, large, at (12,17)

add object, cube, yellow, small, at (8,14)

add object, cube, yellow, small, at (22,12)



Technical challenges of interest to us



Domains

- MNIST Digit Painting
- 3D Scene Construction

Domain 1: MNIST Digit Painting (Task)

Draw 9.

Paint five.

Domain 1: MNIST Digit Painting (Dataset)

- Instructions paired with MNIST images (60K images)
- Instruction template -- <Action> <Class Label>
  - <Action> = “Draw” | “Put” | “Paint” | “Add” | “Create”
  - <Class Label> = numerical (“0”) / word form (“zero”)

Domain 1: MNIST Digit Painting (Environment and Action Space)

- Environment: *libmypaint* – painting library

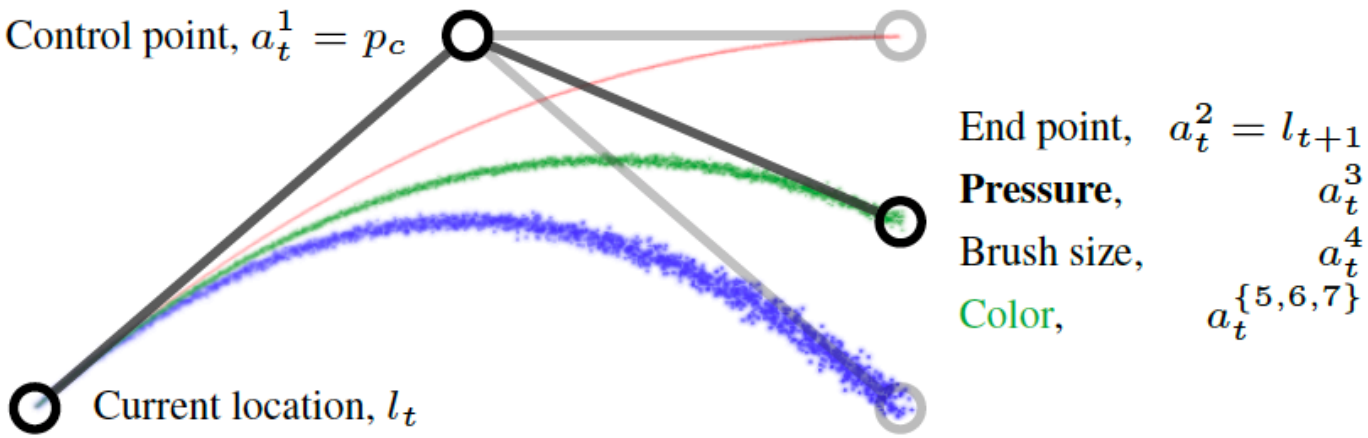


Fig. credit: *Ganin et al., ICML18*

Domain 1: MNIST Digit Painting (Environment and Action Space)

- Action Space:
  - end point of the brush (on 32 x 32 grid),
  - control point of the brush (on 32 x 32 grid),
  - pressure applied to the brush (10 levels),
  - brush size (4 levels),
  - binary flag -- draw stroke / skip
- Size of the action space -- 83,886,080

Domain 2: 3D Scene Construction (Task)

There is a green cylinder.

There is a large sphere.

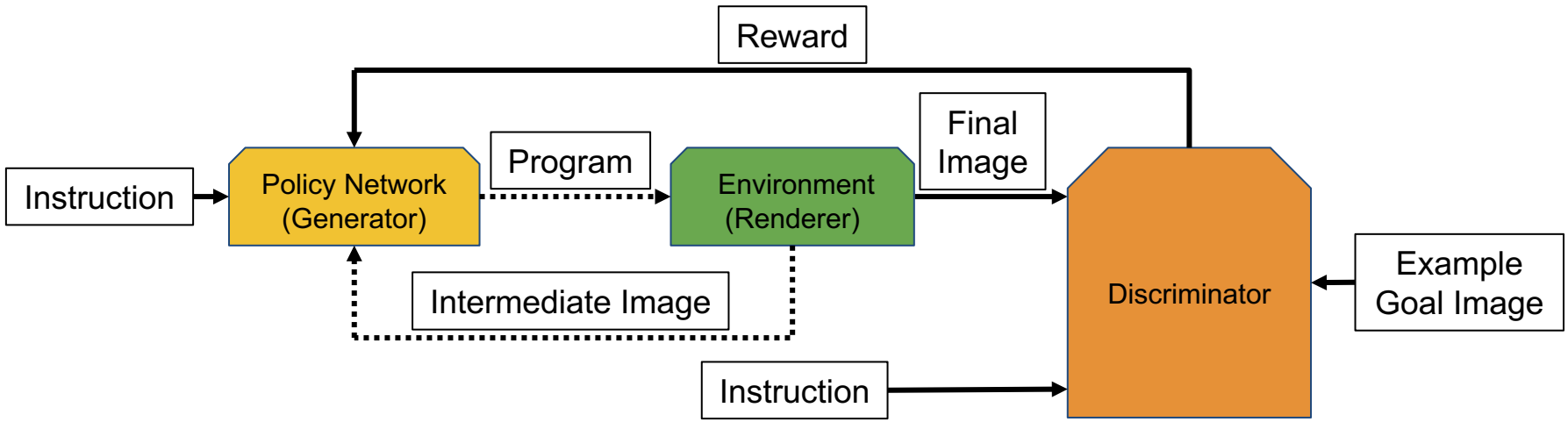
Domain 2: 3D Scene Construction (Dataset)

- Instructions paired with 3D scene images (16,159 images)
- Instruction template: “There is a” <Attribute> <Shape>
  - Attribute: any color (8), any size (large, small)
  - Shape: any shape (sphere, cube, cylinder)
- Total possible unique instructions = (8+2)*(3) = 30

Domain 2: 3D Scene Construction (Environment and Action Space)

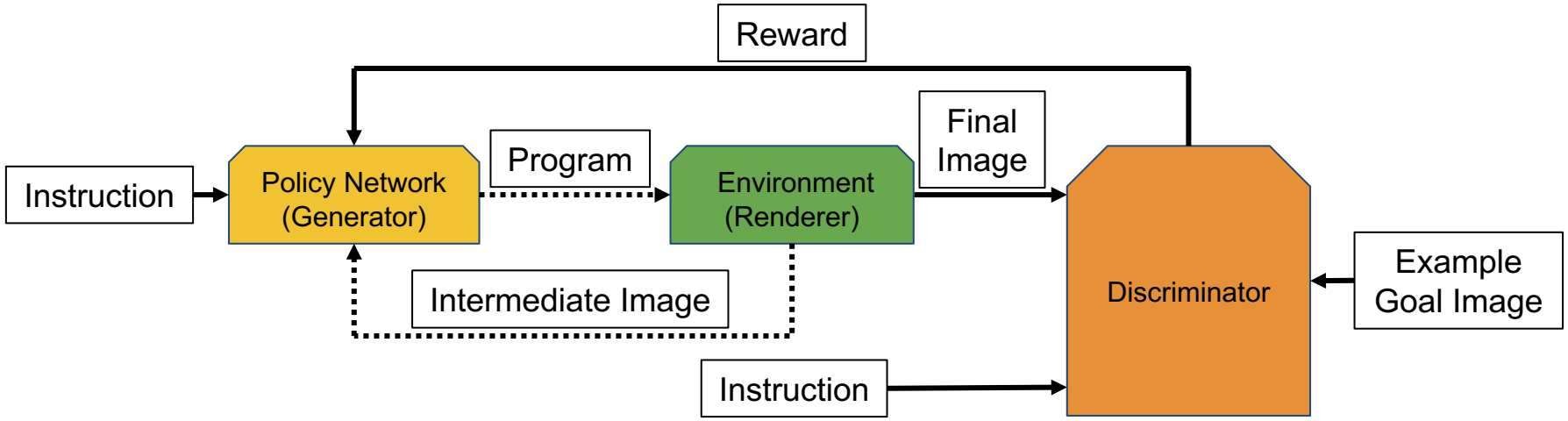
- Environment: 3D Editor
- Action Space:
  - location of the object (on 32 x 32 grid),
  - object shape (3 shapes),
  - object size (2 sizes),
  - object color (8 colors),
  - flag -- add object / modify object / skip
- Size of the action space -- 147,456

Overview of the approach



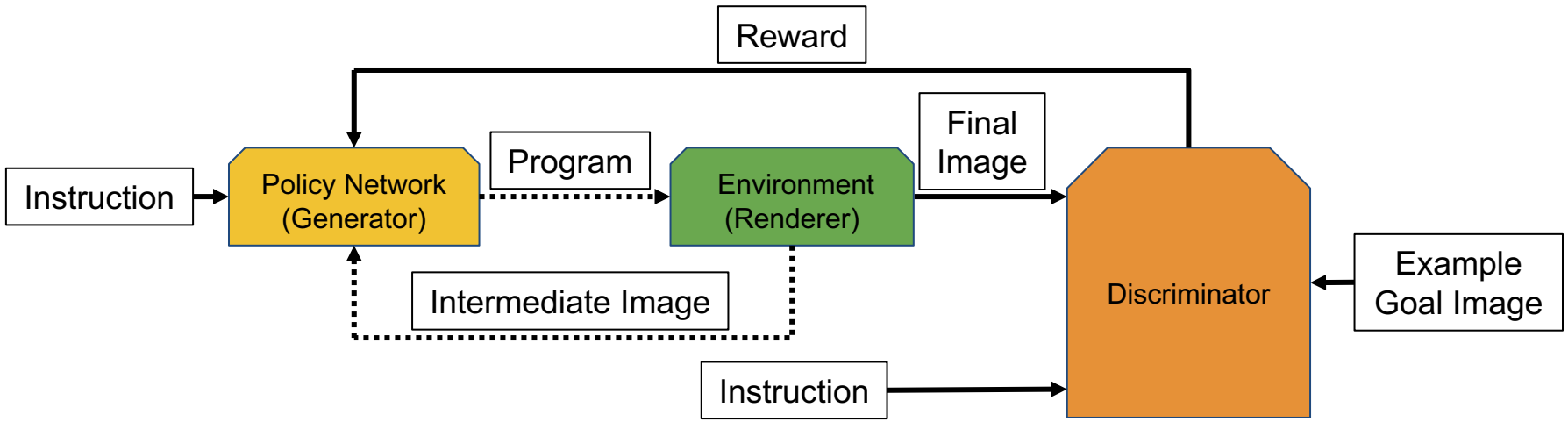
Extending *Ganin et al., ICML18*

Overview of the approach

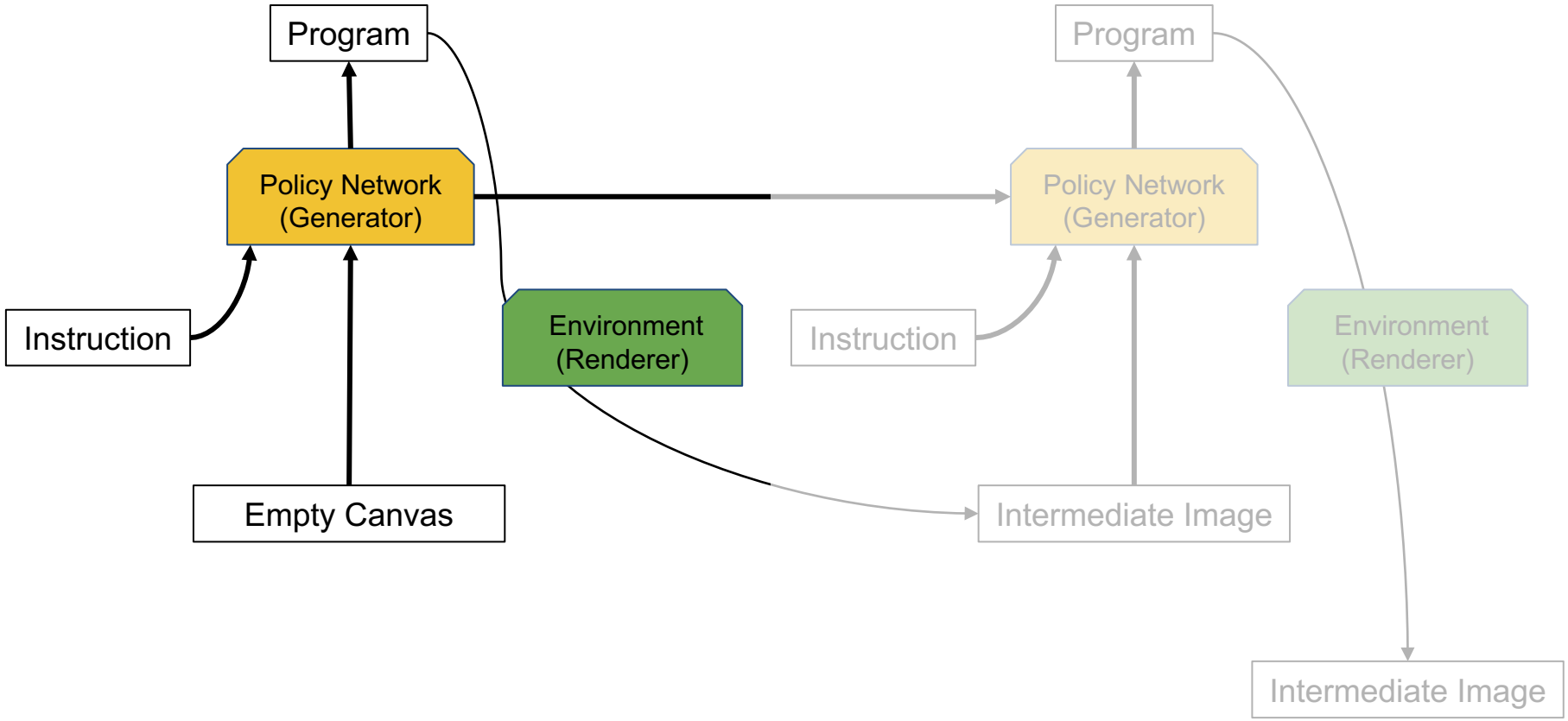


- **Reward learning:** discriminator is learning consistency between instruction and image
- **Diversity:** Action sampling from non-peaky distribution

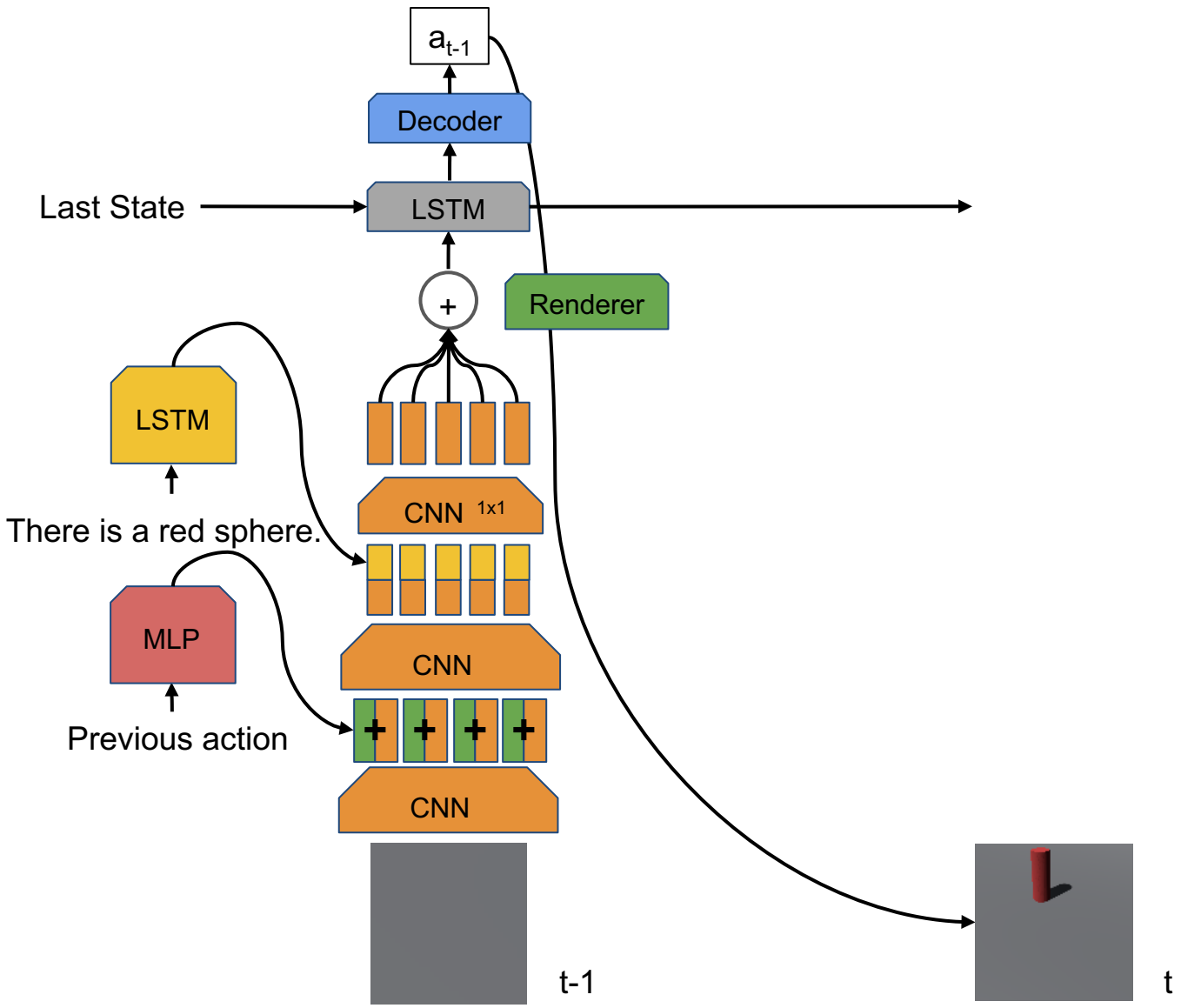
Overview of the approach



Overview of the approach

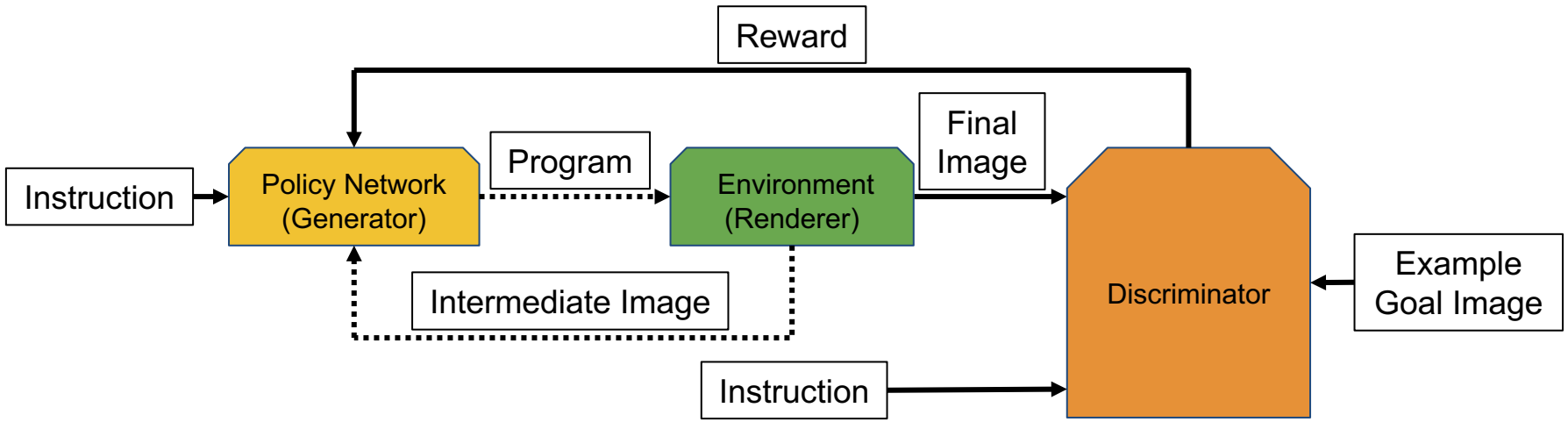


Policy Network

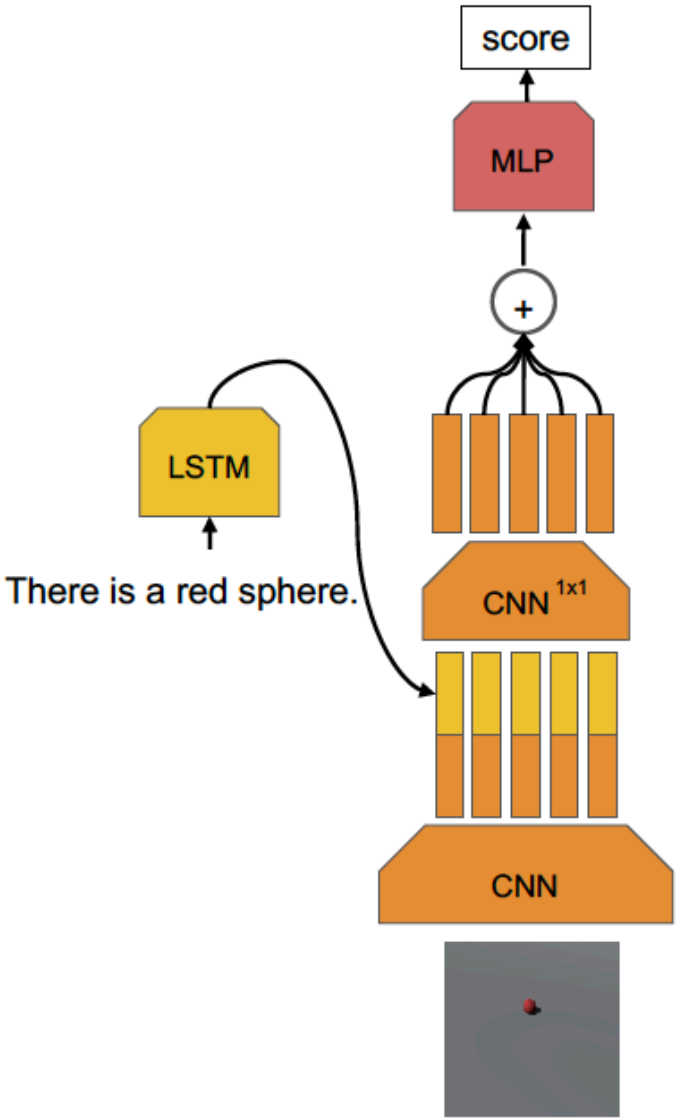


Decoder from *Ganin et al., ICML18*

Overview of the approach



Discriminator



Domain 1: MNIST Digit Painting

Discriminator

Create zero

Put 1

Paint two

Draw 3

Add four

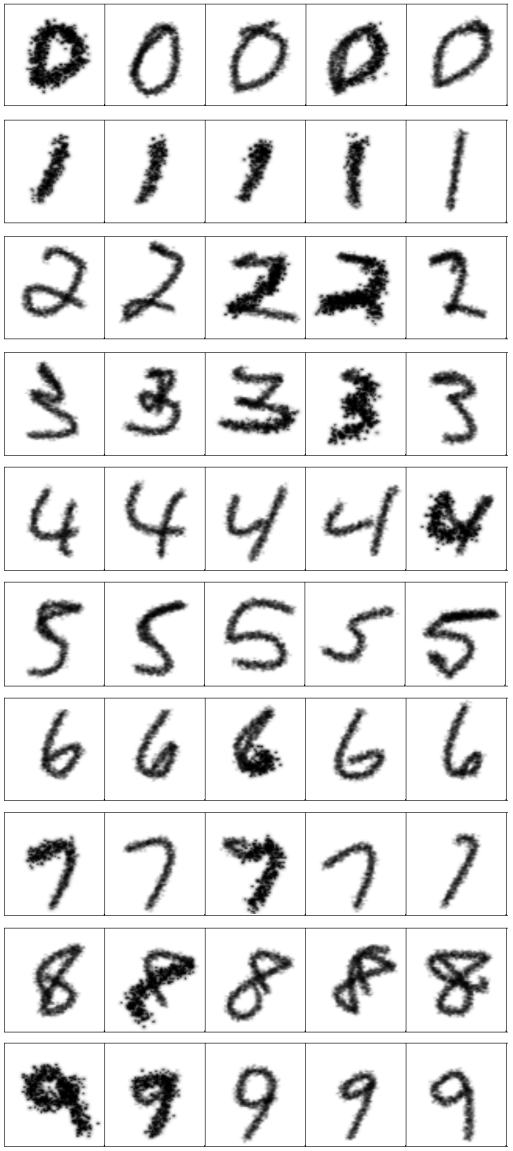
Draw 5

Paint six

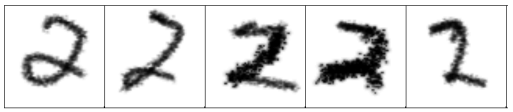
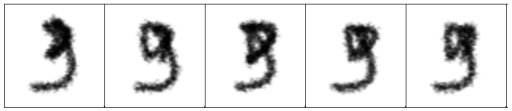
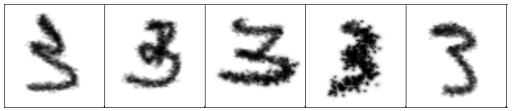
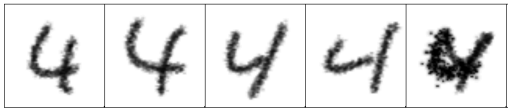
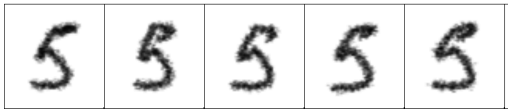
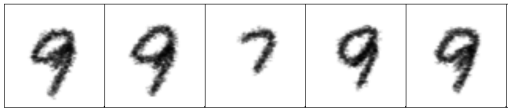
Put 7

Create eight

Add 9

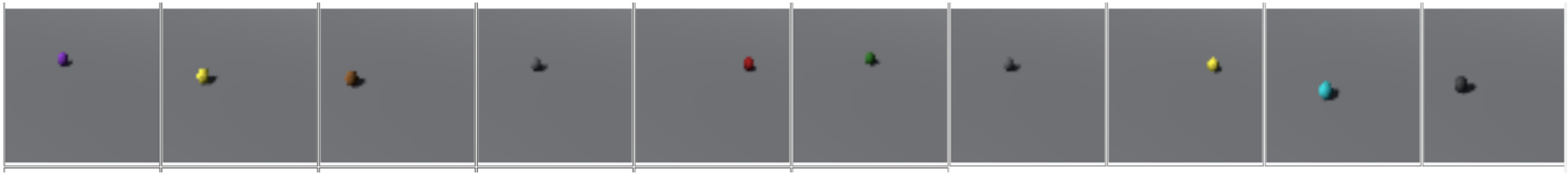


Domain 1: MNIST Digit Painting

	L2	Discriminator
Create zero		
Put 1		
Paint two		
Draw 3		
Add four		
Draw 5		
Paint six		
Put 7		
Create eight		
Add 9		

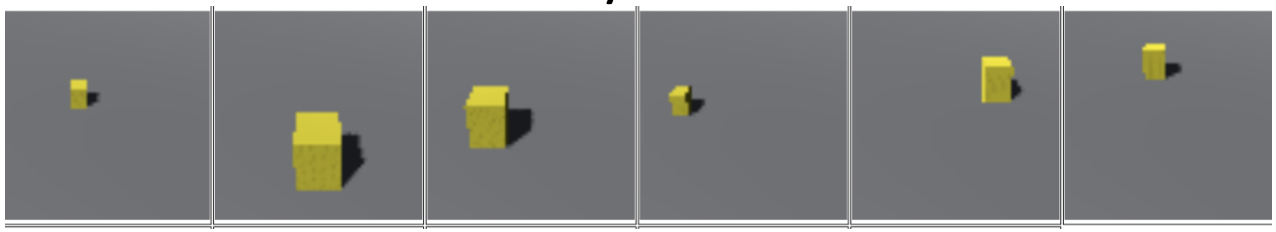
Domain 2: 3D Scene Construction (Discriminator)

There is a small sphere.



There is a large cylinder.

There is a yellow cube.

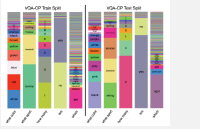


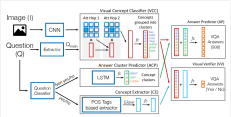
Outline

Overview of VQA
[ICCV'15, IJCV'16, AI Mag'16]

Problem with existing setup + models
[EMNLP'16]

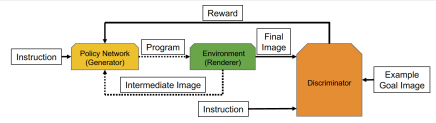
Overcoming priors
- A novel split [CVPR'18]
- A novel architecture [CVPR'18]
- A novel objective function [NIPS'18]





$$\min_{f,g,h} \max_{f_Q} L_{VQA}(f,g,h) - \lambda_Q \mathcal{L}_Q(f_Q, g) - \lambda_H \mathcal{L}_H(f,g,h,f_Q)$$

Beyond VQA



Stanislaw Antol
Traptic, Inc.

Jiasen Lu
Georgia Tech

Sainandan Ramakrishnan
Georgia Tech

Akrit Mohapatra
Virginia Tech → eBay

Yash Goyal
Georgia Tech

Stefan Lee
Georgia Tech

Meg Mitchell
Google Research

Mateusz Malinowski
DeepMind

Felix Hill
DeepMind

Ali Eslami
DeepMind

Oriol Vinyals
DeepMind

Tejas Kulkarni
DeepMind

Ani Kembhavi
AI2

Larry Zitnick
FAIR

Devi Parikh
Georgia Tech / FAIR

Dhruv Batra
Georgia Tech / FAIR

Thanks Rama!

A man holding a beer bottle with two hands and looking at it.
A man in a white t-shirt looks at his beer bottle.
A man with black curly hair is looking at a beer.
A man holds a bottle of beer examining the label.

⋮

A guy holding a beer bottle.
A man holding a beer bottle.
A man holding a beer.
A man holds a bottle.
Man holding a beer.

What color are her eyes?
What is the mustache made of?

Thanks!

Questions?