

Don't Just Assume; Look and Answer

Overcoming Priors for Visual Question Answering



Aishwarya Agrawal
(Georgia Tech)



Dhruv Batra
(Georgia Tech /
FAIR)



Devi Parikh
(Georgia Tech /
FAIR)



ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE



Ani Kembhavi
(AI2)

Outline

Overview of VQA

Problem with existing
setup + models

A novel VQA Split

A novel VQA model

Conclude

Outline

Overview of VQA

Problem with existing
setup + models

A novel VQA Split

A novel VQA model

Conclude

What is Visual Question Answering (VQA)?

VQA Task

VQA Task



VQA Task

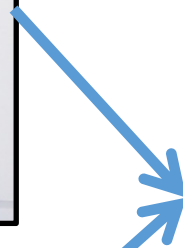


What is the mustache
made of?

VQA Task



What is the mustache
made of?

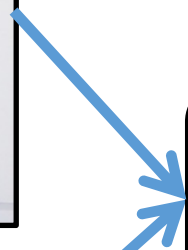


AI System

VQA Task



What is the mustache
made of?



AI System



bananas

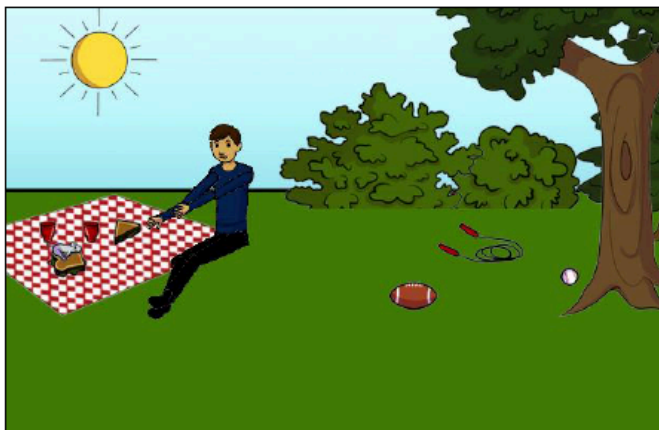
VQA v1 Dataset



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?

Papers using VQA

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson^{1*} Xiaodong He² Chris Buehler³ Damien Teney⁴
Mark Johnson⁵ Stephen Gould¹ Lei Zhang³

¹Australian National University ²JD AI Research ³Microsoft Research ⁴University of Adelaide ⁵Macquarie University
¹firstname.lastname@anu.edu.au, ²xiaodong.he@jd.com, ³{chris.buehler, leizhang}@microsoft.com
⁴damien.teney@adelaide.edu.au, ⁵mark.johnson@mq.edu.au

Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding

Akira Fukui^{*1,2} Dong Huk Park^{*1} Daylen Yang^{*1}
Anna Rohrbach^{*1,3} Trevor Darrell¹ Marcus Rohrbach¹

¹UC Berkeley EECS, CA, United States
²Sony Corp., Tokyo, Japan
³Max Planck Institute for Informatics, Saarbrücken, Germany

Compositional Memory for Visual Question Answering

Aiwen Jiang^{1,2}

Fang Wang²

Fatih Porikli²

Yi Li^{* 2,3}

¹Jiangxi Normal University

²NICTA and ANU

³Toyota Research Institute North America

¹aiwen.jiang@nicta.com.au

²{fang.wang, fatih.porikli}

³fatih.porikli@tri.global

Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Huijuan Xu
UMass Lowell

¹xu@cs.uml.edu

Kate Saenko
UMass Lowell

¹saenko@cs.uml.edu

... and many more

Deep Compositional Question Answering with Neural Module Networks

Jacob Andreas Marcus Rohrbach Trevor Darrell Dan Klein

Department of Electrical Engineering and Computer Sciences

University of California, Berkeley

{jda, rohrbach, trevor, klein}@{cs, eeecs, eecs, cs}.berkeley.edu

Where To Look: Focus Regions for Visual Question Answering

Kevin J. Shih, Saurabh Singh, and Derek Hoiem

University of Illinois at Urbana-Champaign

{kjshih2, ssl, dhoiem}@illinois.edu

ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering

Kan Chen
University of Southern California
kanchen@usc.edu

Jiang Wang
Baidu Research - IDL
wangjiang03@baidu.com

Liang-Chieh Chen
UCLA
lcchen@cs.ucla.edu

Haoyuan Gao
Baidu Research - IDL
gaohaoyuan@baidu.com

Wei Xu
Baidu Research - IDL
wei.xu@baidu.com

Ram Nevatia
University of Southern California
nevatia@usc.edu

Stacked Attention Networks for Image Question Answering

Zichao Yang¹, Xiaodong He², Jianfeng Gao², Li Deng², Alex Smola¹

¹Carnegie Mellon University, ²Microsoft Research, Redmond, WA 98052, USA
zichaoy@cs.cmu.edu, {xiaohe, jfgao, deng}@microsoft.com, alex@smola.org

VQA Challenges

2016

	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony ^[14]	83.24	39.47	58	66.47
Naver Labs ^[10]	83.31	38.7	54.62	64.79
DLAIT ^[5]	83.25	40.07	52.09	63.68
snubi-naverlabs ^[25]	83.16	39.14	51.33	63.18
POSTECH ^[11]	81.67	38.16	52.79	63.17
Brandeis ^[3]	82.11	37.73	51.91	62.88
VTCComputerVison ^[19]	79.95	38.22	51.95	62.06
MIL-UT ^[7]	81.98	37.56	49.75	61.77
klab ^[23]	81.53	39.27	49.61	61.69
SHB_1026 ^[13]	82.07	36.81	47.77	60.76
MMCX ^[8]			48.33	60.36
VT_CV_Jiasen ^[20]			47.87	60.33
LV-NUS ^[6]			46.1	59.54
ACVT_Adelaide ^[1]	81.07	37.12	45.83	59.44
UC Berkeley (DNMN) ^[15]	80.98	37.48	45.81	59.44
CNNAtt ^[4]	81.04	36.44	45.76	59.33
san ^[24]	79.11	36.41	46.42	58.85
UC Berkeley (NMN) ^[16]	81.16	37.7	44.01	58.66
global_vision ^[22]	78.24	36.27	46.32	58.43
vqateam-deeperLSTM_NormizeCNN ^[27]	80.56	36.53	43.73	58.16
Mujtaba hasan ^[9]	80.28	36.92	42.24	57.36
RJT ^[12]	78.82	35.97	42.13	56.61
Bolei ^[2]	76.76	34.98	42.62	55.89
UPV_UB ^[18]	78.88	36.33	40.27	55.77
att ^[21]	78.1	35.3	40.27	55.34
vqateam-lstm_cnn ^[28]	79.01	35.55	36.8	54.06
UPC ^[17]	78.05	35.53	36.7	53.62
vqateam-nearest_neighbor ^[29]	71.73	24.31	22	42.73
vqateam-prior_per_qtype ^[30]	71.17	35.63	9.32	37.55
vqateam-all_yes ^[26]	70.53	0.43	1.26	29.72

25 teams

2017

	By Answer Type			Overall
	Yes/No	Number	Other	
Adelaide-Teney ACRV MSR ^[2]	85.54	47.45	59.82	69.13
DLAIT ^[6]	83.17	46.66	60.15	68.22
HDU-USYD-UNCC ^[7]	84.5	45.39	59.01	68.09
LV_NUS ^[9]	81.92	48.38	59.63	67.64
Athena ^[3]	82.88	43.17	57.95	66.67
lonely_shepherd ^[18]	82.32	43.06	56.71	65.84
UPMC-LIP6 ^[14]	82.07	41.06	57.12	65.71
JuneflowerIvaNipr ^[8]	81.09	41.56	57.83	65.7
yudf2001 ^[28]	82.1	45.56	55.43	65.41
Adelaide-Teney ^[1]	82.37	41.09	55.86	65.3
coral2017 ^[17]			55.3	65.05
usyd_zju ^[20]			55.82	64.79
POSTECH ^[12]	79.32	40.67	55.3	63.66
yahia zakaria ^[27]	79.77	40.53	54.75	63.57
anon_team ^[16]	76.52	39.29	57.31	63.31
VQAMachine ^[15]	79.82	40.91	53.35	62.97
vqa_hack3r ^[21]	79.88	38.95	53.58	62.89
vqahhi_drau ^[22]	78.86	39.91	53.76	62.66
DCD_ZJU ^[5]	79.85	38.64	52.95	62.54
vqateam_mcb_benchmark ^[25]	78.82	38.28	53.36	62.27
CRCV_REU ^[4]	74.08	36.43	54.84	60.81
neural-vqa-attention ^[19]	69.77	35.65	47.18	55.28
vqateam_deeper_LSTM_Q_norm_l ^[23]	73.46	35.18	41.83	54.22
MIC_TJ ^[10]	69.22	34.16	35.97	49.56
vqateam_language_only ^[24]	67.01	31.55	27.37	44.26
UPC ^[13]	66.97	31.38	25.81	43.48
MultiLab ^[11]	62.98	29.97	16.68	37.33
vqateam_prior ^[26]	61.2	0.36	1.17	25.98

24 teams

2018

	By Answer Type			Overall
	Yes/No	Number	Other	
FAIR-A* ^[38]	87.82	51.59	63.43	72.25
HDU-UCAS-USYD ^[8]	87.61	51.92	63.19	72.09
SNU-BI ^[12]	87.22	54.37	62.45	71.84
casia_iva ^[27]	86.98	51.05	62.31	71.31
MIL-UT ^[9]	87	52.6	61.62	71.16
Tohoku CV Lab ^[13]	87.29	53.25	61.13	71.12
graph-attention-msm ^[33]	86.54	51.65	61.42	70.77
ut-swk ^[36]	86.34	54.26	60.8	70.68
vqabyte ^[39]	86.93	49.93	61.11	70.6
fs ^[31]	86.18	50.36	61.37	70.46
DCD_ZJU ^[5]	86.21	48.82	61.58	70.4
Adelaide-Teney ^[1]			60.57	70.34
UTS_YZZD ^[20]			60.64	70.23
VQA-ReasonTensor ^[41]	86.81	49.17	60.76	70.17
wyvernba1 ^[45]	86.3	48.9	60.49	69.93
UPMC-LIP6 ^[15]	85.71	48.25	61.05	69.88
VQA_NTU ^[22]	86.03	48.65	60.4	69.74
nagizero ^[34]	85.89	48.54	60.45	69.7
CFM-UESTC ^[3]	85.71	47.96	60.72	69.69
caption_vqa ^[26]	86.2	47.26	60.41	69.67
cvqa ^[29]	86.1	47.42	60.38	69.63
yudf2010 ^[47]	85.42	50.24	59.66	69.31
nmlab612 ^[35]	85.61	47.74	59.85	69.21
TsinghuaCVLab ^[14]	85.42	48.92	59.65	69.16
CIST-VQA ^[4]	85.76	48.4	59.43	69.14
VLC Southampton ^[18]	83.56	51.39	59.11	68.41
University of Guelph MLRG ^[17]	84.64	47.65	58.07	67.95
RelVQA ^[11]	83.98	46.77	58.79	67.92
zhi-smile ^[48]	82.52	50.28	57.86	67.26
NTU_ROSE_USTC ^[10]	83.66	45.16	57.95	67.22
VQA-Machine+ ^[20]	82.42	48.66	57.48	66.86

40 teams

Outline

Overview of VQA

Problem with existing
setup + models

A novel VQA Split

A novel VQA model

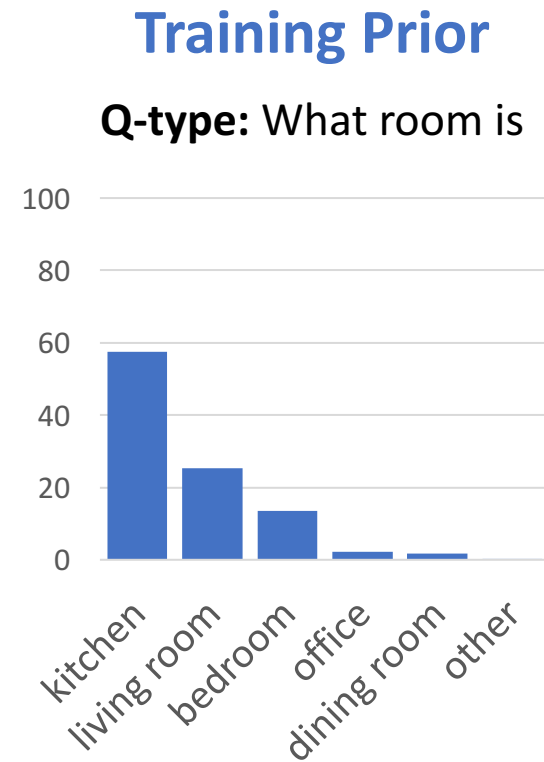
Conclude

Problem with existing setup + models

Today's VQA models –

- are driven by superficial correlations in training data
- lack sufficient image grounding

Problem with existing setup + models



Problem with existing setup + models

Train

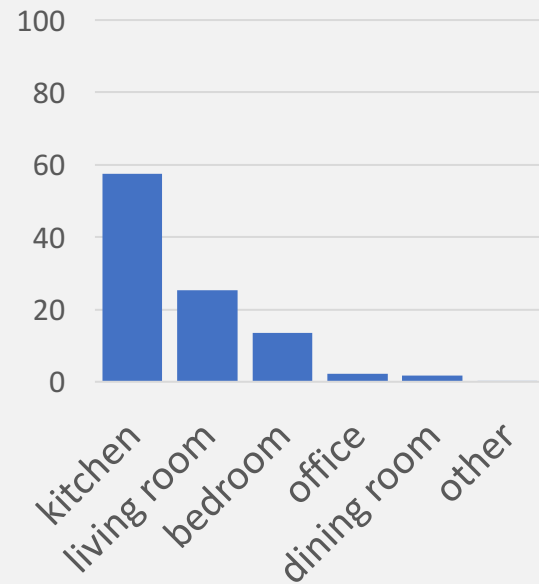
Q: What room is this?

A: Kitchen



Training Prior

Q-type: What room is



Problem with existing setup + models

Train

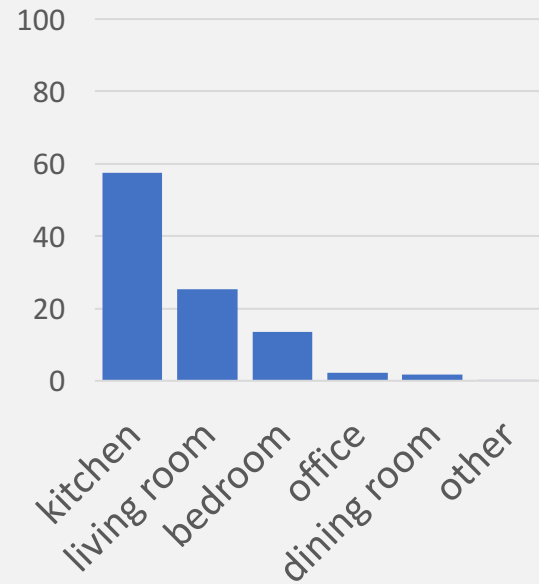
Q: What room is this?

A: Kitchen



Training Prior

Q-type: What room is



Test

Q: What room is this?

A: Bathroom



Problem with existing setup + models

Train

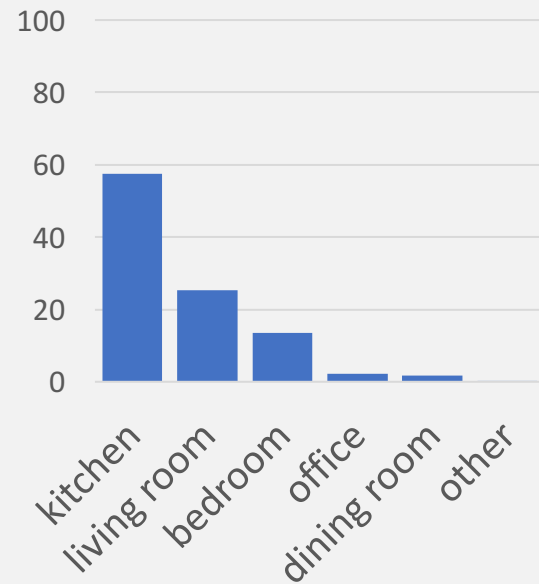
Q: What room is this?

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

Outline

Overview of VQA

Problem with existing
setup + models

A novel VQA Split

A novel VQA model

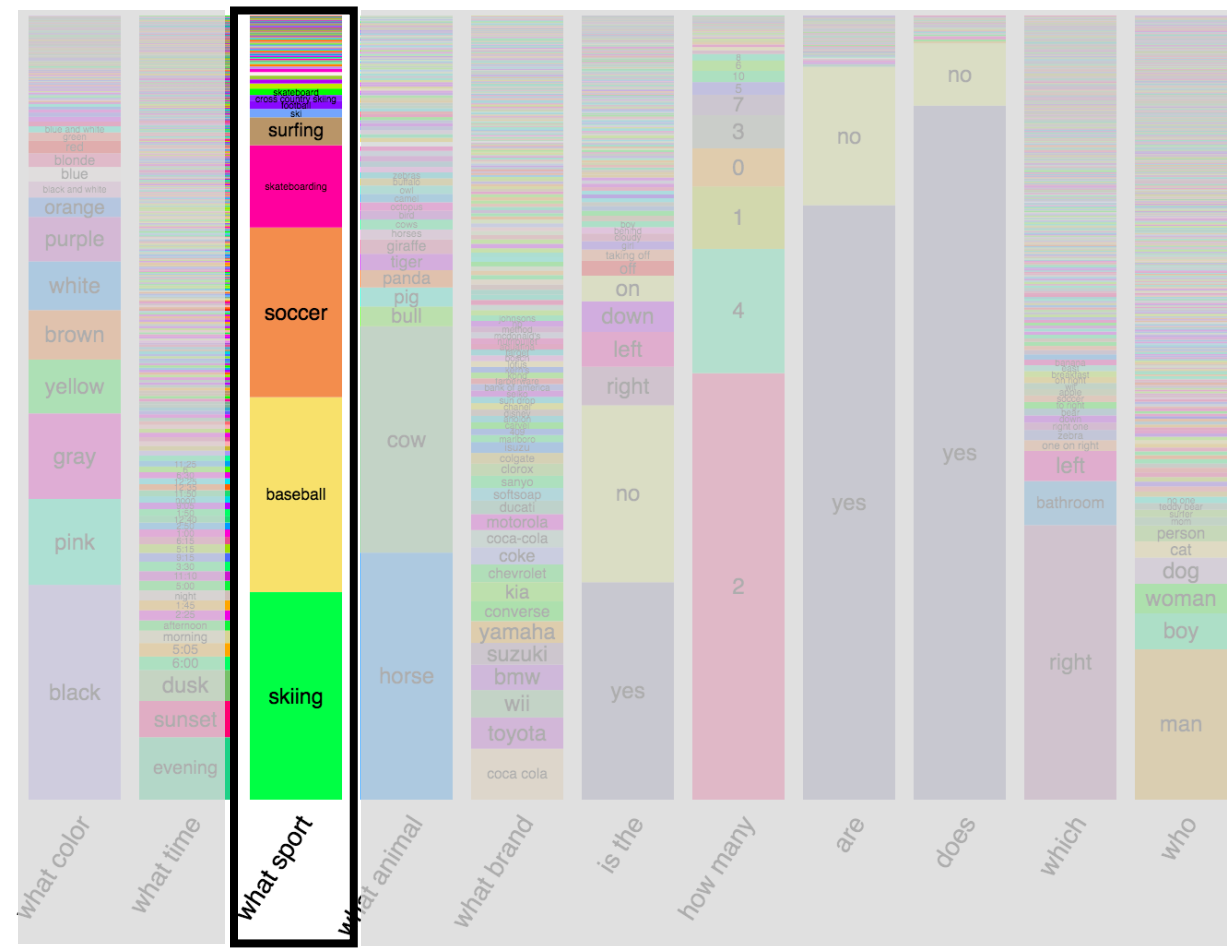
Conclude

Meet VQA-CP!

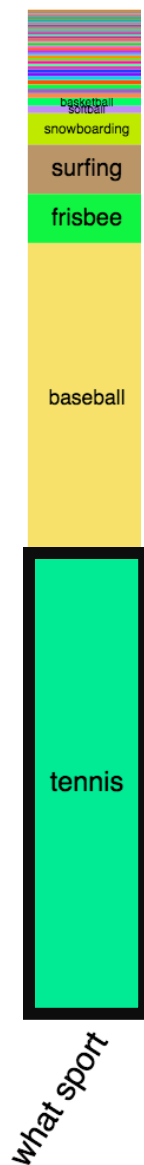
- New splits of the VQA v1 and VQA v2 datasets
- Visual Question Answering under Changing Priors (VQA-CP v1/v2)

VQA-CP Train Split

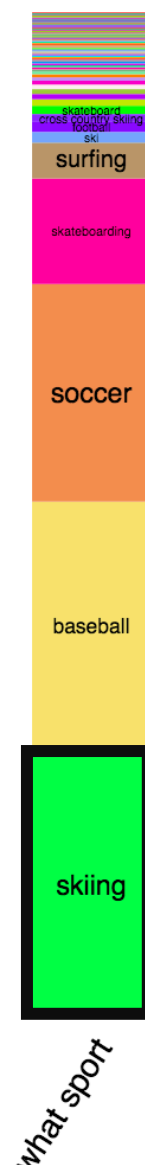
VQA-CP Test Split



VQA-CP Train Split



VQA-CP Test Split



VQA-CP creation

Train

("What color is the dog?", "white")

("What color is the cat?", "black")

Test

("What color is the plate?", "white")

("What color is the bag?", "black")

VQA-CP creation

Train

("What color is the dog?", "white")

("What color is the cat?", "black")

("What color is the plate?", "white")

Test ("What color is the bag?", "black")

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

Problem with existing
setup + models

A novel VQA Split

A novel VQA model

Conclude

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

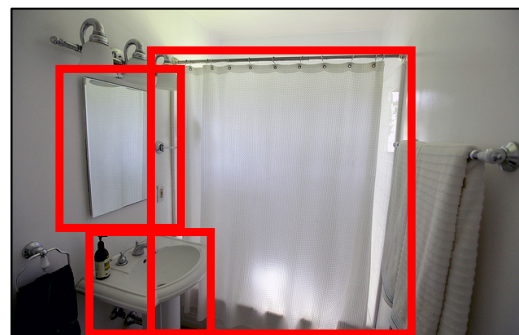
- Inductive biases in model architecture to prevent relying on priors
- Designed to disentangle:
 - What can be said?
 - What should be recognized?

Q: What **room** is this?

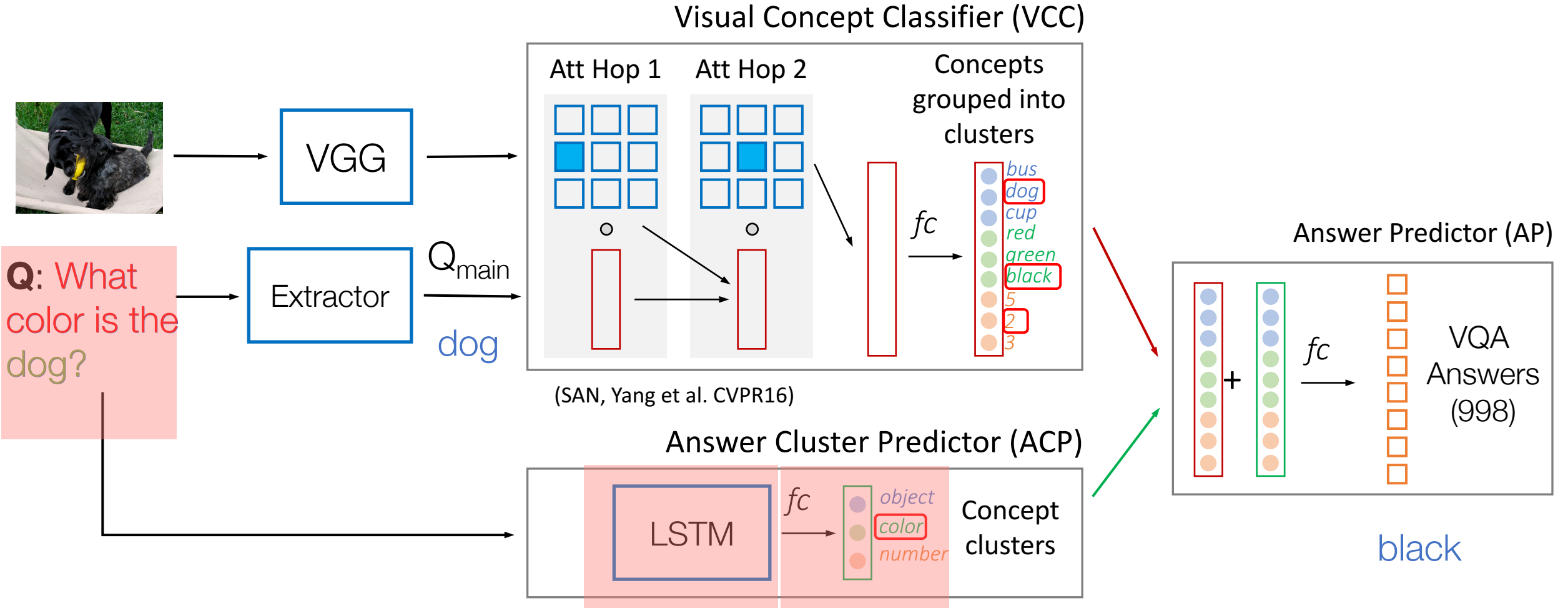
Grounded Visual Question Answering (GVQA) Model

- Inductive biases in model architecture to prevent relying on priors
- Designed to disentangle:
 - What can be said?
 - What should be recognized?

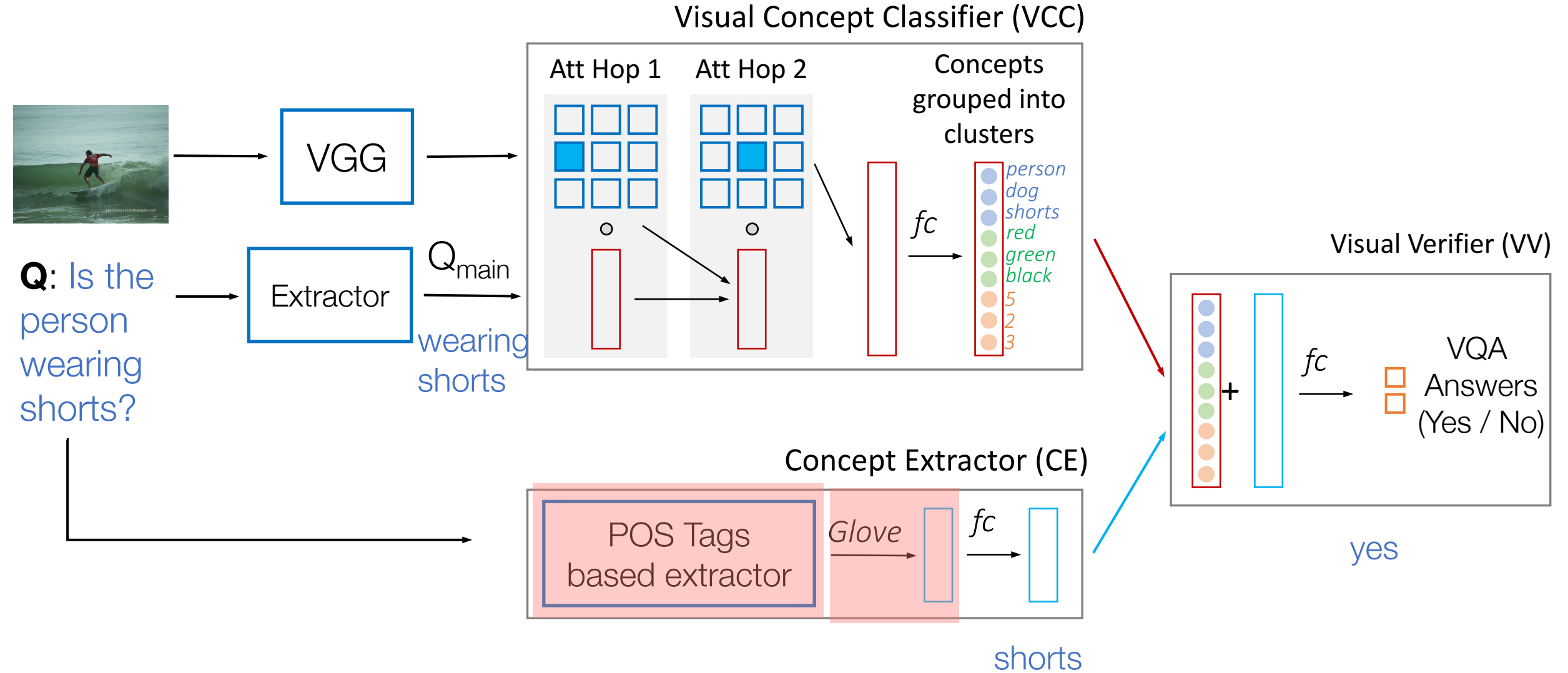
Q: What **room** is this?



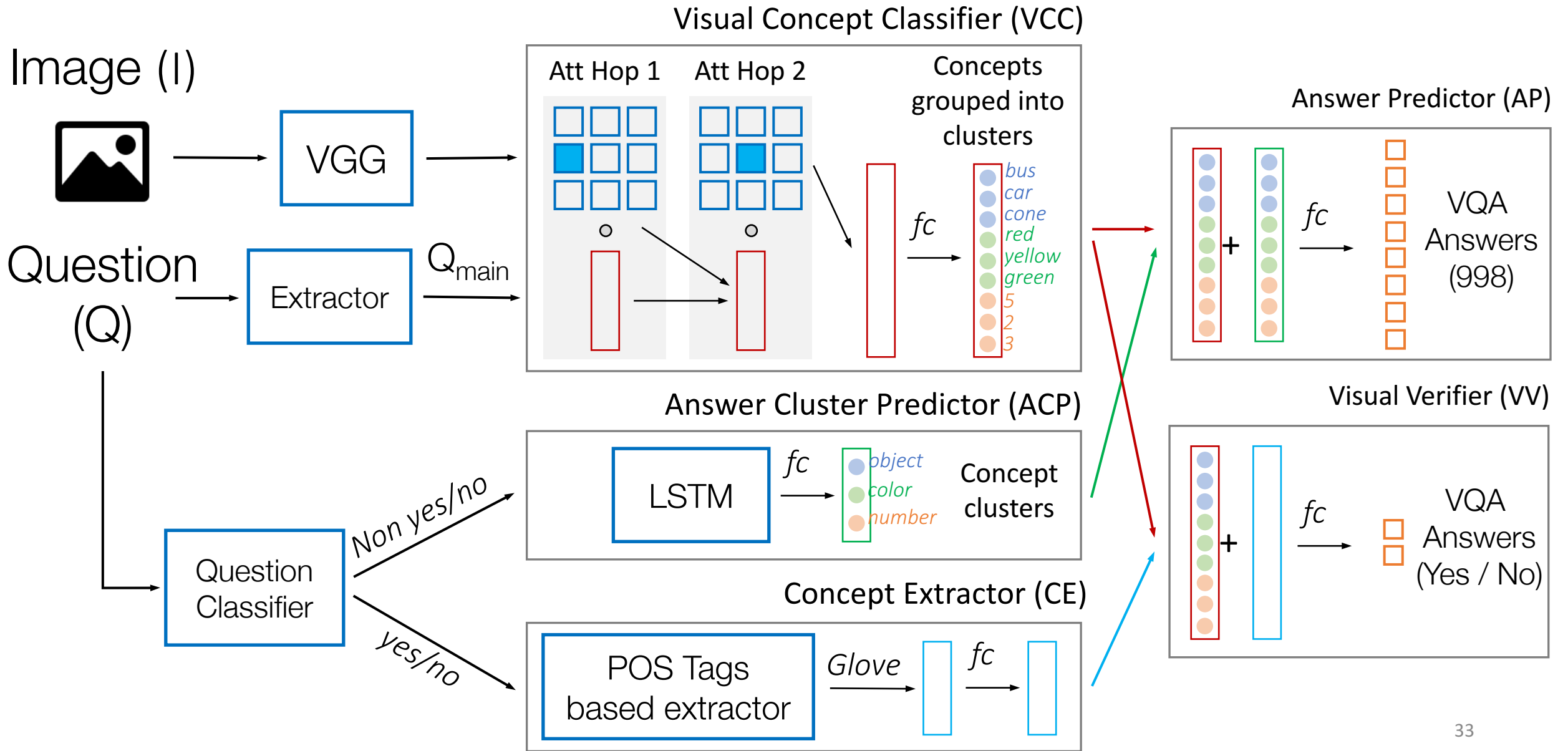
GVQA



GVQA



GVQA



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	} ↑ +12%
	SAN (Yang et al. CVPR16)	26.88	
VQA-CP v2	GVQA (Ours)	31.30	} ↑ +6%
	SAN (Yang et al. CVPR16)	24.96	

GVQA's output

Train

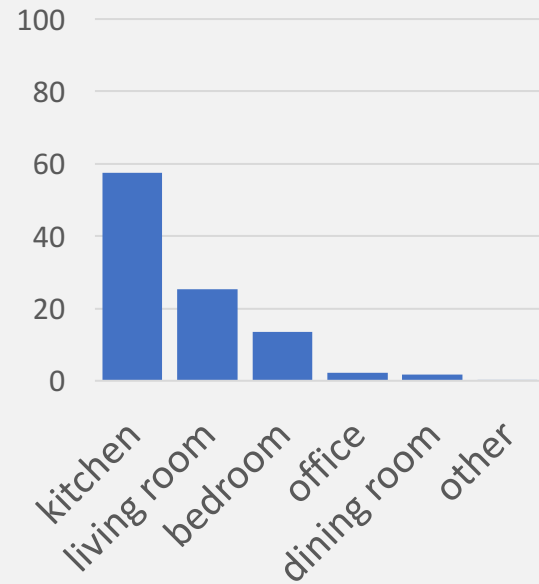
Q: What room is this?

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

non yes/no

ACP

color

VCC

bananas

green

many

food

Answer

green



GVQA's output

Q: Is the person smiling?



Q-classifier

yes/no

CE

smiling

VCC

smiling

woman

Answer

yes



GVQA's output

Q: What color are his pants?



Q-classifier

ACP

VCC

Answer

non yes/no

color

black

pants

black

1

dirt



GVQA's output


Q: What color are his pants?



Results on the Original Splits

Results on the Original Splits

Model	VQA v1
SAN	55.86
GVQA	51.12

 -5%

Results on the Original Splits

Model	VQA v1
SAN	55.86
GVQA	51.12

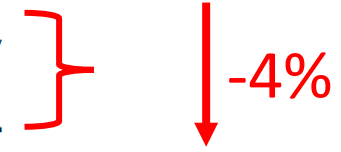
↓ -5%

Dataset	Model	Overall
VQA-CP v1	GVQA	39.23
	SAN	26.88

↑ 12% gain

Results on the Original Splits

Model	VQA v1	VQA v2
SAN	55.86	52.02
GVQA	51.12	48.24



A red bracket groups the VQA v2 scores for SAN (52.02) and GVQA (48.24). A red arrow points downwards from the SAN score to the GVQA score, with the text "-4%" next to it, indicating a 4% decrease in performance.

Results on the Original Splits

Model	VQA v1	VQA v2
Oracle (GVQA, SAN)	63.77	61.96

Diagram illustrating the percentage improvement in results:

- Improvement from VQA v2 (61.96) to VQA v1 (63.77) is $+(3-5\%)$ (blue arrow).
- Improvement from a baseline (implied to be 57.5) to VQA v1 (63.77) is $+(8-10\%)$ (green arrow).

Results on the Original Splits

Model	VQA v1	VQA v2
Oracle (GVQA, SAN)	63.77	61.96
Oracle (SAN, SAN)	60.85	56.68
Ensemble (GVQA, SAN)	56.91	52.96
SAN	55.86	52.02

Diagram illustrating performance differences between the Ensemble (GVQA, SAN) and SAN models:

- Ensemble (GVQA, SAN) VQA v1: 56.91
- SAN VQA v1: 55.86
- Improvement in VQA v1: $+1\%$
- Ensemble (GVQA, SAN) VQA v2: 52.96
- SAN VQA v2: 52.02
- Improvement in VQA v2: $+(0.4-0.5\%)$

Transparency

- GVQA has interpretable intermediate outputs
- Insights into why it is predicting what it is predicting
- Enables system designer to identify causes of error

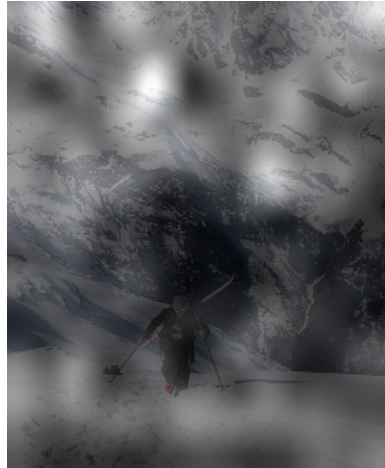


Q: What season is it?

A: winter

SAN answers
summer ❌

SAN's attention map



GVQA answers
winter ✅

GVQA's attention map



VCC says:

white
skiing
winter
mountains

ACP says answer
should be a season



Q: What is the most prominent ingredient?

A: pasta

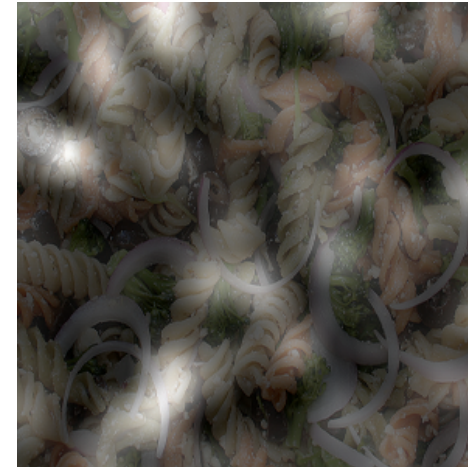
SAN answers
carrots ❌

SAN's attention map



GVQA answers
carrots ❌

GVQA's attention map



VCC says:

carrots
pasta
green
plate

ACP says answer
should be a
vegetable

Outline

Overview of VQA

Problem with existing
setup + models

A novel VQA Split

A novel VQA model

Conclude

Summary

- Models largely driven by superficial correlations
- A new split of the VQA dataset -- VQA-CP
- A novel Grounded VQA model – GVQA
- Moving forward: best of both worlds – priors + grounding

Dataset, model and code at:

www.cc.gatech.edu/~aagrwal307/vqa-cp/

Thank you!

Questions?