Visual Recognition beyond Appearances, and its Robotic Applications

'YZ' Yezhou Yang, Assistant Professor, CIDSE, ASU

Group Lead, Active Perception Group, Arizona State University

Tech co-Lead, The Institute of Automated Mobility (IAM), AZ

@Yezhou_Yang

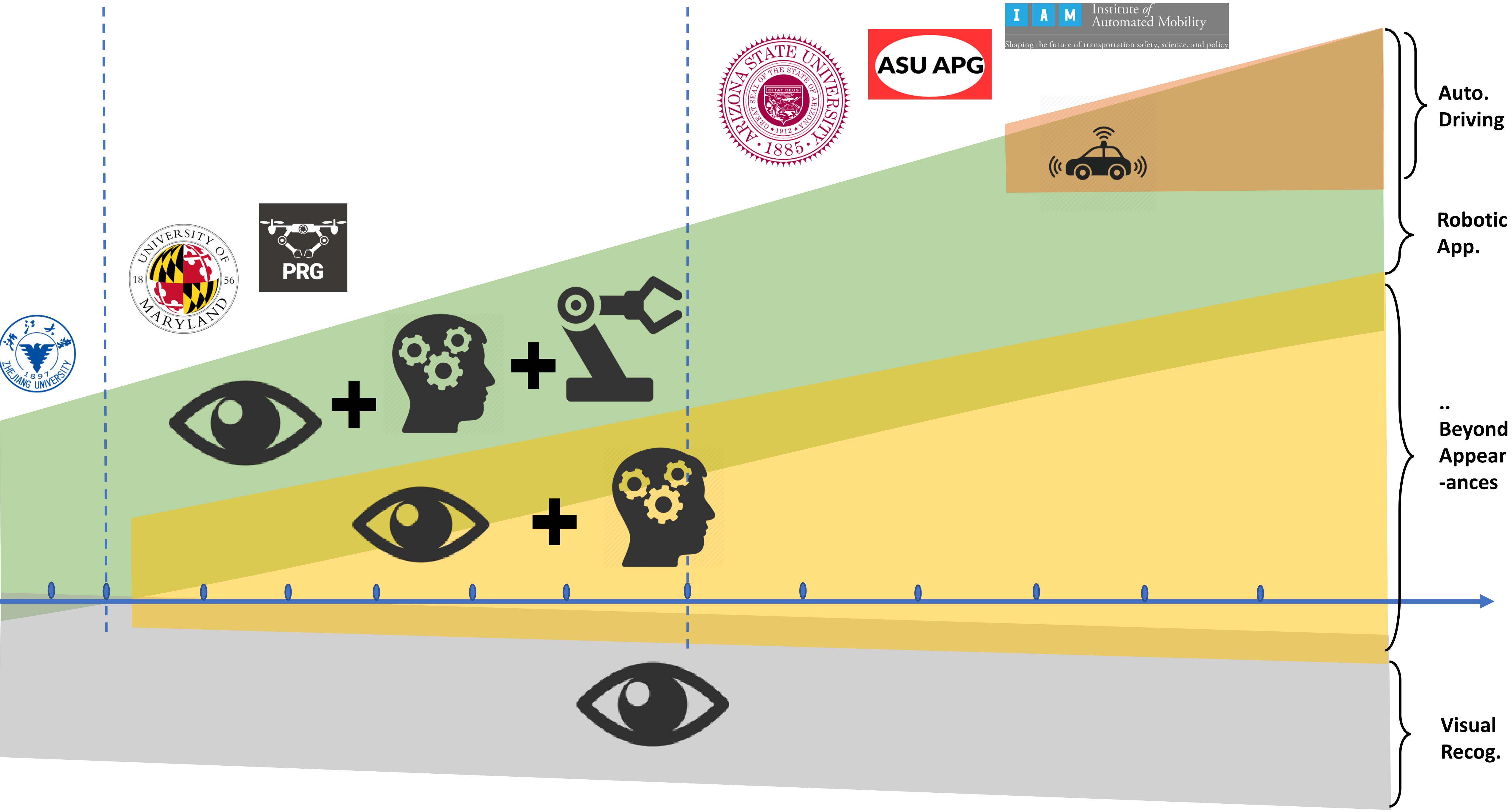
Arizona State University

ASU APG

IAM Institute of Automated Mobility

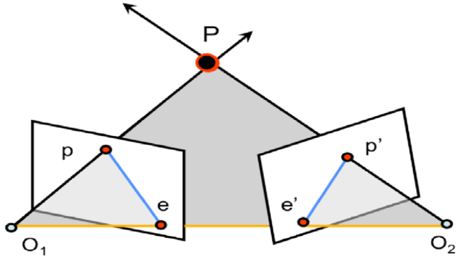
Shaping the future of transportation safety, science, and policy

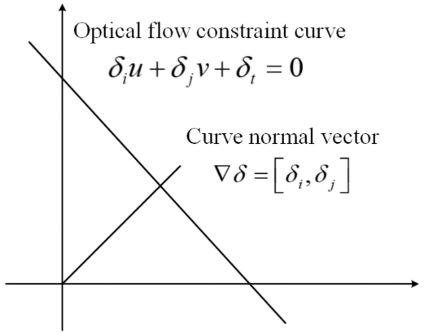
March 30th 2021 @ Microsoft Research

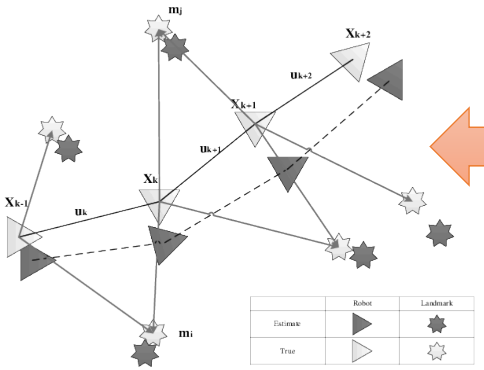


Institute of Automated Mobility
Shaping the future of transportation safety, science, and policy
ASU APG
PRG
Auto. Driving
Robotic App.
.. Beyond Appear -ances
Visual Recog.

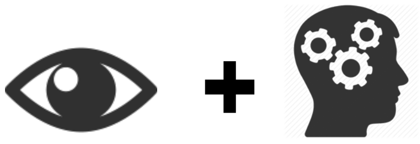
Physical Constraints!







from Internet

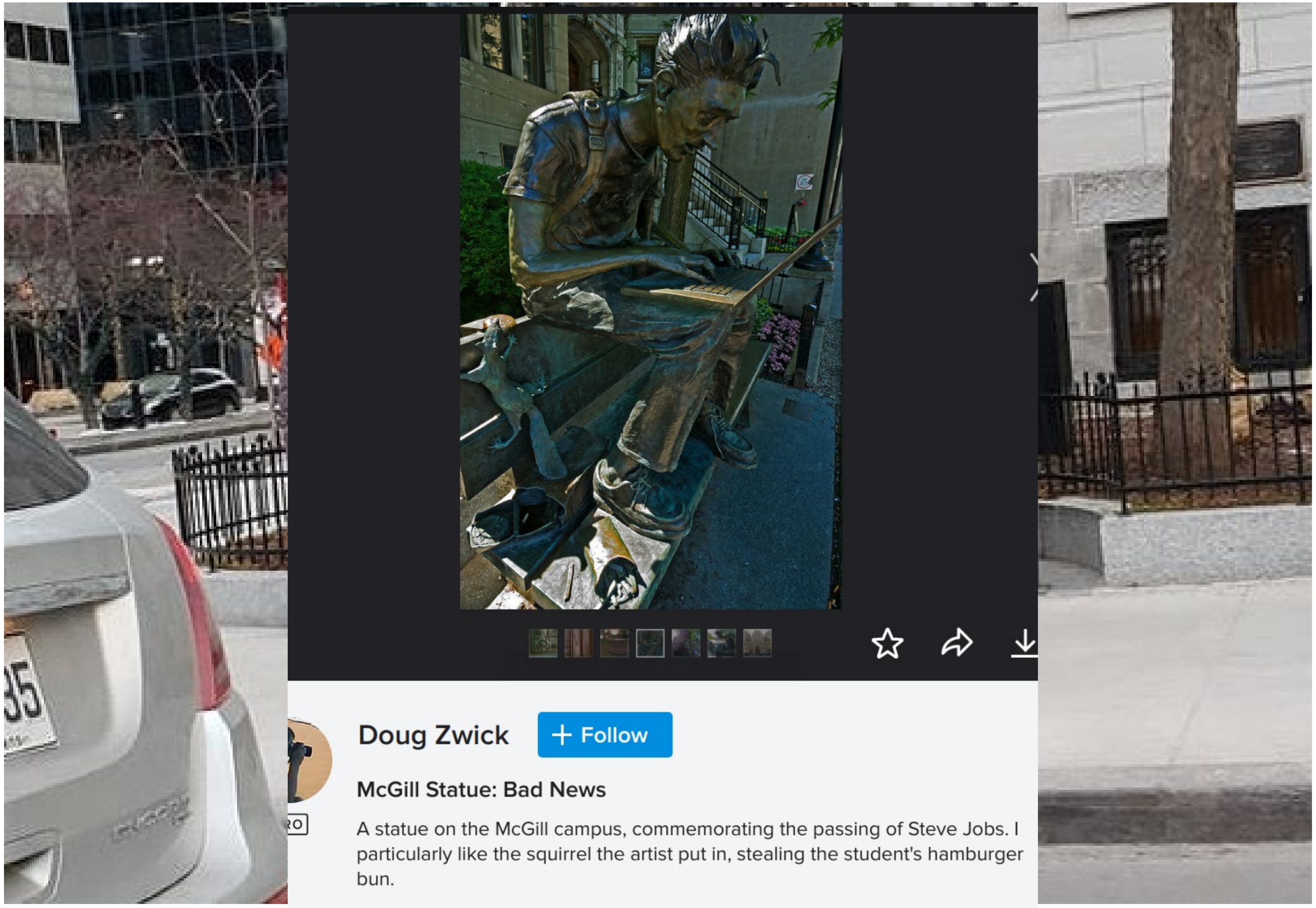


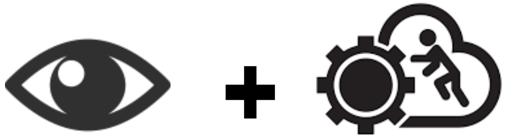
Semantic Constraints???

Visual Question Answering

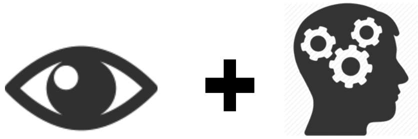
Q: how many people are waiting for bus?

A: Two? or Three?



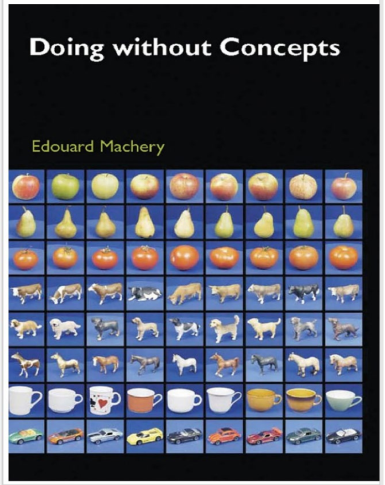


≠



Visual Recognition as Pattern Matching:

"Visual recognition is a cognitive process that involves identification of a visible CATEGORY from previous encounters "



Visual Recognition as it is:

"Visual recognition is a cognitive process that involves identification of a visible CONCEPT from previous encounters or KNOWLEDGE."

What is a concept?

"… A theory of concepts should describe the kind of knowledge stored in concepts, the way they are used in agents' cognitive processes, their format, their acquisition, and their neural localization… "

Categories ≠ Concepts

BUT, before we move on… we need benchmarking tasks… to validate our ideas… From the community:

Image Captioning (Flickr 8k, MSCOCO, etc.)

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Video Captioning (MSR-VTT, VATEX, etc.)

English:
a young girl does a cartwheel in her homes living room .
Ground Truth:
一个 年轻 女孩 在 她家 的 起居室 里 做 侧手 翻 。
NMT:
一个 年轻 女孩 在 她 的 房间 里 做 车轮 。
VMT:
一个 年轻 女孩 在 她 的 房间 里 翻筋斗 。

English:
a boy hits his head on a wall and knocks himself out .
Ground Truth:
一个 男孩 的 头 撞 在 墙上，把 自己 撞 倒 了 。
NMT:
一个 男孩 撞 他 的 头 在 墙上， 然后 敲 自己 出去 。
VMT:
一个 男孩 他 的 头 撞 在 墙上， 然后 自己 撞倒 了 。

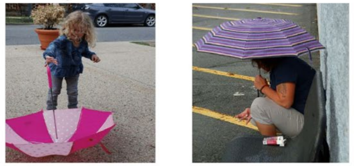
English:
a girl shows how to apply eyeliner, describing how to use strokes .
Ground Truth:
一个 女孩 展示 了 如何 使用 眼线笔 , 讲述 如何 画眉 。
NMT:
一个 女孩 展示 了 如何 使用 眼线笔 , 描述 了 如何 使用 笔画 。
VMT:
一个 女孩 展示 了 如何 使用 眼线笔 , 描述 了 如何 画眼线 。

Visual Question Answering (VQA, VQA-CP, etc.)

Who is wearing glasses?
man woman

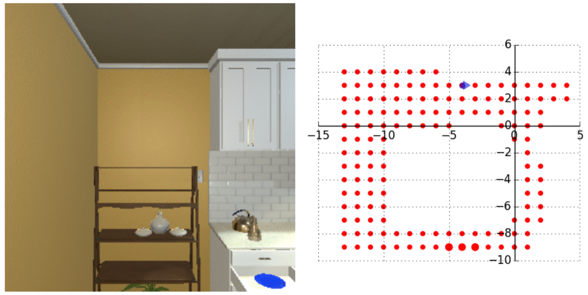
Where is the child sitting?
fridge arms

Is the umbrella upside down?
yes no

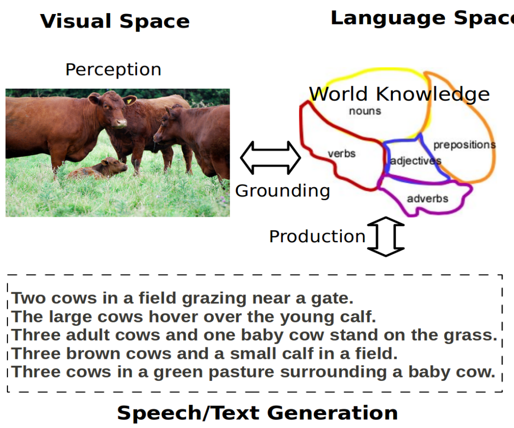


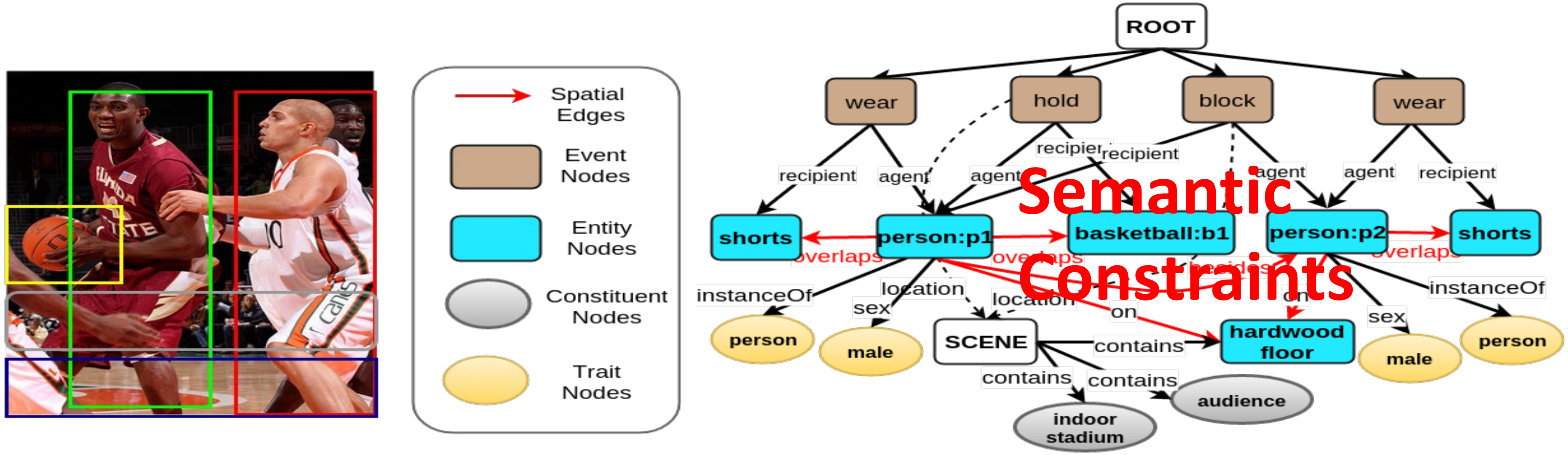
How many children are in the bed?
2 1

Visual Navigation (House3D (FAIR), AI2THOR, RxR etc.)



linguistic contextual information -> Explicit Knowledge Representation: Scene Description Graphs (SDGs)





EMNLP 11'
Sen. Gen. from Img, Captioning

ACS 16'
DeepIU Scene Description Graph (SDGs)

CVIU 17'
Image Under. w/ SDG

SDGs project webpage:
https://adityasomak.github.io/publication/sdg_cviu/

Experiment	BRNN-Karpathy	Our Method	Gold Standard
$R \pm D(8k)$	2.08 ± 1.35	$\mathbf{2.82 \pm 1.56}$	4.69 ± 0.78
$T \pm D(8k)$	2.24 ± 1.33	$\mathbf{2.62 \pm 1.42}$	4.32 ± 0.99
$R \pm D(30k)$	1.93 ± 1.32	$\mathbf{2.43 \pm 1.42}$	4.78 ± 0.61
$T \pm D(30k)$	2.17 ± 1.34	$\mathbf{2.49 \pm 1.42}$	4.52 ± 0.93
R±D(COCO)	$\mathbf{2.69 \pm 1.49}$	2.14 ± 1.29	4.71 ± 0.67
T±D(COCO)	$\mathbf{2.55 \pm 1.41}$	2.06 ± 1.24	4.37 ± 0.92

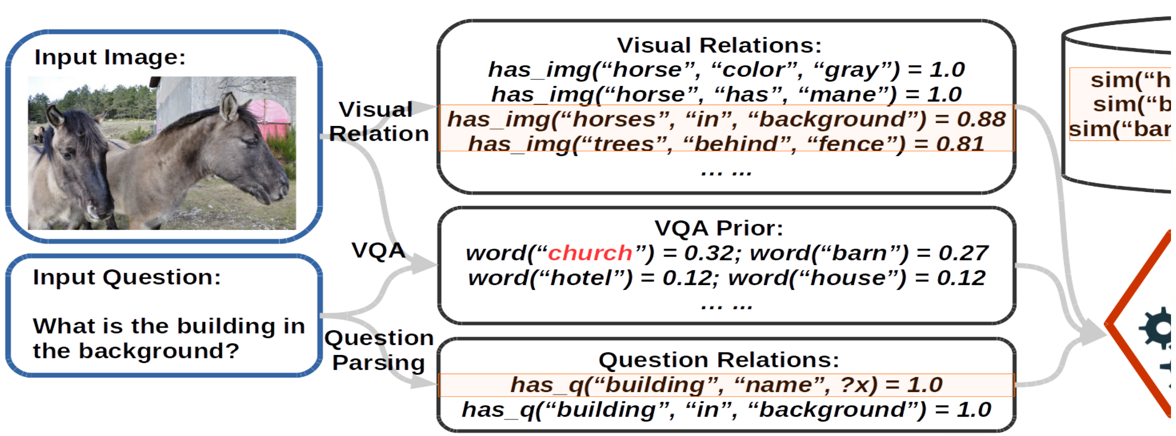
Table 1: Sentence generation relevance (R) and thoroughness (T) human evaluation results with gold standard and BRNN-Karpathy on Flickr 8k, 30k and MS-COCO datasets. D: Standard Deviation.

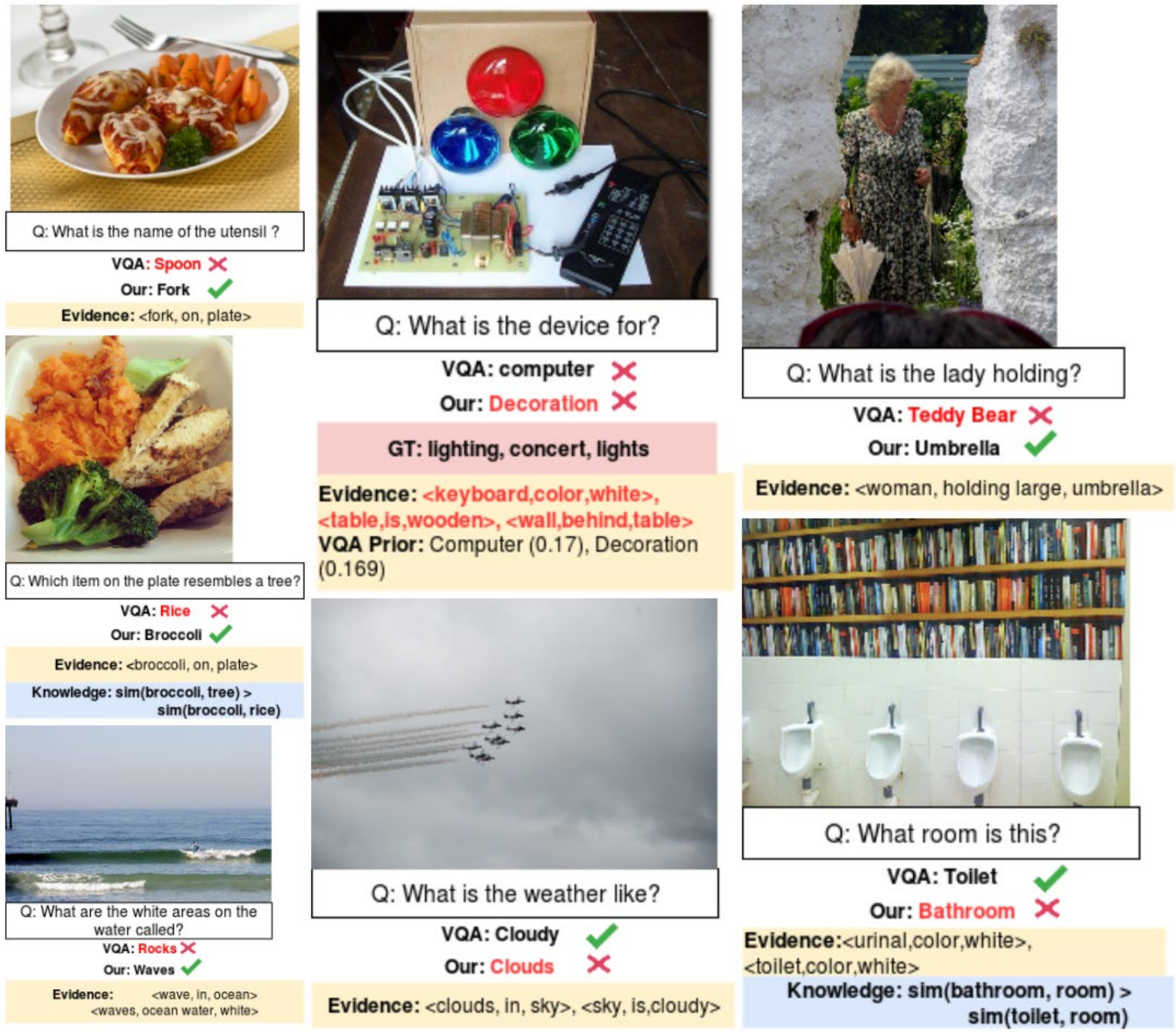
	Flickr8k			
Model	**R@1**	**R@5**	**R@10**	**Med r**
BRNN-Karpathy	11.8	32.1	44.7	12.4
Our Method-SDG	**18.1**	**39.0**	**50.0**	**10.5**
	Flickr30k			
BRNN-Karpathy	15.2	37.7	50.5	9.2
Our Method-SDG	**26.5**	**48.7**	**59.4**	**6.0**
	MS-COCO			
BRNN-Karpathy (1k)	**20.9**	**52.8**	**69.2**	**4.0**
Our Method-SDG (1k)	19.3	35.5	49.0	11.0
Our Method-SDG (2k)	15.4	32.5	42.2	17.0

Table 2: Image-Search Results: We report the recall@K (for $K = 1, 5$ and 10) and Med r (Median Rank) metric for Flickr8k, 30k and COCO datasets. For COCO, we experimented on first 1000 (1k) and random 2000 (2k) validation images.

Explicit Knowledge Representation Pros:

- Compatible with explicit reasoning over multiple knowledge resources;
- A direct decoding yields explicit explanations for end users (for explainable AI).





AAAI 18'
Explicit Reasoning f/ VQA

UAI 18'
Knowledge & Reasoning for Image Puzzles

Explicit Knowledge Representation Limitations:

- Even with soft reasoning engines such as (Probabilistic Soft Logic), the **lingering inconsistencies** among multiple knowledge resources could still hurt the overall performance.
- High fidelity requirement and **low error tolerance** towards knowledge sources. Especially when dealing with noisy detection inputs from the visual pathway.
- **Computationally expensive** (even with an accelerated & approximate PSL engine), the inferencing time is still comparatively much slower than end-to-end approaches.

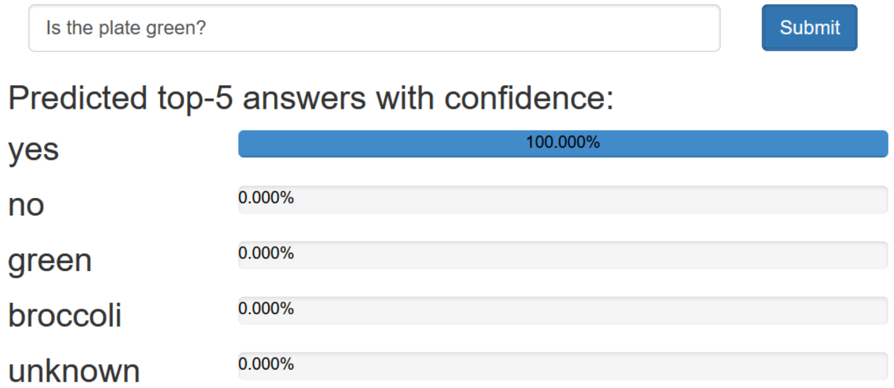
IJCAI 19’
Integrating Know. & Rea.
f/ Image Under.

	Categories	CoAttn	PSLDVQ	PSLDVQ-+CN
Speci-fic	what animal is (516)	65	**66.22**	**66.36**
	what brand (526)	38.14	37.51	37.55
	what is the man (1493)	54.82	**55.01**	54.66
	what is the name (433)	8.57	8.2	7.74
	what is the person (500)	54.84	**54.98**	54.2
	what is the woman (497)	45.84	**46.52**	45.41
	what number is (375)	4.05	**4.51**	**4.67**
	what room is (472)	88.07	87.86	**88.28**
	what sport is (665)	89.1	**89.1**	89.04
	what time (1006)	22.55	22.24	22.54
Sum-mary	**Other**	57.49	**57.59**	57.37
	Number	2.51	**2.58**	**2.7**
	Total	48.49	**48.58**	48.42
Color Related	what color (791)	48.14	47.51	47.07
	what color are the (1806)	56.2	55.07	54.38
	what color is (711)	61.01	58.33	57.37
	what color is the (8193)	62.44	61.39	60.37
	what is the color of the (467)	70.92	67.39	64.03
Gener-al	what (9123)	39.49	39.12	38.97
	what are (857)	51.65	**52.71**	**52.71**
	what are the (1859)	40.92	40.52	40.49
	what does the (1133)	21.87	21.51	21.49
	what is (3605)	32.88	**33.08**	32.65
	what is in the (981)	41.54	40.8	40.49
	what is on the (1213)	36.94	35.72	35.8
	what is the (6455)	41.68	41.22	41.4
	what is this (928)	57.18	56.4	56.25
	what kind of (3301)	49.85	49.81	49.84
	what type of (2259)	48.68	48.53	**48.77**
	where are the (788)	31	29.94	29.06
	where is the (2263)	28.4	28.09	27.69
	which (1421)	40.91	**41.2**	40.73
	who is (640)	27.16	24.11	21.91
	why (930)	16.78	16.54	16.08
	why is the (347)	16.65	16.53	**16.74**

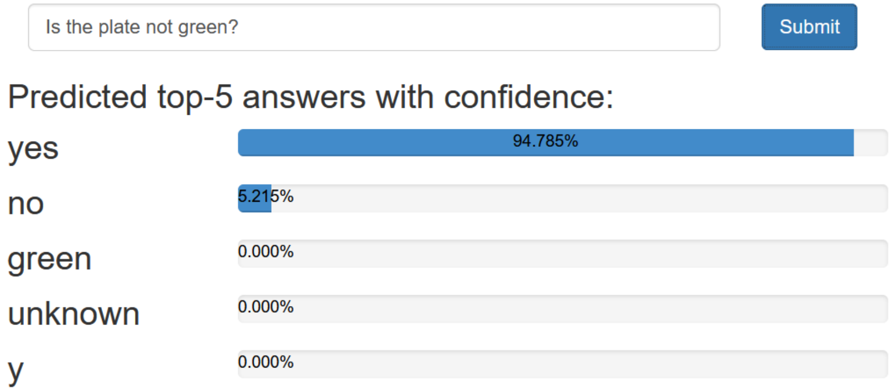
Table 3: Comparative results on the VQA validation questions. We report results on the non-Yes/No and non-Counting question types. Highest accuracies achieved by our system is presented in bold. We report the summary results of the set of “specific” question categories.

Explicit Knowledge Representation has limitations, so what's next?

- Observation: VQA models cannot comprehend *NEGATION, CONJUNCTION, and DISJUNCTION*
- **Solution: Explicit Knowledge Distillation with Data Re-engineering to improve VQA robustness?**



NEGATION



Tan & Bansal, LXMERT: Learning Cross-Modality Encoder Representationsfrom Transformers, EMNLP 2019

Q_1 : *Is there beer?*

VQA Model → YES ✓

Q_2 : *Is the man wearing shoes?*

VQA Model → NO ✓

LOGICAL COMPOSITION

$Q_1 \wedge \neg Q_2$: *Is there beer* **and** *is the man* **not** *wearing shoes?*

VQA Model → NO

VQA-Compose

$$Q^* = \widehat{Q_1} \circ \widehat{Q_2}, \qquad where\ \ \widehat{Q_1} \in \{Q_1, \neg Q_1\},\ \widehat{Q_2} \in \{Q_2, \neg Q_2\}.$$

Compositions of questions from VQA-v2

For each pair of questions, we use 10 propositional formulae to generate logically composed questions, and their ground-truth answer

QF	Question	AF	Answer
Q_1	Is there beer?	A_1	Yes
Q_2	Is the man wearing shoes?	A_2	No
$\neg Q_1$	Is there no beer?	$\neg A_1$	No
$\neg Q_2$	Is the man not wearing shoes?	$\neg A_2$	Yes
$Q_1 \wedge Q_2$	Is there beer and is the man wearing shoes?	$A_1 \wedge A_2$	No
$Q_1 \vee Q_2$	Is there beer or is the man wearing shoes?	$A_1 \vee A_2$	Yes
$Q_1 \wedge \neg Q_2$	Is there beer and is the man not wearing shoes?	$A_1 \wedge \neg A_2$	Yes
$Q_1 \vee \neg Q_2$	Is there beer or is the man not wearing shoes?	$A_1 \vee \neg A_2$	Yes
$\neg Q_1 \wedge Q_2$	Is there no beer and is the man wearing shoes?	$\neg A_1 \wedge A_2$	No
$\neg Q_1 \vee Q_2$	Is there no beer or is the man wearing shoes?	$\neg A_1 \vee A_2$	No
$\neg Q_1 \wedge \neg Q_2$	Is there no beer and is the man not wearing shoes?	$\neg A_1 \wedge \neg A_2$	No
$\neg Q_1 \vee \neg Q_2$	Is there no beer or is the man not wearing shoes?	$\neg A_1 \vee \neg A_2$	Yes

VQA-Supplement

Created using objects, antonyms, and captions

Objects (B)
person, cup, cell phone

Captions (C)
- *a man outside a clothing shop taking a video*
- *a man with a hat and eye glasses holding a cell phone*

QF	AF	Q	A
Q	A	Is he wearing a hat?	Yes
$\neg Q$	$\neg A$	Is he not wearing a hat?	No
$Q \wedge B$	A	Is he wearing a hat and is there a cell phone?	Yes
$Q \vee B$	$\top$	Is he wearing a hat or is there a cell phone?	Yes
$Q \wedge anto\,(B)$	$\perp$	Is he wearing a hat and is there a bowl?	No
$Q \vee anto\,(B)$	A	Is he wearing a hat or is there a bowl?	Yes
$Q \wedge C$	A	Is he wearing a hat and is this a man outside a clothing shop taking a video?	Yes
$Q \vee C$	$\top$	Is he wearing a hat or is this a man outside a clothing shop taking a video?	Yes
$Q \wedge \neg B$	$\perp$	Is he wearing a hat and is there no cell phone?	No

How to design semantic constraints or regularizations that can help leverage the data re-engineering?

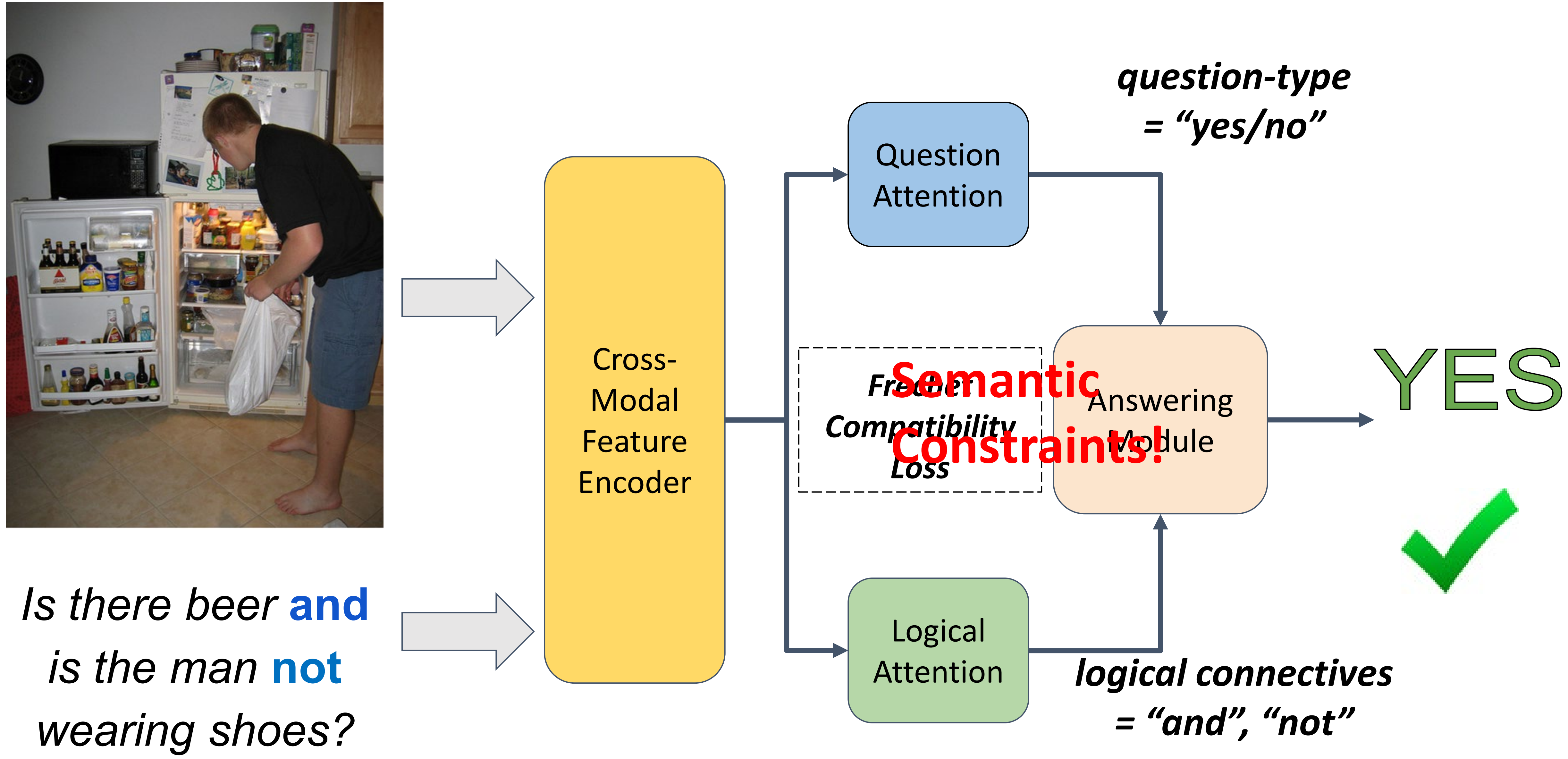
- Fréchet Inequlities bound the probabilities of events involving logical operations [Fréchet, 1935].

$$max(0, p(A_1) + p(A_2) - 1) \leq p(A_1 \wedge A_2) \leq min(p(A_1), p(A_2)).$$
$$max(p(A_1), p(A_2)) \leq p(A_1 \vee A_2) \leq min(1, p(A_1) + p(A_2)).$$

- In our case, we can use Fréchet Inequalities, with events being the answers to the questions.

- We define *Fréchet Mean* $\mathbf{m_A}$ to be the average of the left and right *Fréchet bounds*; $\mathbf{m_A = (b_L + b_R)/2}$.

- **Then, the Fréchet-Compatibility Loss is given by** $\mathscr{L}_{FC} = (p(A) - 1(m_A > 0.5))^2$


Is there beer and is the man not wearing shoes?
Cross-Modal Feature Encoder
Question Attention
question-type = "yes/no"
Frechet Compatibility Loss
Semantic Constraints!
Answering Module
YES
Logical Attention
logical connectives = "and", "not"

Visual Question Answering under the Lens of Logic
VQA-LOL

	Question		Predicted Answer	Accuracy (%) SOTA	Accuracy (%) LOL
VQA					
Q_1:	Is there beer?		YES (0.96)	88.20	86.55
Q_2:	Is the man wearing shoes?		NO (0.90)		
VQA-Compose				50.69	82.39
$\neg Q_2$:	Is the man *not* wearing shoes?		NO (0.80)		
$\neg Q_2 \wedge Q_1$	Is the man *not* wearing shoes *and* is there beer?		NO (0.62)		
$Q_1 \wedge C$	Is there beer and does this seem like a man bending over to look inside of a fridge?		NO (1.00)		
VQA-Supplement				50.61	87.80
$\neg Q_2 \vee B$	Is the man not wearing shoes or is there a clock?		NO (1.00)		
$Q_1 \wedge anto(B)$	Is there beer and is there a wine glass?		YES (0.84)		

Comparison with Baseline models

on VQA test-set and logical samples

Model	Parser	**Training Data**	**Test-Std. Accuracy (%) ↑**				**Val. Accuracy (%) ↑**		
			Yes-No	Number	Other	Overall	Compose	Supplement	Overall
MCAN	None	VQA [47]	86.82#	53.26#	60.72#	70.90	52.42	*	*
LXMERT	None	VQA [44]	**88.20**	**54.20**	**63.10**	**72.50**	50.79	50.51	50.65
LOL (qATT)	None	VQA	87.33	54.03	62.40	72.03	48.99	50.54	49.77
LXMERT	Oracle	VQA	88.20	54.20	63.10	72.50	86.38	74.29	80.33
LXMERT	Trained	VQA	88.20	54.20	63.10	72.50	86.35	68.75	77.55
LOL (full)	Oracle	VQA+Ours	86.55	53.42	61.58	71.04	85.79	88.51	87.15
LOL (full)	Trained	VQA+Ours	86.55	53.42	61.58	71.04	82.13	84.17	83.15
LXMERT	None	VQA+Ours	85.23	51.25	60.58	69.78	75.31	85.25	80.28
LOL (qATT)	None	VQA+Ours	86.79	52.66	61.85	71.19	79.88	87.12	83.50
LOL (full)	None	VQA+Ours	86.55	53.42	61.58	71.04	**82.39**	**87.80**	85.10

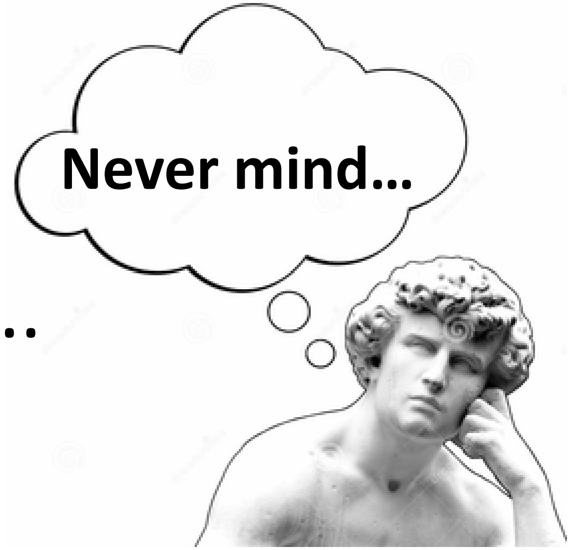
Explicit Knowledge Representation has limitations, so what’s next?

- Observation: VQA models cannot comprehend *NEGATION, CONJUNCTION, and DISJUNCTION*
- Solution: Explicit Knowledge Distillation with Data Re-engineering to improve VQA model robustness?
- **A continuation: VQA-LOL is with linguistic re-engineering, how about image re-engineering to improve model robustness?**

A: Green

What is the color of the frisbee?

A: I think it is still green?...



- *What color is the banana?*

Yellow (coz dataset, duh)

- *What sport are the men playing?*

Tennis (coz dataset, duh)

Concept of Input Mutations

Enable the mutation of inputs (questions and images) to expose the VQA model to perceptually similar, yet semantically dissimilar samples.

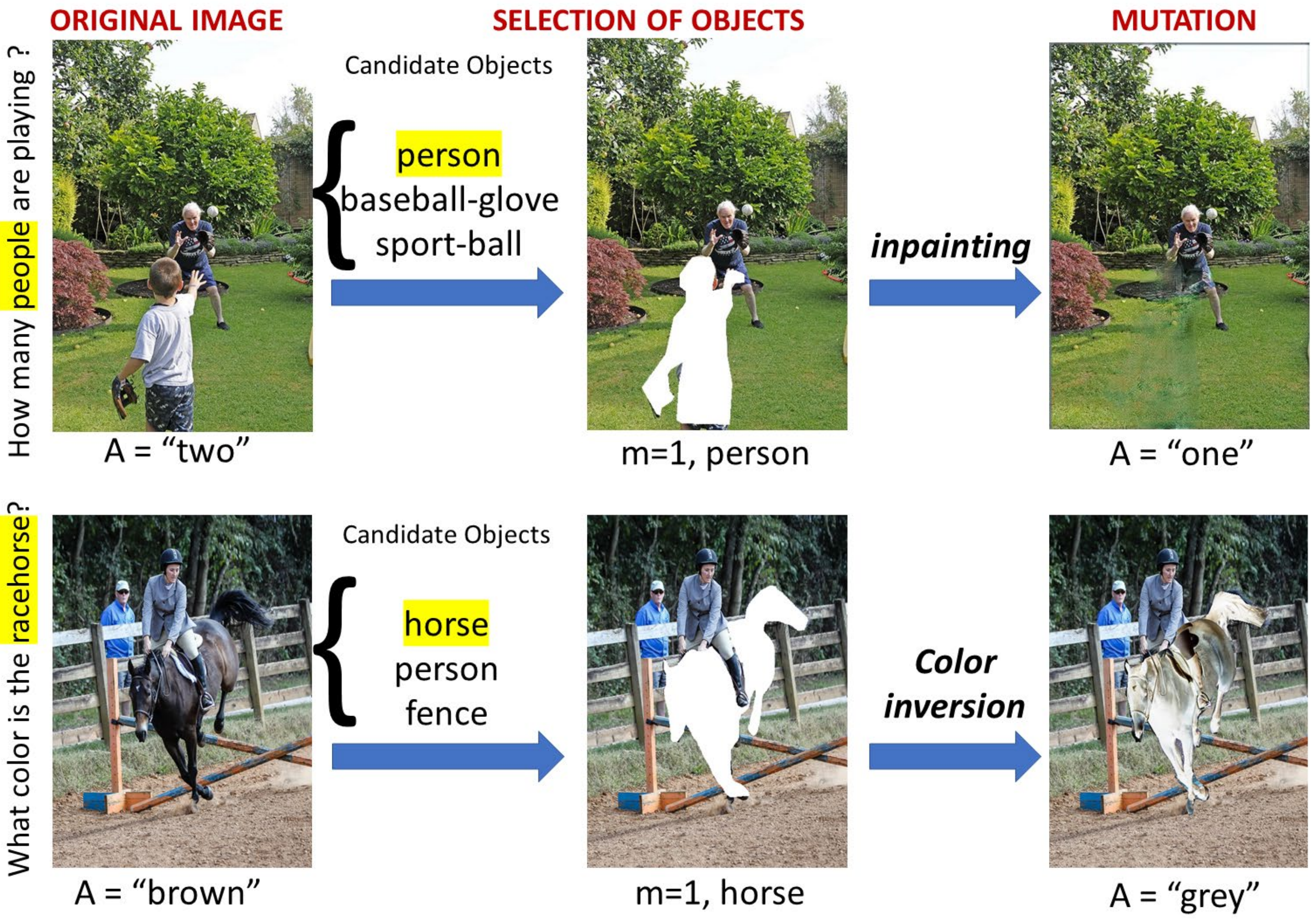
Let X = (Q, I) denote an input to a VQA system with a ***true answer*** "a".

A ***mutant*** input X* = ***(Q, I*)*** , or X* = ***(Q*, I)*** leads to a ***new answer*** "a*".

Image Mutations: removal of objects, morphing of object colors

Question Mutations: word-masking, word-substitution, negation

Generating Input Mutations



Generating Input Mutations

Mutation Type	Question	Answer
Original	Is the lady holding the baby?	Yes
Substitution (Negation)	Is the lady not holding the baby?	No
Substitution (Adversarial)	Is the cat holding the baby?	No
Original	How many people are there?	Three
Deletion (Masking)	How many [MASK] are there?	"Number"
Original	What is the color of the man's shirt?	Blue
Substitution (Negation)	What is not the color of the man's shirt?	Magenta
Deletion (Masking)	Is the [MASK] holding the baby?	Can't say
Original	What color is the umbrella ?	Pink
Deletion (Masking)	What color is the [MASK]?	"color"

Table 1: Examples of our question mutation. The image is shown on the left, and the original question is in the first row of the table. Examples of the two types of mutation are shown in the table.

VQA-MUTANT. Loss Functions

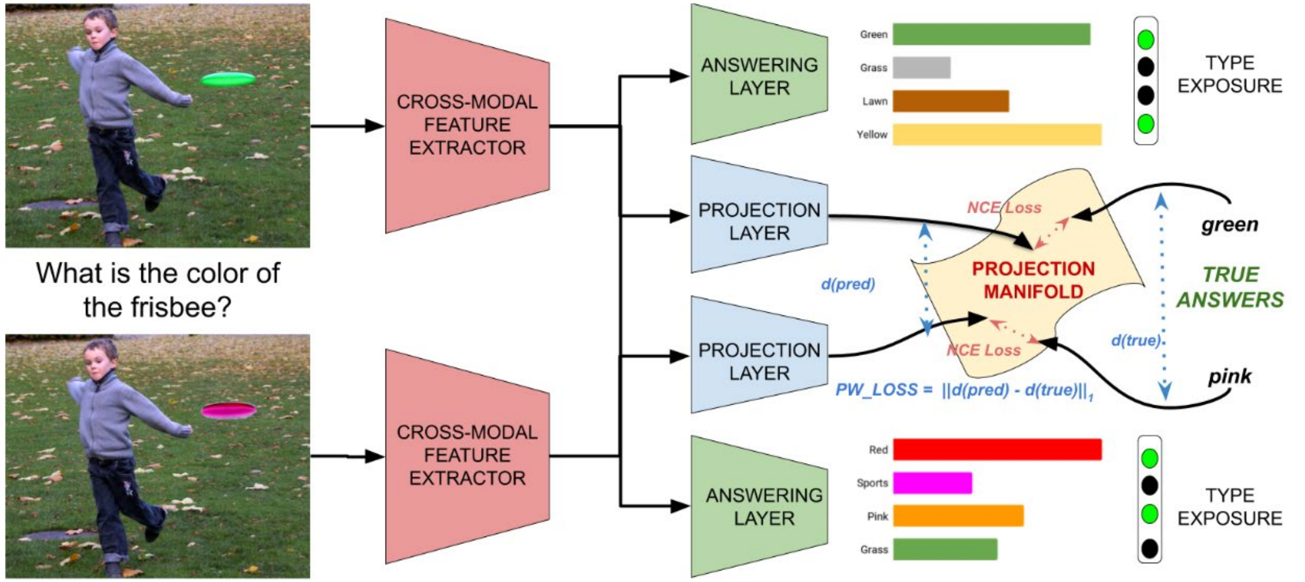
Traditional VQA Loss:

$$\mathcal{L}_{VQA} = \frac{-1}{N}\sum_{i=1}^{N} log(softmax(f_{VQA}(X_i), a_i)). \quad (1)$$

Answer Projection:

$$\mathcal{L}_{NCE} = -log(\frac{e^{cos(z_{feat}, z_a)}}{\sum_{a_i \in \mathcal{A}} e^{cos(z_{feat}, z_{a^i})}}), \quad (2)$$

$$z_{feat} = f_{proj}(z) \text{ and } z_a = f_{proj}(glove(a))$$



VQA-MUTANT. Loss Functions

Pair-wise Consistency: $$\mathcal{L}_{PW} = ||cos(z_{a_{GT}}, z^m_{a_{GT}}) - cos(z_{a_{pred}}, z^m_{a_{pred}})||_1.$$

“distance between predictions for mutant sample and original sample, must be consistent with the distance between true answers for mutant and original samples”

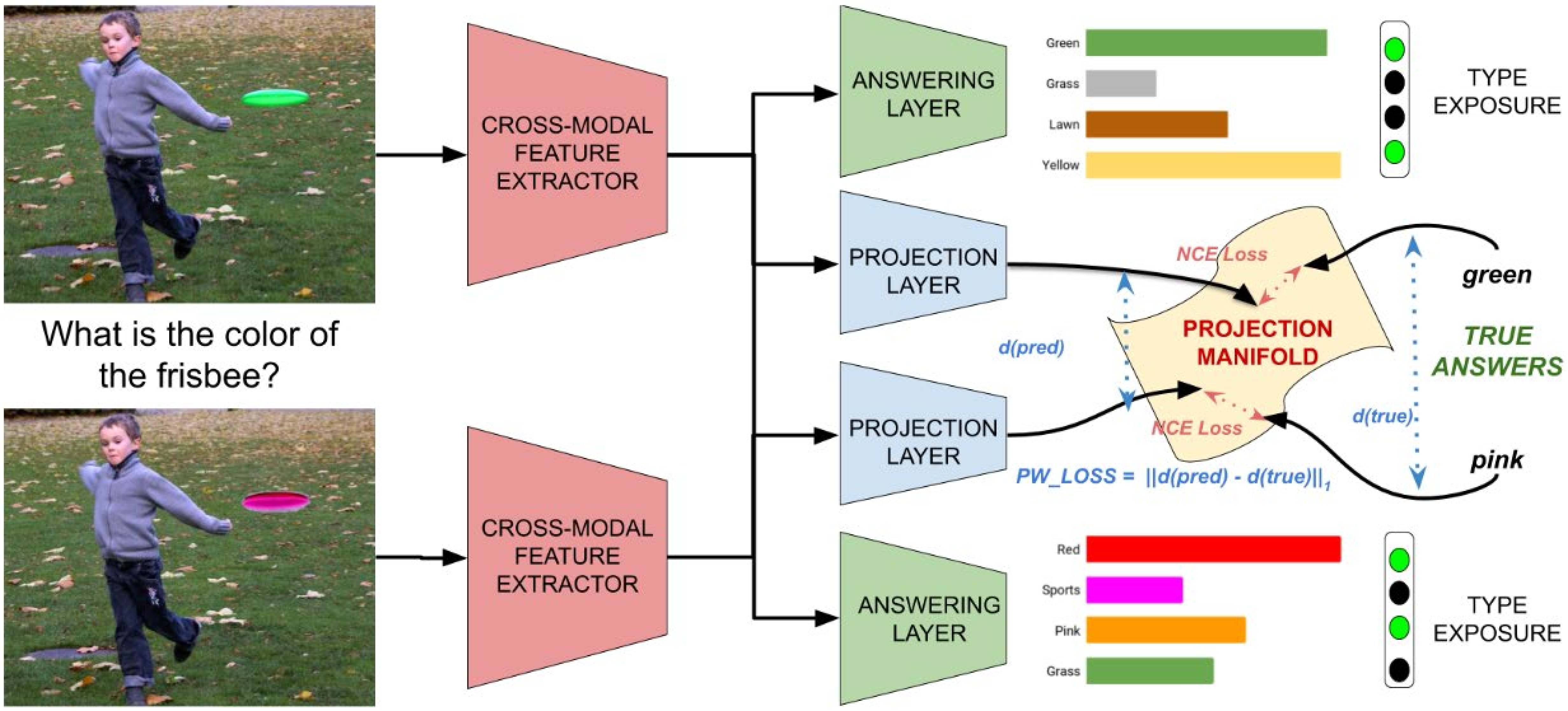
Semantic Constraints!

X  → z_{a_pred}

X^m  → $z^m_{a\_pred}$

A *“three”* → z_{a_GT}

A^m “zero” → $z^m_{a\_GT}$



What is the color of the frisbee?
CROSS-MODAL FEATURE EXTRACTOR
CROSS-MODAL FEATURE EXTRACTOR
ANSWERING LAYER
PROJECTION LAYER
PROJECTION LAYER
ANSWERING LAYER
Green
Grass
Lawn
Yellow
TYPE EXPOSURE
NCE Loss
PROJECTION MANIFOLD
NCE Loss
d(pred)
d(true)
green
pink
TRUE ANSWERS
PW_LOSS = ||d(pred) - d(true)||1
Red
Sports
Pink
Grass
TYPE EXPOSURE

UpDn: Anderson et al. CVPR 2018
LXMERT: Tan et al. EMNLP 2019

Results: VQA-CP Accuracy

Model	VQA-CP v2 test (%) ↑				VQA-v2 val (%) ↑				Gap (%)
	All	Yes/No	Num	Other	All	Yes/No	Num	Other	
GVQA (Agrawal et al., 2018b)	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65	16.94
AReg (Ramakrishnan et al., 2018)	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16	21.58
RUBi (Cadene et al., 2019)	47.11	68.65	20.28	43.18	63.10	-	-	-	14.05
SCR (Wu and Mooney, 2019)	48.47	70.41	10.42	47.29	62.30	77.40	40.90	56.50	13.83
LMH (Clark et al., 2019)	52.45	69.81	44.46	45.54	61.64	77.85	40.03	55.04	9.19
CSS (Chen et al., 2020a)	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11	0.96
UpDn (Anderson et al., 2018)	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66	23.74
UpDn + Ours	61.72	88.90	49.68	50.78	62.56	82.07	42.52	53.28	0.84
LXMERT (Tan and Bansal, 2019)	46.23	42.84	18.91	55.51	**74.16**	**89.31**	**56.85**	**65.14**	27.97
LXMERT + Ours	**69.52**	**93.15**	**67.17**	**57.78**	<u>70.24</u>	<u>89.01</u>	<u>54.21</u>	<u>59.96</u>	**0.72**

Table 3: Accuracies on VQA-CP v2 test and VQA-v2 val set. *"Ours"* represents the final model with Answer Projection, Type Exposure and Pairwise Consistency. Overall best scores are **bold**, our best are <u>underlined.</u>

UpDn: Anderson et al. CVPR 2018
LXMERT: Tan et al. EMNLP 2019

Analysis: Effect of Mutant Samples

Model	Data	VQA-CP v2 test ↑ (%)			
		All	Yes/No	Num	Other
UpDn	VQA-CP	39.74	42.27	11.93	46.05
UpDn	VQA-CP + Mutant	50.16	61.45	35.87	50.14
Increase in Accuracy		*10.42*	*19.18*	*23.94*	*4.09*
LXMERT	VQA-CP	46.23	42.84	18.91	55.51
LXMERT	VQA-CP + Mutant	59.69	73.19	32.85	59.29
Increase in Accuracy		*13.46*	*30.35*	*13.94*	*3.78*
LXM + Ours	VQA-CP + Img. Mut.	64.85	85.68	66.44	53.80
LXM + Ours	VQA-CP + Que. Mut.	67.92	91.64	65.73	56.09
LXM + Ours	VQA-CP + Both Mut.	**69.52**	**93.15**	**67.17**	**57.78**

Comparison of Backbone models (UpDn, LXMERT) trained with VQA-CP data augmented with MUTANT samples.

Comparison of our best model when trained with: image mutations, question mutations, and both types of mutations.

VQA-CP Leaderboard

A collections of papers about the VQA-CP dataset and a benchmark / leaderboard of their results. VQA-CP is an out-of-distribution dataset for Visual Question Answering, which is designed to penalize models that rely on question biases to give an answer.

Notes:

- All reported papers do not use the same baseline architectures, so the scores might not be directly comparable. This leaderboard is only made as a reference of all bias-reduction methods that were tested on VQA-CP.
- We mention the presence or absence of a validation set, because for out-of-distribution datasets, it is very important to find hyperparameters and do early-stopping on a validation set that has the same distribution as the training set. Otherwise, there is a risk of overfitting the testing set and its biases, which defeats the point of the VQA-CP dataset. This is why we **highly recommand** for future work that they build a **validation set** from a part of training set.

VQA-CP v2

Name	Base Arch.	Conference	All	Yes/No	Numbers	Other	Validation
MUTANT	LXMERT	EMNLP 2020	69.52	93.15	67.17	57.78	No valset
MUTANT	UpDown	EMNLP 2020	**61.72**	**88.90**	**49.68**	**50.78**	No valset
CL	UpDown + LMH + CSS	EMNLP 2020	59.18	86.99	49.89	47.16	No valset
RMFE	UpDown + LMH	NeurIPS 2020	54.55	74.03	49.16	45.82	No Valset
Loss-Rescaling	UpDown + LMH	Preprint 2020	53.26	72.82	48.00	44.46	
GradSup	Unshuffling	ECCV 2020	46.8	64.5	15.3	45.9	**Valset**
VGQE	S-MRL	ECCV 2020	50.11	66.35	27.08	46.77	No valset
CSS	UpDown + LMH	CVPR 2020	**58.95**	**84.37**	**49.42**	**48.21**	No valset

SimCLR: Chen et al. ICML 2020
AugMix: Hendrycks et al. ICLR 2020

Explicit Knowledge Representation has limitations, so what's next?

- Observation: VQA models cannot comprehend *NEGATION, CONJUNCTION, and DISJUNCTION*
- Solution: Explicit Knowledge Distillation with Data Re-engineering to improve VQA model robustness? Yes.
- **A continuation: VQA-LOL is with linguistic re-engineering, how about image re-engineering to improve model robustness? Yes.**

- We distinguish LOL and MUTANT from data-augmentation, **because the mutations can inform the design of semantic constraints or regularizations that can help leverage a pair of related inputs.**
- Recent work in image classification (SimCLR, AugMix) shows that carefully designed input manipulations can benefit generalization.

Robustness in VQA has become an active area of research within the past few years, with many challenges and benchmarks being established

- Challenges such as VQA-CP aim to achieve generalization w.r.t. distributional shift in the answer-space.
- *Selvaraju et al, CVPR 2020* tackle robustness to sub-questions.
- *Ray et al, EMNLP 2019* tackle robustness to entailed questions.
- *Ribeiro et al, ACL 2019* work on robustness to implied questions.
- *Shah et al, CVPR 2019* use cycle-consistency for rephrased questions.

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

Linjie Li, Zhe Gan, Jingjing Liu
Microsoft Dynamics 365 AI Research
{lindsey.li, zhe.gan, jingjl}@microsoft.com

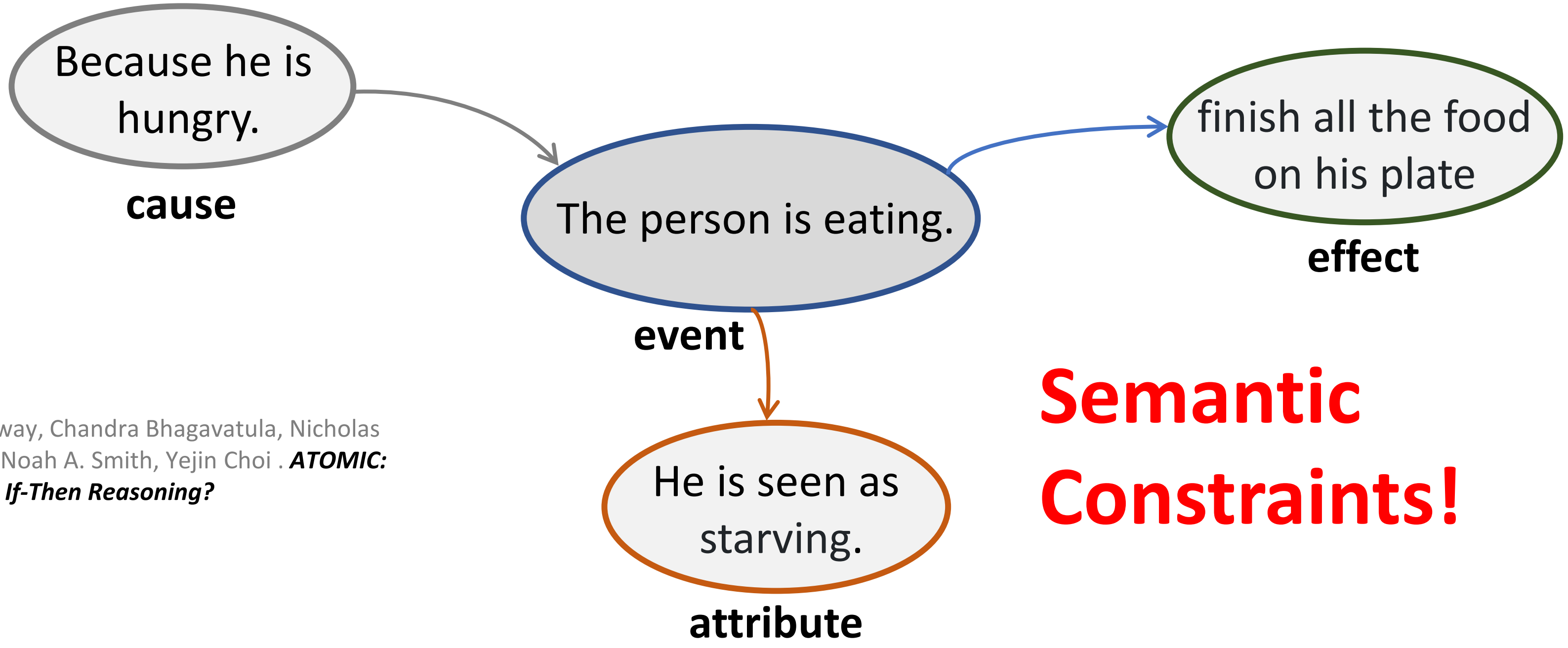
Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models

Linjie Li, Jie Lei, Zhe Gan, Jingjing Liu

Explicit Knowledge Representation has limitations, so what’s next?

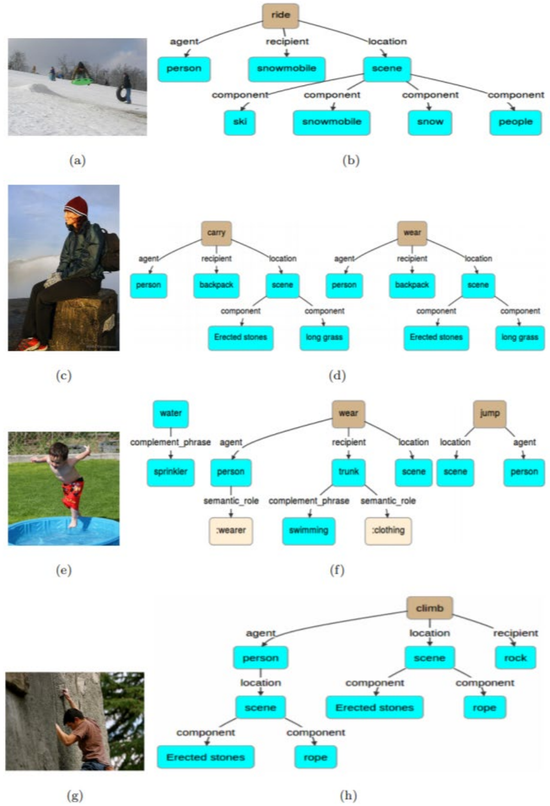
- Observation: VQA models cannot comprehend *NEGATION, CONJUNCTION, and DISJUNCTION*
- Solution: Explicit Knowledge Distillation with Data Re-engineering to improve VQA model robustness? Yes. VQA-LOL.
- VQA-LOL is with linguistic re-engineering, how about image re-engineering to improve model robustness? Yes. VQA-MUTANT, because the mutations can inform the design of semantic constraints or regularizations that can help leverage a pair of related inputs.
- **Can we distill explicit knowledge into a model to enrich generated outputs? (such as video captions).**

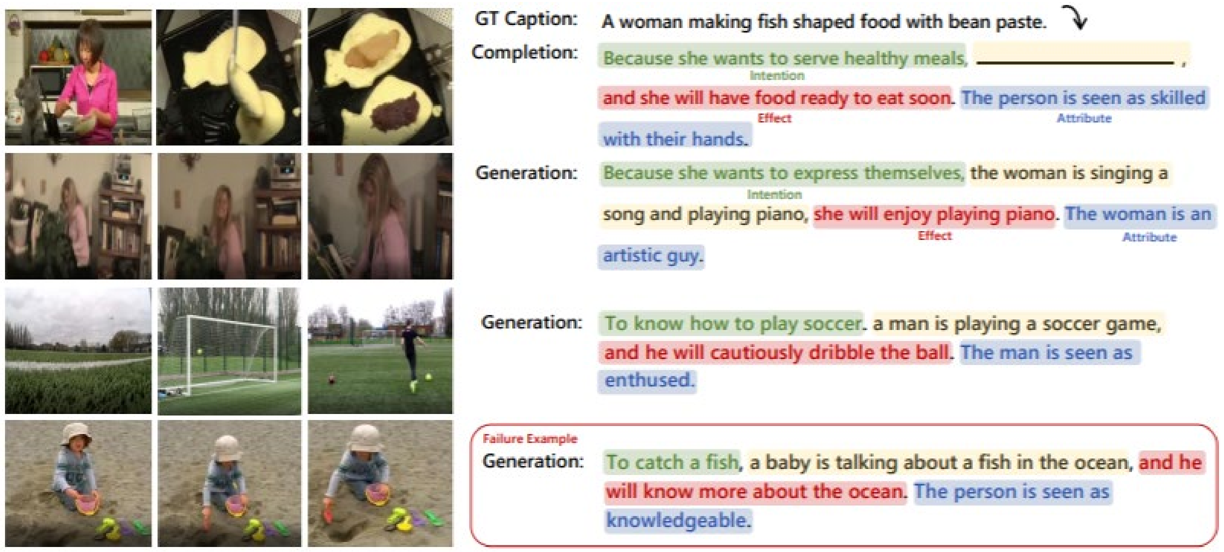
How does human understand the observed event? [1]



Semantic Constraints!

1. Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, Yejin Choi . ***ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning?***





EMNLP 11'
Sen. Gen. from Img, Captioning

ACS 16'
DeepIU
Scene
Description
Graph (SDG)

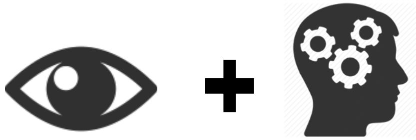
CVIU 17'
Image Under.
w/ SDG

EMNLP 20'
V2C: Video to
Commonsense

The 2020 Conference on Empirical Methods in Natural Language Processing
16th – 20th November 2020

https://asu-active-perception-group.github.io/Video2Commonsense/index.html

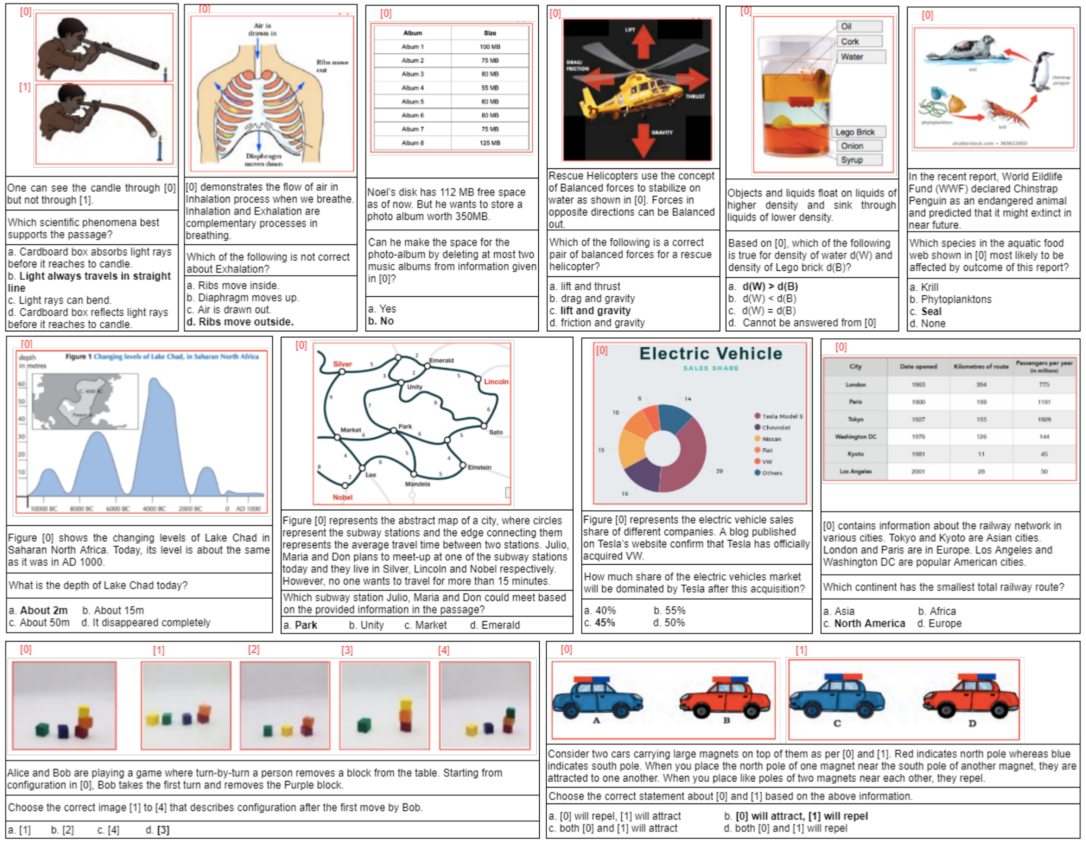
Our Datasets and Benchmarking tasks for

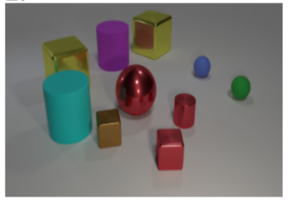


Question: “What word connects these images?”. Answer is “Fall”. The first image shows the season fall, the second and third image respectively has waterfall and rainfall in it and in the fourth image, a statue is “fall”-ing.

https://imageriddle.wordpress.com/imageriddle/

UAI 2018





1. T_A: Paint the small green ball with cyan color.
   Q_H: Are there equal number of yellow cubes on left of purple object and cyan spheres? (A: yes)
2. T_A: Add a brown rubber cube behind the blue sphere that inherits its size from the green object.
   Q_H: How many things are either brown or small? (A: 6)
3. T_A: John moves the small red cylinder on the large cube that is to the right of purple cylinder.
   Q_H: What color is the object that is at the bottom of the small red cylinder? (A: yellow)

Figure 2: Three examples from CLEVR_HYP dataset: given image (I), action text (T_A), question about hypothetical scenario (Q_H) and corresponding answer (A). The task is to understand possible perturbations in I with respect to various action(s) performed as described in T_A. Questions test various reasoning capabilities of a model with respect to the results of those action(s).

https://shailaja183.github.io/vlqa/

EMNLP 2020 findings

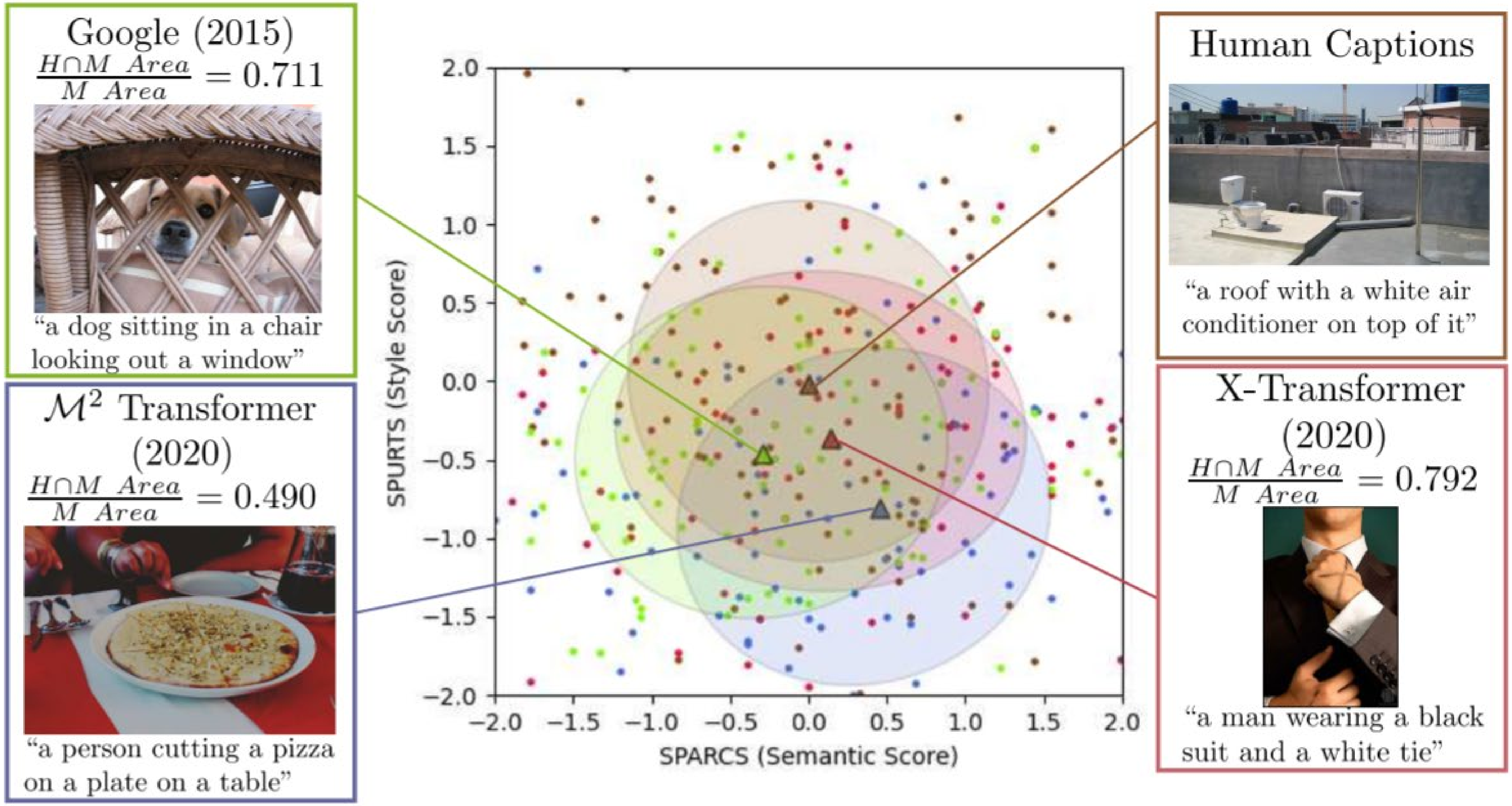
https://github.com/shailaja183/clevr_hyp

NAACL 2021 to appear

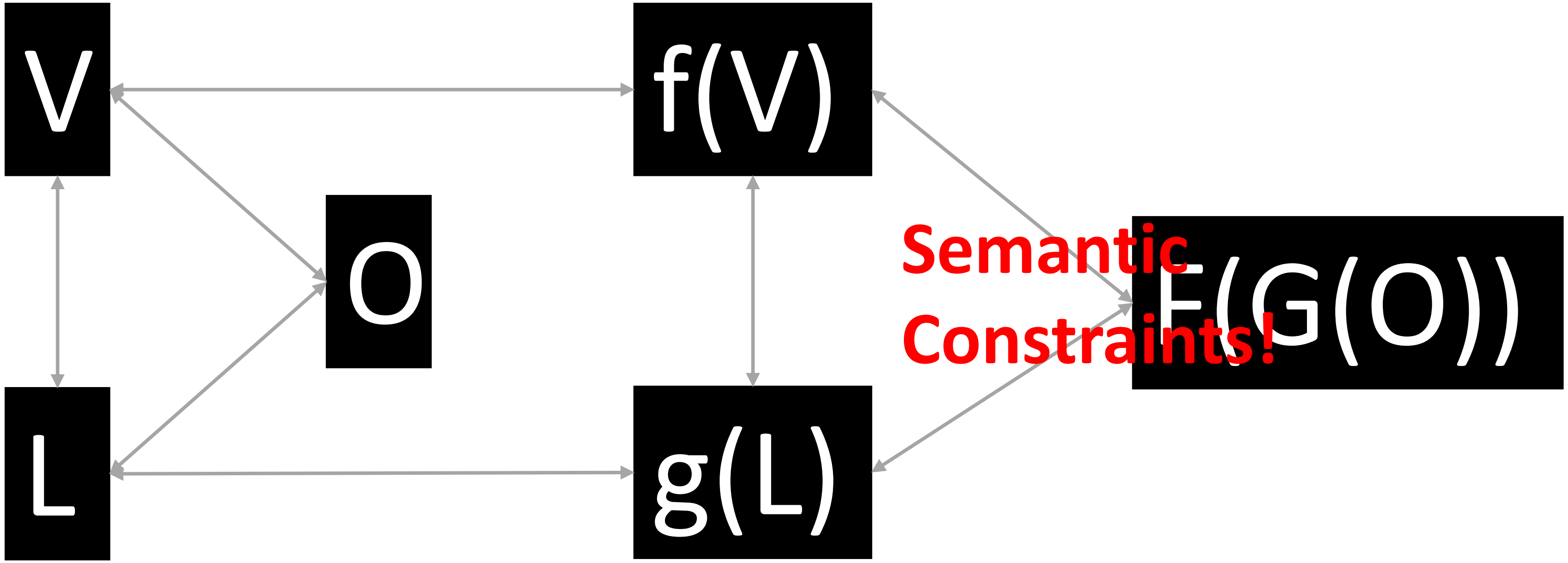
Shailaja Sampat

Lacks definite ground truth, thus evaluation is challenging…

https://github.com/JoshuaFeinglass/SMURF
SMURF; J. Feinglass and Y. Yang, ACL 2021



LOL, MUTANT, V2C… A common semantic augmentation service?



f: CV (AI) g: NLP (AI)

https://github.com/JoshuaFeinglass/SMURF

WeaQA; P. Banerjee, et.al, ACL 2021 Findings

Captions

- *A car that seems to be parked illegally behind a legally parked car*
- *A couple of cars parked in a busy street sidewalk*
- *Cars try to maneuver into parking spaces along a densely packed street.*
- *two cars parked on the sidewalk on the street*

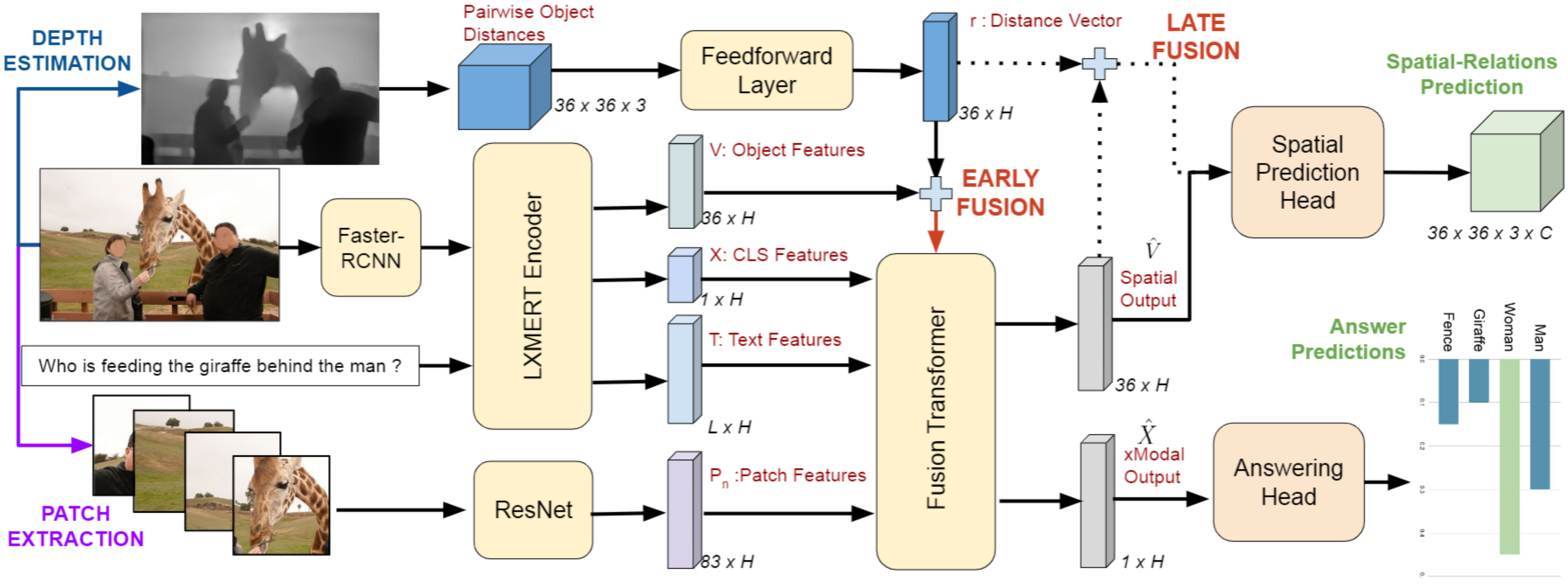
	Question	Answer(Confidence)
VQA-v2	1. How many doors does the gray car have ?	4 (1.0)
	2. Why does the windshield look opaque ?	Clear (0.6), No (0.3), Reflection (0.9)
Synthetic (Ours)	1. How is something parked ?	Illegally (1.0)
	2. Is there a truck ?	No (1.0)
	3. Is it a couple of cars parked in a busy street sidewalk?	Yes (1.0)
	4. Where does something maneuver?	Into Parking Spaces (1.0)

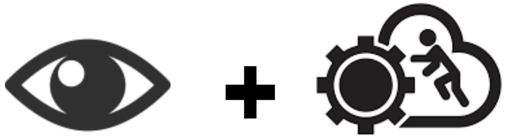
- *A man in skies is coming up the hill*
- *A skier is passing a competition race marker*
- *A man takes a picture of a skier*
- *A cross-country skier is competing at night in snow*

More examples can be found in the Appendix.

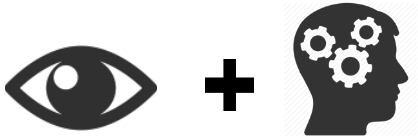
	Question	Answer(Confidence)
GQA	1. Is the man on the left or on the right ?	Right (1.0)
	2. Who is wearing the jersey ?	Man (1.0)
Synthetic (Ours)	1. What is someone passing ?	A competition race marker (1.0)
	2. When is someone competing ?	At night (1.0)
	3. Who is coming ?	A man in skiis (1.0)
	4. Is that a man in skateboard coming up the hill ?	No
	5. Where is someone coming?	Up the hill (1.0)

Geometry cues (depth, surface normal, etc.) guided
semantic constraints…
Under Review



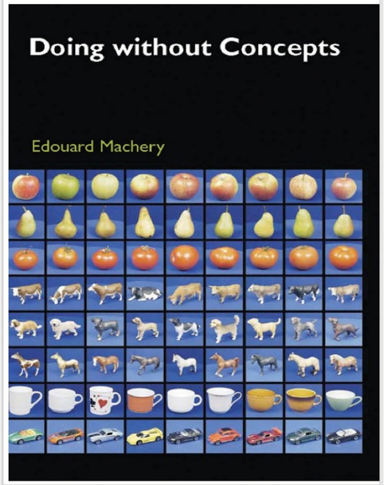


≠



Visual Recognition as Pattern Matching:

“Visual recognition is a cognitive process that involves identification of a visible CATEGORY from previous encounters ”



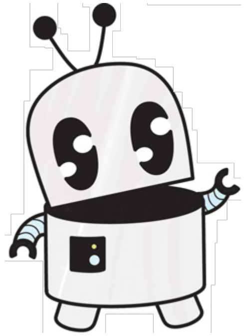
Visual Recognition as it is:

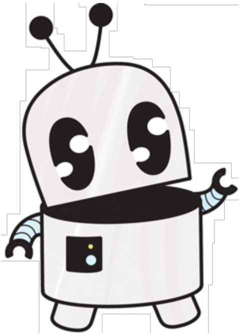
“Visual recognition is a cognitive process that involves identification of a visible CONCEPT from previous encounters or KNOWLEDGE.”

What is a concept?

“… A theory of concepts should describe the kind of knowledge stored in concepts, the way they are used in agents’ cognitive processes, their format, their acquisition, and their neural localization… ”

Categories ≠ Concepts ← Agents





A

Goal: *Locate Music Instrument*

Visual Navigation Model

Action: MOVE_FORWARD()

Goal: *Locate Coffee Mug*

Visual Navigation Model

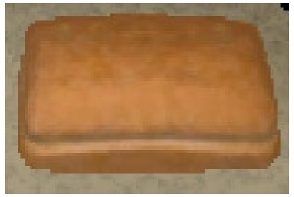
Action: TURN_RIGHT (45 degree)

BUT, before we move on… we STILL need benchmarking tasks… to validate our ideas…

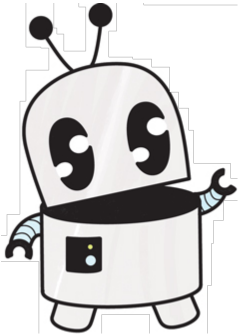
Visual Navigation (Robotic Object Search) as an Active Object Perceiver:

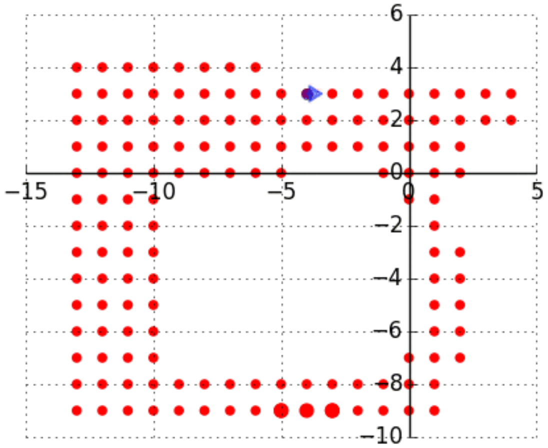
Motivation & Task:

Robot with vision that finds objects



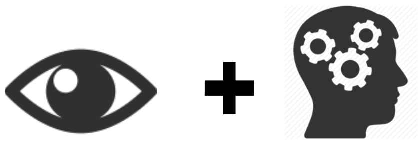
Target Object

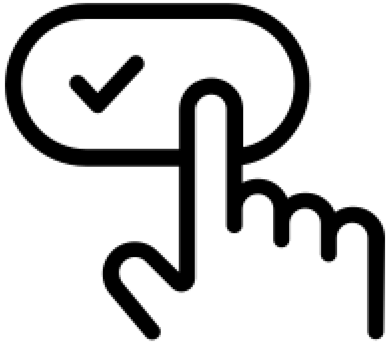




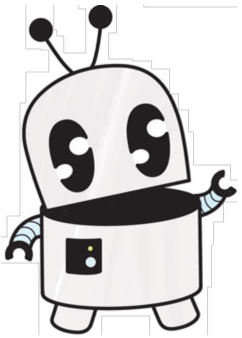
*E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, “AI2-THOR: An Interactive 3D Environment for Visual AI,” arXiv,2017.

Why Robotic Object Search?

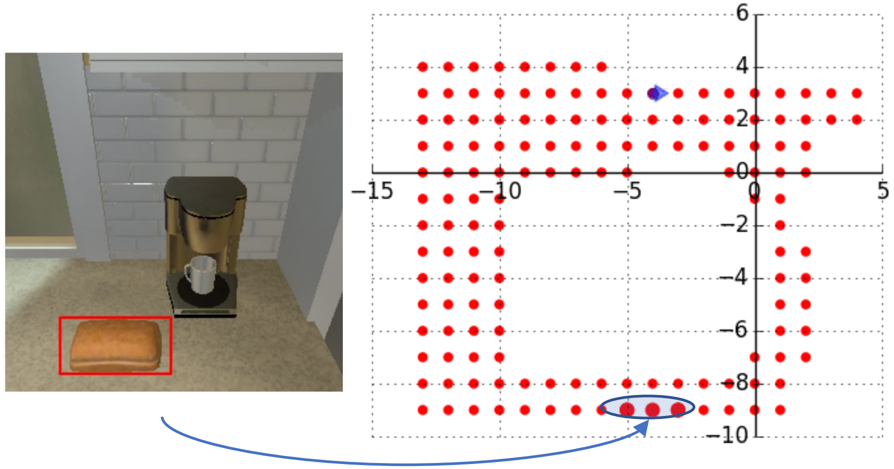


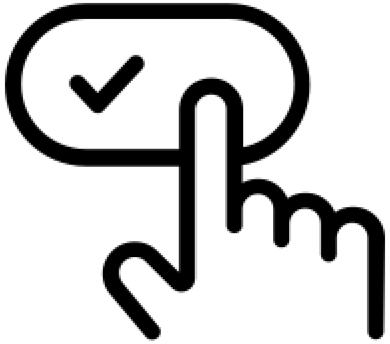


Captioning; Dense Captioning;
Visual Question Answering;
Image/Video understanding;
Visual Commonsense Reasoning;
…



…

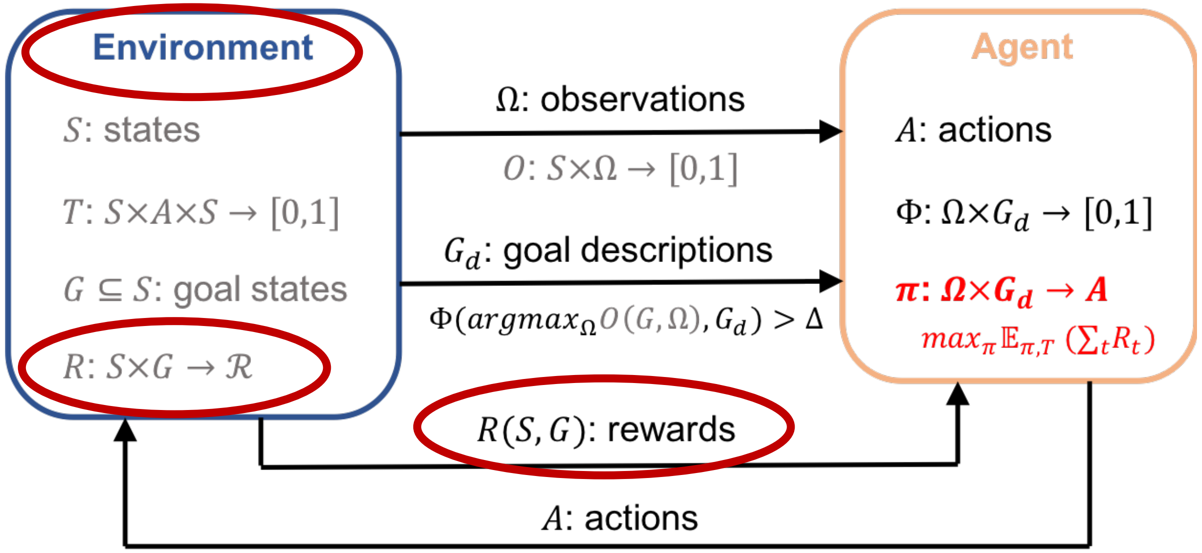




Visual Navigation;
Visual Language Navigation;
Embodied Visual QA;
Embodied Commonsense Reasoning
…

Vision-guided Policy Learning for Robotic Object Search

- How to define a good reward function?
  - ✓ Reward Functions via Visual Understanding [1]
- How to learn in a sparse reward setting?
  - ✓ Efficient Exploration with Hierarchical Policy [2]
- How to generalize across various instances?
  - ✓ Task-relevant Features from State Observations [3]
  - ✓ Goal Representation with Goals Relational Graph [4]
  - ➢ Data-efficient Neural-symbolic Modeling



[1] Active Object Perceiver: Recognition-Guided Policy Learning for Object Searching on Mobile Robots. IROS 2018.

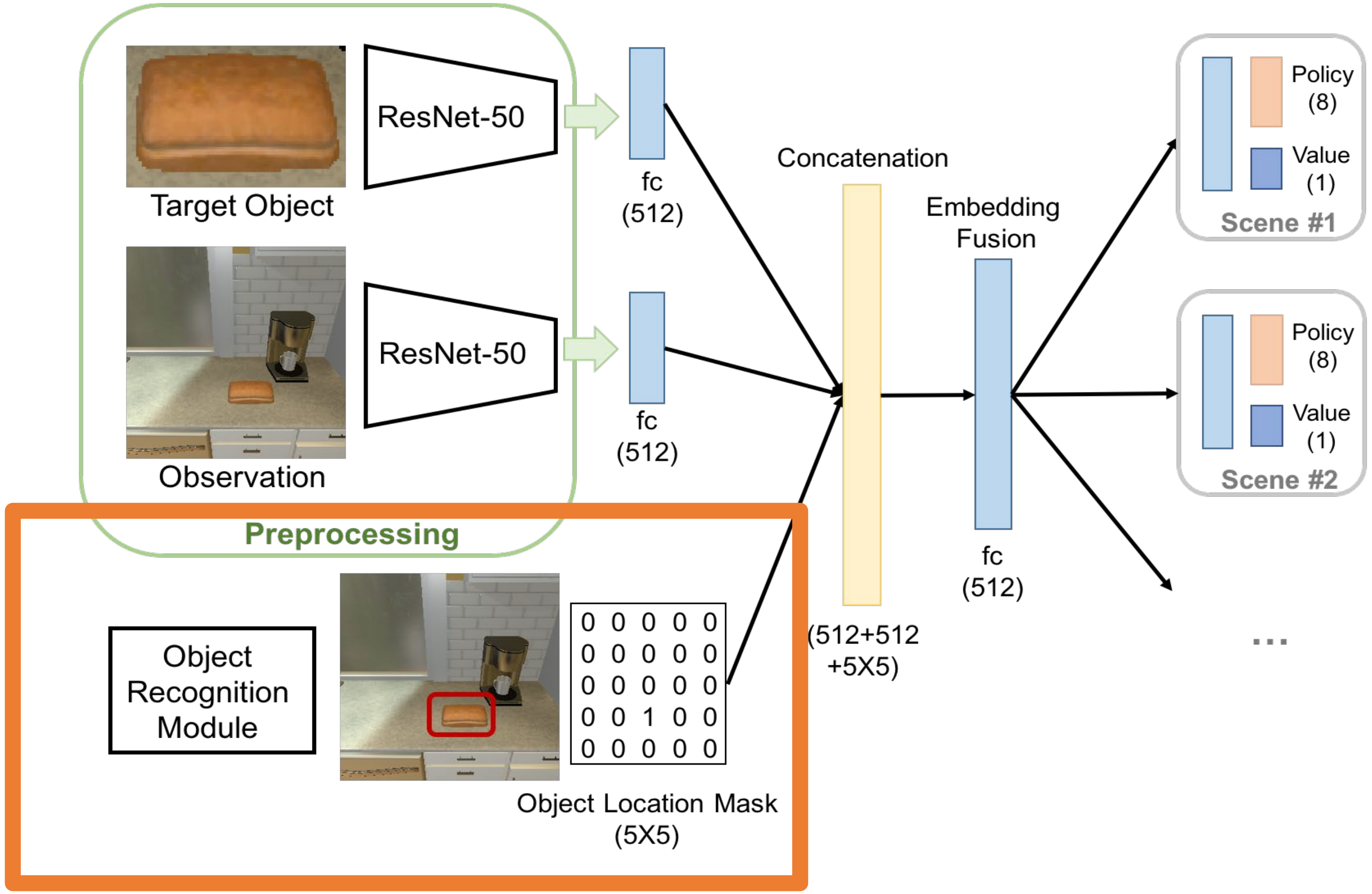
[2] Efficient Robotic Object Search via HIEM: Hierarchical Policy Learning with. Intrinsic-Extrinsic Modeling. RA-L & ICRA 2021

[3] GAPLE: Generalizable Approaching Policy LEarning for Robotic Object Searching in Indoor Environment. RA-L & IROS 2019.

[4] Hierarchical and Partially Observable Goal-driven Policy Learning with Goals Relational Graph. CVPR 2021, to appear.

Active Object Perceiver:

Recognition-guided Action Policy Learning



ASU APG

IROS 18'
Active Object
Perceiver

Reward Functions via Visual Understanding

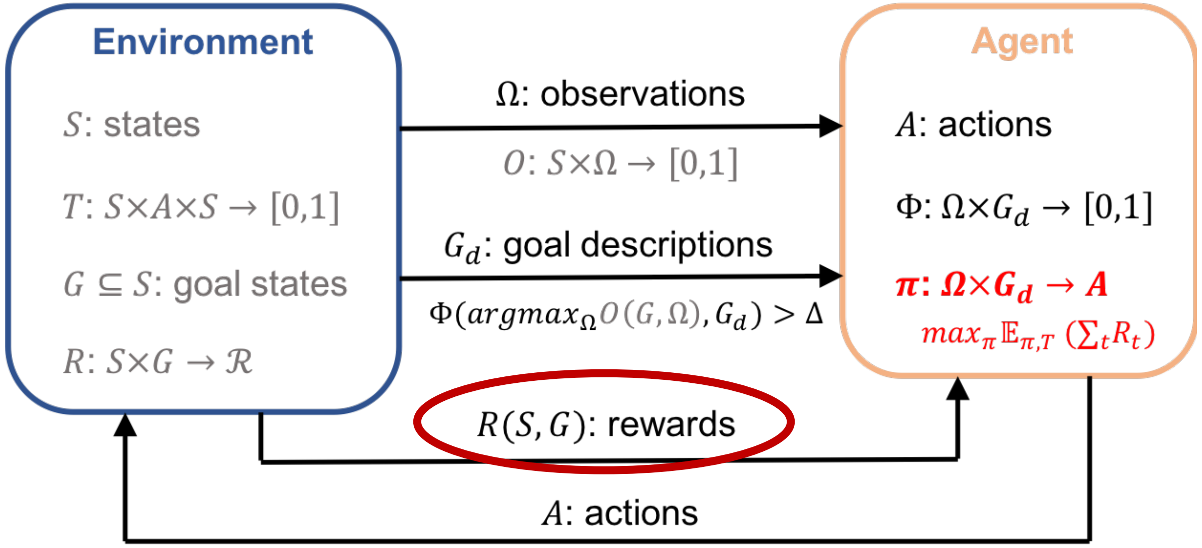
- Qualitative Examples

Reward Func. 2: the area of the target object bounding box

Ours

Vision-guided Policy Learning for Complex tasks

- How to define a good reward function?
  - ✓ Reward Functions via Visual Understanding [1]
- How to learn in a sparse reward setting?
  - ✓ Efficient Exploration with Hierarchical Policy [2]
- How to generalize across various instances?
  - ✓ Task-relevant Features from State Observations [3]
  - ✓ Goal Representation with Goals Relational Graph [4]
  - ➢ Data-efficient Neural-symbolic Modeling



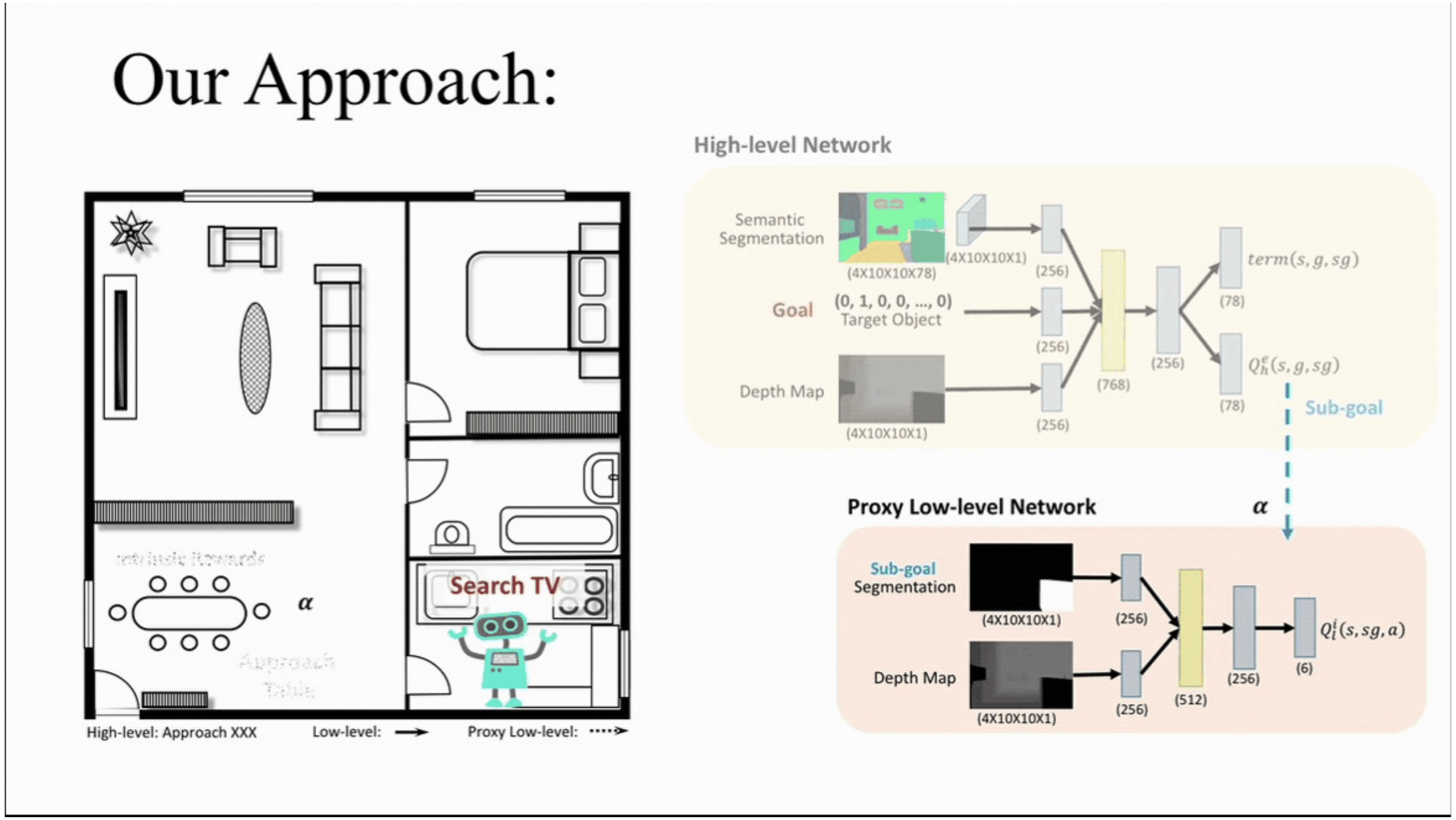
[1] Active Object Perceiver: Recognition-Guided Policy Learning for Object Searching on Mobile Robots. IROS 2018.

[2] Efficient Robotic Object Search via HIEM: Hierarchical Policy Learning with. Intrinsic-Extrinsic Modeling. RA-L & ICRA 2021

[3] GAPLE: Generalizable Approaching Policy LEarning for Robotic Object Searching in Indoor Environment. RA-L & IROS 2019.

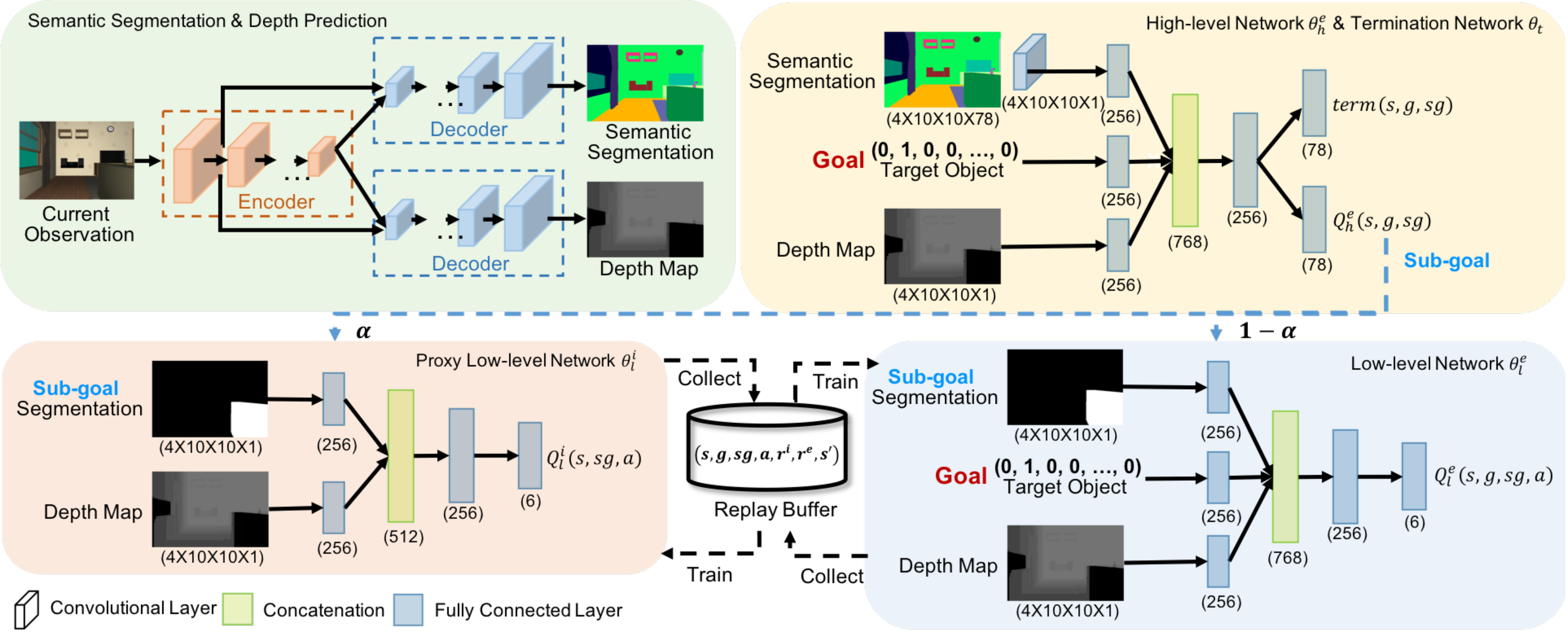
[4] Hierarchical and Partially Observable Goal-driven Policy Learning with Goals Relational Graph. CVPR 2021.

Hierarchical Policy Learning:



HIEM: Hierarchical Policy Learning:

RA-L ICRA 21
HIEM



Efficient Robotic Object Search via HIEM: Hierarchical Policy Learning with. Intrinsic-Extrinsic Modeling. RA-L & ICRA 2021

Low-level approaching policy:

GAPLE: Generalizable Approaching Policy LEarning for Robotic Object Searching in Indoor Environment. RA-L & IROS 2019.

Efficient Exploration with Hierarchical Policy

- Quantitative Results
  - Dataset: House3D*
  - Conclusions:
    - The intrinsic rewards help to explore.
    - Our intrinsic-extrinsic modeling tends to obtain a better performing policy.
    - Early termination of the non-optimal low-level policy is necessary.

Method	SR↑	AS / MS↓	SPL↑	AR↑
ORACLE	1.00	25.63 / 25.63	1.00	0.79
RANDOM	0.19	188.11 / 7.05	0.03	0.08
A3C	0.13	93.23 / 4.00	0.03	0.08
DQN	0.47	120.74 / 16.09	0.20	0.26
OC	0.14	99.29 / 5.14	0.06	0.09
H-DQN	0.74	182.15 / 23.62	0.17	0.23
Ours				
HIEM-proxy	0.40	95.08 / 15.03	0.12	0.22
HIEM-low	0.99	76.81 / 25.55	0.47	0.56
HIEM-term	**1.00**	49.42 / 25.63	0.65	0.66
HIEM	**1.00**	**41.18 / 25.63**	**0.72**	**0.70**

SR: Success Rate;
AS / MS: Average Steps / Minimal Steps over all successful cases;
SPL: Success weighted by inverse Path Length;
AR: Average discounted cumulative extrinsic Rewards.

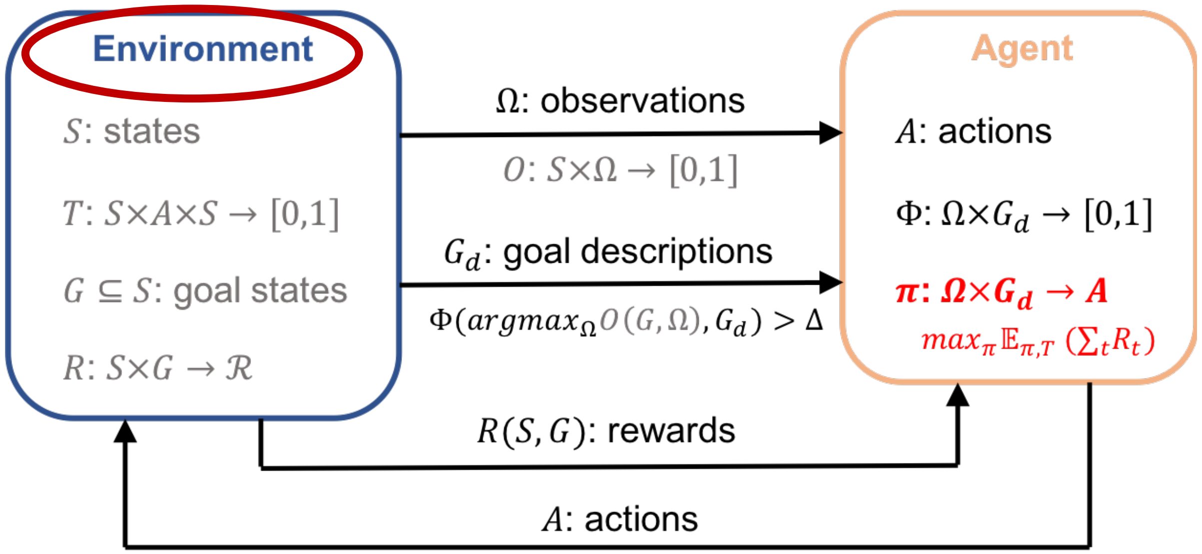
* Wu et al. Building generalizable agents with a realistic and rich 3d environment. arXiv 2018.

Efficient Exploration with Hierarchical Policy

- Qualitative Examples (Ours)

Vision-guided Policy Learning for Complex tasks

- How to define a good reward function?
  - ✓ Reward Functions via Visual Understanding [1]
- How to learn in a sparse reward setting?
  - ✓ Efficient Exploration with Hierarchical Policy [2]
- How to generalize across various instances?
  - ✓ Task-relevant Features from State Observations [3]
  - ✓ Goal Representation with Goals Relational Graph [4]
  - ➢ Data-efficient Neural-symbolic Modeling

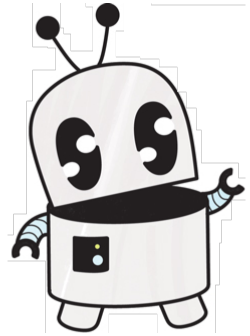


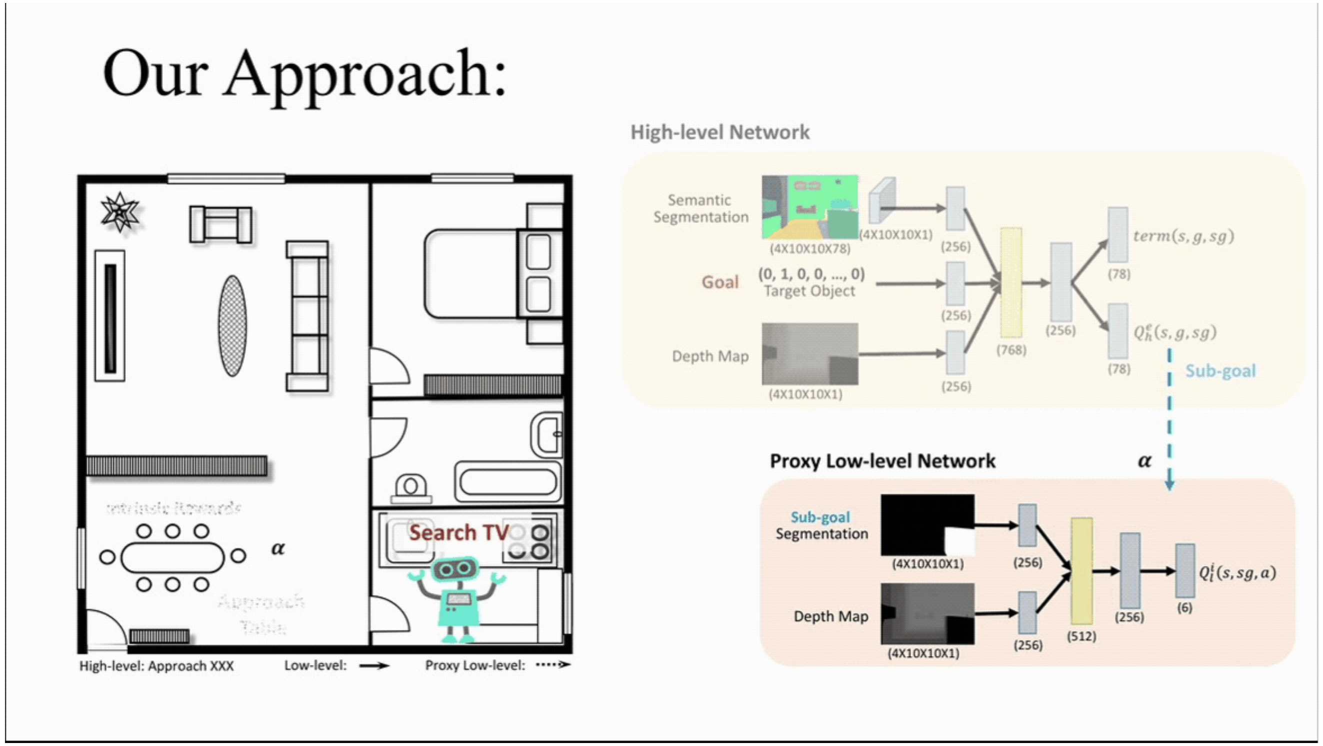
[1] Active Object Perceiver: Recognition-Guided Policy Learning for Object Searching on Mobile Robots. IROS 2018.

[2] Efficient Robotic Object Search via HIEM: Hierarchical Policy Learning with. Intrinsic-Extrinsic Modeling. RA-L & ICRA 2021, under review

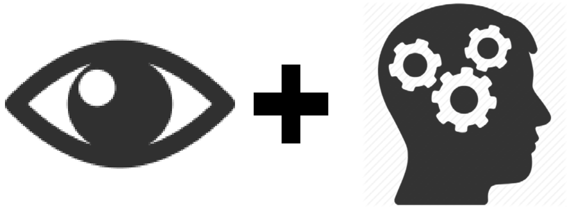
[3] GAPLE: Generalizable Approaching Policy LEarning for Robotic Object Searching in Indoor Environment. RA-L & IROS 2019.

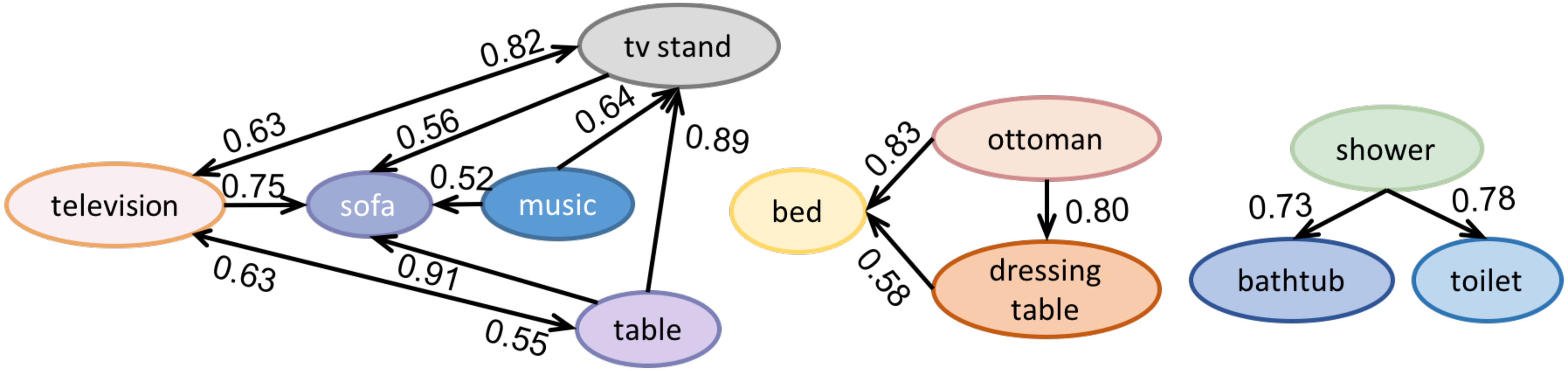
[4] Hierarchical and Partially Observable Goal-driven Policy Learning with Goals Relational Graph. CVPR 2021.





Our Approach:
High-level Network
Semantic Segmentation
(4X10X10X78)
(4X10X10X1)
Goal
(0, 1, 0, 0, ..., 0) Target Object
Depth Map
(4X10X10X1)
(256)
(768)
(256)
term(s, g, sg)
(78)
$Q_h^e(s, g, sg)$
(78)
Sub-goal
Proxy Low-level Network
α
Sub-goal Segmentation
(4X10X10X1)
Depth Map
(4X10X10X1)
(256)
(512)
(256)
(6)
$Q_l^i(s, sg, a)$
Search TV
α
High-level: Approach XXX
Low-level:
Proxy Low-level:

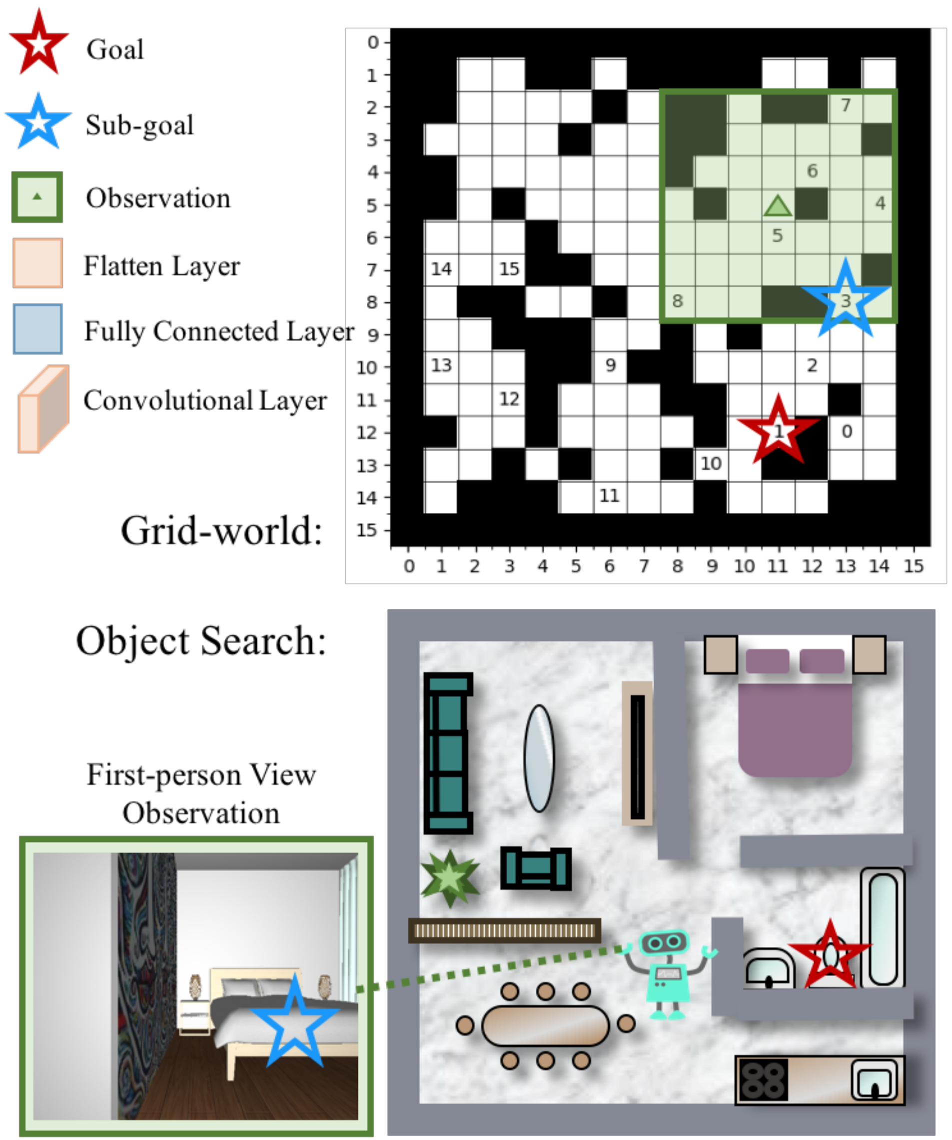


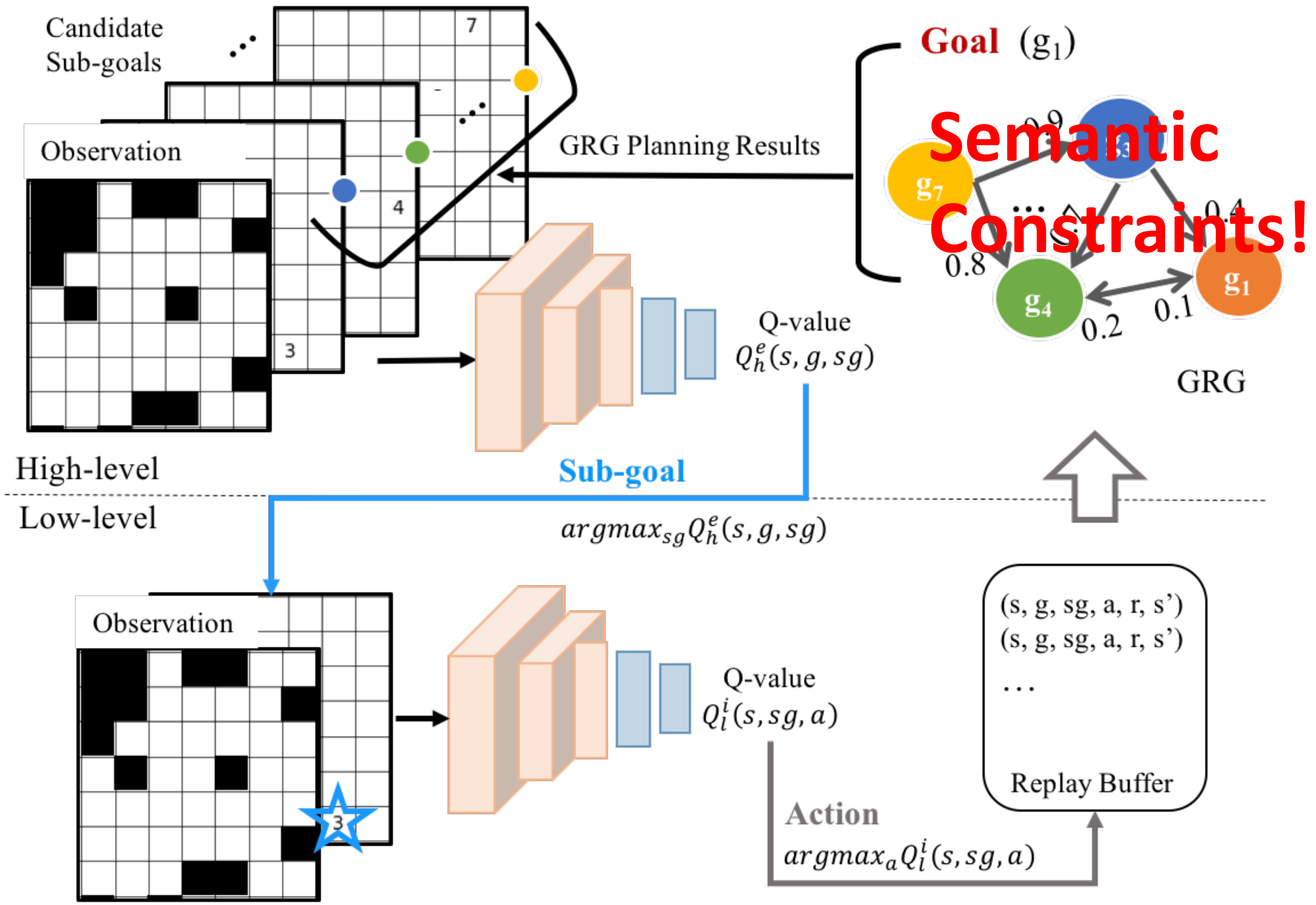


tv stand
television
sofa
music
table
0.82
0.63
0.56
0.64
0.89
0.75
0.52
0.63
0.91
0.55
bed
ottoman
dressing table
0.83
0.80
0.58
shower
bathtub
toilet
0.73
0.78

Hierarchical Policy Learning with Goal Relational Graphs (GRGs)

HRL-GRG
CVPR 21'





Hierarchical and Partially Observable Goal-driven Policy Learning with Goals Relational Graph. CVPR 2021, to appear.

State Representation: Unravelling the Unseen

- Goal Representation with Goals Relational Graph
  - Quantitative Results on Grid-world Domain (goals relations are pre-defined)

Generalize especially well towards unseen goals!

	Seen Goals			Unseen Goals			Overall		
Method	SR↑	AS / MS↓	SPL↑	SR↑	AS / MS↓	SPL↑	SR↑	AS / MS↓	SPL↑
ORACLE	1.00	11.81 / 11.81	1.00	1.00	11.28 / 11.28	1.00	1.00	10.38 / 10.38	1.00
RANDOM	0.16	42.15 / 5.47	0.03	0.15	42.38 / 4.81	0.04	0.18	36.62 / 4.69	0.05
DQN	0.20	20.28 / 5.47	0.13	0.20	11.90 / 4.10	0.15	0.32	16.23 / 5.71	0.23
H-DQN	0.43	**20.25 / 7.95**	0.28	0.19	26.09 / 6.38	0.08	0.45	20.84 / 7.16	0.26
Ours	**0.57**	28.71 / 9.03	**0.33**	**0.70**	**24.19 / 8.73**	**0.45**	**0.74**	**24.02 / 8.65**	**0.46**

The performance of all methods on the unseen gird-world maps.

SR: Success Rate
AS / MS: Average Steps / Minimal Steps over all successful cases
SPL: Success weighted by inverse Path Length

State Representation: Unravelling the Unseen

- Goal Representation with Goals Relational Graph
  - Quantitative Results for Robotic Object Search

Object Relations from Visual Genome
Yang et al. Visual semantic navigation using scene priors. ICLR 2019.

AI2THOR
Kolve et al. AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv 2017.

		Seen Goals		Unseen Goals	
		SR↑	SPL↑	SR↑	SPL↑
Seen Env.	[36]	**+0.49**	**+0.61**	+0.32	+0.23
	Ours	+0.37	+0.24	**+0.33**	+0.23
Unseen Env.	[36]	+0.21	+0.14	+0.24	+0.11
	Ours	**+0.33**	**+0.21**	**+0.38**	**+0.23**

+: Performance boost to the Random method
SR: Success Rate
SPL: Success weighted by inverse Path Length

House3D
Wu et al. Building generalizable agents with a realistic and rich 3d environment. arXiv 2018.

	Single Environment				Multiple Environments			
	Seen Goals		Unseen Goals		Seen Env.		Unseen Env.	
Method	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
RANDOM	0.20	0.05	0.23	0.04	0.39	0.03	0.60	0.05
DQN	0.58	0.27	0.18	0.05	0.42	0.06	0.39	0.04
A3C	0.53	0.18	0.27	0.09	0.48	0.03	0.47	0.03
HRL	0.77	0.15	0.05	0.00	0.43	0.05	0.28	0.02
Ours	**0.88**	**0.33**	**0.79**	**0.21**	**0.76**	**0.20**	**0.62**	**0.10**

State Representation: Unravelling the Unseen

- Goal Representation with Goals Relational Graph
  - Qualitative Results for Robotic Object Search (Unseen Environment Unseen Goal)

AI2THOR

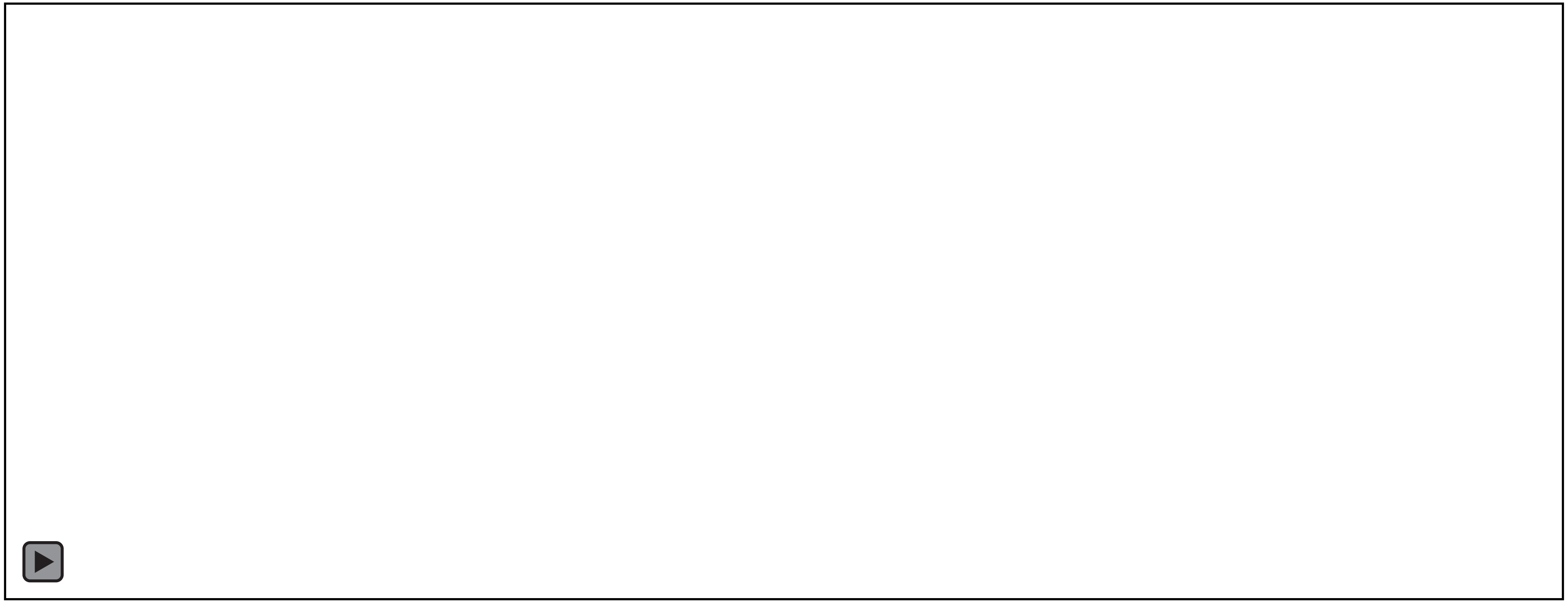
Kolve et al. AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv 2017.

State Representation: Unravelling the Unseen

- Goal Representation with Goals Relational Graph
  - Qualitative Results for Robotic Object Search (Unseen Environment Unseen Goal)

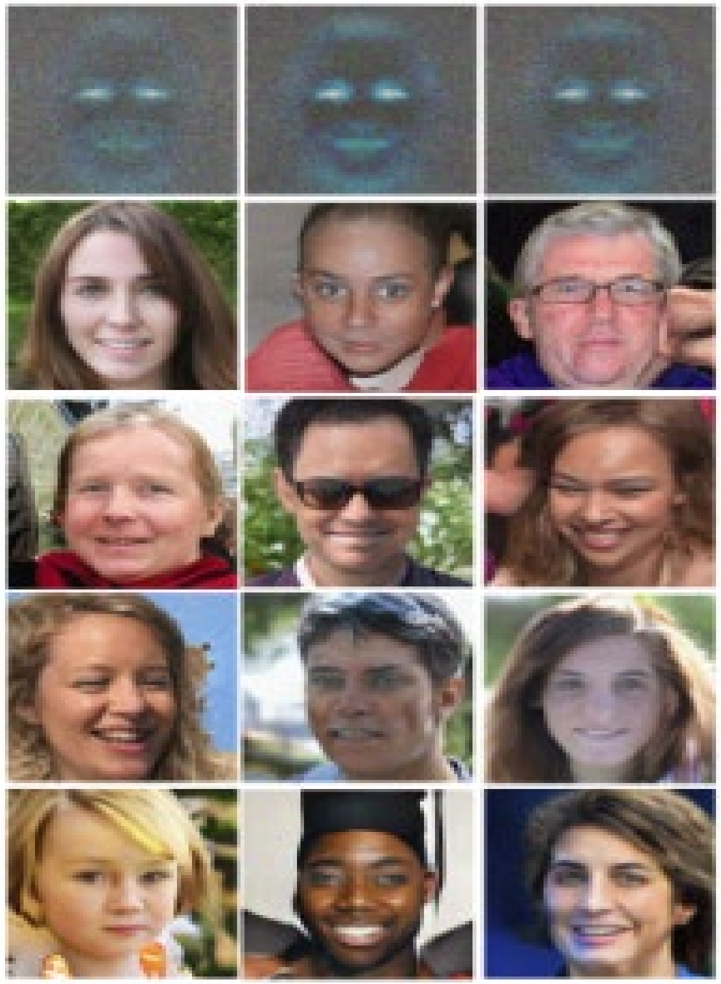
House3D

Wu et al. Building generalizable agents with a realistic and rich 3d environment. arXiv 2018.



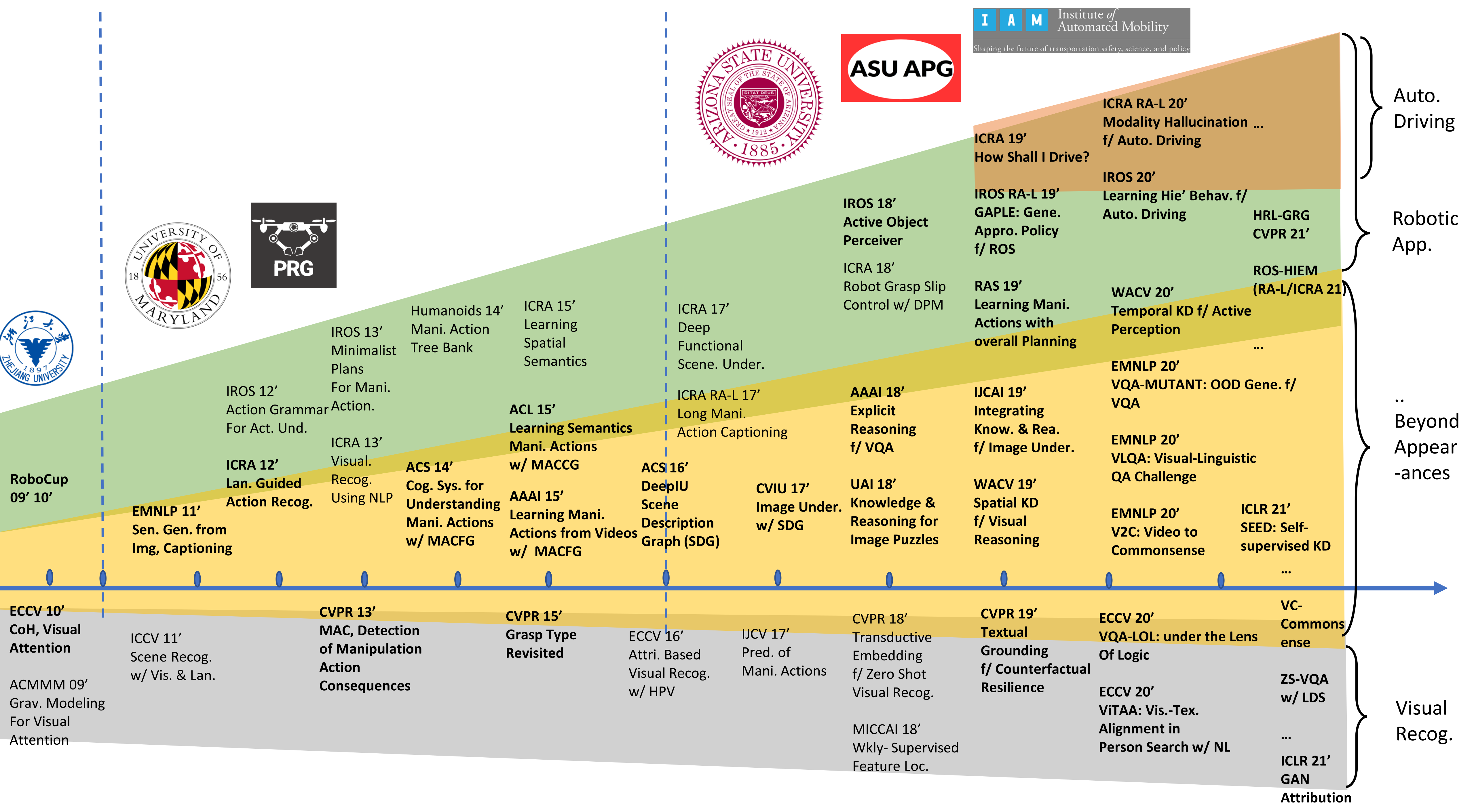
Model Attribution through Watermarking

- We studied the sufficient conditions of watermarks to guarantee attributability.
- I.e., with high probability, contents generated by one model will not be mistaken as by other models



[2]

Decentralized Attribution of Generative Models
Kim, Ren and **Yezhou Yang**. 2021
ArXiv: https://arxiv.org/pdf/2010.13974.pdf
To appear: ICLR 2021 (next talk)



Institute of Automated Mobility
Shaping the future of transportation safety, science, and policy
ASU APG
PRG
RoboCup 09' 10'
ECCV 10' CoH, Visual Attention
ACMMM 09' Grav. Modeling For Visual Attention
EMNLP 11' Sen. Gen. from Img, Captioning
ICCV 11' Scene Recog. w/ Vis. & Lan.
IROS 12' Action Grammar For Act. Und.
ICRA 12' Lan. Guided Action Recog.
IROS 13' Minimalist Plans For Mani. Action.
ICRA 13' Visual. Recog. Using NLP
CVPR 13' MAC, Detection of Manipulation Action Consequences
Humanoids 14' Mani. Action Tree Bank
ACS 14' Cog. Sys. for Understanding Mani. Actions w/ MACFG
ICRA 15' Learning Spatial Semantics
ACL 15' Learning Semantics Mani. Actions w/ MACCG
AAAI 15' Learning Mani. Actions from Videos w/ MACFG
CVPR 15' Grasp Type Revisited
ICRA 17' Deep Functional Scene. Under.
ICRA RA-L 17' Long Mani. Action Captioning
ACS 16' DeepIU Scene Description Graph (SDG)
ECCV 16' Attri. Based Visual Recog. w/ HPV
CVIU 17' Image Under. w/ SDG
IJCV 17' Pred. of Mani. Actions
IROS 18' Active Object Perceiver
ICRA 18' Robot Grasp Slip Control w/ DPM
AAAI 18' Explicit Reasoning f/ VQA
UAI 18' Knowledge & Reasoning for Image Puzzles
CVPR 18' Transductive Embedding f/ Zero Shot Visual Recog.
MICCAI 18' Wkly- Supervised Feature Loc.
ICRA 19' How Shall I Drive?
IROS RA-L 19' GAPLE: Gene. Appro. Policy f/ ROS
RAS 19' Learning Mani. Actions with overall Planning
IJCAI 19' Integrating Know. & Rea. f/ Image Under.
WACV 19' Spatial KD f/ Visual Reasoning
CVPR 19' Textual Grounding f/ Counterfactual Resilience
ICRA RA-L 20' Modality Hallucination f/ Auto. Driving ...
IROS 20' Learning Hie' Behav. f/ Auto. Driving
WACV 20' Temporal KD f/ Active Perception
...
EMNLP 20' VQA-MUTANT: OOD Gene. f/ VQA
EMNLP 20' VLQA: Visual-Linguistic QA Challenge
EMNLP 20' V2C: Video to Commonsense
ECCV 20' VQA-LOL: under the Lens Of Logic
ECCV 20' ViTAA: Vis.-Tex. Alignment in Person Search w/ NL
HRL-GRG CVPR 21'
ROS-HIEM (RA-L/ICRA 21)
ICLR 21' SEED: Self-supervised KD
...
VC-Commonsense
ZS-VQA w/ LDS
...
ICLR 21' GAN Attribution
Auto. Driving
Robotic App.
.. Beyond Appearances
Visual Recog.

Thank you and Acknowledgements

NSF CAREER 18' VR-K

NSF RI SMALL

NSF NRI

NSF CPS

NSF CCRI (planning)

NSF I-Corps

DARPA KAIROS
LESTAT project

ONR Social
Interaction

Machine Learning
Research Award 19'

Institute of Automated Mobility

Shaping the future of transportation safety, science, and policy

ASU APG

… …

and ASU close collaborating groups (Chitta Baral [KR & NLP], Max Yi Ren [Optimization & ML] …)

Moving forward…

- Moving towards a "post-dataset/simulator era"?
  - A. Efros, Imagining a post-dataset era, ICML'20 Plenary Talk.
- Breaking the vicious cycles of vision and language research (or even more general, all AI…) from a macro-historical view?

Exciting new challenge -> Performance saturation -> Repeating flaws identified (like language bias) -> Performance re-saturation again -> …

Captioning -> VQA -> VLN -> VCR…

Adversarial learning (AAAI 2021), self-supervised learning (ICLR 2021), and test-time adaptation, causal reasoning (@Damien Teney) might be (already have been shown to be) the ways to go.

Applying hypothetical actions and reason before and after an action is done?

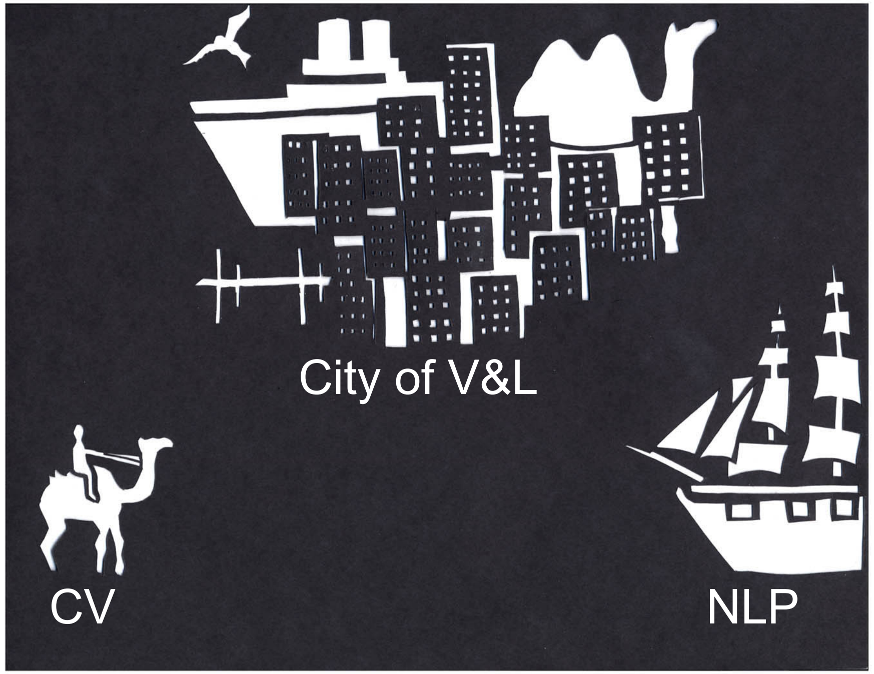
(CLEVR_HYP: A Dataset and Baselines for Visual Question Answering with Hypothetical Actions over Images, Shailaja Sampat, et. al. to appear NAACL-HLT 2021)

- Bridging Low-level Perception (such as Depth) with Visual Reasoning?
- Semantic Augmentation/modeling (LOL, MUTANT, GRGs, etc.) as a general tool set for new challenges, to pose novel semantic constraints? (many under review, maybe for the next talk?)

https://github.com/ASU-Active-Perception-Group/Robust-Vision-Language-Inference-using-STAT

	Category	Original	Transformed
Semantics-Inverting (SI)	Noun-Antonym	The two women are driving on the street with the convertible top down.	The two ***men*** are driving on the street with the convertible top down.
	Verb-Antonym	There are children standing by the door.	There are children **sitting** by the door.
	Comparative-Antonym	There are more monitors in the image on the right than on the left.	There are ***few*** monitors in the image on the right than on the left.
	Number-Substitution	There are three bowls of dough with only one spatula.	There are ***eleven*** bowls of dough with only one spatula.
	Pronoun-Substitution	In one of the images, a woman is taking a selfie.	In one of the images, ***he*** is taking a selfie.
	Subject-Object Swap	The two women are driving on the street with the convertible top down.	The two ***top*** are driving on the street with the convertible ***women*** down.
	Negation	The closet doors on the right are mirrored.	The closet doors on the right are ***not*** mirrored
Semantics-Preserving (SP)	Noun-Synonym	The right image shows three bottles of beer lined up.	The right ***picture*** shows three bottles of beer lined up.
	Verb-Synonym	Someone is using a kitchen utensil	Someone is ***utilizing*** a kitchen utensil.
	Comparative-Synonym	The bottle on the right is larger than the bottle on the left.	The bottle on the right is ***bigger*** than the bottle on the left.
	Number-Substitution	The two white swans are swimming in the canal gracefully.	The ***less than seven*** white swans are swimming in the canal gracefully.
	Pronoun-Substitution	In one of the images, a woman is taking a selfie.	In one of the images, ***she*** is taking a selfie.
	Paraphrasing	A man in a green shirt came on the porch and started knocking on the door.	A man in a green shirt came ***up to*** the porch and started knocking on the door.

Table 1. Illustrative examples for the effect of each of our 13 transformations on input sentences.



Email: yz.yang@asu.edu

Open for ZOOM chat. Shoot me an email.
Or on Twitter

@Yezhou_Yang

ASU APG