Domain-robust VQA with diverse datasets and methods but no target labels

Mingda Zhang Tristan Maidment Ahmad Diab
Adriana Kovashka Rebecca Hwa

University of Pittsburgh

University of Pittsburgh
CVPR VIRTUAL JUNE 19-25

Visual Question Answering Borders AI-Completeness

- Visual Question Answering (VQA) targets the intersection of computer vision and natural language understanding, thus receives much attention as an alternative form of Turing test.
- In the past decade, researchers have collected multiple datasets to enable relevant research, such as VQA v1/v2, VQA Abstract, GQA, VizWiz, COCO QA, Visual 7W, Visual Genome, etc.

What foods are placed on the table?

How many people will dine over there?

Does the curtain near the table have small size and white color?

What time is it?

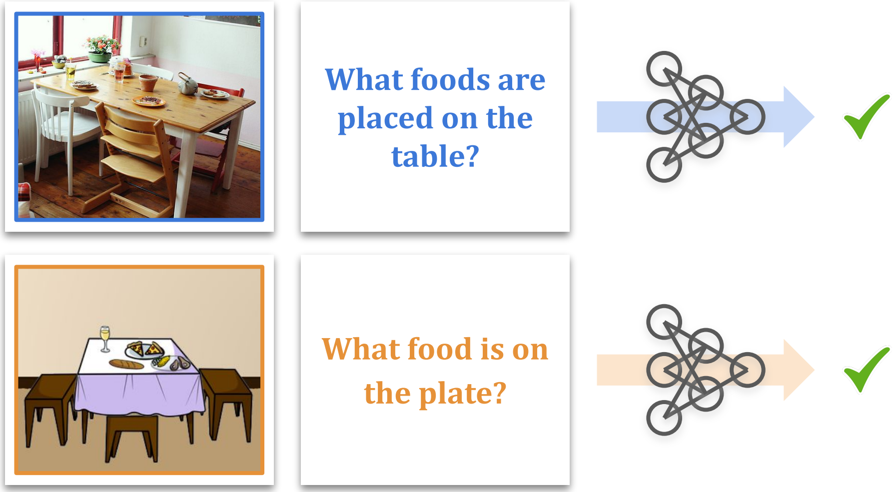
VQA Lacks Domain Robustness

❖ However, a widely known issue with VQA models is that they can easily overfit to the dataset bias, thus hard to generalize across datasets.

VQA Lacks Domain Robustness

- However, a widely known issue with VQA models is that they can easily overfit to the dataset bias, thus hard to generalize across datasets.

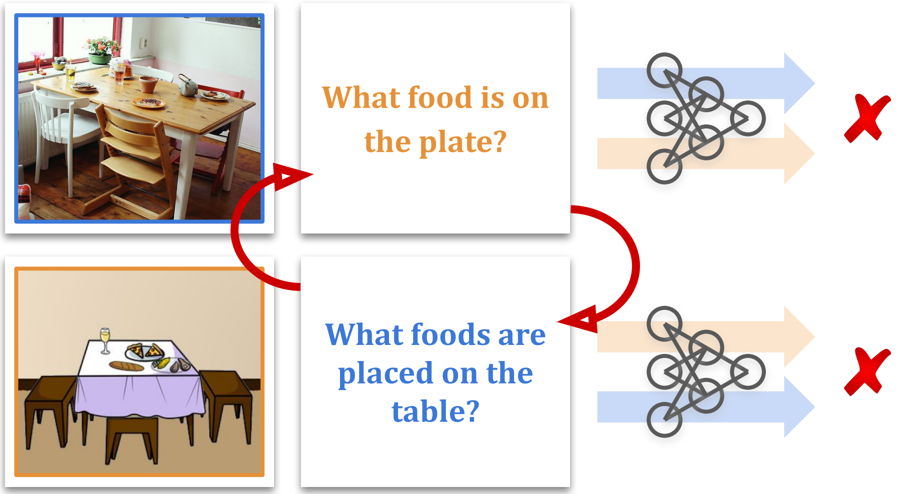
- For example, a model can usually achieve satisfactory performance when it is trained and evaluated on the same dataset, i.e., following similar data distribution.



VQA Lacks Domain Robustness

- However, a widely known issue with VQA models is that they can easily overfit to the dataset bias, thus hard to generalize across datasets.

- But when the model is applied to a different dataset (or even a single modality differs from the original data distribution), the VQA performance usually drops significantly although the desired knowledge is similar.

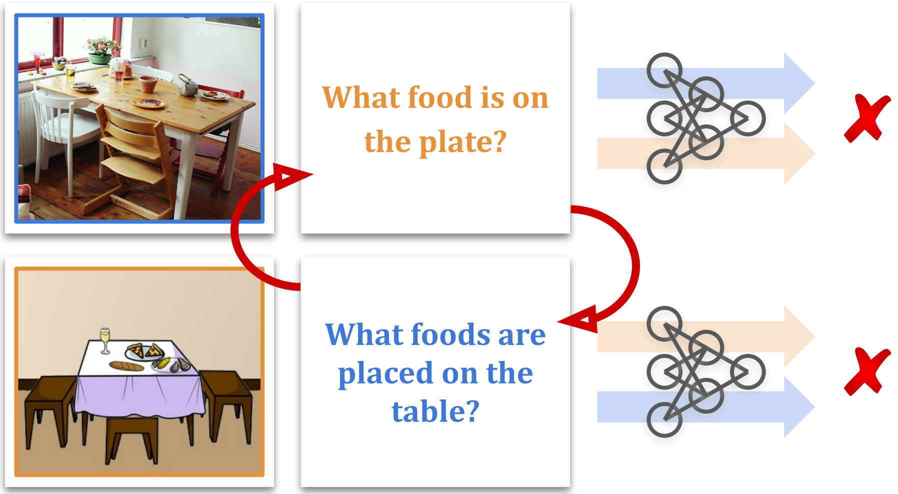


VQA Lacks Domain Robustness

- However, a widely known issue with VQA models is that they can easily overfit to the dataset bias, thus hard to generalize across datasets.

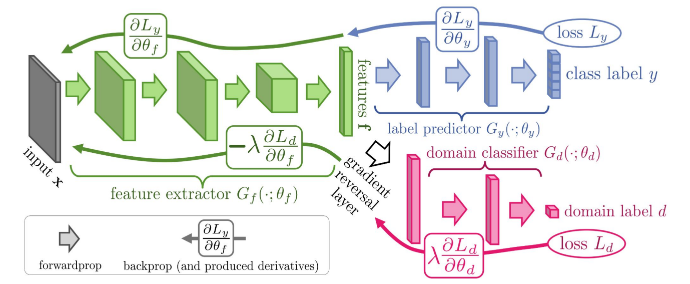
- But when the model is applied to a different dataset (or even a single modality differs from the original data distribution), the VQA performance usually drops significantly although the desired knowledge is similar.

- This issue is known as lack of domain robustness, and it prevents wider application of VQA systems.

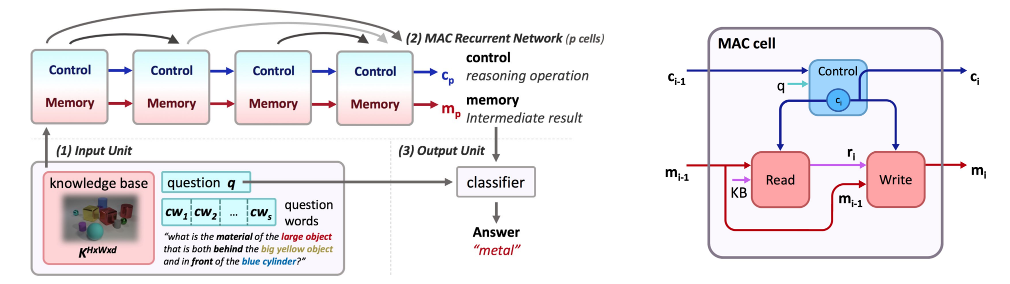


Domain Adaptation – Promising yet Non-trivial in VQA

- Domain adaptation is a popular technique to reduce domain gaps and improve domain robustness, and has been widely studied in visual tasks like object detection.
- However, directly applying domain adaptation techniques in VQA setting achieves limited success.



Domain Adversarial Neural Networks (DANN)



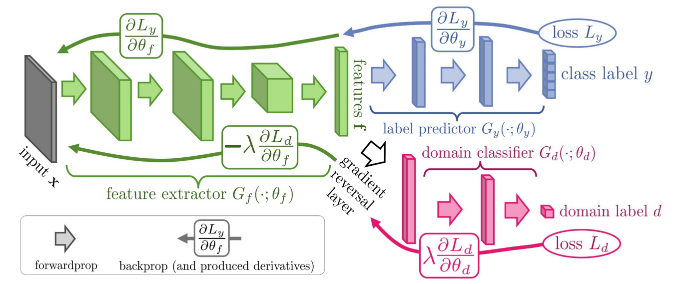
Memory, Attention and Composition Network (MAC)

Domain Adaptation – Promising yet Non-trivial in VQA

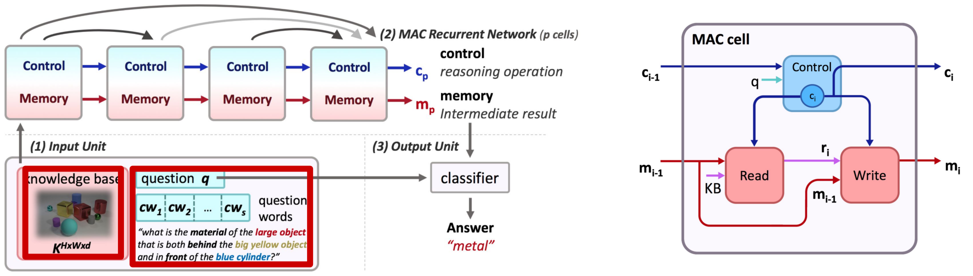
❖ Domain adaptation is a popular technique to reduce domain gaps and improve domain robustness, and has been widely studied in visual tasks like object detection.

❖ However, directly applying domain adaptation techniques in VQA setting achieves limited success.

➢ VQA models need to process multi-modal inputs, e.g., analyzing both image and question.



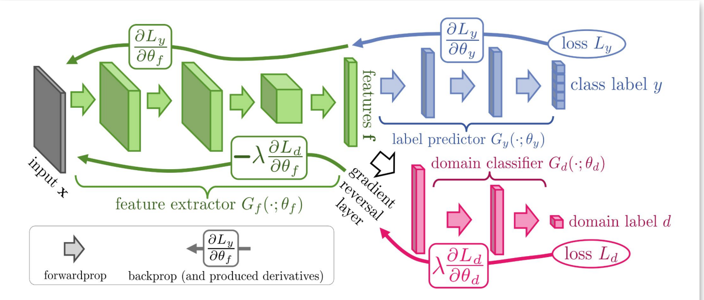
Domain Adversarial Neural Networks (DANN)



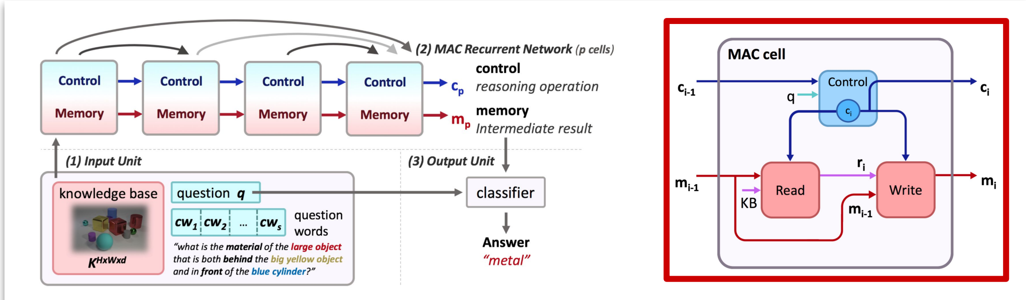
Memory, Attention and Composition Network (MAC)

Domain Adaptation – Promising yet Non-trivial in VQA

- Domain adaptation is a popular technique to reduce domain gaps and improve domain robustness, and has been widely studied in visual tasks like object detection.
- However, directly applying domain adaptation techniques in VQA setting achieves limited success.
  - VQA models need to process multi-modal inputs, e.g., analyzing both image and question.
  - Complicated reasoning modules, which capture the interaction across input modalities, are critical in VQA models but are not directly compatible with domain adaptation.

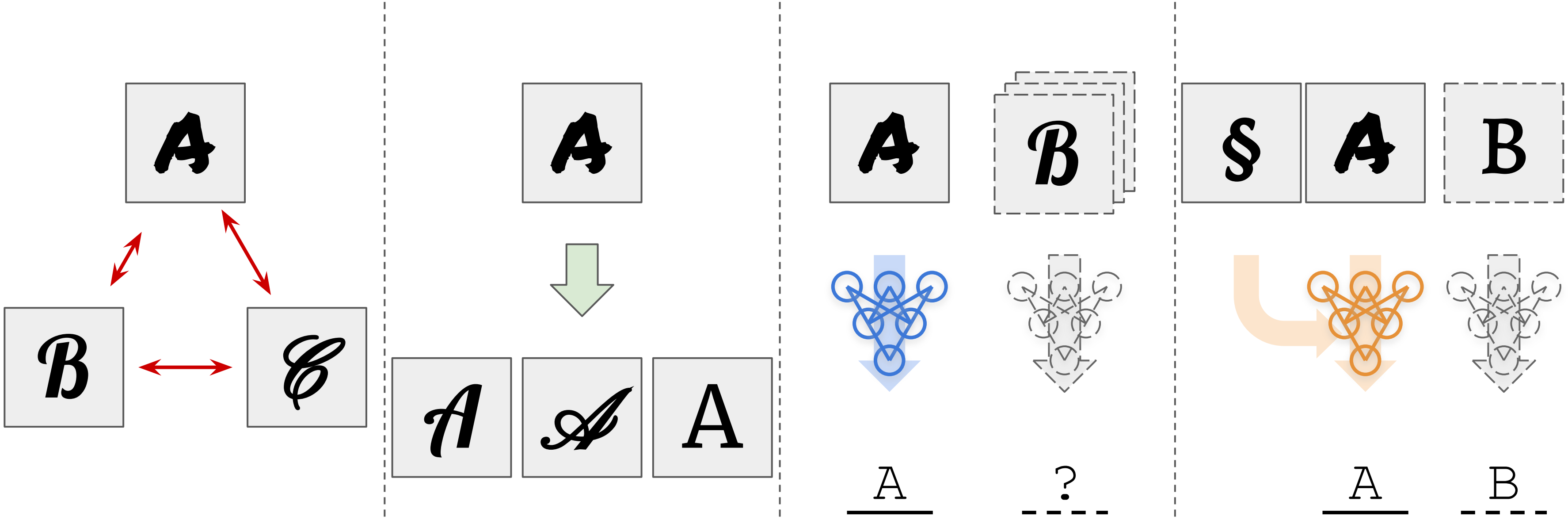


Domain Adversarial Neural Networks (DANN)



Memory, Attention and Composition Network (MAC)

Domain Robust Visual Question Answering



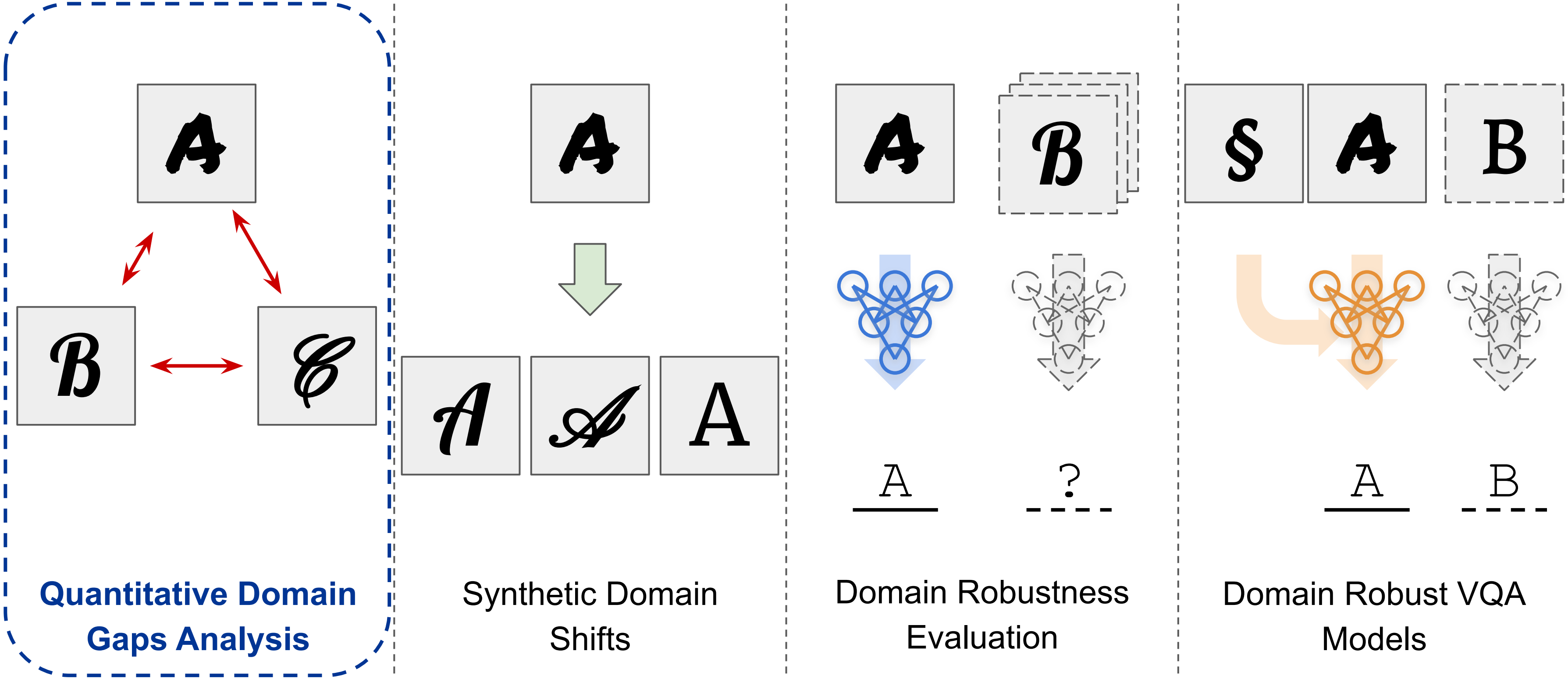
Quantitative Domain Gaps Analysis

Synthetic Domain Shifts

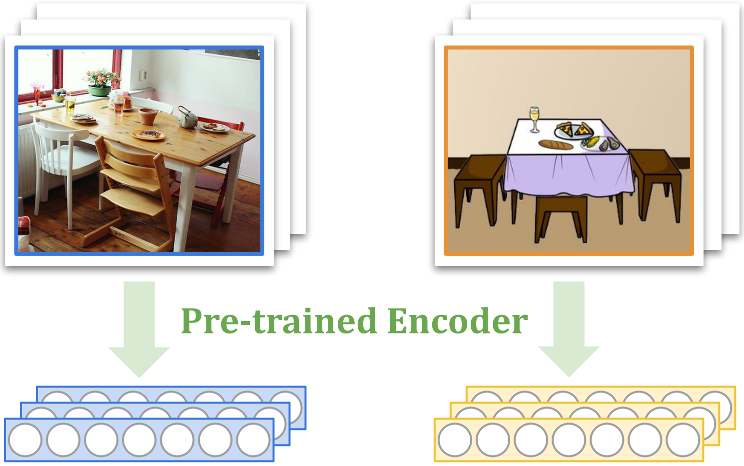
Domain Robustness Evaluation

Domain Robust VQA Models

Domain Robust Visual Question Answering



Domain Gaps in Real Datasets



$$\mathrm{MMD}(\mathcal{D}^S, \hat{\mathcal{D}}^T) = \|\mathbb{E}_{X\sim\mathcal{D}^S}[\varphi(X)] - \mathbb{E}_{Y\sim\hat{\mathcal{D}}^T}[\varphi(Y)]\|_{\mathcal{H}}$$
$$= \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} k(\mathbf{y}_i, \mathbf{y}_j)$$
$$- \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(\mathbf{x}_i, \mathbf{y}_j)$$

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.04	0.18	0.18	0.56	0.88	0.18	0.46	0.25
Vis. Genome	0.01	—	0.16	0.16	0.54	0.87	0.16	0.44	0.27
VQA v1	0.06	0.07	—	0.00	0.44	0.81	0.03	0.34	0.28
VQA v2	0.06	0.07	0.00	—	0.44	0.81	0.03	0.35	0.28
COCO QA	0.20	0.20	0.15	0.15	—	0.69	0.44	0.26	0.58
CLEVR	0.22	0.22	0.17	0.17	0.19	—	0.81	0.58	0.76
VQA Abstract	0.06	0.06	0.02	0.02	0.15	0.19	—	0.34	0.27
GQA	0.10	0.11	0.06	0.06	0.15	0.13	0.07	—	0.43
VizWiz	0.06	0.06	0.10	0.10	0.23	0.22	0.10	0.12	—

Domain Gaps in Real Datasets

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.00	0.01	0.01	0.01	0.10	0.08	0.00	0.04
Vis. Genome	0.01	—	0.00	0.01	0.01	0.10	0.08	0.00	0.04
VQA v1	0.02	0.02	—	0.00	0.00	0.10	0.08	0.01	0.04
VQA v2	0.03	0.02	0.01	—	0.00	0.10	0.08	0.01	0.03
COCO QA	0.04	0.04	0.03	0.03	—	0.10	0.08	0.01	0.03
CLEVR	0.54	0.54	0.54	0.54	0.54	—	0.10	0.10	0.09
VQA Abstract	0.36	0.36	0.36	0.36	0.36	0.59	—	0.08	0.08
GQA	0.03	0.03	0.03	0.03	0.04	0.54	0.36	—	0.04
VizWiz	0.22	0.22	0.21	0.21	0.21	0.52	0.42	0.22	—

Visual Semantic/Appearance Gap

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.04	0.18	0.18	0.56	0.88	0.18	0.46	0.25
Vis. Genome	0.01	—	0.16	0.16	0.54	0.87	0.16	0.44	0.27
VQA v1	0.06	0.07	—	0.00	0.44	0.81	0.03	0.34	0.28
VQA v2	0.06	0.07	0.00	—	0.44	0.81	0.03	0.35	0.28
COCO QA	0.20	0.20	0.15	0.15	—	0.69	0.44	0.26	0.58
CLEVR	0.22	0.22	0.17	0.17	0.19	—	0.81	0.58	0.76
VQA Abstract	0.06	0.06	0.02	0.02	0.15	0.19	—	0.34	0.27
GQA	0.10	0.11	0.06	0.06	0.15	0.13	0.07	—	0.43
VizWiz	0.06	0.06	0.10	0.10	0.23	0.22	0.10	0.12	—

Textual Semantic/Syntactic Gap

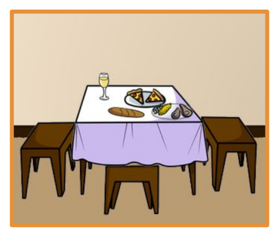
Domain Gaps in Real Datasets

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Vis. Genome	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
VQA v1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
VQA v2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
COCO QA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
CLEVR	0.54	0.54	0.54	0.54	0.54	—	0.10	0.10	0.09
VQA Abstract	0.36	0.36	**0.36**	**0.36**	0.36	0.59	—	0.08	0.08
GQA	0.03	0.03	0.03	0.03	0.04	0.54	0.36	—	0.04
VizWiz	0.22	0.22	0.21	0.21	0.21	0.52	0.42	0.22	—

Visual Semantic/Appearance Gap

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.25
Vis. Genome	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.27
VQA v1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.28
VQA v2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.28
COCO QA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.58
CLEVR	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.76
VQA Abstract	0.06	0.06	**0.02**	**0.02**	0.15	0.19	—	0.34	0.27
GQA	0.10	0.11	0.06	0.06	0.15	0.13	0.07	—	0.43
VizWiz	0.06	0.06	0.10	0.10	0.23	0.22	0.10	0.12	—

Textual Semantic/Syntactic Gap



What foods are placed on the table?

What food is on the plate?

❖ VQA v1/v2 and VQA Abstract have large visual distinctions, however the questions are more similar, probably because they are collected following the same data annotation protocol.

Domain Gaps in Real Datasets

What is next to the pond with green grass all around?

COCO QA

Is the person that is to the left of the ski wearing boots?

GQA

Does this chair look comfortable?

VQA v2

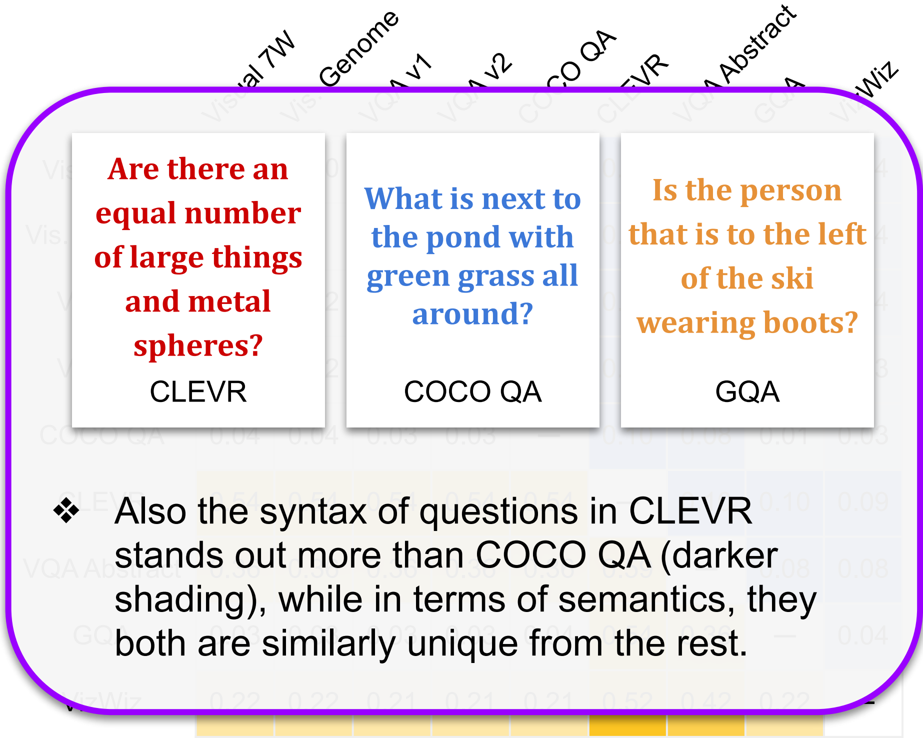
❖ Syntactically, COCO QA is most similar with GQA, as they both use automatic rules to generate questions, rather than relying on crowdsourcing.

Visual Semantic/Appearance Gap

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.04	0.18	0.18	0.56	0.88	0.18	0.46	0.25
Vis. Genome	0.01	—	0.16	0.16	0.54	0.87	0.16	0.44	0.27
VQA v1	0.06	0.07	—	0.00	0.44	0.81	0.03	0.34	0.28
VQA v2	0.06	0.07	0.00	—	0.44	0.81	0.03	0.35	0.28
COCO QA	0.20	0.20	0.15	0.15	—	0.69	0.44	**0.26**	0.58
CLEVR	0.22	0.22	0.17	0.17	0.19	—	0.81	0.58	0.76
VQA Abstract	0.06	0.06	0.02	0.02	0.15	0.19	—	0.34	0.27
GQA	0.10	0.11	0.06	0.06	0.15	0.13	0.07	—	0.43
VizWiz	0.06	0.06	0.10	0.10	0.23	0.22	0.10	0.12	—

Textual Semantic/Syntactic Gap

Domain Gaps in Real Datasets



❖ Also the syntax of questions in CLEVR stands out more than COCO QA (darker shading), while in terms of semantics, they both are similarly unique from the rest.

Visual Semantic/Appearance Gap

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.04	0.18	0.18	0.56	0.88	0.18	0.46	0.25
Vis. Genome	0.01	—	0.16	0.16	0.54	0.87	0.16	0.44	0.27
VQA v1	0.06	0.07	—	0.00	0.44	0.81	0.03	0.34	0.28
VQA v2	0.06	0.07	0.00	—	0.44	0.81	0.03	0.35	0.28
COCO QA	0.20	0.20	0.15	0.15	—	0.69	0.44	0.26	0.58
CLEVR	0.22	0.22	0.17	0.17	0.19	—	0.81	0.58	0.76
VQA Abstract	0.06	0.06	0.02	0.02	0.15	0.19	—	0.34	0.27
GQA	0.10	0.11	0.06	0.06	0.15	0.13	0.07	—	0.43
VizWiz	0.06	0.06	0.10	0.10	0.23	0.22	0.10	0.12	—

Textual Semantic/Syntactic Gap

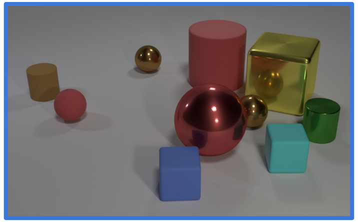
Domain Gaps in Real Datasets

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.00	0.01	0.01	0.01	0.10	0.08	0.00	0.04
Vis. Genome	0.01	—	0.00	0.01	0.01	0.10	0.08	0.00	0.04
VQA v1	0.02	0.02	—	0.00	0.00	0.10	0.08	0.01	0.04
VQA v2	0.03	0.02	0.01	—	0.00	0.10	0.08	0.01	0.03
COCO QA	0.04	0.04	0.03	0.03	—	0.10	0.08	0.01	0.03
CLEVR	0.54	0.54	0.54	0.54	0.54	—	0.10	0.10	0.09
VQA Abstract	0.36	0.36	0.36	0.36	0.36	0.59	—	0.08	0.08
GQA	0.03	0.03	0.03	0.03	0.04	**0.54**	0.36	—	0.04
VizWiz	0.22	0.22	0.21	0.21	0.21	0.52	0.42	0.22	—

Visual Semantic/Appearance Gap

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.04	0.18	0.18	0.56	0.88	0.18	0.46	0.25

Textual Semantic/Syntactic Gap



- On the image domain, GQA is vastly different from CLEVR as GQA features real-world images while CLEVR is built with synthetic images.

Domain Gaps in Real Datasets

What food is in the bowl the chopsticks are to the right of?

What size is the cylinder that is left of the brown metal thing that is left of the big sphere?

- For questions, since both GQA and CLEVR are designed to investigate compositionality, the gaps are greatly reduced to a similar level with GQA and Visual 7W/Visual Genome.

Visual Semantic/Appearance Gap

	Visual 7W	Vis. Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.04	0.18	0.18	0.56	0.88	0.18	0.46	0.25
Vis. Genome	0.01	—	0.16	0.16	0.54	0.87	0.16	0.44	0.27
VQA v1	0.06	0.07	—	0.00	0.44	0.81	0.03	0.34	0.28
VQA v2	0.06	0.07	0.00	—	0.44	0.81	0.03	0.35	0.28
COCO QA	0.20	0.20	0.15	0.15	—	0.69	0.44	0.26	0.58
CLEVR	0.22	0.22	0.17	0.17	0.19	—	0.81	0.58	0.76
VQA Abstract	0.06	0.06	0.02	0.02	0.15	0.19	—	0.34	0.27
GQA	0.10	0.11	0.06	0.06	0.15	**0.13**	0.07	—	0.43
VizWiz	0.06	0.06	0.10	0.10	0.23	0.22	0.10	0.12	—

Textual Semantic/Syntactic Gap

Domain Robust Visual Question Answering

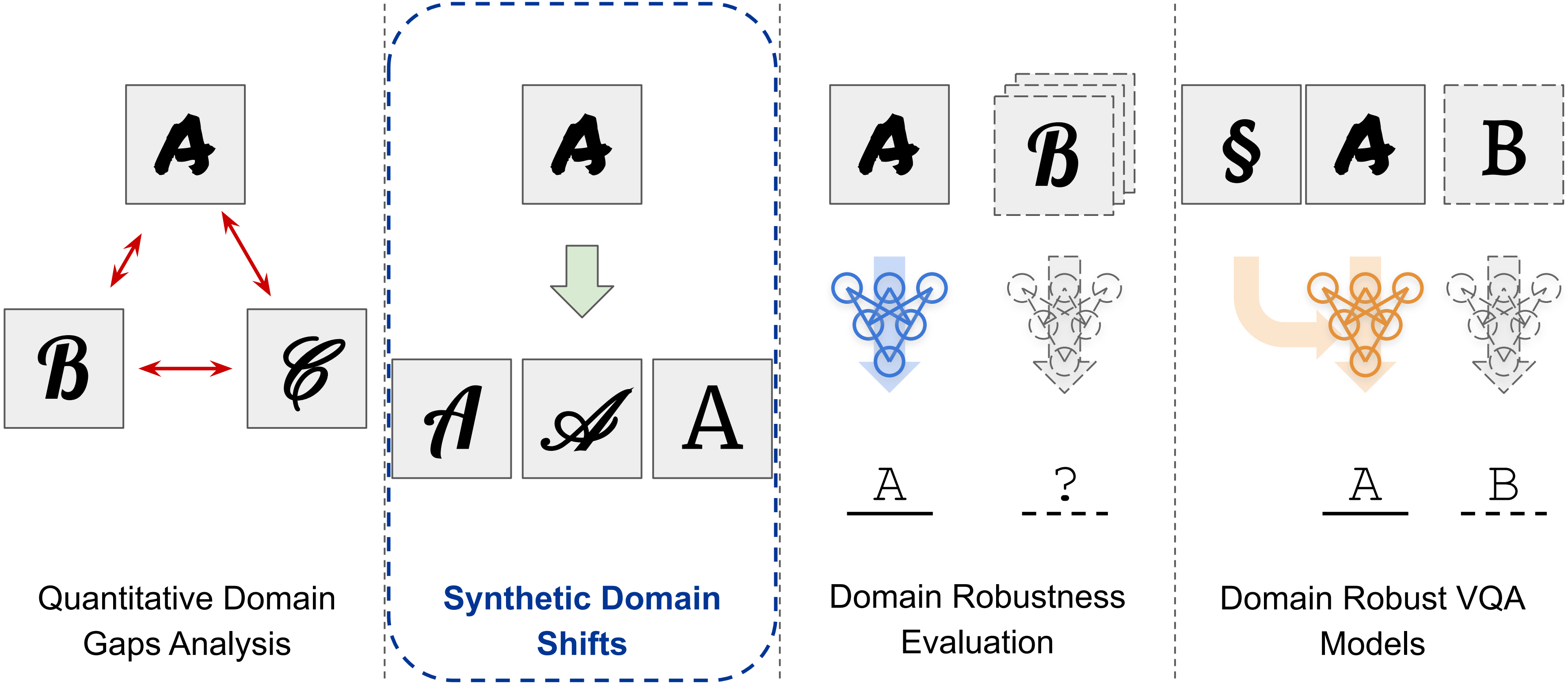
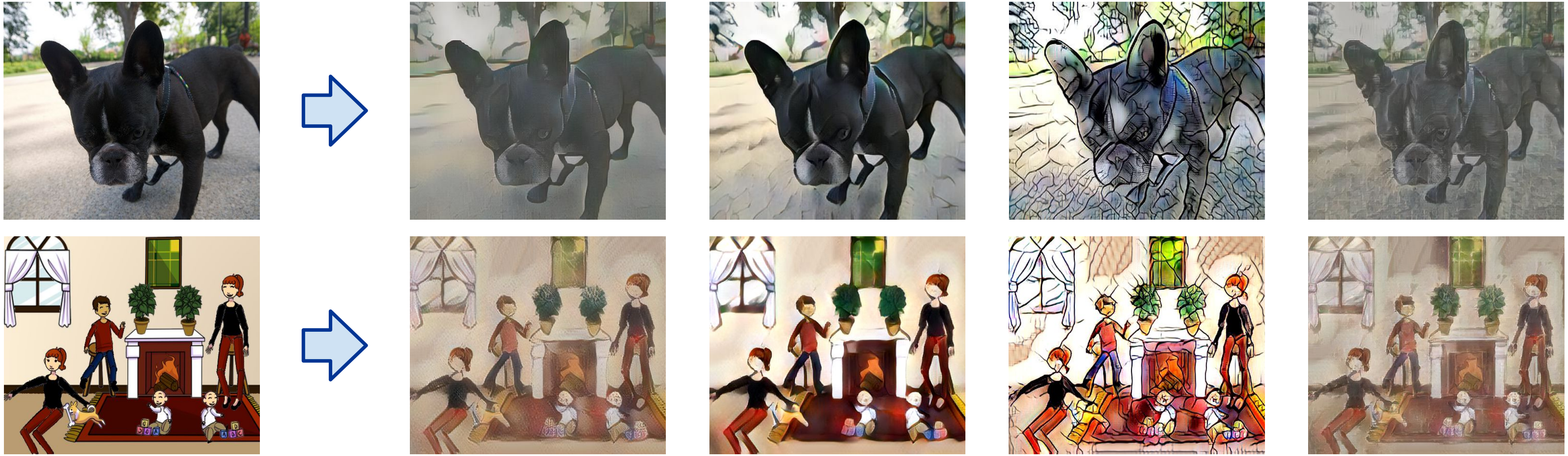


Image Style Transfer

- As we have seen, most datasets have domain gaps in both image and text space, thus it is very difficult to explain the performance degradation.

Image Style Transfer

- As we have seen, most datasets have domain gaps in both image and text space, thus it is very difficult to explain the performance degradation.
- By generating an artistic illustration while preserving the contents of original image, image style transfer provides an opportunity to control the visual shifts under appearance level only.

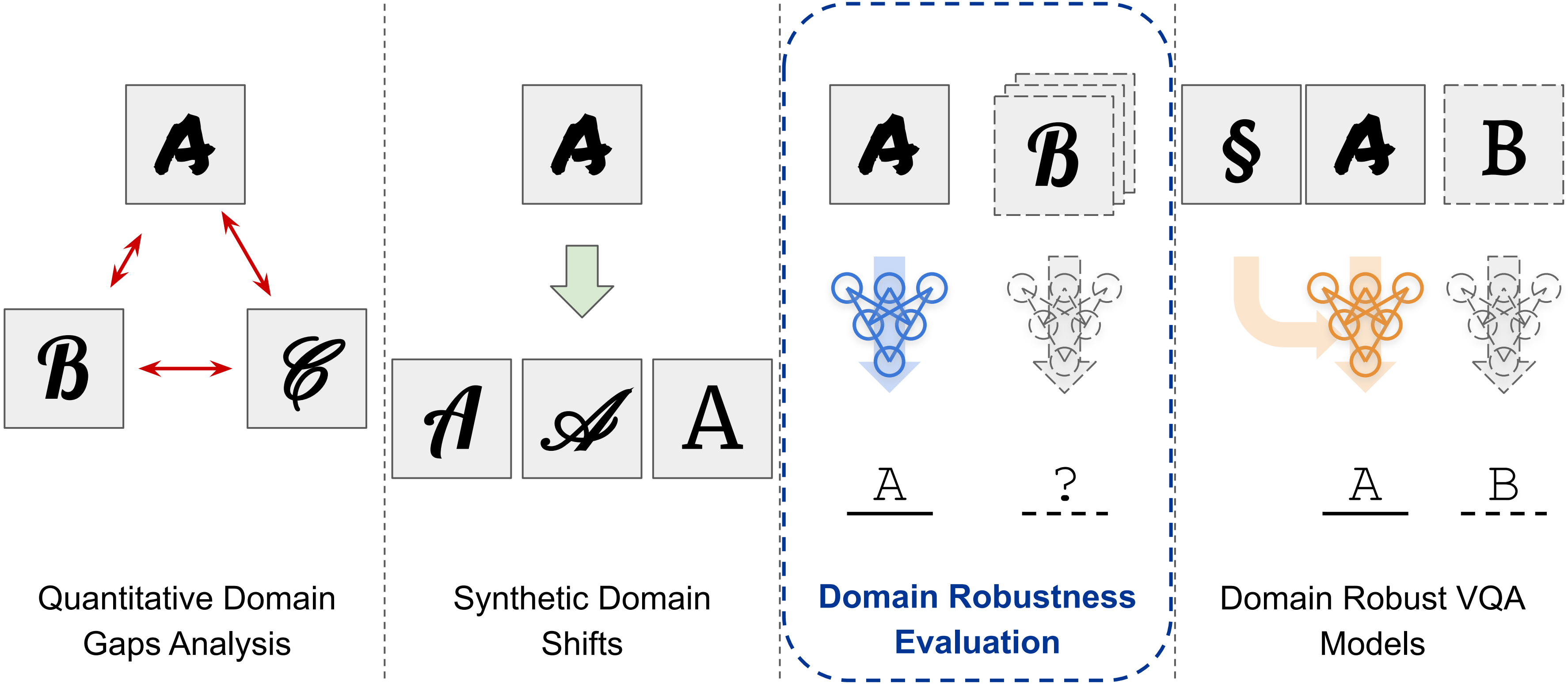


Paraphrase Generation

❖ Similarly, sequence-to-sequence language generation models can be used to create paraphrases, which keep the core message of question but in a different writing style.

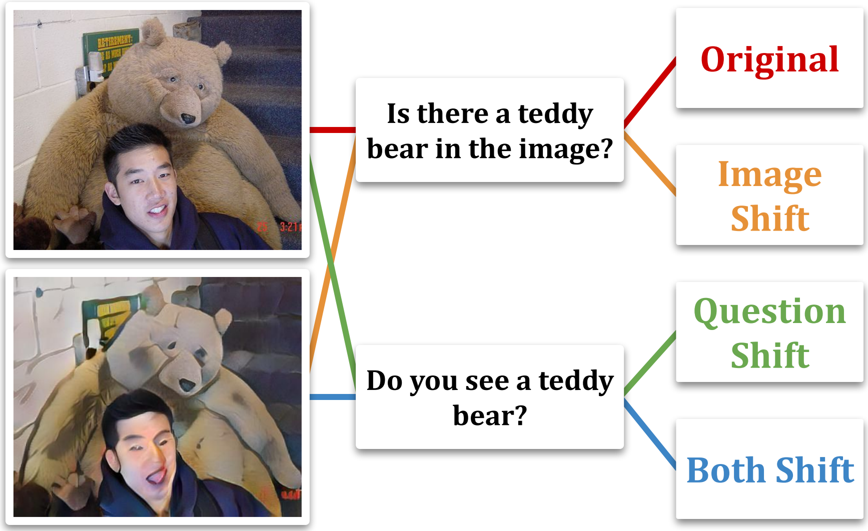
Original Question	Paraphrased Question
What is the weather like?	What weather is it?
What is written on the white square on the bus?	What does the white square say on the bus?
What shape is the bench seat?	What is the shape of the bench?
What number of red spheres are behind the shiny object that is on the left of the tiny matte cylinder?	How many red spheres are behind the shiny object on the left side of the small dull cylinder?

Domain Robust Visual Question Answering



Create Datasets with Synthetic Domain Shifts

- By composing style transferred images and paraphrased questions, we manually create datasets with synthetic domain shifts in individual modality.
- We test most recent mainstream VQA models on the precisely controlled testbed to better analyze the sensitivity of different models regarding domain shifts.



Methods	Original	Image Shift	Question Shift	Both Shift
NSCL (NS)	98.0	**71.0**	—	—
MAC (NS/CL)	93.4	45.9	52.2	28.1
TbD (NS/CL)	**99.1**	55.7	**52.9**	**36.1**
RelNet (CL)	93.7	20.5	49.6	19.1
LXMERT (TR)	94.8	50.6	**53.4**	**36.6**

Domain Robustness on Real Datasets

- We test MAC (NS/CL) and LXMERT (TR) on real datasets, and find the performance degradation generally aligns with our analysis on domain gaps.
- The shading in last two columns represents normalized transfer effectiveness, the darker the better.

Datasets		Image		Question	
A	B	App.	Sem.	Syn.	Sem.
VQA v2	CLEVR	High (0.10)	High (0.54)	High (0.81)	High (0.17)
	GQA	Low (0.01)	Low (0.03)	Med H (0.35)	Medium (0.06)
	VQA Abstract	Med H (0.08)	Med H (0.36)	Low (0.03)	Low (0.02)
	Visual Genome	Low (0.01)	Low (0.02)	Med L (0.16)	Medium (0.07)

		Accuracy (%)			
	B	A → A	B → B	A → B	B → A
MAC (NS/CL)	CLEVR	53.3	95.9	29.8	18.7
	GQA		44.4	32.0	35.6
	VQA Abs		48.3	33.6	31.7
	VG		33.3	26.2	23.1
LXMERT (TR)	CLEVR	67.6	84.9	31.6	34.8
	GQA		58.2	50.5	51.5
	VQA Abs		56.3	34.3	34.6
	VG		41.0	36.7	31.4

Domain Robustness on Real Datasets

❖ The large domain gaps between VQA v2 and CLEVR make it very difficult for both models to successfully transfer knowledge from one to the other.

Datasets		Image		Question	
A	B	App.	Sem.	Syn.	Sem.
VQA v2	CLEVR	High (0.10)	High (0.54)	High (0.81)	High (0.17)
	GQA	Low (0.01)	Low (0.03)	Med H (0.35)	Medium (0.06)
	VQA Abstract	Med H (0.08)	Med H (0.36)	Low (0.03)	Low (0.02)
	Visual Genome	Low (0.01)	Low (0.02)	Med L (0.16)	Medium (0.07)

		Accuracy (%)			
	B	A → A	B → B	A → B	B → A
MAC (NS/CL)	CLEVR	53.3	95.9	29.8	18.7
	GQA		44.4	32.0	35.6
	VQA Abs		48.3	33.6	31.7
	VG		33.3	26.2	23.1
LXMERT (TR)	CLEVR	67.6	84.9	31.6	34.8
	GQA		58.2	50.5	51.5
	VQA Abs		56.3	34.3	34.6
	VG		41.0	36.7	31.4

Domain Robustness on Real Datasets

❖ We test MAC (NS/CL) and LXMERT (TR) on real datasets, and find the performance degradation generally aligns with our analysis on domain gaps.

❖ From VQA v2 ↔ VQA Abstract and VQA v2 ↔ Visual Genome, we can see that both MAC and LXMERT are more sensitive to image domain gaps, as the knowledge transfer is more effective when image domain gaps is small, even if there exists relatively large question domain shifts.

Datasets		Image		Question	
A	B	App.	Sem.	Syn.	Sem.
VQA v2	CLEVR	High (0.10)	High (0.54)	High (0.81)	High (0.17)
	GQA	Low (0.01)	Low (0.03)	Med H (0.35)	Medium (0.06)
	VQA Abstract	Med H (0.08)	Med H (0.36)	Low (0.03)	Low (0.02)
	Visual Genome	Low (0.01)	Low (0.02)	Med L (0.16)	Medium (0.07)

		Accuracy (%)			
	B	A → A	B → B	A → B	B → A
MAC (NS/CL)	CLEVR	53.3	95.9	29.8	18.7
	GQA		44.4	32.0	35.6
	VQA Abs		48.3	33.6	31.7
	VG		33.3	26.2	23.1
LXMERT (TR)	CLEVR	67.6	84.9	31.6	34.8
	GQA		58.2	50.5	51.5
	VQA Abs		56.3	34.3	34.6
	VG		41.0	36.7	31.4

Domain Robustness on Real Datasets

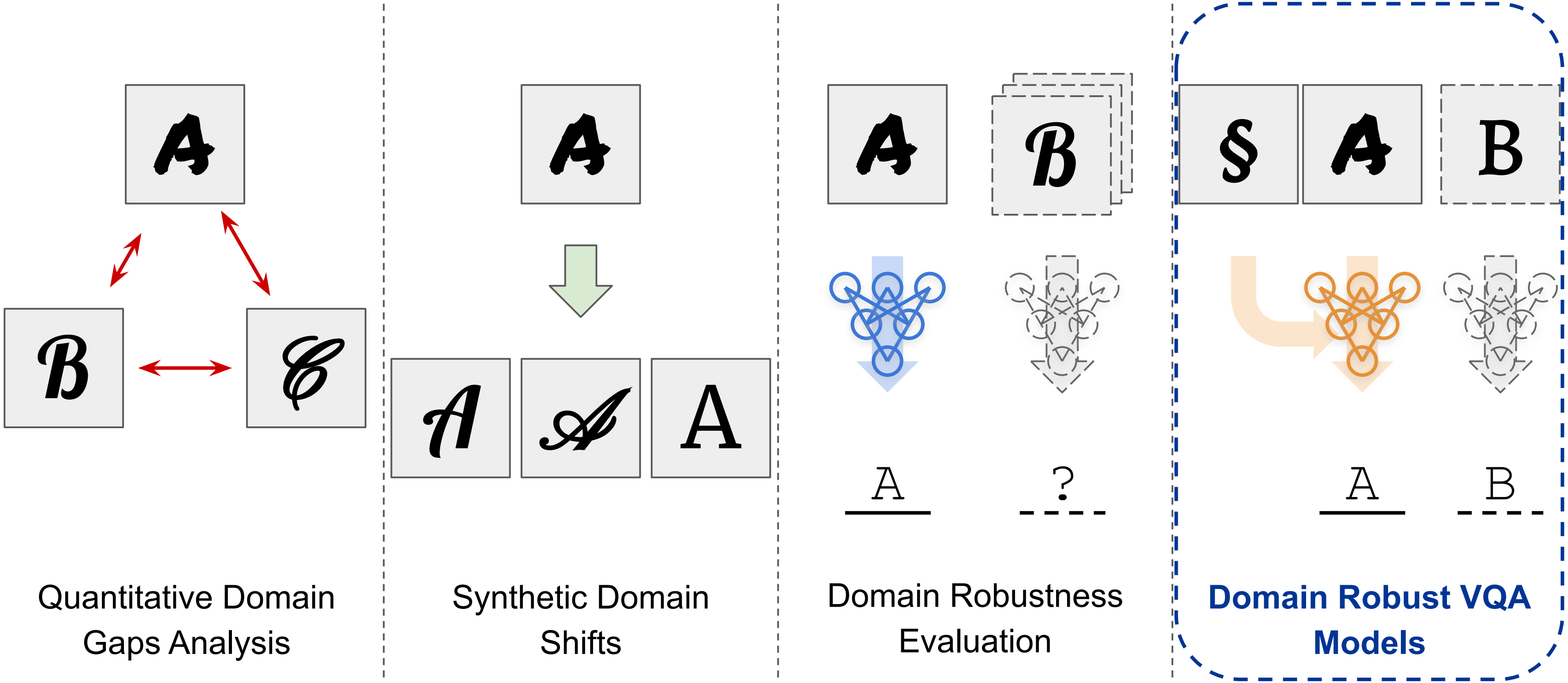
❖ We test MAC (NS/CL) and LXMERT (TR) on real datasets, and find the performance degradation generally aligns with our analysis on domain gaps.

❖ By comparing VQA v2 ↔ GQA and VQA v2 ↔ Visual Genome, we also see that syntactic domain shifts are less troublesome for LXMERT model compared to MAC, probably because the pre-trained question encoder in LXMERT is more robust against linguistic variations.

Datasets		Image		Question	
A	B	App.	Sem.	Syn.	Sem.
VQA v2	CLEVR	High (0.10)	High (0.54)	High (0.81)	High (0.17)
	GQA	Low (0.01)	Low (0.03)	Med H (0.35)	Medium (0.06)
	VQA Abstract	Med H (0.08)	Med H (0.36)	Low (0.03)	Low (0.02)
	Visual Genome	Low (0.01)	Low (0.02)	Med L (0.16)	Medium (0.07)

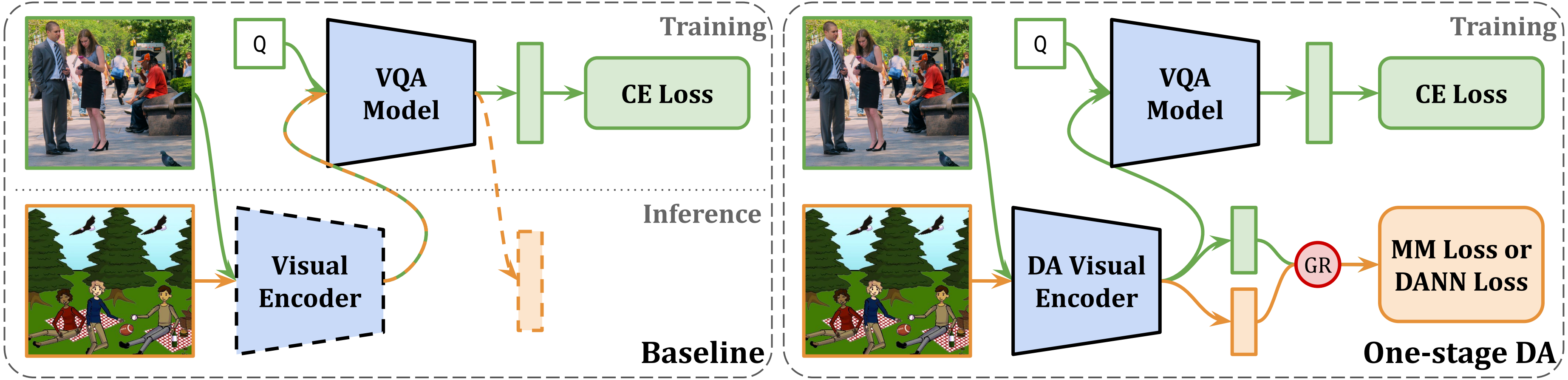
		Accuracy (%)			
	B	A → A	B → B	A → B	B → A
MAC (NS/CL)	CLEVR	53.3	95.9	29.8	18.7
	GQA		44.4	32.0	35.6
	VQA Abs		48.3	33.6	31.7
	VG		33.3	26.2	23.1
LXMERT (TR)	CLEVR	67.6	84.9	31.6	34.8
	GQA		58.2	50.5	51.5
	VQA Abs		56.3	34.3	34.6
	VG		41.0	36.7	31.4

Domain Robust Visual Question Answering



Introducing Domain Adaptation to VQA

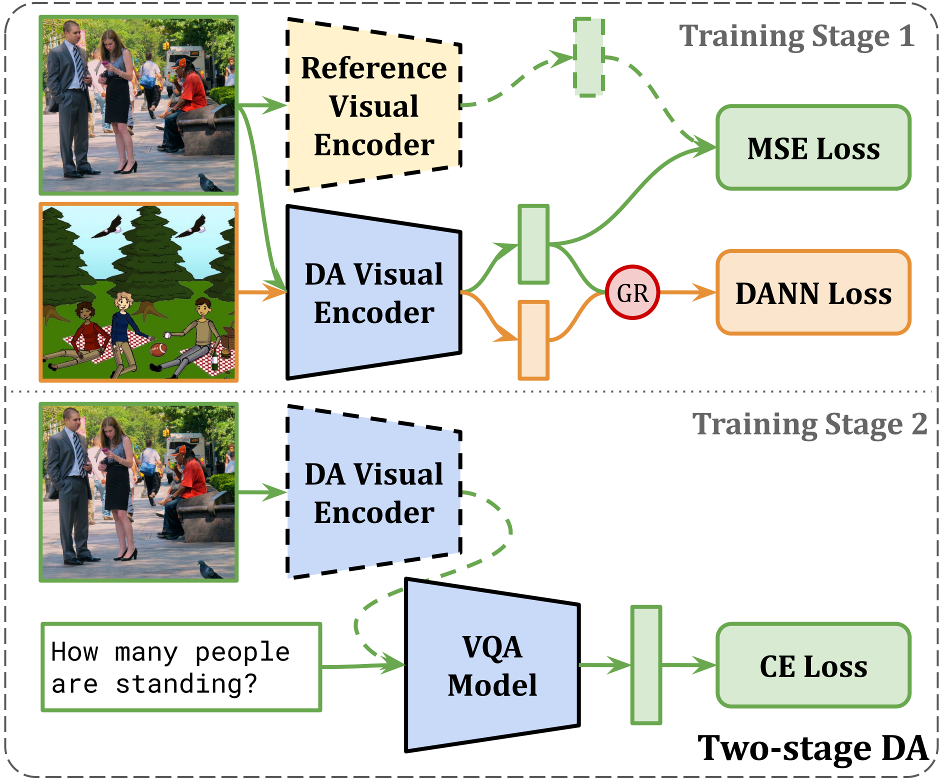
- ❖ We also try introducing several successful visual domain adaptation techniques in VQA.
  - ➢ DANN: Reverse gradients of visual encoder to interfere with domain discriminators.
  - ➢ Moment Matching: Applies moment matching regularizers to diminish domain discrepancies.
- ❖ We experiment with mainstream VQA models on a synthetic dataset pair (CLEVR and its visual style domain shifted counterpart) to analyze the effectiveness of domain adaptations.



Two-stage Domain Adaptation

- Similar with neuro-symbolic approaches, we find that recognition and reasoning modules can be *disentangled*, and a two-stage training strategy that separately optimizes the two achieves strongest performance thus most effective domain adaptation.

	VQA v2	CLEVR	GQA
Source Acc.	54.0	95.8	44.6
Target (Direct)	41.0	45.9	37.3
1-stage DANN	42.2	45.7	37.4
1-stage MM	42.6	46.6	**38.6**
2-stage DANN	**42.8**	**46.7**	38.5
Target (Full)	49.1	90.0	42.1



University of Pittsburgh

CVPR VIRTUAL JUNE 19-25

Domain-robust VQA with diverse datasets and methods but no target labels

Mingda Zhang Tristan Maidment Ahmad Diab
Adriana Kovashka Rebecca Hwa

University of Pittsburgh