Q: How to Specialize Large Vision-Language Models to Data-Scarce VQA Tasks? A: Self-Train on Unlabeled Images!

Zaid Khan[1], Vijay Kumar BG[2], Samuel Schulter[2], Xiang Yu[3], Yun Fu[1], Manmohan Chandraker[2,4]
[1]Northeastern University, [2]NEC Laboratories America, [3]Amazon, [4]UC San Diego

Northeastern University

NEC
NEC Laboratories America

When facing data scarcity, can large vision-language exploit unlabeled images to self-improve?

Key Ideas

- Apply self-training to visual question answering.
- Treat visual question generation as a direct image-conditional text-generation task.

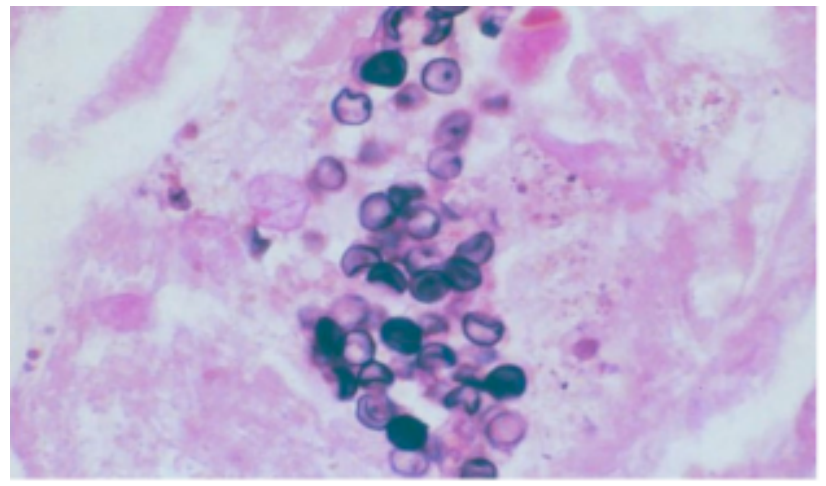
Results

- Increased performance on data scarce VQA tasks.
- Resistance to adversarial questions, reduced shortcut learning, and increased consistency of answers.
- Improved domain generalization.
- Reduced catastrophic forgetting of numerical reasoning.

What happens when you want to apply your vision-language model to a specific task?

Finetuning on small datasets is problematic.

Source: VQAv2



Q: What are these GMS-stained organisms?
A1: *Blastomyces dermatitidis.*
A2: *Cryptococcus neoformans.*
A3: *Pneumocystis jiroveci*.
A4: trophozoites of *Entamoeba histolytica*.
A5: yeasts of *Candida* species.

Q: What was the name of the first cloned type of this animal?
A: Dolly

- Question types can be very different.
- Images are often from a different domain.
- Heavily overparameterized model on very small dataset (200m+ vs <10k datapoint).
- Not enough data to learn the task well.
- Catastrophic forgetting of already learned skills (e.g. numerical reasoning).

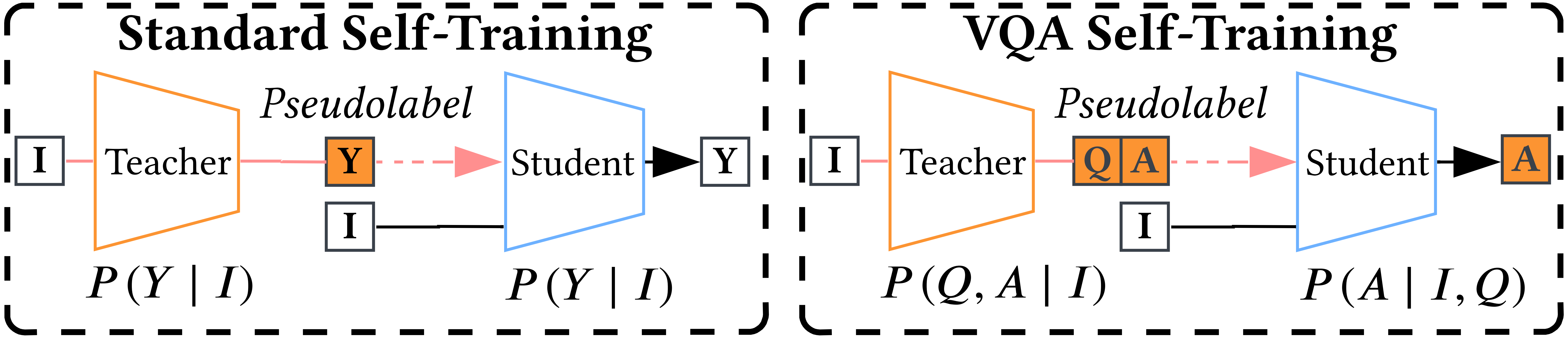
Can we take advantage of unlabeled images?

- Acquiring more annotations for complex tasks is expensive and time consuming.
- But unlabeled images are cheap and plentiful.

Self-training looks promising…

- Train on self-predictions on unlabeled images.
- Shown to be successful in object detection and image classification.

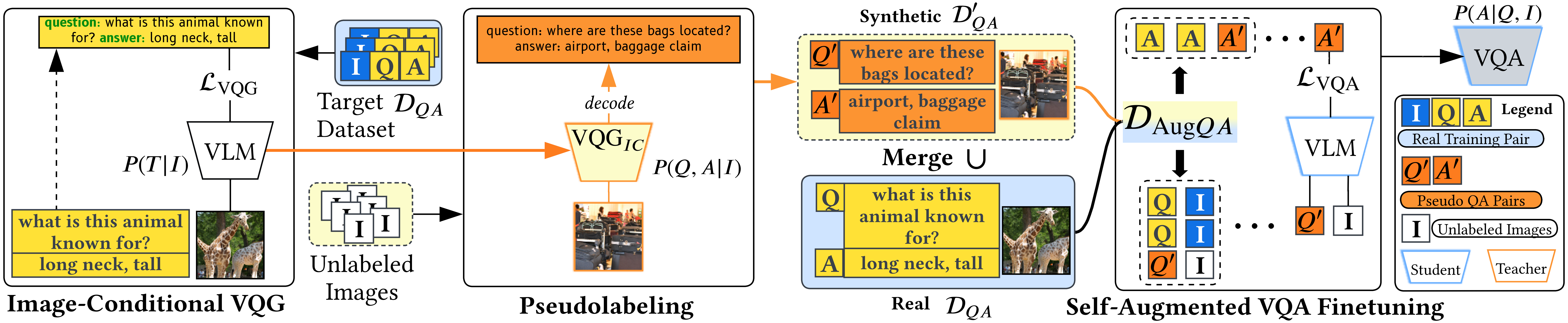
How can we apply self-training in VQA?



- Challenge 1: Task of student and teacher is **identical** in standard self-training, but teacher and student have **different tasks** in VQA self-training.

- Challenge 2: Our pseudo labels are visual questions, but approaches for visual question generation *require* dense annotations to generate questions.
  - Existing paradigms **can't work with unlabeled images!**

SelTDA: Self-Taught Data Augmentation



Selling Points

- Modular (no specialized architectures needed).
- Straightforward treatment of pseudolabeling as text generation.
- Offline and decoupled from training.

Images		Questions						
Labeled	Unlabeled	Real	Synthetic	Total	Multiplier	Accuracy	% Gain	Questions/Image
17,000	0	17,000	0	17,000	1x (baseline)	57.11		N/A
17,000	0	17,000	17,000	34,000	2x	57.85	+0.74	1 / 1
17,000	0	17,000	34,000	51,000	3x	**60.01**	**+2.90**	2 / 1
17,000	0	17,000	51,000	68,000	4x	59.73	+2.62	3 / 1
17,000	0	17,000	0	17k	1x (baseline)	57.11		N/A
17,000	8,500	17,000	17,000	34,000	2x	60.69	+3.57	2 / 1
17,000	17,000	17,000	34,000	51,000	3x	**62.09**	**+4.98**	2 / 1
17,000	25,500	17,000	51,000	68,000	4x	61.31	+4.20	2 / 1

Self-Taught Data Augmentation Improves Performance

- On a data-scarce task (A-OKVQA).
- Can work even *without extra images*, just by generating more questions.
- Not overly sensitive to hyperparameters.
- There’s a saturation point.

	# of Real + Synthetic QA Pairs			Robustness Test Sets				
	Real	Synthetic	Multiplier	AdVQA	VQA-CE	VQA-Rephrasings	Avg. % Increase	Robustness Total
(a)	17,000	0	×1	31.06	51.43	65.88	0	148.37
(b)	17,000	2,000	×1.1	37.09	52.96	67.94	+3.21	157.99
(c)	17,000	4,500	×1.3	36.99	53.15	**67.98**	+3.25	158.12
(d)	17,000	8,000	×1.5	37.34	**53.33**	67.57	**+3.29**	**158.24**
(e)	17,000	12,000	×1.7	**37.43**	52.62	67.35	+3.01	157.4
(f)	17,000	17,000	×2	36.95	52.05	66.95	+2.53	155.95
(g)	17,000	34,000	×3	36.89	51.00	65.64	+1.72	153.53
(h)	17,000	51,000	×4	36.06	50.25	64.78	+0.91	151.09
	Max % increase on each dataset			+6.03	+1.9	+2.1		+9.87

Self-Taught Data Augmentation Improves Robustness

- Adversarial Questions (AdVQA)
- Multimodal Shortcut Learning (VQA Counterexamples)
- Self-Consistency (VQA Rephrasings)

	Target (0-shot)		
Model	ArtVQA	PathVQA	RSVQA
Baseline (BLIP)	31.65	25.09	37.78
BLIP + *SelTDA*	38.03	26.76	38.99
% gain w.r.t baseline	+6.38	+1.67	+1.1

Self-Taught Data Augmentation Improves Domain Generalization

- ArtVQA (fine art images)
- PathVQA (medical images)
- RSVQA (remote sensing images)
- Note: only in-domain images were used!

Initialization	# Training Pairs Real	Synth	Numerical Reasoning VQAv2	VQA-Rephrasings
BLIP_{VQAv2}	17000	0	13.49	13.06
BLIP_{VQAv2}	17000	2000	38.73	33.74
BLIP_{VQAv2}	17000	4500	40.4	35.91
BLIP_{VQAv2}	17000	8000	42.9	36.5
BLIP_{VQAv2}	17000	12000	**43.3**	**37.77**
max % gain w.r.t baseline			+29.81	+24.71
BLIP	17000	0	1.42	1.29
BLIP	17000	17000	4.53	11.44
BLIP	17000	34000	**5.05**	11.77
BLIP	17000	51000	4.26	**11.86**
max % gain w.r.t baseline			+3.63	+10.57

Mitigation of Catastrophic Forgetting

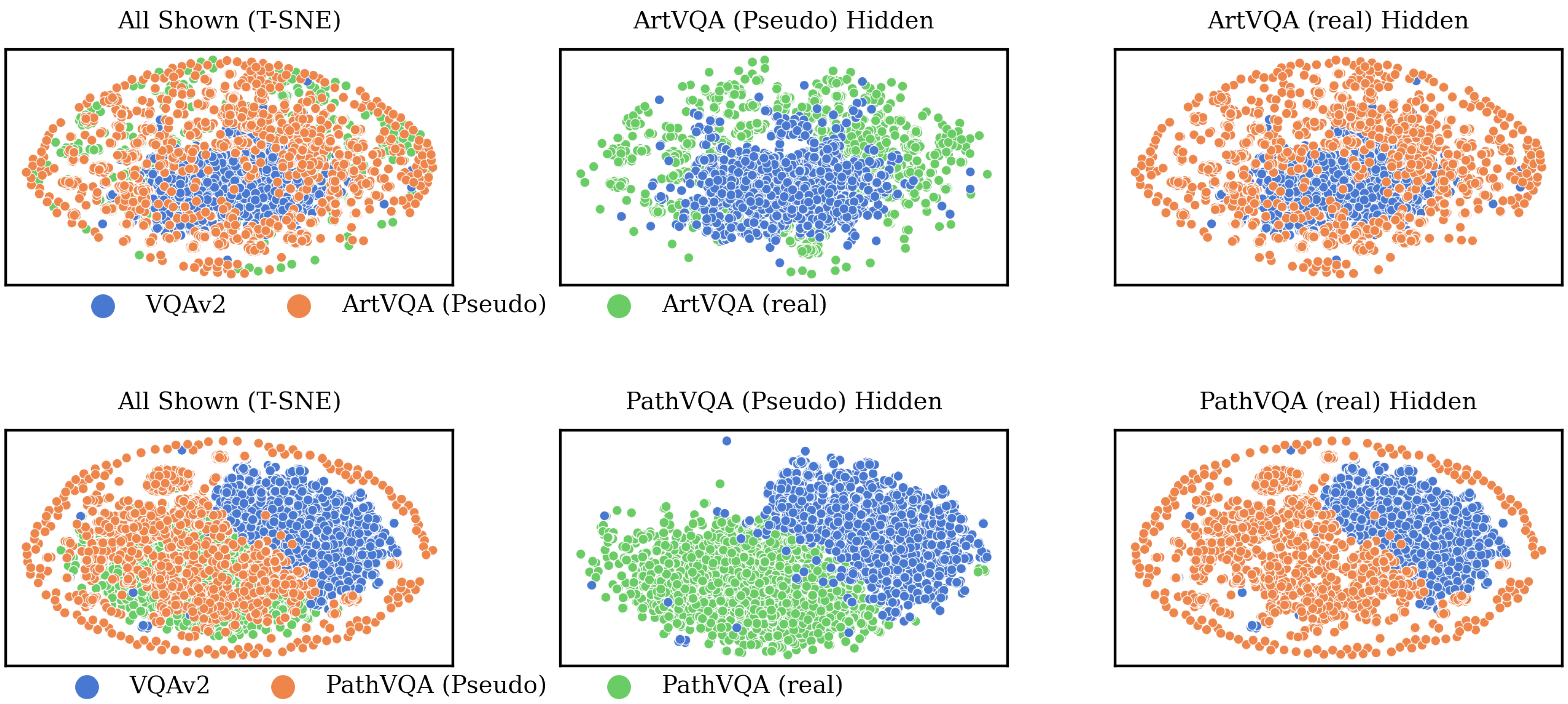
- Finetuning on small tasks really hurts numerical reasoning ability.
- Using self-taught data augmentation helps to retain it.
- Can even induce numerical reasoning ability when original model did not have it.

How good are the generated questions?

Question Type	Well-Posed Question	Answers Correct	Answerable	% of Total (95% CI)
External Knowledge	73%	62%	70%	29.6% - 50.00%
Visual Identification	94%	88%	94%	11.18% - 27.65 %
Visual Reasoning	83%	70%	80%	32.54% - 53.17%
Overall (95% CI)	71.16% - 87.96%	59.77% - 78.98%	68.83% - 86.22%	

- Human evaluation (~100 questions).
- Plenty of noise, but not too far away from annotator agreement (~80%) on real datasets.
- Question quality stratified by type of questions.
    - Model has competencies.

How good are the generated questions?



- Generated questions (orange) are diverse, covering:
  - real task/domain-specific areas (green)
  - generic post-training areas (blue).

Why does it work?

- Pseudolabels can act as regularization
- Distillation of dark knowledge from pretraining
  - Subtle difference in conditioning

VQA $P(A \mid I, Q)$ vs $P(T \mid I)$ Pretraining

model has lot more experience with one

Where do we go from here?

- Language capacity in VLMS has been increasing over time.
  - Makes self-improvement more promising.
  - More pre-existing knowledge about the world to draw on.
- Can we start correcting *specific errors* with self-training?
  - *Your answer is wrong, think about the problem 'till you get it right.*