Research Statement: Tejas Gokhale

tejasgokhale.com

My mission is to research and develop robust and reliable AI systems by leveraging the complex interactions between vision and language. I work at the wonderful intersection of machine learning, computer vision, and natural language processing, with two central goals: (a) to **design algorithms to improve robustness**, interpretability, and reliability of AI systems, powered by **semantic data engineering**, and (b) to develop benchmarks and evaluation protocols to **discover, quantify, and mitigate failure modes**. This mission is directly aligned with the clarion call for safe and robust AI from government agencies (DARPA[1], White House OSTP[2]), and academia (ACL[3], AAAI[4]). AI has undergone a paradigm shift in the past decade – the connection between vision and language (V+L) is now an integral part of AI, with deep impact beyond vision and NLP – robotics, graphics, cybersecurity, and HCI are utilizing V+L tools and there are direct industrial implications for software, arts, and media *(eg. I used text-to-image generators to generate pictures in the figure below!)*

Leverage Complex Relationships between Vision and Language (V + L)

- *Identify new types of distribution shift, biases, and threats to the reliability of V+L models*
- *Robust Optimization Algorithms powered by semantic data engineering*
- *Benchmarks for evaluation based on logical, linguistic, geometric, and physical knowledge*

Discover and Design Data Transformations for Improving Robustness and Reliability

- *Semantic data engineering guided by logical and semantic features of natural language*
- *Data engineering cannot be static – develop model-in-the-loop augmentations*
- *Knowledge-guided Adversarial training for robustness and domain generalization*

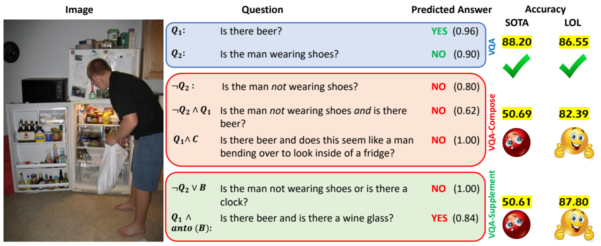
As V+L models are being widely adopted, new types of challenges and failure modes are emerging due to the multimodal and non-trivial relationships between images and text (as I have shown through my research). This means that we will need to simultaneously (1) discover failure by rigorous testing and benchmarking and (2) develop exciting new functionalities and capabilities with improved accuracy. The biggest challenge in robust multimodal learning is the scarcity of task-specific and functionality-specific data. While recent pre-trained models use millions of image-text pairs from the web to learn representations – they often fail when reasoning capabilities and fine-grained understanding is required. My research identifies these performance gaps and offers the unique combination of **semantic data engineering** and **knowledge-guided adversarial training** as a solution.

The findings from my research together show that active design and discovery of data transformations and adversarial training algorithms is the key for improving robustness, in multimodal (V+L) tasks as well as robust image classification. This work has been published in premier AI, vision, and NLP conferences, and has served as the foundation of grant proposals that I helped write (eg. a funded NSF Robust Intelligence grant[5] and ongoing proposals to IARPA and DARPA). I have led collaborative projects with ASU, Lawrence Livermore National Laboratory, Microsoft Research, Carnegie Mellon, and Adobe Research.

(A) Robust Multimodal (Vision+Language) Perception

▶ Robustness to Logical Transformations in Visual Question Answering [ECCV'20] [1]

Multi-modal tasks involving both vision and language (V&L) inputs, such as visual question answering (VQA), open up intriguing domain discrepancies that can affect model performance of test time. For the VQA task, models are trained to predict the answers to questions about images. My paper VQA-LOL [1], discovered that existing VQA models fail when logical transformations such as negation, conjunction, and disjunction are introduced in the questions. I built on this surprising finding to develop a data augmentation tool that produces logical combinations of multiple questions, I designed a logic-inspired training objective based on Frechet inequalities to guide the predicted probabilities of answers to questions with logical connectives. VQA-LOL was instrumental as a reality check for VQA performance and was included (by other researchers) as part of a compendium of datasets for testing VQA robustness [2]. VQA-LOL led to a series

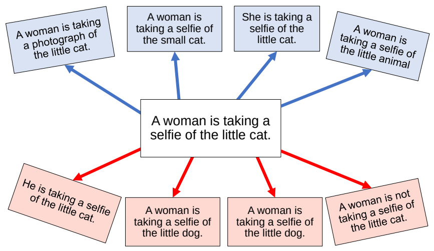


	Question	Predicted Answer	
Q_1:	Is there beer?	YES (0.96)	VQA
Q_2:	Is the man wearing shoes?	NO (0.90)	VQA
$\neg Q_2$:	Is the man *not* wearing shoes?	NO (0.80)	VQA-Compose
$\neg Q_2 \land Q_1$	Is the man *not* wearing shoes *and* is there beer?	NO (0.62)	VQA-Compose
$Q_1 \land C$	Is there beer and does this seem like a man bending over to look inside of a fridge?	NO (1.00)	VQA-Compose
$\neg Q_2 \lor B$	Is the man not wearing shoes or is there a clock?	NO (1.00)	VQA-Supplement
$Q_1 \land anto(B)$:	Is there beer and is there a wine glass?	YES (0.84)	VQA-Supplement

	Accuracy SOTA	Accuracy LOL
VQA	88.20 ✓	86.55 ✓
VQA-Compose	50.69	82.39
VQA-Supplement	50.61	87.80

VQA-LOL revealed VQA models' surprising failure to answer logical compositions of multiple questions.

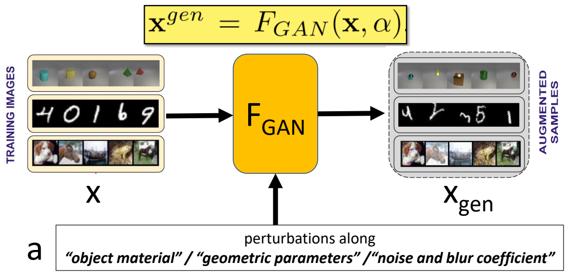
of papers [EMNLP'20][3], [ACL'22][4], [ACL'22][5] that adopted linguistic and semantic transformations for image–text alignment, video–text reasoning, and natural language inference. With collaborators, we further expanded our approach for synthetic data generation that enabled design of weakly-supervise VQA models for limited-data settings [NAACL'21][6], [ACL'21][7], and for creating video QA benchmarks for reasoning about physical properties of objects [EMNLP'22][8] and commonsense reasoning about people's actions [EMNLP'20][9].


A woman is taking a photograph of the little cat.
A woman is taking a selfie of the small cat.
She is taking a selfie of the little cat.
A woman is taking a selfie of the little animal
A woman is taking a selfie of the little cat.
He is taking a selfie of the little cat.
A woman is taking a selfie of the little dog.
A woman is taking a selfie of the little dog.
A woman is not taking a selfie of the little cat.

Training with SISP transforms improves accuracy, robustness, calibration of V+L models.

▶ Semantically Distributed Robust Optimization Improves Vision–Language Inference [ACL'22] [4] I identified that knowledge of linguistic transformations can inform the algorithm design for V+L tasks. This led to the development of the "SISP transformation" suite – a controlled method to semantically manipulate text to generate augmented data that is semantics-inverting (SI) or semantics-preserving (SP). I showed that these SISP transformations can be leveraged to train robust models by developing a new knowledge-guided adversarial training algorithm called *Semantically Distributed Robust Optimization (SDRO)*. The combination of SISP (data engineering) and SDRO (robust optimization) led to improvements on image-based reasoning, video-based reasoning, and visual question answering, along several dimensions of robustness – in-domain and out-of-domain accuracy, adversarial robustness, and calibration, and also on my previous VQA-LOL benchmark [1].

(B) Robust Image Classification and Domain Generalization


$\mathbf{x}^{gen} = F_{GAN}(\mathbf{x}, \alpha)$
TRAINING IMAGES
F_{GAN}
AUGMENTED SAMPLES
x
x_{gen}
a
perturbations along
"object material" / "geometric parameters" /"noise and blur coefficient"

AGAT, an adversarial training algorithm, supports a wide-range of attribute specifications and improves robustness.

▶ Robustness under Attribute Shift [AAAI'21] [10] Previous work on robust image classification focuses on pixel-level adversarial attacks. However, in real world scenarios, test examples can vary along known attributes such as size, shape, colors, and geometric transformation. Unfortunately, these cannot be covered by methods that utilize norm-bounded and additive pixel-level perturbations. We consider a setting where information about the target domain is available only in terms of a set of attributes that are known to differ at test time – there is no access to a target validation set, or knowledge about the magnitudes and combinations of attributes at test time. As such, standard data augmentation and pixel-level adversarial training is ineffective. I developed a new form of adversarial training: *Attribute-Guided Adversarial Training (AGAT)* that parameterizes the input space in terms of attributes, and adversarially perturbs image attributes to maximize exposure of the classifier to previously unobserved variations. AGAT supports a wide-range of attribute specifications, which we demonstrate with large gains in three different use-cases: (1) object-level attribute-shift (2) geometric transformations (3) common natural corruptions.

▶ Improving Diversity with Adversarially Learned Transformations [WACV'23] [11] *Single source domain generalization (SSDG)* is a challenging setting, where the model has access only to a single training domain (eg. real photos), and is expected to generalize to multiple testing domains with domain shift (eg. sketches and cartoons). Unlike the setting for AGAT, there is no access to attributes or external knowledge about the nature or magnitude of domain shift. Success of SSDG depends on maximizing diversity of training data; this naturally implies that data augmentation is one of the main sources of diversity! But what augmentation method should we choose? We found that pre-specified augmentations [12, 13] cannot model large domain shift in SSDG effectively. This led to our novel framework that discovers adversarially learned transformations (ALT), by perturbing the parameter space of an "adversity" network to model plausible yet hard image transformations. ALT offers a synergy between diversity and adversity, exposing the model to increasingly unique, challenging, and semantically diverse examples – ideally suited for SSDG. ALT's ability of improving the training diversity resulted in performance gains over all existing techniques, on multiple domain generalization benchmarks.

Future Research Agenda

Over the last decade, the nature of AI research has changed considerably. Research communities that were largely isolated are now actively leveraging the connecting elements between the visual world and human-assigned meaning (language). However the link between V and L goes beyond image–text similarity. Language is ideally suited for developing reasoning capabilities beyond the visible – these reasoning capabilities are key for allowing V+L models to interact with humans. My goal is to integrate vision, language, and human collaboration together for active decision making, complex reasoning, and for learning novel concepts, without sacrificing robustness.

► **Short-Term Goals: Towards Reliable Visual Reasoning.** In the short term, my focus will be on developing reasoning capabilities that are geared towards *correctness* of outputs, for instance reasoning about spatial relationships and scene geometry and reasoning about everyday actions.

- **Spatial reasoning** is a fundamental aspect of computer vision. In WeaQA **[ICCV'21]**[14] we showed that VQA models lacked this understanding, but their performance can be improved via weak geometric supervision. My ongoing work involves investigating the spatial understanding of text-to-image synthesis (T2I) – I am developing an evaluation framework called *"VISOR"* for quantifying the fidelity of T2I models in generating spatial relationships between objects. VISOR reveals the surprising finding that although recent SOTA models like DALLE exhibit high photorealism, they are ineffective in composing images with two or more distinct objects, especially when a spatial relationship such as left/right/above/below is specified. I plan to explore this direction further and develop prompting and finetuning techniques to improve spatial reasoning of image generators. T2I models are ideally poised to serve a crucial purpose in computer vision research – my research will investigate how the ability of generating images corresponding to text prompts can be leveraged for data generation, augmentation, and transformation for low-resource settings and to enhance the robustness in V+L.
- **Reasoning about Actions.** In our previous work **[EMNLP'20]**[9], we developed commonsense video captioning to speculate about effect of actions. However, can V+L models reason about unlikely and atypical actions (eg. people often kick footballs, but rarely kick walls)? I plan to investigate how V+L models like CLIP [15] can be used for reasoning about everyday actions and commonsense aspects, for both typical and atypical situations, by using counterfactual text-based image manipulation to reflect atypical situations. This study is expected to reveal that V+L models are biased towards spurious correlations between actions and objects. My research will develop debiasing techniques and constraint-based learning for reasoning about actions and their consequences.

► **Long-Term Goals.**

Human-Computer Collaborative Reasoning. I am convinced that the use of language has immense potential in changing the way we interact with AI and for democratizing and simplifying access to graphics and robotics. While we have begun to develop visual grounding and reasoning frameworks, how can we improve these abilities and embed dialog and cooperation with humans into the reasoning process? I am excited about starting this new research program of "Human-Computer Collaborative Reasoning", an under-explored direction, which will bring together reasoning and human-aware AI research, to improve visual reasoning capabilities of computer vision models. My previous experience in discovering and mitigating failure models will continue to be a core element of this agenda, by studying how human collaboration and feedback can help avoid such failures.

Connections between Adversarial and Distributional Robustness. While standard notions of distribution shift in ML are limited to single modalities and static train–test splits, the theoretical investigation of effect of interactions between different modalities remains unexplored. I am interested in understanding fundamental connections between adversarial and distributional robustness, especially when multiple modalities and data formats are involve. This will be particularly challenging effort for models that will interact and continually learn and reason with human collaboration. My recent empirical investigation **[ACL'22]** [16] found that data filtering methods with good intentions of removing spurious correlations, can hurt adversarial robustness. This finding has been recently corroborated by other researchers [17, 18]. I plan to pursue this direction and expect it to lead to actionable design considerations for building robust V+L models.

In sum, I will pursue knowledge-guided, human-aware, and robust learning and reasoning about V+L.

References

[1] **Tejas Gokhale**, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*. Springer, 2020.

[2] Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020.

[3] **Tejas Gokhale**, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 878–892, Online, 2020.

[4] **Tejas Gokhale**, Abhishek Chaudhary, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Semantically distributed robust optimization for vision-and-language inference. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1493–1513, 2022.

[5] Neeraj Varshney, Pratyay Banerjee, **Tejas Gokhale**, and Chitta Baral. Unsupervised natural language inference using phl triplet generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2003–2016, 2022.

[6] Pratyay Banerjee, **Tejas Gokhale**, and Chitta Baral. Self-supervised test-time learning for reading comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1200–1211, Online, 2021. Association for Computational Linguistics.

[7] Pratyay Banerjee, **Tejas Gokhale**, Yezhou Yang, and Chitta Baral. Weaqa: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3420–3435, 2021.

[8] Maitreya Patel, **Tejas Gokhale**, Chitta Baral, and Yezhou Yang. Counterfactual reasoning about implicit physical properties via video question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[9] Zhiyuan Fang, **Tejas Gokhale**, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2Commonsense: Generating commonsense descriptions to enrich video captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860, Online, 2020. Association for Computational Linguistics.

[10] **Tejas Gokhale**, Rushil Anirudh, Bhavya Kailkhura, Jayaraman J Thiagarajan, Chitta Baral, and Yezhou Yang. Attribute-guided adversarial training for robustness to natural perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7574–7582, 2021.

[11] **Tejas Gokhale**, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[12] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[13] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *International Conference on Learning Representations*, 2020.

[14] Pratyay Banerjee, **Tejas Gokhale**, Yezhou Yang, and Chitta Baral. Weakly supervised relative spatial reasoning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1908–1918, October 2021.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[16] **Tejas Gokhale**, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2705–2718, 2022.

[17] Mazda Moayeri, Kiarash Banihashem, and Soheil Feizi. Explicit tradeoffs between adversarial and natural distributional robustness. *NeurIPS 2022*, 2022.

[18] Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. *arXiv preprint arXiv:2209.00613*, 2022.

[1] https://www.darpa.mil/work-with-us/ai-next-campaign

[2] https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[3] https://2023.aclweb.org/calls/main_conference/#theme-track-reality-check

[4] https://aaai.org/Conferences/AAAI-23/safeandrobustai/

[5] https://nsf.gov/awardsearch/showAward?AWD_ID=2132724