Proactive Human-Robot Interaction using Visuo-Lingual Transformers and Object Interaction Graphs

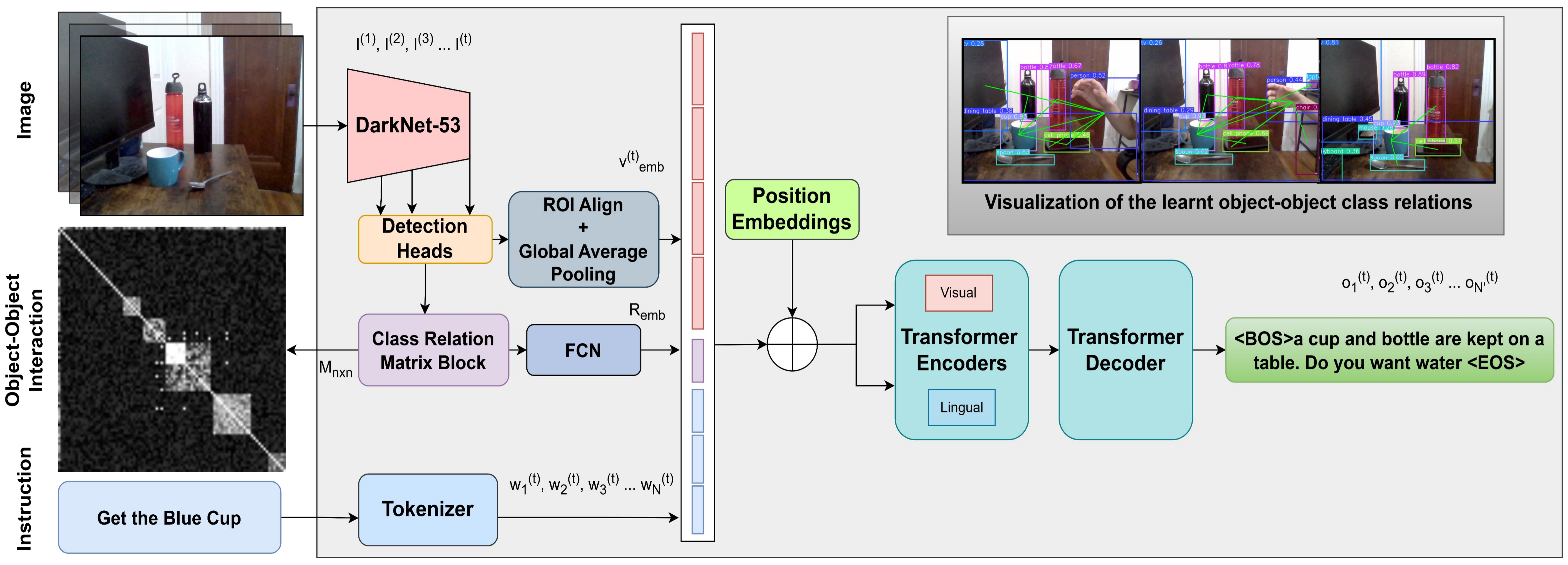
Pranay Mathur
Georgia Institute of Technology

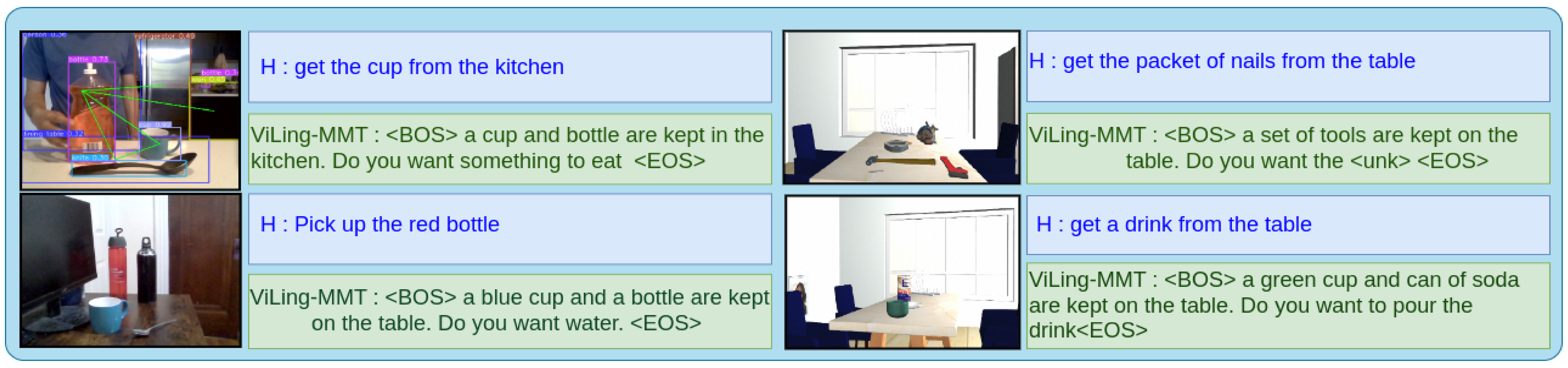
Motivation

Humans possess innate ability to extract latent visuo-lingual cues to infer context through observation and human interaction.

Enables proactive prediction of the underlying intention engendering an intuitive method for task agnostic collaboration

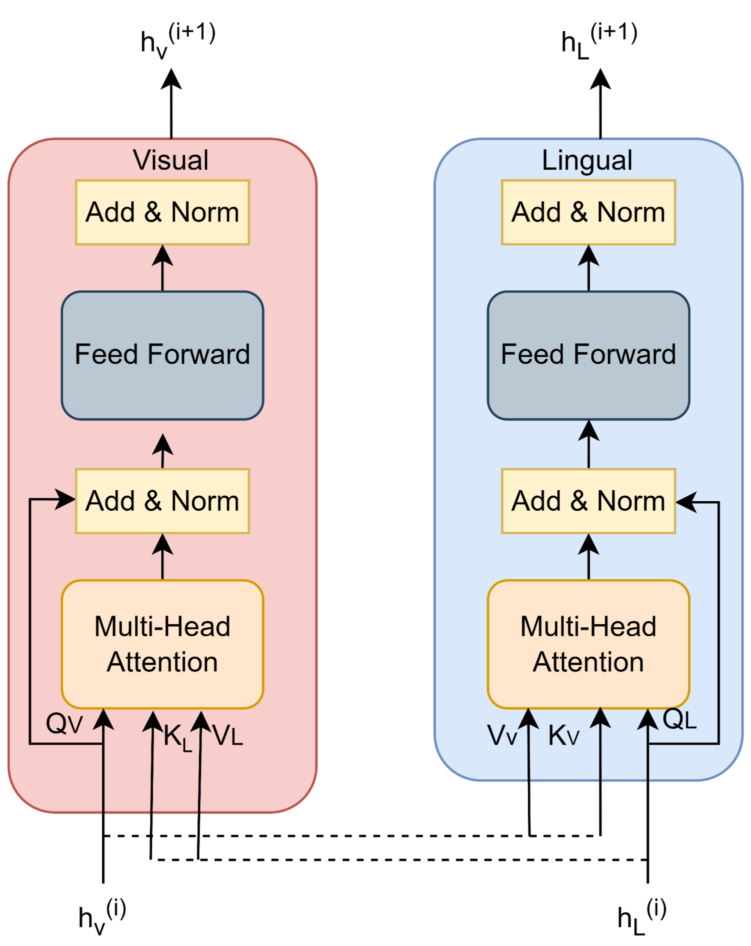
Goal: Endow social robots with the ability to reason about the end goal and proactively predict intermediate tasks without hand-crafted triggers which are specific to a scene





Contributions

- End-to-end multimodal transformer architecture **ViLing-MMT** that uses visual cues from the scene and intermediate task instructions to initiate pro-active behavior
- Incorporating graphical representation of learnt prior object-object relations in an unsupervised manner



Co-attentional Transformer Layer

Method

Vision Encoder

- Incorporate visual context awareness using an encoder based upon the Darknet-53 neural network architecture [1]
- Generate image region features by extracting bounding-boxes and their visual features
- Apply RoIAlign pooling to normalize the sizes of feature maps as well as global average pooling (GAP) to reduce the feature representation dimension

Graph Encoder

- Use class occurrences of objects to form a graph encoding historical object-object relations
- Each class c_n is represented as a vertex $\mathbf{v_{cn}} \in \mathbf{V}$, N (V) = $\boldsymbol{n}$ and a relation is denoted by an edge
- The weight $\mathbf{w_{c1c2}}$ of the edge is a measure of the extent to which the object classes $c_1 - c_2$ are

$$\mathbf{w_{c1c2}} = \frac{N(c_1 \cap c_2)}{N(c_1) \cdot N(c_2)}$$

Transformer Encoder-Decoder

Transformer Encoder

- Lingual instructions represented by a sequence of tokens
- Combined with vision embeddings and object-object interaction graph embedding to create $\mathbf{vRL}^{(t)}_{emb}$

$$\mathbf{f}_{\theta_{enc}}(\mathbf{w}_{1:N}, \mathbf{I}^{(t)}, M_{n\times n}) = \mathbf{vRL}^{(t)}_{emb}$$

- Encoder shares architectural similarities with ViLBERT [3] which uses multi-modal streams of data that interact through co-attentional transformer layers
- Novel cross-modal key and value communication allows variable individual modality-specific depths and promotes cross-modal connections at various depths

Transformer Decoder

$$\begin{aligned}
&\mathbf{p}_{\theta_{enc},\theta_{dec}}(\mathbf{o}_{1:N'}|\mathbf{w}_{1:N}, \mathbf{I}^{(t)}, M_{n\times n}) \\
&= \prod_{i=1}^{N'} \mathbf{p}_{\theta_{enc},\theta_{dec}}(\mathbf{o}_i|\mathbf{o}_{0:i-1}, \mathbf{w}_{1:N}, \mathbf{I}^{(t)}, M_{n\times n}) \forall i \in 1,\cdots,N' \\
&= \prod_{i=1}^{N'} \mathbf{p}_{\theta_{dec}}(\mathbf{o}_i|\mathbf{o}_{0:i-1}, \mathbf{vRL}^{(t)}_{emb}) \forall i \in 1,\cdots,N'
\end{aligned}$$

- Defines the conditional probability distribution of target sequence given the contextual encoding sequence

Training

Datasets:

- Flickr8K [4] and MSCOCO [5] for pre-training
- Flickr8K annotations augmented with reference captions and task suggestions along with trigger variable

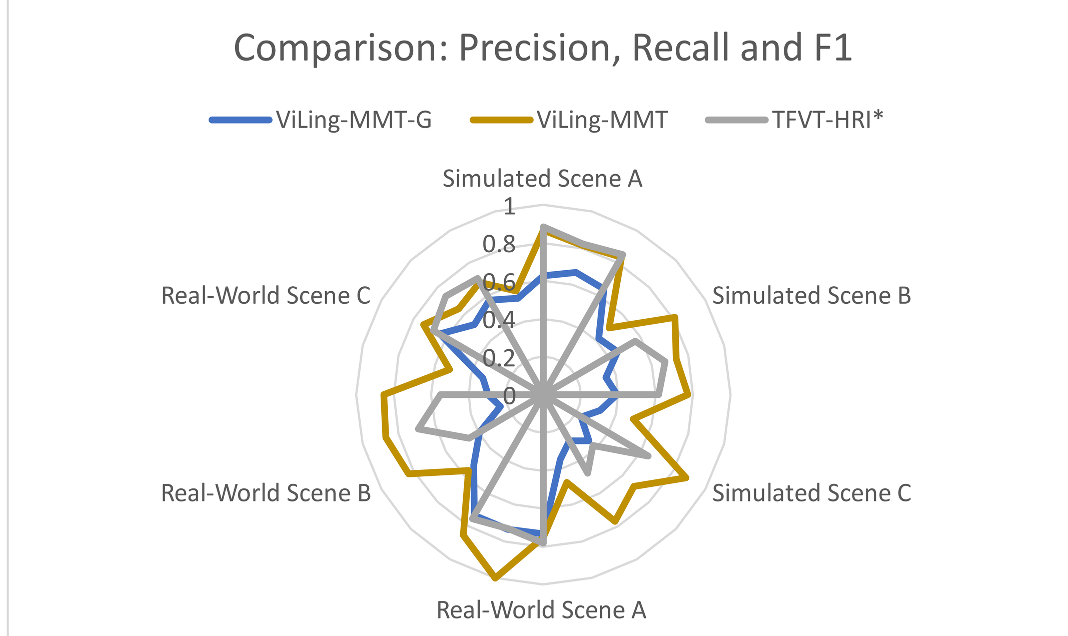
Loss:

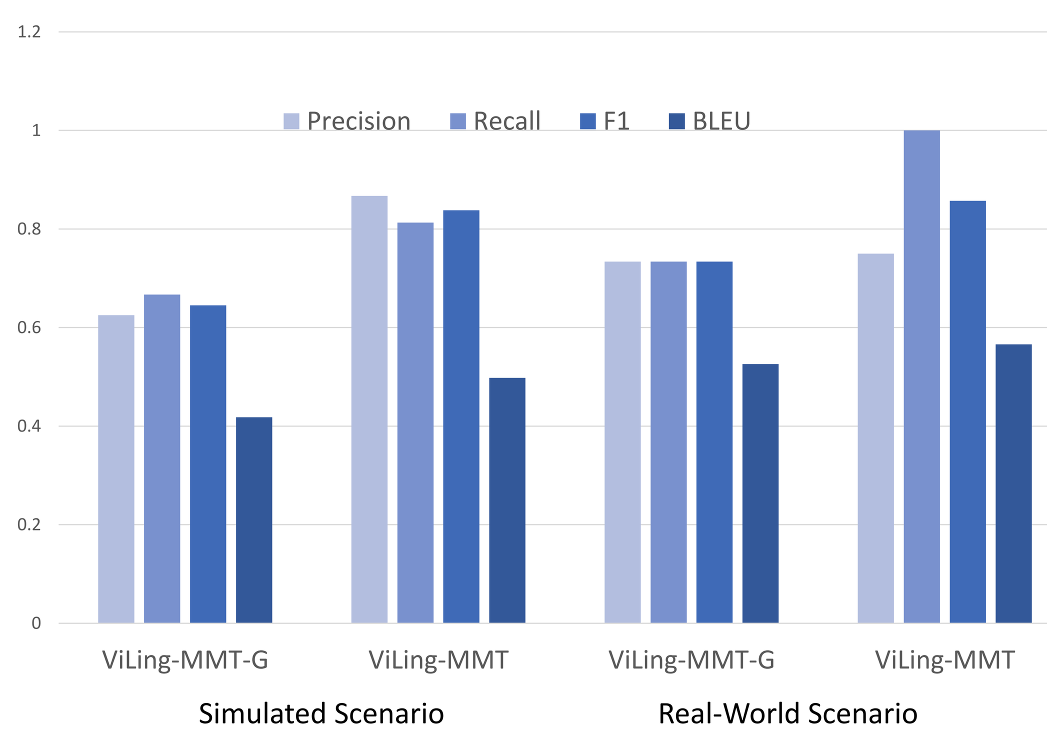
Minimize cross-entropy loss of action triggering and sum of the negative log-likelihood of the word provided in the ground truth description

$$\mathcal{L}_d = -log(p(O_t))$$

$$\mathcal{L} = \mathcal{L}_{ce}(\hat{i}_t, i_t) + \Sigma_t i_t \Sigma_{i=1}^{N'} \mathcal{L}_d$$

Evaluation





Ablation Studies: Precision, Recall, F1 and BLEU score

Model	Simulated Scene			
	Precision	Recall	F1	BLEU
ViLing-MMT-G	0.625	0.667	0.645	0.418
ViLing-MMT	**0.867**	**0.813**	**0.838**	**0.498**
Model	**Real-World Scene**			
	Precision	**Recall**	**F1**	**BLEU**
ViLing-MMT-G	0.734	0.734	0.734	0.526
ViLing-MMT	**0.75**	**1**	**0.857**	**0.566**

Ablation Studies: Precision, Recall, F1 and BLEU score

References

1. Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
3. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretrain-ing task-agnostic visiolinguistic representations for vision-and-language tasks, 2019
4. Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of AI Research, 2013
5. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dolľar, and C. Zitnick. Microsoft coco: Common objects in context. 2014
6. Y. Xue et al., "Proactive Interaction Framework for Intelligent Social Receptionist Robots," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021