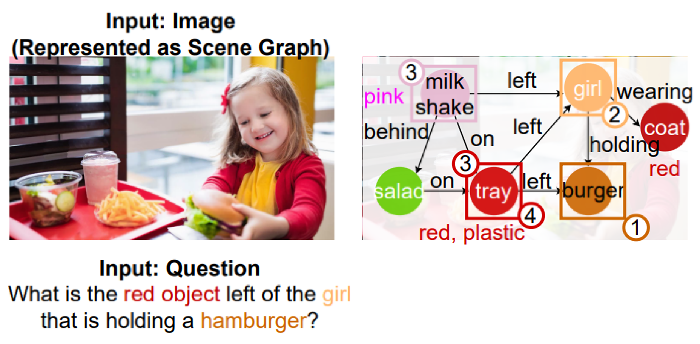
SelfGraphVQA

Bruno Souza ⚽ Marius Aasan ❃ Helio Pedrini ⚽ Adín Ramírez Rivera ❃

⚽ University of Campinas (UNICAMP) ❃ University of Oslo (UiO)

Our Problem

VQA...



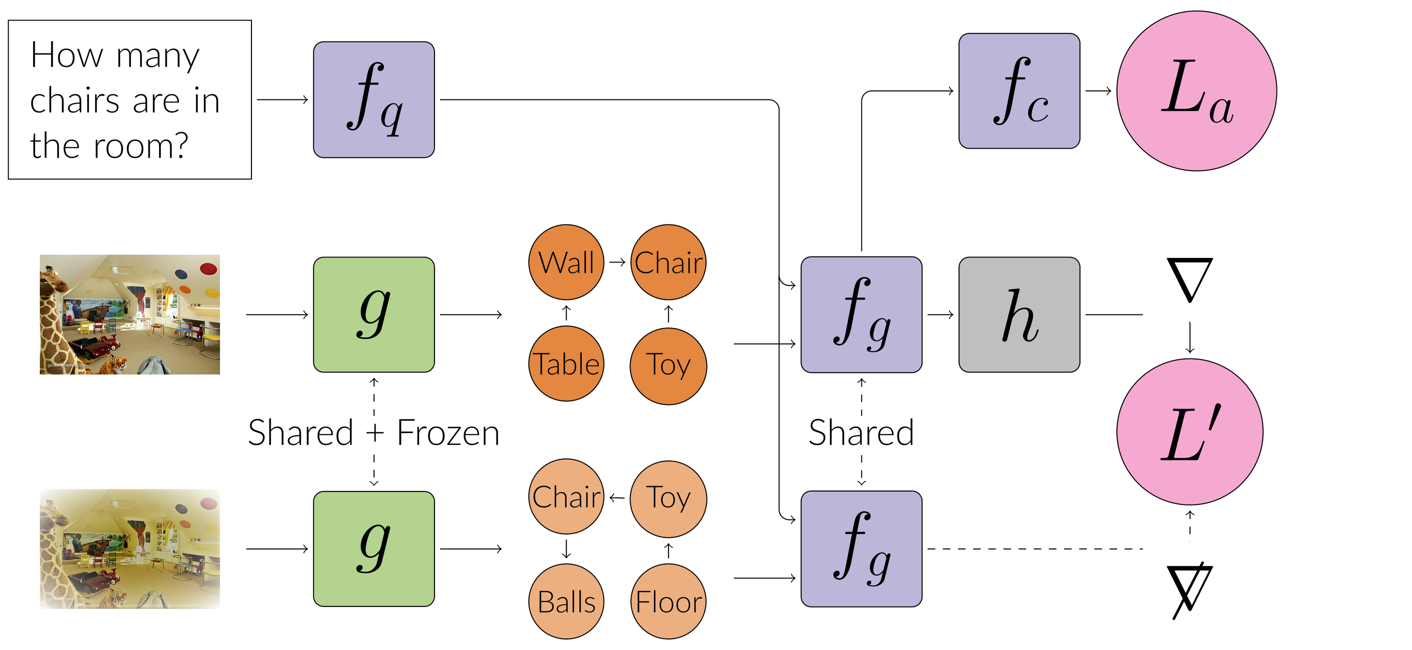
...does not work for non-ideal data!

Method	Eval. Data	Acc (%)
Human	–	89.3
GraphVQA	Annotated/SGG	94.8
LRTA	Annotated/SGG	93.1
CRF	Annotated	72.1
LXMERT	Extracted	59.8

Motivation: Research Questions

- Do scene graph VQA models work with **non-ideal** generated scene graphs? **No.**
- Does un-normalized contrastive learning enhance visual information in the VQA task? **Yes. (Sometimes)**

SelfGraphVQA Architecture



Three distinct maximization strategies:

1. **Local Similarity**. A localized node representation (i.e., object-wise):

$$L^*_\ell(p_1, z_2) = \frac{1}{O}\sum_i^O \arg\min_{z_{2,j}} D(p_{1,i}, z_{2,j}), \tag{1}$$

where O is the number of object in the scene. Symmetrically, we compute $L^*_\ell(p_2, z_1)$,

$$L_\ell(z_1, z_2) = \frac{1}{2}\Big(L^*_\ell(p_1, z_2) + L^*_\ell(p_2, z_1)\Big). \tag{2}$$

2. **Global Similarity**. A global pooled graph representation (i.e., scene-wise):

$$L_g(z_1, z_2) = \frac{1}{2}\Big(D(p_1, z_2) + D(p_2, z_1)\Big). \tag{3}$$

3. **Regularization for Permutation Equivariance**. Align similar nodes and encourage regularization. The anchors' similarity $s_{1,i} = \arg\min_{z_{1,j}} D(z_{1,i}, z_{1,j})$ and similarities of augmented views $s_{2,ij} = D(z_{2,i}, z_{2,j})$. We compute cross entropy (CE) between anchors and augmentations

$$J(z_1, z_2) = \mathrm{CE}(s_1, s_2), \tag{4}$$

which acts as a regularizer to constrain permutation equivariance for the augmentations in addition to the local loss, yielding

$$L_s(z_1, z_2) = L_\ell(z_1, z_2) + J(z_1, z_2), \tag{5}$$

Ablations

Change in accuracy under potentially disruptive augmentations and perturbations.

Question Type	Augmentation	Baseline	Global	Local	SelfSim
Relation	Flip	−1.6	−3.4	−3.2	−3.9
Attribute	Strong Color Jitter	+1.14	−3.7	−0.8	−1.2
Global	Gaussian Noise + Crop	−5.6	−7.7	−5.5	−8.1

Results (%) of the Aug. Baseline and SelfSim.

Method	Binary	Open	Validity	Plausibility	Acc
Baseline Aug	65.1	28.7	94.6	90.1	50.1
SelfSim	68.4	31.3	94.9	90.7	54.0

Sensitivity of accuracy (%) for bias question analyzes of SelfGraphVQA and SelfGraphVQABERT.

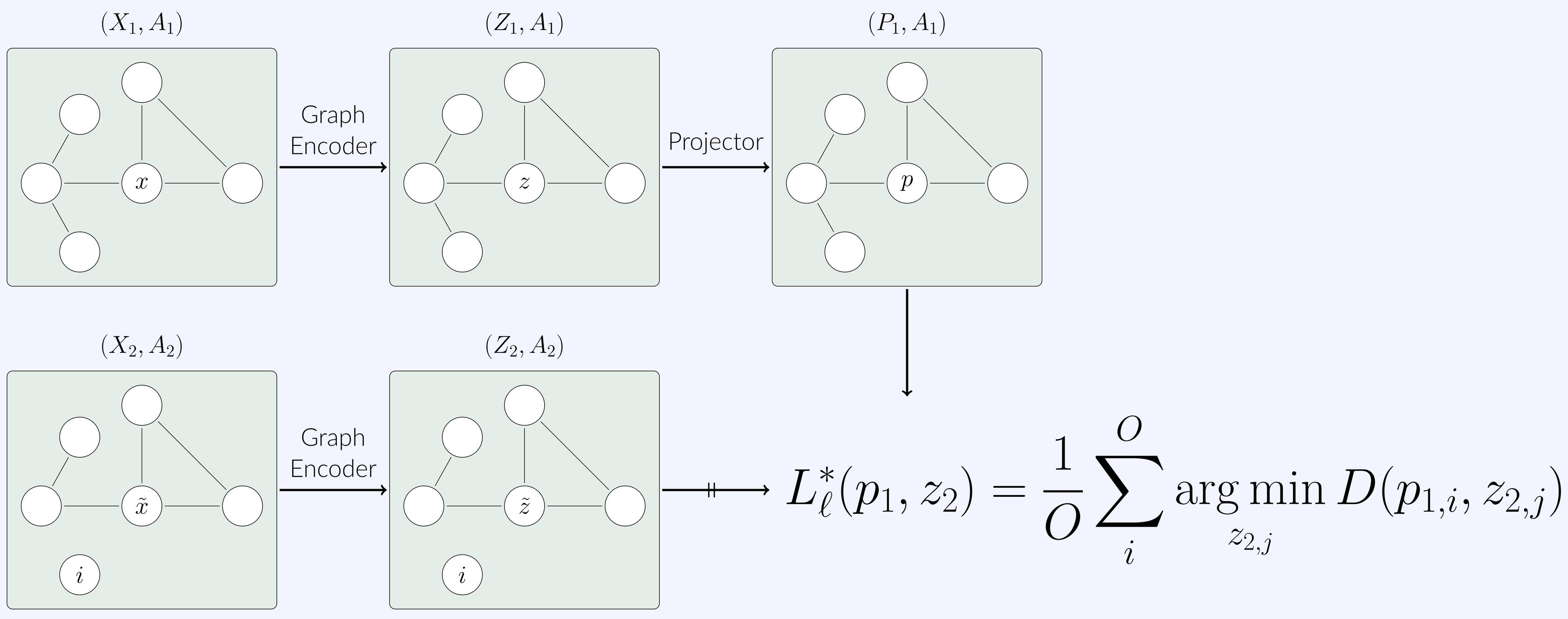
Setup	Methods			
Scene Graph + Question	Baseline	Local	Global	SelfSim
Noise + SG	16.2	16.6	28.6	26.6
Question + Noise	39.9	38.3	37.4	39.8
Noise + Noise	12.7	14.6	18.9	21.0
Question + Scene Graph	BERT Baseline	BERTGlobal+link	BERTSelfSim+link	
Noise + SG	21.0	23.2	24.5	
Question + Noise	42.4	41.8	42.8	
Noise + Noise	19.8	21.7	21.3	

Acknowledgments

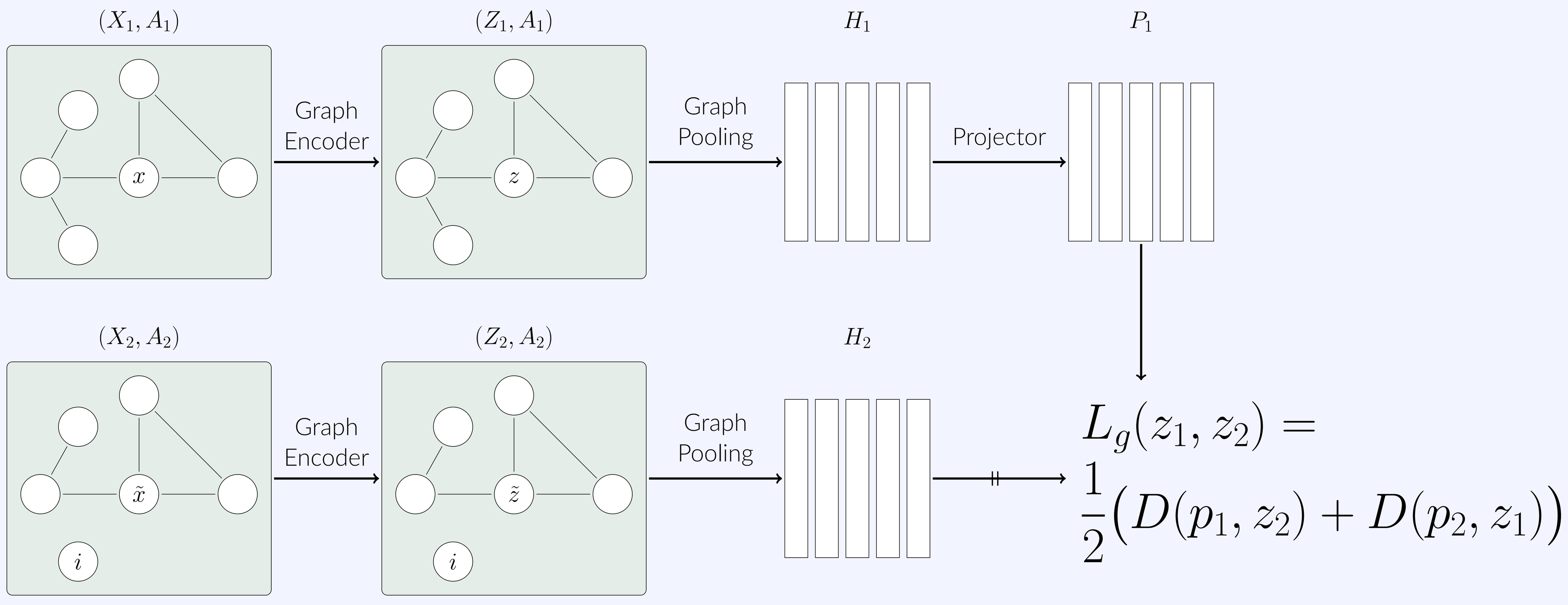
This work was partially funded by the FAPESP (São Paulo Research Foundation). The computations were partially performed on resources provided by Sigma2. This work was partially performed at the Artificial Intelligence Lab., Recod.ai. This work was partially funded by the Research Council of Norway, via the Visual Intelligence Centre for Research-based Innovation.

Maximization Similarity Strategies

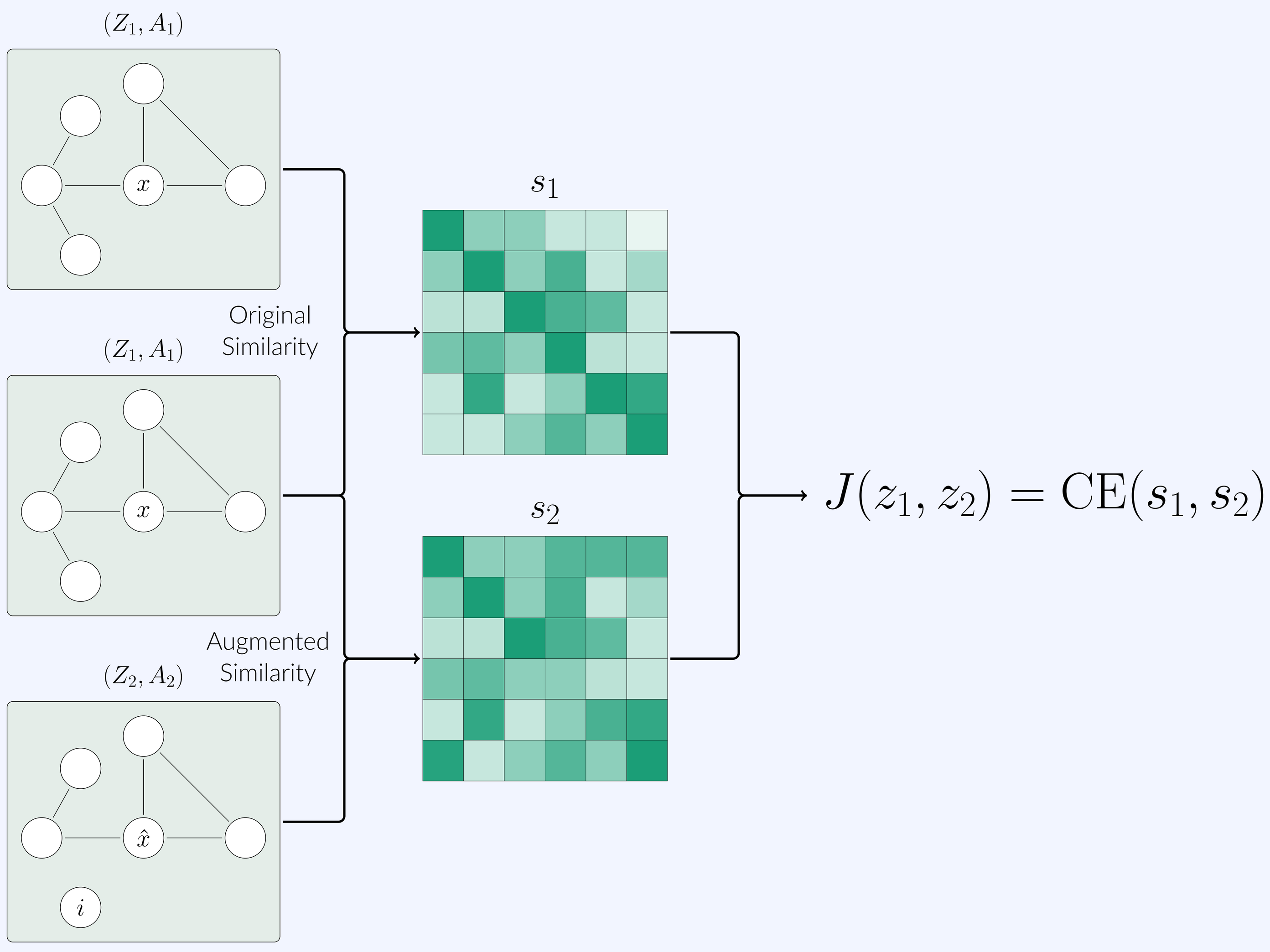
Local Similarity. Contrast local repr. to enhance object's visual cues



Global Similarity. Contrast global repr. to enhance global visual cues



Regularization for Permutation Equivariance (SelfSim). Align similar nodes and regularize

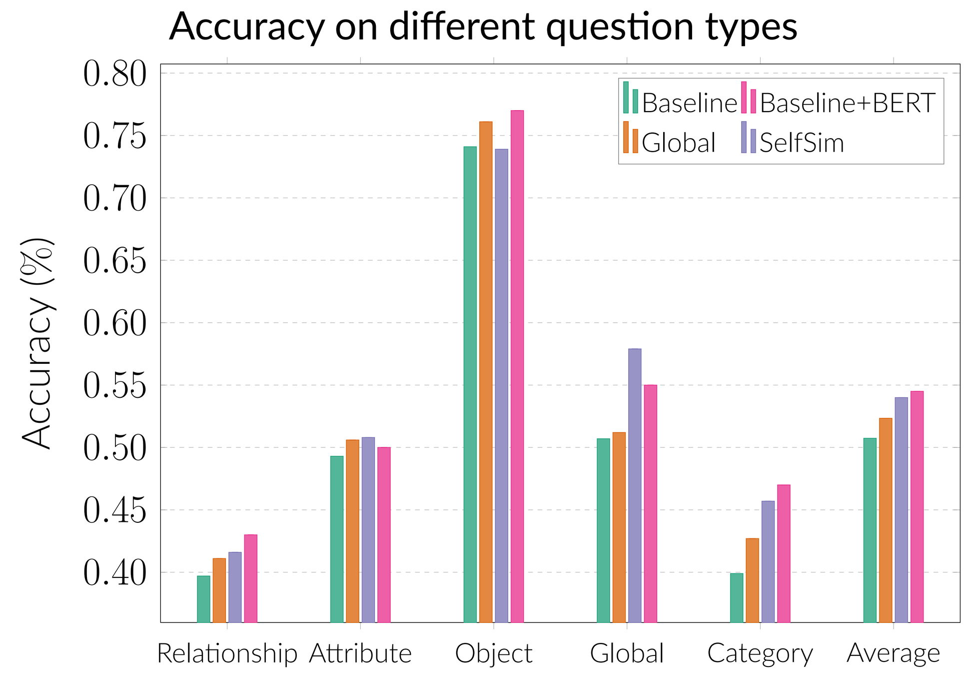


Results

Results (%) on GQA by standard metrics

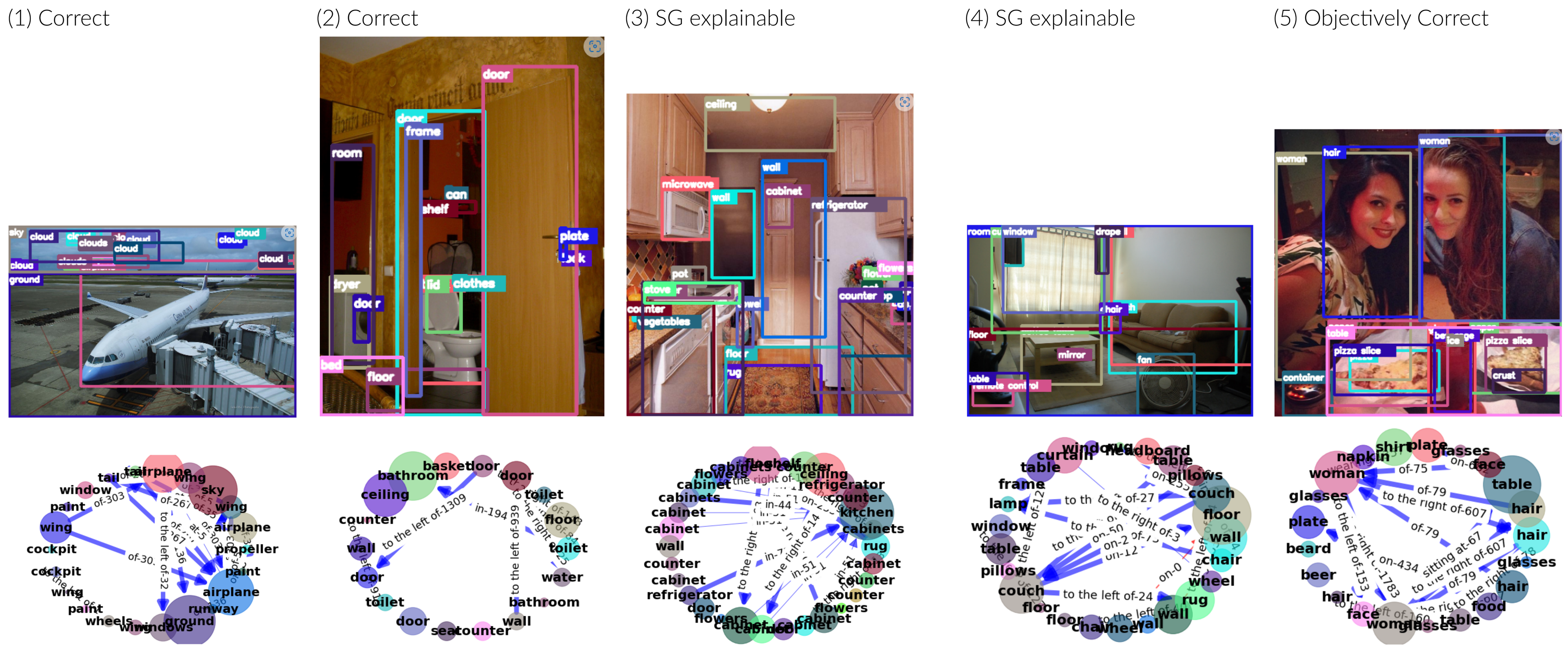
Method	Binary (↑)	Open (↑)	Consist. (↑)	Validity (↑)	Plausab. (↑)	Distr. (↓)	Acc (↑)
Baseline	65.8	29.7	58.2	94.9	90.5	11.7	50.1
Baseline+BERT	68.0	32.2	62.6	95.0	90.9	7.7	53.8
Local	66.8	30.2	59.4	94.9	90.6	8.8	51.5
Global	67.7	30.8	62.5	94.9	90.6	6.7	52.3
SelfSim	**68.4**	31.3	**65.9**	94.9	90.7	**2.1**	54.0
Global+BERT+link	68.0	**33.0**	63.9	95.0	**91.2**	8.9	**54.5**
SelfSim+BERT+link	68.2	32.8	64.3	**95.0**	91.0	8.0	**54.5**

Accuracy on different question types



Examples

(1) Correct (2) Correct (3) SG explainable (4) SG explainable (5) Objectively Correct



Q: What is the aircraft on the ground?
Answer: Airplane
Prediction: Airplane

Q: Are there any parachutes or bags?
Answer: No
Prediction: No

Q: What is the white pot holding?
Answer: Flower
Prediction: Flowers

Q: Which kind of furniture is right of the curtains?
Answer: Chair
Prediction: Chair

Q: What is in the red glass??
Answer: Beverage
Prediction: Liquid