

Compositional Substitutivity of Visual Reasoning for Visual Question Answering - Supplementary Material

1 VQA-SPS v2 Dataset

We construct the VQA-SPS v2 dataset based on the val split of the VQA v2 dataset [2], where questions are human-written rather than programmatically generated. For Word SPS and Visual Entity SPS, we use the same step as for the GQA-SPS dataset. For Referent SPS, we use a different method than the GQA-SPS dataset to generate new questions, as the VQA v2 dataset does not provide the ground-truth of scene graphs. Firstly, we collect some common relationships as a relationship database. For each original sample in the val split of VQA v2, we randomly select five relationships from the database for a noun in the question, and try to artificially annotate a new question by referring the five relationships and the corresponding image. In doing so, the three rules we mentioned in the main manuscript are difficult to meet at the same time, resulting in a small sample size. Finally, we obtain 62556, 9170 and 813 samples for Word SPS, Visual Entity SPS and Referent SPS, respectively.

2 Experiments

Implementation Details For our reimplemented methods that use object features as visual input, including MAC, MAC+SUPS, LXMERT, LXMERT+SUPS and VL-T5, we extracted object features by the official released code of LXMERT. Inspired by the warm start mechanism, we do not apply our framework into the baseline method until the E_s epoch of training. For training MAC+SUPS, we set $E_s = 10$. For finetuning LXMERT+SUPS, ViLT+SUPS, mPLUG+SUPS and BEiT-3+SUPS, we set $E_s = 0$ as they are pretrained. For all five baseline methods incorporated with our framework, we set the loss weight λ_q , λ_v and λ_c as 1, 1 and 0.1, respectively.

2.1 Evaluation on VQA v2, VQA-CP v2 and VQA-SPS v2

The experimental results on VQA v2 [2], VQA-CP v2 [1] and VQA-SPS v2 are shown in Tab. 1, which show that our framework is beneficial for in-distribution (ID) setting (*e.g.*, VQA v2) apart from the OOD setting (*e.g.*, VQA-CP v2, VQA-SPS v2), compared to most existing methods that provide performance gains in the OOD testing at the expense of ID performance, which can be observed from [4]. We mainly focus on the performance gains on compositional substitutivity, which is a special case of OOD generalization, and experimental results on VQA-CP v2 and VQA-SPS v2 have already demonstrated improvements.

Table 1: Accuracy (%) and Consistency (%) on VQA v2 [2], VQA-CP v2 [1] and VQA-SPS v2.

Method	VQA v2	VQA-CP v2	VQA-SPS v2		
			<i>Cons</i> (W)	<i>Cons</i> (V)	<i>Cons</i> (R)
LXMERT	73.01	46.23	58.37	50.80	52.77
+ Ours	73.21	51.43	59.43	51.71	54.74

2.2 Ablation Studies about Support Question Generation

To explore whether mBART [3] is better than instruction-tuned models for generating support questions, we perform experiments with LLaMA3¹. Experimental results in Tab. 2 show that mBART outperforms instruction-tuned models (LLaMA3). The reason is that the questions generated by mBART are more diverse than that generated by LLaMA3, even though we set a high temperature when using LLaMA3, as evidenced by our observations.

Table 2: Ablation studies about support question generation on GQA-SPS, where we use LXMERT [5] as the baseline method.

SUPS(Q)	Word SPS			Visual Entity SPS			Referent SPS		
	<i>Acc1</i>	<i>Acc2</i>	<i>Cons</i>	<i>Acc1</i>	<i>Acc2</i>	<i>Cons</i>	<i>Acc1</i>	<i>Acc2</i>	<i>Cons</i>
LLaMA3	80.6	74.2	70.1	81.2	79.7	73.6	80.4	63.9	58.1
mBART	80.7	74.7	70.9	82.1	80.6	74.8	80.5	64.1	58.4

2.3 Parameter Analysis about $K_q \neq K_v$

We conduct experiments about $K_q \neq K_v$ with LXMERT [5] on our GQA-SPS dataset, to analyse what the influence of the question support set is versus the image support set. As shown in Tab. 3, we can observe the accuracy and consistency fluctuate more when K_q changes, which suggests that the question support set has a greater impact on performance than the image support set.

2.4 Qualitative Analysis

Fig. 1 depicts several qualitative examples that show the effectiveness of our framework for improving compositional substitutivity. The examples come from our GQA-SPS dataset, and we visualize two qualitative examples for each type of synonymous primitive substitution. For the first example in Fig. 1 (a), the original question is “Is the person on the edge of the step wearing a cap?”, LXMERT makes a correct prediction for the original question but makes a wrong prediction when the word “cap” is replaced by its synonym “hat”. By using our

¹ <https://github.com/meta-llama/llama3>

Table 3: Parameter analysis about $K_q \neq K_v$ with LXMERT as baseline on GQA-SPS.

K_q	K_v	Word SPS			Visual Entity SPS			Referent SPS		
		<i>Acc1</i>	<i>Acc2</i>	<i>Cons</i>	<i>Acc1</i>	<i>Acc2</i>	<i>Cons</i>	<i>Acc1</i>	<i>Acc2</i>	<i>Cons</i>
10	6	80.6	74.6	70.7	81.8	80.5	74.7	80.4	63.8	58.1
	10	80.7	74.7	70.9	82.1	80.6	74.8	80.5	64.1	58.4
	14	80.5	74.8	71.1	81.7	80.5	74.7	80.5	63.6	57.9
10	6	80.6	75.1	71.5	81.1	79.5	73.5	80.3	74.4	58.3
	10	80.7	74.7	70.9	82.1	80.6	74.8	80.5	74.4	58.4
	14	80.5	74.7	70.8	81.7	80.4	74.5	80.4	63.8	57.9

method, LXMERT+SUPS makes predictions accurately for both the original question and the question after synonymous word substitution. These qualitative examples prove that our framework is effective in improving the consistency of visual question answering models for synonymous substitutions.

3 Examples from Our Dataset







We visualize several examples from the proposed GQA-SPS dataset in Fig. 2. The GQA-SPS dataset consists of three types of samples generated via synonymous word substitution, synonymous visual entity substitution, and synonymous referent substitution. We provide three examples for each type of sample.

References







1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don’t just assume; look and answer: Overcoming priors for visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4971–4980 (2018)
2. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6904–6913 (2017)
3. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
4. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12700–12710 (2021)
5. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. pp. 5100–5111 (2019)









Fig. 1: Qualitative comparisons between LXMERT+SUPS (Ours) and LXMERT on GQA-SPS. (a) Qualitative examples under word SPS. (b) Qualitative examples under visual entity SPS. (c) Qualitative examples under referent SPS. The red words in questions and the red boxes in images denote synonymous primitives.

Val-A	Val-B
<p>Q: What is the food that is on the chocolate donut? (GT: chocolate)</p> 	<p>Q: What is the food that is on the chocolate doughnut? (GT: chocolate)</p> 
<p>Q: Which side of the photo is the orange box on, the right or the left? (GT: left)</p> 	<p>Q: Which side of the picture is the orange box on, the right or the left? (GT: left)</p> 
<p>Q: What does the boy stand on? (GT: skis)</p> 	<p>Q: What does the male child stand on? (GT: skis)</p> 

(a) Examples of Word SPS

Val-A	Val-B
<p>Q: Is the color of the tree the same as the street sign? (GT: no)</p> 	<p>Q: Is the color of the tree the same as the street sign? (GT: no)</p> 
<p>Q: Is there a pot in the image? (GT: yes)</p> 	<p>Q: Is there a pot in the image? (GT: yes)</p> 
<p>Q: Is there any can to the left of the chair? (GT: yes)</p> 	<p>Q: Is there any can to the left of the chair? (GT: yes)</p> 

(b) Examples of Visual Entity SPS

Val-A	Val-B
<p>Q: How large do you think is the bicycle? (GT: small)</p> 	<p>Q: How large do you think is the ride to the right of the boat? (GT: small)</p> 
<p>Q: Is the pizza black or red? (GT: red)</p> 	<p>Q: Is the food sitting on the plate black or red? (GT: red)</p> 
<p>Q: What is the giraffe licking? (GT: leaf)</p> 	<p>Q: What is the animal to the right of the hand licking? (GT: leaf)</p> 

(c) Examples of Referent SPS

Fig. 2: Examples from GQA-SPS. The red words in questions and the red boxes in images denote synonymous primitives.