We thank all the reviewers for their efforts and constructive comments! Most reviewers (R1, R2 and R4) think our method is novel, while having concerns on clarity and evaluation. Below we address the important and common issues.

1. Supervision and ablation study (R1, R2, R3). • Our method is self-supervised in the sense that it does not rely on any ground-truth visual relationship annotations, avoiding the challenging manual annotation of visual relationships; however, we do not restrict supervision for other learning components. • Our self-supervised relationship probing is based on the assumption that relations mentioned in image descriptions are visually observable. To model it effectively, we use dependency trees on the language side. • Although our framework consists of several modules and multiple losses, it is not complex at the conceptual level. Essentially, it contains two training stages, where the first stage resembles conventional BERT training, and the second stage is the proposed relationship probing. • In addition to Table 3, which shows the effectiveness of data augmentation and the proposed relation probing, we conduct additional ablation study to show the effect of the matching loss $\mathcal{L}_{\text{Match}}$ and the probe loss $\mathcal{L}_{\text{Probe}}$ in Table A. We can see that the matching loss is critical since it can help the model learn meaningful alignments between vision and language entities during training. On the other hand, the probing loss can further help improve the performance.

Table A: Ablation study on NLVR2 Dev set. ✔/✘ indicates presence/absence.

Method	Stage 1, Aug.(✘)		Stage 1, Aug.(✔)		Stage 1+2, Aug.(✔)	
	$\mathcal{L}_{\text{Match}}$(✘)	$\mathcal{L}_{\text{Match}}$(✔)	$\mathcal{L}_{\text{Match}}$(✘)	$\mathcal{L}_{\text{Match}}$(✔)	$\mathcal{L}^{S}_{\text{Probe}}$(✘)	$\mathcal{L}^{S}_{\text{Probe}}$(✔)
$\text{SSRP}_{\text{Share}}$	50.86	60.53	51.69	61.67	62.78	64.25
$\text{SSRP}_{\text{Visual}}$	52.09	69.92	52.51	70.75	71.41	72.03
$\text{SSRP}_{\text{Cross}}$	57.91	74.35	58.54	74.48	75.11	75.71

2. Framework contribution (R1, R3). • While $\text{SSRP}_{\text{Cross}}$ resembles LXMERT, we would like to emphasize that the goal of this paper is not to design a new BERT model for V&L. Instead, we learn visual relationships by self-supervised learning. Particularly, the designed relationship probe with its training process is novel and unique. As mentioned by R4, "this paper introduces a new and fresh idea compared to a lot of minor ablations that we are seeing in V&L pretraining domain". • Besides, our other two methods: $\text{SSRP}_{\text{Share}}$ and $\text{SSRP}_{\text{Visual}}$ are different from LXMERT. LXMERT and other V&L BERT-based models cannot be directly applied (without fine-tuning) to single-modality vision tasks such as image captioning due to the cross-attention used in pretraining, while our $\text{SSRP}_{\text{Share}}$ and $\text{SSRP}_{\text{Visual}}$ can.

3. Analysis of implicit graphs and Fig. 4 (R1, R2, R4). The reviewers raised a good point that it would be better to provide quantitative results for image retrieval. However, it is hard to do so on MSCOCO since there are no officially annotated positive/negative pairs. As suggested by R4, we now report the results by calculating the sentence-level BLEU between captions (query image) and reference captions (retrieved images) in Table B. We see that 'Obj.+Rel.' outperforms 'Obj.', which shows the effectiveness of our implicit visual relationships for single-modality visual tasks.

4. Language augmentation (R2). We use two pivot languages, German (De) and Russian (Ru), and also different beam sizes to achieve diversity. As shown in supp., we can generate diverse captions and preserve the semantic meanings at the same time. We did explore other text augmentations, such as word substitution using BERT, *etc*. However, we found that they severely corrupt the original meanings. *E.g*., given a sentence 'A large passenger airplane flying through the air', the back-translation provides 'A large passenger plane flying in the air', while the BERT-based word substitution gives 'a large passenger airplane flying through at night'.

Table B: Results on 1K query images randomly sampled from MSCOCO. We compute the BLEU scores based on all the associated captions from Top-$\{1, 5, 10\}$ retrieved images.

Method	Top-1		Top-5		Top-10	
	B@1	B@4	B@1	B@4	B@1	B@4
'Obj.'	38.28	6.11	45.37	6.18	48.46	6.28
'Obj.+Rel.'	40.17	6.31	48.84	6.70	52.67	7.10

5. Sampling strategies and larger corpora (R3). For each training iteration, we sample a minibatch of image-caption pairs instead of individual images, as mentioned in Sec.3.3.2. The setting of this work is different from VL-BERT* trained on both visual-linguistic corpus and text-only corpus. This work is for discovering rich implicit visual relations directly from their textual descriptions. Thus, we can only use image-captions as pretraining corpus.

6. Clarity and other issues. R1) • The long-tail distribution of visual relationships mainly comes from the human annotations. Our models are not trained on such annotation. • We will improve the algorithm description and draw a simplified version for Fig. 2. • For the notations, we do not distinguish the input and output contextual features for simplicity. We will use different notations in the revised paper. • The main reason we compare with BUTD is that BUTD is a relatively well-tested framework, and also takes the object detection features as input similar to ours. Meanwhile, our best cider score (126.7) is close to that of SGAE[5] (129.1) which uses ground-truth visual relationships during training. During the submission period, UNITER[33] was not accepted to any venues, and thus we did not compare with it in Tab. 4. We will include it in our revised paper. • As for reproducibility, we will consider releasing the source code later. **R2)** • It is true that we use the object label predicted by the pre-trained object detectors, which can force the model infer the label by exploiting the linguistic clues when the corresponding RoI is masked out. $g(\cdot)$ is a non-linear mapping (learnable) layer (see the supplementary). • 'Outputs the unmasked visual feature' – it means predicting the unmasked features of the masked RoI input. • The text encoder is initialized with BERT pretrained model. • For unaligned image and text inputs, the MLM loss is still applied to each modality, but the alignment prediction training label is set to zero. **R4)** • We will consider low resource tasks and related papers in the revised version as per R4's suggestions. • The MLM loss helps to learn contextualized multi-modal representations in a self-supervised manner via self- and cross-attentions, our relationship probing generates relationship graphs in each modality from the encoded contextual representations. • The relative parameter size differences between the 3 variants are around $\pm 1\%$.