Towards Grounded Visual Spatial Reasoning in Multi-Modal Vision Language Models

Navid Rajabi NRAJABI@GMU.EDU
Department of Computer Science
George Mason University
Fairfax, VA 22030, USA

Jana Košecká KOSECKA@GMU.EDU
Department of Computer Science
George Mason University
Fairfax, VA 22030, USA

Reviewed on OpenReview: *https://openreview.net/forum?id=gXprdL2EIW*

Abstract

Large vision-and-language models (VLMs) trained to match images with text on large-scale datasets of image-text pairs have shown impressive generalization ability on several vision and language tasks. Several recent works, however, showed that these models lack fine-grained understanding, such as the ability to count and recognize verbs, attributes, or relationships. The focus of this work is to study the understanding of spatial relations. This has been tackled previously using image-text matching (e.g., Visual Spatial Reasoning benchmark) or visual question answering (e.g., GQA or VQAv2), both showing poor performance and a large gap compared to human performance. In this work, we show qualitatively (using explainability tools) and quantitatively (using object detectors) that the poor object localization "grounding" ability of the models is a contributing factor to the poor image-text matching performance. We propose an alternative fine-grained, compositional approach for recognizing and ranking spatial clauses that combines the evidence from grounding noun phrases corresponding to objects and their locations to compute the final rank of the spatial clause. We demonstrate the approach on representative VLMs (*such as* LXMERT, GPV, and MDETR) and compare and highlight their abilities to reason about spatial relationships.

Keywords: Visual Spatial Reasoning, Vision-and-Language Models (VLMs), Multi-Modal Transformers, Fine-grained Grounding, Compositional Understanding.

1 Introduction

Visual reasoning is among the general goals of vision-and-language models (VLMs). Recent advances in multi-modal vision-language models were enabled by self-supervised pre-training objectives on web-scale datasets of image-text pairs. The evaluation methodologies usually resort to measuring performance on downstream tasks that include visual question answering (VQA), referring expression comprehension, image-to-text/text-to-image retrieval, image-text matching, or generative/auto-regressive tasks, like image captioning. Multiple recent studies have shown that vision-language models lack fine-grain understanding of verbs Hendricks and Nematzadeh (2021), spatial relationship Liu et al. (2022); Kamath et al. (2023a), word order Thrush et al. (2022), and lack general visio-linguistic compositionality Yuk-

sekgonul et al. (2022) that is critical for compositional reasoning and generalization. In

Figure 1: Although the ground-truth caption for this image is: "The **bed** is **next to** the **bicycle** ", multiple different *spatial clauses* from the language domain can be inferred from the visual domain (image) as to fill out the spatial clauses correctly, like *behind*, *left of*, *near*, *close to*, *touching*, etc. This type of intrinsic ambiguity from the language side makes formulating the spatial reasoning task more challenging.

this work, we focus on visual spatial reasoning task considered in Visual Spatial Reasoning (VSR) benchmark Liu et al. (2022) and show qualitatively, using explainability tools Chefer et al. (2021), and quantitatively, using the performance of object detectors, that a significant weakness of multi-modal models is the ability of grounding subjects and objects of the spatial relationships. We then propose an alternative compositional ranking-based approach for spatial clause understanding built on the top of General Purpose Vision (GPV) multi-modal model Gupta et al. (2022) pre-trained with object-level bounding box supervision and classification, visual question answering, and captioning. Our contributions can be summarized as follows:
i) We analyze the top-performing LXMERT model Tan and Bansal (2019)) on VSR benchmark Liu et al. (2022) and identify problems with noun grounding and the image-text matching (ITM) evaluation methodology.
ii) We decompose spatial clauses into simpler primitives by grounding and localizing *subject* and *object* using an encoder-decoder GPV model Gupta et al. (2022) and train a separate, lightweight model for predicting their *spatial relation*. This yields a more traditional structured approach for combining the evidence about objects and spatial relations and ranking the spatial clauses.
iii) The approach is evaluated on a subset of the VSR benchmark, demonstrating the effectiveness of our strategy for spatial clause understanding and the suitability of our ranking-based approach that avoids the biases of VQA/GQA setting and achieves better relative performance compared to random chance.

2 Probing LXMERT on VSR

We start by establishing LXMERT's image-text matching baselines on the VSR dataset by conducting the fine-tuning experiments in two different settings: (1) Fine-tuning the binary classification head only by freezing the pre-trained LXMERT weights and (2) Fine-tuning

both LXMERT and its classification head end-to-end, similar to what was done in Liu et al. (2022). According to Table 1, the second setting, indicated as `End-to-End`, achieves the

Model	**ZS**	**Rand**
Random Chance	50	50
LXMERT Fine-tuning (`CLS-Head`)	54.0	60.0
LXMERT Fine-tuning (`End-to-End`)	**65.6**	**74.1**

Table 1: VSR's test set performance after fine-tuning on VSR train set. **ZS** and **Rand** stand for Zero-shot and Random splits in the VSR benchmark, respectively.

closest performance to what is reported in Liu et al. (2022). This rather low performance is an indicator of the difficulty in efficient adaptation of the LXMERT-style VLMs for more complex reasoning tasks. Other models have similar difficulties. More recently, authors in Herzig et al. (2023) proposed alternative pre-training of BLIP-2 Li et al. (2023) with additional structured supervision provided by scene graph annotations. While this pre-training helped to achieve the accuracy of 63.4% on the VSR dataset, it still underperformed the LXMERT end-to-end fine-tuning accuracy. Additional advancements in the pre-training objectives and data reported in Bugliarello et al. (2023) using larger X-VLM model with object localization head Zeng et al. (2021), still underperforms smaller LXMERT model on VSR Random split test set.

Quantitative Noun Grounding Analysis. Due to the straightforward grammatical structure of the captions in the VSR dataset, we can extract *subject*, *relationship*, and *object* from the captions with the minimal effort. Next, we quantify the performance of the subject/object detector on the VSR dataset. The rows marked 1. to 3. in Table 2 focus on cases when LXMERT's image-matching prediction is right, given that one, both, or none of the subject & object is/are being detected by Faster-RCNN. For example, row 1. refers to the cases in which the binary ITM is Successful (indicated as **S**). but only in 27.08% of the examples both subject and object phrases were detected category labels of Faster-RCNN outputs and 30.41% of them were WordNet Miller (1995); Bird (2006) synonyms, not the exact word match. The same applies to the Rand columns, while in row 2, in 56.66% of the successful binary ITM predictions, only one of the subject and object phrases were among the detections. A similar analysis for the incorrect image-text matching prediction can be

Case	**S**	**One**	**Both**	**ZS**	**Rand**
Rand	–	–	–	50	50
Acc	–	–	–	65.66	74.11
1.	✓	✓	✓	27.08 / 30.41	27.93 / 32.93
2.	✓	✓	✗	56.66 / 55.41	53.13 / 50.73
3.	✓	✗	✗	16.25 / 14.16	18.93 / 16.33

Table 2: LXMERT's results on original VSR test set Liu et al. (2022) and Faster-RCNN error analysis. ZS refers to the zero-shot setting in VSR, in which there is no concept overlap between the train/dev/test splits, while in Rand, all the data is distributed randomly. *Acc* stands for accuracy as % of correctly predicted ITM binary labels.

found in Table 5 in the Appendix. While these numbers are not directly indicative of the

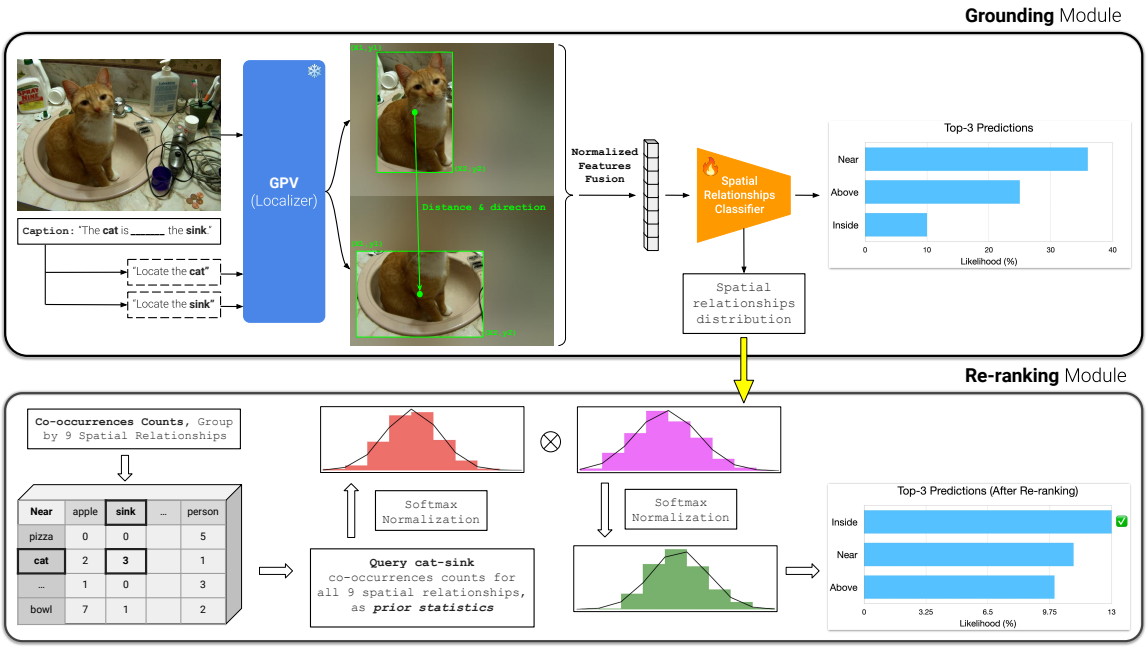


Figure 2: **Our Approach** consists of two main modules: (1) **Grounding Module** predicts the locations of objects along with their confidences, and MLP takes the bounding box coordinates and predicts the distribution of spatial relationships. These are then combined to compute the initial ranking of spatial clauses. (2) **Re-ranking Module** adjusts the ranking given the co-occurrence priors. This example shows the effectiveness of the Re-ranking Module in adjusting the spatial clause distribution, which brings *inside* to the 1st rank, while the initial top-3 predictions were semantically correct, anyway.

ultimate ability of the model to ground subject/object nouns with ROI feature tokens, the weak correlation between them is inferred, as it was also studied in previous works like Frank et al. (2021) and suggests challenges with subject and object grounding in VSR. Additional qualitative analysis using the explainability method of Chefer et al. (2021) can be found in Figure 5 in the Appendix. This analysis also confirms that a significant factor in understanding spatial relationships in images is the lack of noun grounding. Furthermore, the fact that image and text can be matched successfully in the absence of noun phrase grounding skews the conclusions made by ITM probing tasks.

3 Approach

We propose a ***compositional*** approach for spatial reasoning by ***decoupling*** the process of (1) grounding `Subject`, and `Object` of the <`Subject`, `Relationship`, `Object`> triplet extracted from the ground-truth caption and (2) predicting the `Relationship` by training a multi-class classification model using location features obtained by the detector. This approach is motivated by earlier structured Conditional Random Fields (CRF) methods for modeling image content using scene graphs, like in Johnson et al. (2015) that applied it to image retrieval tasks. We used encoder-decoder vision-language models, MDETR Kamath

et al. (2021) and GPV Gupta et al. (2022), and their ability to query the locations of objects. Our approach is outlined in Figure 2.

Grounding Module takes as input an image and spatial clause in the form of <`Subject`, `Relation`, `Object`> and queries $GPV_{Localizer}$ with the prompt of Q_i = "*Locate the Subject*", getting back normalized bounding box coordinates as $l_i = [x_i, y_i, h_i, w_i]$ and confidence score $p(i) = Pr(o_i|I, Q_i)$ of the most confident prediction given the query Q_i. Similarly for $p(j) = Pr(o_j|I, Q_j)$ with Q_j = "*Locate the Object*". The concatenation of the bounding box coordinates is fed to the MLP to generate the initial probability distribution over all r_k spatial relationships[1]. $\left\{Pr^1_{ij}, ..., Pr^k_{ij}\right\} = MLP([l_i, l_j])$, where $Pr^k_{ij} = Pr(r_k|l_i, l_j)$. The score of the spatial clause is then calculated using a simple scoring function where score $S_k(I, T)$ is computed as follows:

$$S_k(I, T) = p(i)Pr(r_k|l_i, l_j)p(j).$$

Re-ranking Module then uses the prior probability distribution $Pr[R_k(i, j)]$ of two objects i and j appearing in certain spatial relation k, computed by counting all prior co-occurrences of i and j appearing in relation k. The final relation probability is then ranked by $r_k(i, j) = Pr(r_k|l_i, l_j)Pr[R_k(i, j)]$ yielding the final ranking function:

$$S_k(I, T) = p(i)r_k(i, j)p(j).$$

4 Experiments

In terms of data pre-processing, we first discarded the orientation-based clauses like *"facing away"* and *"parallel to"*, which require the pose or 3D information. VSR also reported this class of spatial clauses as the worst-performing one, even below the random chance, and the depth-based clauses like *"in front of"*, *"behind"*, that require depth signal. Then, we grouped the semantically similar spatial clauses in VSR together and ended up with nine merged classes as *below*, *above*, *far from*, *right of*, *left of*, *inside*, *outside*, *near*, and *contains*, to have more controlled experiments over primitive spatial clauses. The details of spatial clauses merging/grouping can be found in Table 6. In total, we ended up with 3895 data samples, which we divided randomly into 3116 (80%) and 779 (20%) splits using `StratifiedSampling` to be used as the gold train and test sets for our experiments. We also experimented with feeding additional geometric features of normalized relative distance & direction from the center of subject towards the center of object. Data augmentation by swapping subjects and objects for symmetric clauses, and swapping accompanied by negating asymmetric spatial clauses has been tried as well as the additional feature input fusions to the MLP for spatial relationship classification.

Results Analysis. Table 3 shows the performance of our ranking approach, trained and tested on a subset of VSR. Our top-performing settings of 54.04% and 88.96% perform 42.93% and 55.63% above the random chances reflecting the % of times the correct spatial relationship is the *top-1* or among the *top-3* ranked spatial clauses.

1. $k = 9$ in our case.

Model	Top-1	Top-3
Random Chance$_{Multiclass}$	11.11	33.33
Bbox w/o Re-ranking	41.97	83.05
+ *geo*	40.17	82.54
+ *geo* + *aug*	42.36	80.61
Bbox w/ Re-ranking	53.78	**88.96**
+ *geo*	**54.04**	88.31
+ *geo* + *aug*	**54.04**	87.54
Our Best Δ	**42.93**	**55.63**

Table 3: Ablation study of our approach results on our test set, comparing to the random chance performance similar to the Thrush et al. (2022) evaluation style. "*geo*" means adding three geometry-based features to MLP, and "*aug*" means data augmentation for our training split of bounding box locations.

VQA/GQA Baselines Comparison In order to have a fair comparison on the VSR benchmark of the state-of-the-art VLM's with different architectures and pre-training strategies we chose LXMERT from *Cross-Modal*, MDETR from *Modulated*, and GPV from *Encoder-Decoder* Transformers categories. Since VSR benchmark is curated as image-text pairs with binary labels for ITM, we modified the image-text matching into a "yes/no" VQA/GQA task by turning the captions into questions using the same triplet information. In this way, "yes" is considered as **1** and "no" as **0** binary labels. Therefore, we can run the experiments on models that do not have the ITM classification head by design. The results of golden test set (both for Zero-shot and Random split of VSR) and are shown in Table 4.

Model	Binary	ZS	Random
Random Chance$_{Binary}$	50	50	50
GPV VQA Gupta et al. (2022)	58.27	52.25	54.69
MDETR GQA Kamath et al. (2021)	**95.63**	53.35	54.54
LXMERT GQA Tan and Bansal (2019)	95.37	**57.71**	**55.23**
SOTA Best Δ	**45.63**	**7.71**	**5.23**

Table 4: Pre-trained SOTA results on our test set (indicated as **Binary** on the left) vs. on VSR test set (indicated as **ZS** and **Random** on the right).

Although these models, specifically their GQA classification heads, are already fine-tuned on multiple V+L tasks, they still perform slightly above the random chance in the zero-shot inference. MDETR *pre-trained* on GQA Hudson and Manning (2019) achieves the highest performance on our test set, performing 45.63% above the binary random chance. Note that the biggest improvement was marked by MDETR GQA and LXMERT GQA which are already pre-trained on GQA. We hypothesize that GQA contains more spatial clauses that resemble our modified dataset. Comparing our best performance after re-ranking, which is shown as **Our Best Δ = 55.63** in Table 3, with the best performing model of the baselines, shown as **SOTA Best Δ = 45.63** in Table 4, it demonstrates that our approach

outperforms the best SOTA (MDETR GQA) by 10%, in terms of the relative accuracy over the random chance, on the same test set.

5 Related Works

Recent years have witnessed a fast-paced emergence of different multi-modal vision and language architectures and pretraining strategies. We first review the representative vision-language models and briefly discuss their strengths, weaknesses, and limitations pertaining to the study of the visual spatial reasoning task.

Cross-Modal Encoder Transformers follow the architecture of ViLBERT Lu et al. (2019), LXMERT Tan and Bansal (2019), UNITER Chen et al. (2020), OSCAR Li et al. (2020), and VinVL Zhang et al. (2021). The visual input in these models is typically tokenized using object proposals and their associated region of interest features (and bounding box coordinates) obtained by pretrained object detectors such as Faster-RCNN. While these models achieve impressive performance using smaller amounts of training data (∼9.7 million image-text pairs) for pretraining, they are not end-to-end trainable, and their performance on downstream tasks is affected by the quality of the detected regions.

Several probing studies using specially curated datasets demonstrated that these models lack the understanding of attribution, relations, order, verbs, and counting Yuksekgonul et al. (2022); Hendricks and Nematzadeh (2021); Thrush et al. (2022); Kamath et al. (2023a). Since LXMERT was the best-performing model on the VSR Liu et al. (2022) benchmark, we include this model in our experiments as the representative baseline from this class of architectures.

Dual-Encoder Transformers were first introduced by CLIP Radford et al. (2021) use contrastive learning and a large dataset of image-text pairs (∼400 million) and to align holistic image and text representations, as well as ALIGN Jia et al. (2021) (using ∼1.8 billion image-text pairs). While CLIP demonstrated high performance on image, scene, and action classification benchmarks, multiple probing studies of fine-grained understanding have demonstrated poor performance on tasks that require more compositional reasoning. For example, Subramanian et al. (2022) showed on the CLEVR dataset that if a spatial clause is added to the caption, the performance of image-text matching is at the level of random chance. Follow-up works combine using holistic representations of Dual-Encoders with tokenized representations of Cross-Modal Encoders such as in ALBEF Li et al. (2021), or adding single modality pretraining to this new architecture like FLAVA Singh et al. (2022). The lack of compositionality and fine-grained understanding still remains, according to the findings of Thrush et al. (2022).

Modulated Transformers are trained end-to-end and can predict bounding box coordinates associated with noun phrases in the text description but require correspondences between bounding boxes and noun phrases for the training stage. MDETR Kamath et al. (2021) is representative of this category, which is built on top of DETR Carion et al. (2020) and trained for fine-grained alignment between regions of interest in the image and associated noun phrases using the Hungarian matching algorithm. Additional pre-training tasks

include VQA and GQA with their own classification heads. GLIP Li et al. (2022b) follows a similar approach but uses a different architecture using a series of cross-attention layers from a language encoder and transformer-based object detection module. Due to the specific capabilities of MDETR (i.e. grounding and pre-trained GQA head), we used this model as another baseline for our experiments.

Encoder-Decoder Transformers were introduced by the General Purpose Vision (GPV) Gupta et al. (2022) model. This model has object localization functionality, can do image classification, captioning, and VQA tasks. The task is specified as input to the model in natural language, enabling simultaneous multi-task training. The noun phrase localization task has bounding box ground truth available during training, while other tasks such as classification, question answering, and captioning have ground truth text associated with images. Other task-agnostic vision-language models that are end-to-end trainable on multiple tasks and have generative decoder branches for auto-regressing text include Li et al. (2022a); Wang et al. (2021); Yu et al. (2022). The localization and VQA capabilities of GPV make this model suitable for our experiments. We built our ranking model on top of the GPV Localization module, which is demonstrated in Figure 2.

Spatial Relationship Understanding. Previous works on spatial relationship understanding use the synthetic CLEVR Johnson et al. (2017) dataset, focus on simpler spatial relationships and neglect the challenges in real images posed by object detection and representation learning. More general approaches study spatial relationships as part of VQA Antol et al. (2015) or GQA Hudson and Manning (2019) tasks. According to the GQA benchmark, only 8% and 22.4% of VQA and GQA are allocated to *Spatial Reasoning* questions, respectively. The existing GQA questions that probe spatial understanding typically have binary YES/NO answers and hence inherit the well-known biases of the VQA/GQA task. VSR Liu et al. (2022) is the most recent dataset curated for studying spatial clause understanding in a more visually realistic setting using MSCOCO Lin et al. (2014) images. The authors report performance on image-text matching tasks by fine-tuning the existing models (LXMERT, VisualBERT, and ViLT) both in zero-shot setting and after fine-tuning.

6 Conclusions and Future Works

In this work, we demonstrated the low zero-shot performance of several state-of-the-art VLMs on the spatial reasoning task and demonstrated that this is partly caused by the lack of grounding of noun phrases in models pre-trained with image-text matching using holistic representations of each modality, like in LXMERT Tan and Bansal (2019). We then proposed a compositional approach for spatial reasoning using outputs of the GPV encoder-decoder model with explicit quantification of subject and object grounding and spatial relationship classification/ranking. This approach outperformed the SOTA models in terms of the relative increase over the random chance for each setting, which is a similar evaluation strategy used in multiple probing papers like Thrush et al. (2022) and Kamath et al. (2023b). Another advantage of our approach is the modularity that makes it possible to replace the localization module with any upcoming SOTA Vision-and-Language model with the grounding/localization capability in the future and just re-train the lightweight spatial relationship classification head. In the presented approach, the reasoning is done

purely in 2D with decoupled locations and region ROI features. Further disambiguation of more complex relations requires knowledge of 3D, which can be either revealed by depth and object pose estimates or novel larger datasets, additional means of supervision, or variations in model architectures. Another drawback of the VSR dataset is the lack of ground-truth bounding box annotations for subjects and objects. Densely-annotated datasets like Li et al. (2024) could be considered in curating future benchmarks for studying spatial reasoning. In this work, we considered the GPV localization `relevance scores` as silver labels for the grounding confidences, as the majority of scores fall between 0.8 and 1.0 for the entire dataset. The distribution histogram of the relevance scores can be found in Figure 7 in the Appendix.

Acknowledgments

Most of the experiments of this work were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University.

References

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

S. Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

E. Bugliarello, A. Nematzadeh, and L. A. Hendricks. Weakly-supervised learning of visual relations in multimodal pretraining. *arXiv preprint arXiv:2305.14281*, 2023.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

H. Chefer, S. Gur, and L. Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

S. Frank, E. Bugliarello, and D. Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

T. Gupta, A. Kamath, A. Kembhavi, and D. Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16399–16409, June 2022. URL `https://openaccess.thecvf.com/content/`

CVPR2022/html/Gupta_Towards_General_Purpose_Vision_Systems_An_End-to-End_Task-Agnostic_Vision-Language_Architecture_CVPR_2022_paper.html.

L. A. Hendricks and A. Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

R. Herzig, A. Mendelson, L. Karlinsky, A. Arbelle, R. Feris, T. Darrell, and A. Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*, 2023.

D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Kamath_MDETR_-_Modulated_Detection_for_End-to-End_Multi-Modal_Understanding_ICCV_2021_paper.pdf.

A. Kamath, J. Hessel, and K.-W. Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023a.

A. Kamath, J. Hessel, and K.-W. Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023b.

J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022a.

J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022b.

X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

Y. Li, N. Rajabi, S. Shrestha, R. Alimoor, and J. Košecká. Labeling indoor scenes with fusion of out-of-the-box perception models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 578–587, 2024.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

F. Liu, G. Emerson, and N. Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022. URL `https://arxiv.org/pdf/2205.00363.pdf`.

J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022.

H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Y. Zeng, X. Zhang, and H. Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

Appendix A.

A.1 Implementations Details

For the LXMERT reproduction results, we used the `huggingface` Wolf et al. (2019) checkpoint for the first setting and VSR codebase for the second setting, respectively. For the first setting, pass the 768-dim `Cross-Modality CLS Token` into a 3-layer MLP with the size of (768, 512, 64, 1), followed by `sigmoid` activation for binary classification. Then, the model is trained for 200 epochs and saved the checkpoint based on the early stopping of the validation loss. For the localization part of our model, we used `GPV-1-Loc` checkpoint, pre-trained on `COCO` split. For spatial relationship reasoning, our MLP model consists of (1) an input size of 8, (2) two hidden layers of 16 and 32, and (3) an output layer with nine neurons, respectively. Both hidden layers are followed by `BatchNorm1d` layers, then `ReLU` activation. Training has been done for 100 `epochs` until the convergence, with the `batch-size` of 12 and `learning-rate` of 1e-5, using the `CrossEntropyLoss` criterion and `Adam` optimizer. All the above scripts are mainly developed using `Pytorch` and `transformers` libraries.

A.2 LXMERT Zero-shot vs. Fine-Tuning Settings

Provided in Figure 3.

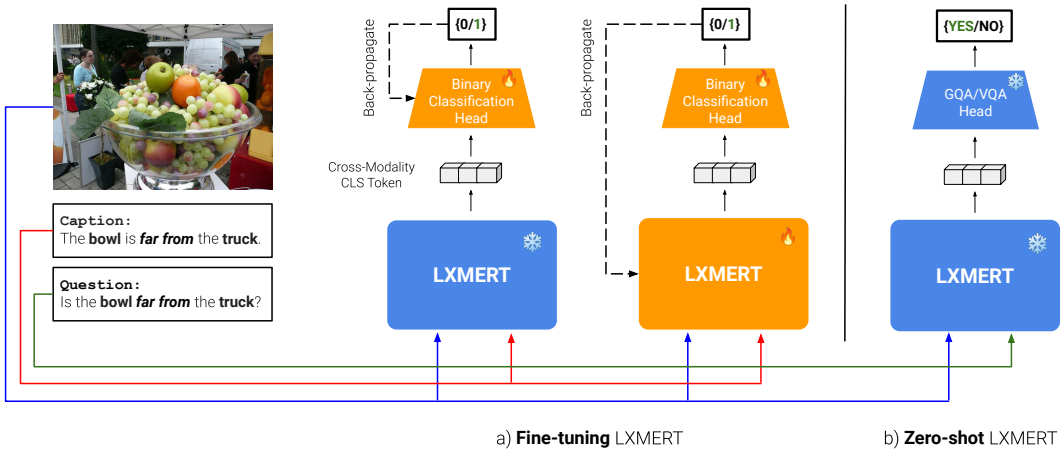


Figure 3: LXMERT fine-tuning and zero-shot performance reproduction experiments design.

A.3 Learning Curves of Fine-tuning LXMERT on VSR (CLS-Head Only)

Provided in Figure 4.

A.4 Full LXMERT Quantitative Results

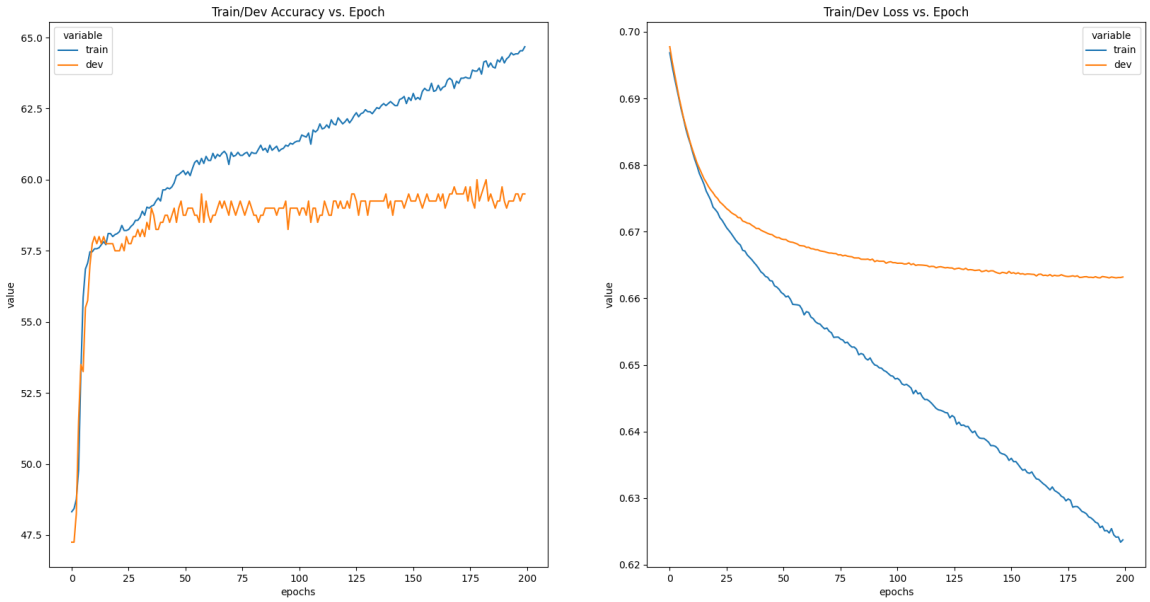
Provided in Table 5.

Case	S	One	Both	ZS	Rand
Rand	–	–	–	50	50
Acc	–	–	–	65.66	74.11
1.	✓	✓	✓	27.08 / 30.41	27.93 / 32.93
2.	✓	✓	✗	56.66 / 55.41	53.13 / 50.73
3.	✓	✗	✗	16.25 / 14.16	18.93 / 16.33
1-*Acc*	–	–	–	34.34	25.89
4.	✗	✓	✓	29.48 / 34.26	25.19 / 28.43
5.	✗	✓	✗	58.56 / 54.98	54.19 / 52.86
6.	✗	✗	✗	11.95 / 10.75	20.61 / 18.70

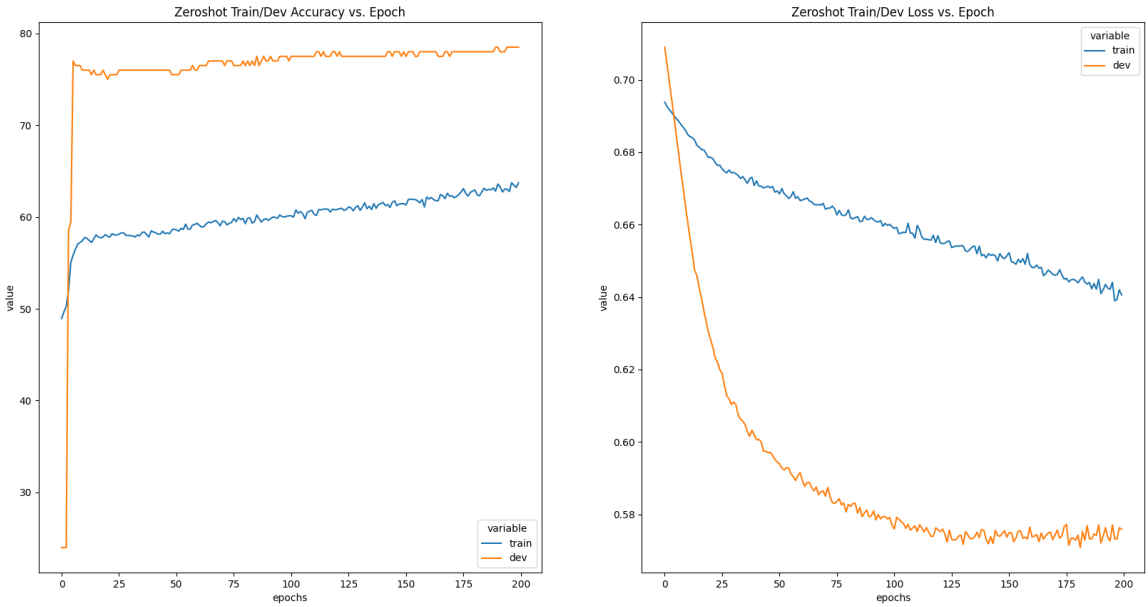
Table 5: LXMERT's results on original VSR test set Liu et al. (2022) and Faster-RCNN error analysis. ZS refers to the zero-shot setting in VSR, in which there is no concept overlap between the train/dev/test splits, while in Rand, all the data is distributed randomly. (*Acc* stands for accuracy as % of correctly predicted ITM binary labels.

A.5 LXMERT Explainability Results Visualization

Provided in Figure 5. Four different cases for which LXMERT predicts image-text labels successfully in all of them. The first column shows the original image-text pairs from the VSR dataset. The second and third columns demonstrate the regions with the highest relevancy score computed from attention weights using Chefer et al. (2021), for *subject* and *object*, respectively.



(a) Random Split



(b) Zero-shot Split

Figure 4: Learning curves for LXMERT classification head fine-tuning on VSR train set.

A.6 Spatial Relationship Groupings

Provided in Table 6.

(a) The cup is to the left of the hot dog. (b) Locate the **cup** (c) Locate the **hot dog**

(d) The potted plant is far away from the bus. (e) Locate the **potted plant** (f) Locate the **bus**

(g) The bowl is beneath the spoon. (h) Locate the **bowl** (i) Locate the **spoon**

(j) The cup is touching the bed. (k) Locate the **cup** (l) Locate the **bed**

Figure 5: **LXMERT Relevancy Scores**: The first row shows an example that both *subject* and *object* attentions imply successful grounding. The second row demonstrates relevant activations for the subject (*potted plant*) but irrelevant attention weights for *bus*. The third and fourth rows depict irrelevant attention weights for both subjects and objects, demonstrating inconsistency in LXMERT's fine-grained grounding while predicting the binary labels correctly.

Merged	Original
below	*below*, *beneath*, *under*
above	*above*, *on*, *on top of*, *over*
far from	*away from*, *far away from*, *far from*
right of	*at the right side of*, *right of*
left of	*at the left side of*, *left of*
inside	*in*, *in the middle of*, *inside*, *part of*, *within*
outside	*outside*
near	*adjacent to*, *at the edge of*, *at the side of*, *attached to*, *beside*, *by*, *close to*, *connected to*, *near*, *next to*, *touching*
contains	*contains*

Table 6: Our spatial clauses merging/mapping. In our setting, 32 spatial relationships are covered, and merged into 9 coarse-grained spatial clauses to avoid confusion.

A.7 Confusion Matrix of Spatial Relations Classifier

Provided in Figure 6.

(a) **Before** re-ranking

(b) **After** re-ranking

Figure 6: Sample confusion matrices are provided to demonstrate the re-ranking effect, for the setting where only BBoX coordinates are used.

A.8 GPV Localization Confidences (`Multimodal Relevance/Objectness Scores`)

Provided in Figure 7.

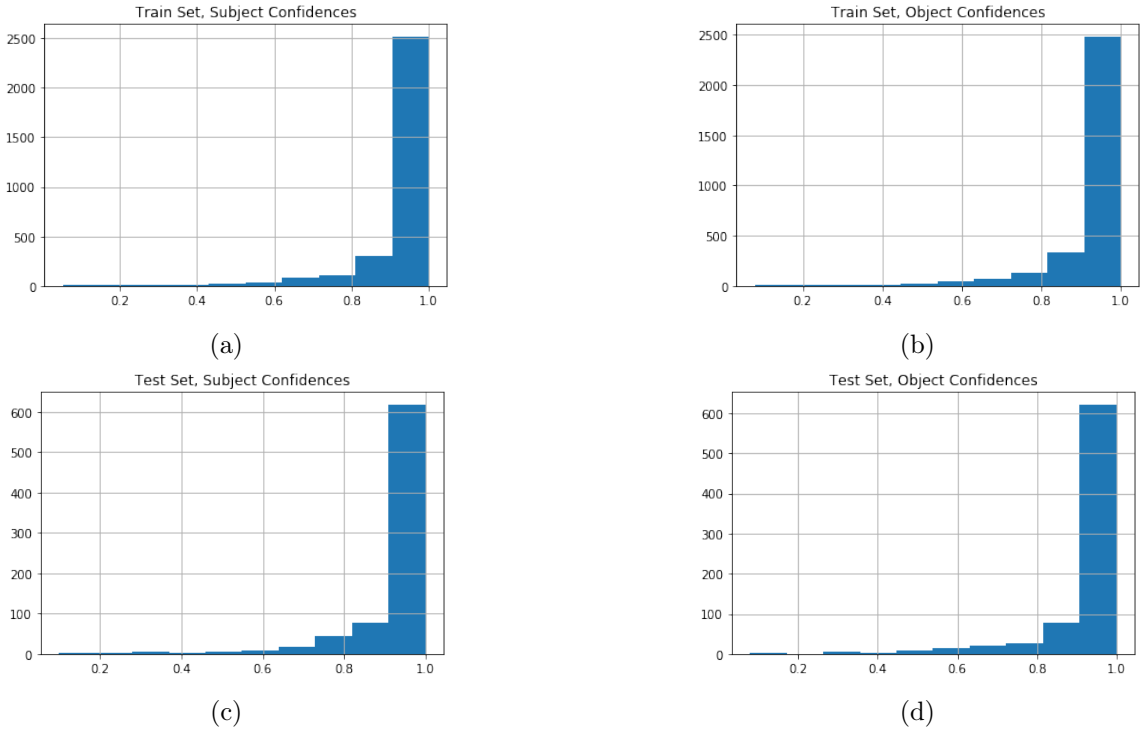


Figure 7: GPV localizer subject and object relevance scores for our train and test splits.