Lost in Space: Probing Fine-grained Spatial Understanding in Vision and Language Resamplers

Georgios Pantazopoulos[1,2] **Alessandro Suglia**[1,2] **Oliver Lemon**[1,2] **Arash Eshghi**[1,2]

[1]Heriot-Watt University; [2]Alana AI

{gmp2000, a.suglia, o.lemon, a.eshghi}@hw.ac.uk

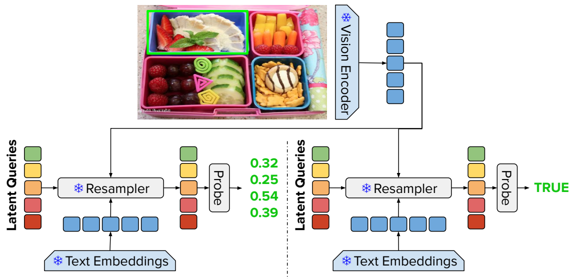
Abstract

An effective method for combining frozen large language models (LLM) and visual encoders involves a *resampler* module that creates a 'visual prompt' which is provided to the LLM, along with the textual prompt. While this approach has enabled impressive performance across many coarse-grained tasks like image captioning and visual question answering, (Alayrac et al., 2022; Dai et al., 2023), more fine-grained tasks that require spatial understanding have not been thoroughly examined. In this paper, we use *diagnostic classifiers* to measure the extent to which the visual prompt produced by the resampler encodes spatial information. Our results show that this information is largely absent from the resampler output when kept frozen during training of the classifiers. However, when the resampler and classifier are trained jointly, we observe a significant performance boost. This shows that the compression achieved by the resamplers can in principle encode the requisite spatial information, but that more object-aware objectives are needed at the pretraining stage to facilitate this capability[1].

1 Introduction

Recent approaches for developing Vision and Language (V&L) models leverage existing vision (Radford et al., 2021; Fang et al., 2023b,a), and language experts (Touvron et al., 2023a; Zhang et al., 2022; Touvron et al., 2023b) and try to learn a mapping between them (Alayrac et al., 2022; Li et al., 2023b; Dai et al., 2023; You et al., 2023; Liu et al., 2023c,b). In most cases, the experts are kept frozen while the only learnable component is the mapping between the visual and the language expert.

The simplest approach uses a linear projection layer that matches the dimensionality of the visual and textual embeddings before feeding them to the LLM (Liu et al., 2023c,b). A more sophisticated method is to use a *resampler* to compress the visual embeddings into a compact 'visual prompt' that is then fed to the LLM either at the input level along with the text prompt (Li et al., 2023b; Dai et al., 2023) or via cross attention layers (Alayrac et al., 2022; Li et al., 2023a). From a practical standpoint, the resampler may accelerate training and inference as it significantly reduces the sequence length, but also facilitates in-context learning capabilities since additional examples can fit into the context window of the LLM. As a result, these approaches have demonstrated impressive performance across multiple 'coarse-grained' tasks such as image captioning, and visual question answering.



Locate the region that is described by: *'Strawberries and cream in an fruit/snacks tray'.*

The phrase *'Strawberries and cream in an fruit/snacks tray'* refers to the **top left** part of the image.

Figure 1: *Explicit* (left) and *implicit* (right) probing for spatial understanding. In the explicit setting, we probe for region localization, while in the implicit setting, the probe is trained to classify whether a description involving an image region is true of the image.

However, fine-grained tasks such as visual grounding and spatial understanding are relatively underexplored. Resamplers are usually pretrained on pairs of image-text data using contrastive learning (Li et al., 2023b; Dai et al., 2023), and/or multimodal masked language modeling (Laurençon et al., 2023; Alayrac et al., 2022), without relying on object-aware objectives. Given the importance of resamplers for the development of V&L models, we ask whether this compression preserves

[1] Code available here

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 540–549
June 16-21, 2024 ©2024 Association for Computational Linguistics

fine-grained spatial information. Do the contrastive and language modeling objectives retain the overall scene structure, or is this information lost due to the absence of object-aware pretraining objectives?

To address these questions, we train diagnostic classifiers to probe two different resampler modules for *explicit* and *implicit* spatial understanding — see Figure 1. Our results indicate that the multimodal resamplers do not facilitate spatial understanding. Nevertheless, in all settings, jointly fine-tuning the diagnostic classifiers and the resamplers significantly boosts performance, demonstrating that the compression achieved by the resamplers can in principle encode the requisite spatial information, but that more object-aware pretraining objectives are needed to facilitate this.

2 Related Work

Resamplers The idea of the resampler is inspired primarily by computer vision, where an attention mechanism is used to compress visual features into learnable queries (often referred to as slots) (Carion et al., 2020; Kamath et al., 2021; Locatello et al., 2020). More recently, resamplers have been applied to more multimodal tasks. Flamingo (Alayrac et al., 2022) and subsequent open-source variants (Laurençon et al., 2023; Li et al., 2023a) are based on the Perceiver Resampler (Jaegle et al., 2022), with cross-attention between the latent queries and the visual embeddings followed by a stack of self-attention blocks that operate on the latent queries. In the Q-Former (Li et al., 2023b; Dai et al., 2023), the latent queries are also informed by the input text and, therefore, create a more 'linguistically informed' visual prompt.

Probing Probing is a class of methods for interpreting neural models by assessing whether the model representations encode specific kinds of information at different processing stages (Belinkov, 2022). The concept of probing is straightforward; we extract representations from a model that is already trained on some task(s), and use a lightweight *diagnostic classifier* on top of these representations to solve a probing task that reflects the information that we seek to find. The classifier's performance is then taken to correlate with the extent to which that information is encoded by the model (Conneau et al., 2018; Hupkes et al., 2018). Many within (multimodal) NLP have thus adopted probing to interpret model behavior (Kajic and Nematzadeh, 2022; Salin et al., 2022; Lindström et al., 2020).

3 Experiments

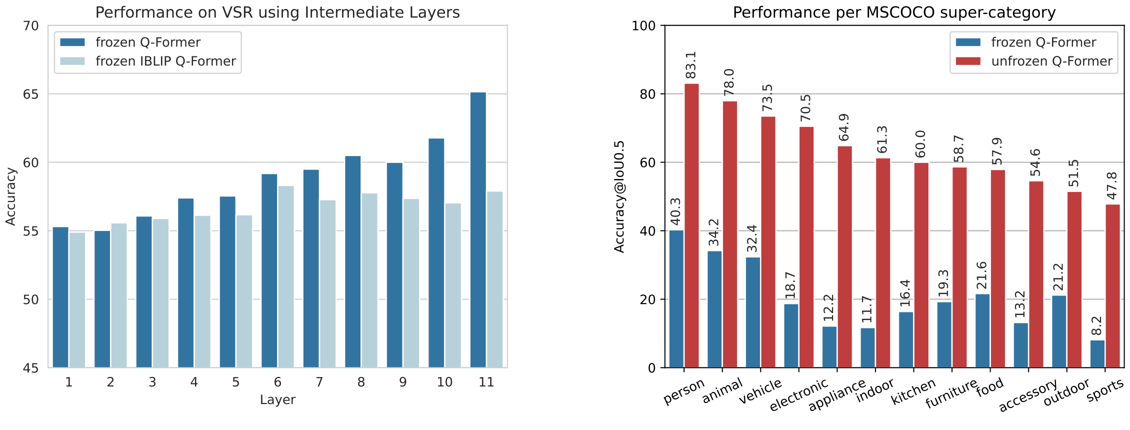
Is spatial understanding a property of V&L resamplers? We experiment with three different spatial understanding tasks. In RefCOCOg (Mao et al., 2016), the objective is to predict the coordinates of the region that is described by the input phrase. Secondly, we use the 'random split' from the VSR dataset (Liu et al., 2023a), where the model has to assess the validity of a caption describing a spatial relationship between two entities. Finally, we introduce the Region Cell Matching (RCM) task, which follows the VSR formulation but is designed to test for a more rudimentary form of spatial understanding regarding the location of one entity in the image. Inspired by CAPTCHAs, an image is divided into a 3x3 grid, and each grid cell is assigned a location description (such as top left or middle). We generate synthetic captions by combining RefCOCOg descriptions with the cell location as shown in the implicit probing example of Figure 1. To ensure that performance is not influenced by frequency biases, we balanced the distribution of positive and negative examples. Appendix A contains further details about the dataset.

In our experiments, we use the Q-Former from the first pretraining stage of BLIP2 (Li et al., 2023b) and InstructBLIP (Dai et al., 2023). To probe the resamplers, we follow past work (Belinkov, 2022) and use a single linear layer after flattening the embeddings of the query tokens. For RefCOCOg, the linear layer predicts the normalized coordinates of the region that matches the referring expression. We use the bounding box loss from (M)DETR (Carion et al., 2020; Kamath et al., 2021): a weighted sum of the Generalised IoU and L1 losses. Similarly, for VSR and the RCM task, we use a linear layer that predicts the probability that the query matches the image trained using binary cross entropy. We tune the learning rate, number of epochs, and loss weights (only for RefCOCOg) using Bayesian hyperparameter optimization (Bergstra et al., 2013) for at least ten iterations. For further implementation details, see Appendix B. In all cases, we evaluate the best model in terms of validation performance.

We compare the two resamplers against similarly-sized models that employ patch representations. We avoid comparison against models with object-centric visual encoding because the task of visual grounding is significantly easier in these models as they need to select the correct can-

	RefCOCOg		VSR random		RCM	
	Validation	Test	Validation	Test	Validation	Test
Random	-	-	-	50.00	50.00	50.00
Human	-	-	-	95.40	-	92.29
MDETR (Kamath et al., 2021)	83.35	83.31	-	-	-	-
CLIP* (Radford et al., 2021)	-	-	-	56.0	-	-
Unitab (Yang et al., 2022)	84.58	84.70	-	-	-	-
ViLT (Kim et al., 2021)	69.14	68.93	71.38	71.53	83.16	83.25
❄ Q-Former	30.39	30.26	66.91	64.97	70.12	69.49
🔥 Q-Former	71.47	71.72	80.86	80.50	81.68	81.35
❄ IBLIP Q-Former	20.00	19.92	58.07	55.72	64.58	63.08
🔥 IBLIP Q-Former	68.89	69.34	78.40	76.99	83.11	80.86

Table 1: Linear probing results. ❄/🔥 denotes that the resampler is frozen/unfrozen. * results from Liu et al. (2023a).



(a)

(b)

Figure 2: Performance on (a) VSR per intermediate layer, (b) RefCOCOg per MSCOCO super-category.

didate bounding box provided from the detector as opposed to explicit image region prediction. Additionally, we provide results where the linear classifier is jointly trained along with the resampler as an upper bound for the performance with frozen representations.

Table 1 shows the results for the models that we are considering. We observe that both resamplers perform poorly on RefCOCOg when kept frozen, and, therefore, are unable to perform explicit visual grounding. A possible counter-argument could be that predicting raw coordinates within the image is too difficult to solve with a single linear layer. However, we observe similar trends with VSR and RCM, which test for spatial understanding in an easier binary classification setup. While the resamplers perform better than random baselines in these tasks, there is a significant gap between the performance of the frozen and fine-tuned backbones. We believe this is an outcome of the pretraining objectives of the Q-Former that do not explicitly facilitate fine-grained object-centric representations. This is in line with previous work, which found that V&L models trained with contrastive objectives act as bag-of-words and do not preserve spatial information (Yuksekgonul et al., 2022). On the other hand, the significant boost achieved by unfreezing the resamplers shows that the compression of the input embeddings is, in principle, able to capture spatial information and, therefore, that the resampler as an architectural choice does not necessarily constitute a bottleneck.

Is spatial information encoded in earlier layers but discarded in deeper layers? We previously observed that resamplers have poor performance in

Category	Adjacency	Directional	Orientation	Projective	Proximity	Topological	Unallocated
❄ Q-Former	61.94	42.05	56.93	62.87	60.15	74.56	68.42
🔥 Q-Former	68.86	75.00	67.15	78.29	81.95	83.94	72.37
❄ IBLIP Q-Former	57.44	38.64	58.39	54.21	40.60	66.14	52.63
🔥 IBLIP Q-Former	62.98	68.18	67.88	74.61	78.95	83.15	77.63

Table 2: VSR results per model for different categories of spatial relationships. ❄/🔥 denotes that the resampler is frozen/unfrozen.

spatial understanding tasks when using representations from the last layer. Next, we examine if the representations from intermediate layers better encode spatial information. Intuitively, representations from earlier layers could lead to greater probing performance as they are closer to the visual encoder's output. Figure 2a shows the results on VSR after probing representations from intermediate layers. Overall, intermediate layer representations do not provide performance gains. There is a clear upward trend regarding the performance of the Q-Former from BLIP2, whereas for InstructBLIP we observe fluctuations within a small range across layers. A similar trend is observed in the RefCOCOg results which are included in Appendix C.

Scaling the Probing Classifier Additionally, we experiment with scaling the probe classifier by introducing non-linearities. In particular, we use 2-layer and 4-layer classifiers with SwiGLU activation functions. We refrain from using more complex classifiers because they may infer features that are not actually used by the underlying model (Hupkes et al., 2018). For training, we used the same setup as with our previous experiments.

Table 3 illustrates the results with increasing prompt complexity. While we observe a common trend of increasing performance when we make the probe more complex, the accuracy of the non-linear probes does not indicate that the resampler encodes spatial information which can be easily retrieved. Additionally, the performance gap between the simplest and the most complex probe in the case of InstructBLIP indicates that fine-grained spatial understanding is 'built-up' within the probe and is not necessarily a property of the resampler component.

Model	#Layers	RefCOCOg	VSR random	RCM
❄ Q-Former				
	1	30.26	64.97	69.49
	2	32.08	65.15	69.98
	4	34.49	65.01	70.71
❄ IBLIP Q-Former				
	1	19.92	55.72	63.08
	2	25.01	58.09	68.66
	4	34.49	59.09	69.29

Table 3: Probing results by scaling the probing classifier.

3.1 Discussion

Performance analysis per object category Figure 2b illustrates the Q-Former's performance on RefCOCOg per MSCOCO (Lin et al., 2014) super-category. We observe that the frozen/unfrozen resamplers behave differently but also have significant variation between object categories. To further understand the possible reasons for this variation, we computed the Kendall coefficient (Kendall, 1938) between the performance of each super-category and 1) the distribution of train examples, 2) the area of each bounding box, 3) and the distance of the bounding box from the center of the image (Table 5). Interestingly, the main factor that correlates positively with the performance per category is the area of the bounding box. We also observe that the further the bounding box deviates from the center, the more the performance drops. These two observations imply that the Q-Former constructs the visual prompt by 'summarizing' the most central entities within an image, ignoring positional outliers.

Which spatial relationships are difficult to capture? In Table 2, we break down the VSR results according to the spatial relationship type. Both resamplers perform the best in topological relations across frozen/unfrozen conditions. Directional relations seem challenging for out-of-the-box resamplers, though this relation can be captured during fine-tuning. Finally, captions describing adjacency or orientation properties are difficult even for fine-tuned resamplers.

Effect of learning objectives We showed that multimodal resamplers pretrained with contrastive learning and multimodal language modeling objectives do not capture spatial information well. These are undoubtedly important objectives as they enable large-scale pretraining, however, on their own, they are not sufficient for enabling fine-grained spatial understanding.

Finally, we observed that BLIP-2's Q-Former consistently outperformed the one from InstructBLIP. However, as shown in Figure 2a, the performance of the two resamplers is comparable for early layers. We hypothesize that during instruction tuning, the InstructBLIP Q-former may get away with providing even less fine-grained information since the language modeling loss is already low due to the high-quality LLM, leading to a forgetting effect (McCloskey and Cohen, 1989).

4 Conclusion

In this paper, we explored to what degree multimodal resamplers preserve spatial information. While previous work has demonstrated the effectiveness of resamplers across a variety of V&L tasks, our investigation revealed their limitations when applied to spatial understanding tasks. In particular, we probed two resamplers and showcased that grounding natural language descriptions in image regions is not an inherent ability of these modules. Furthermore, probing experiments showed limited spatial understanding in two easier settings. These involved image-text matching with captions referencing the absolute location of an entity, or spatial relationships between two entities. Nevertheless, our results showcased that when the resampler is fine-tuned, the compression of the visual encoding induced by the resampler can be effective. We believe that this is due to the lack of an object-aware pretraining objective that would encourage the resamplers to encode spatial information. Future work should build upon our findings and design objectives that incentivize disentangled representations (Bengio et al., 2013).

Limitations

This study centered on exploring some architectural components of current V&L models with regard to their ability to encode spatial information. For the purpose of our study, it is necessary that the visual and textual representations are already fused. Models adopting unimodal resamplers do not facilitate this because 1) the fusion happens only in the successive cross-attention layers of the LLM (Alayrac et al., 2022), or 2) the visual embeddings are concatenated with the text embeddings at the input of the LLM (Bai et al., 2023). While we could extract representations from intermediate layers from a model like IDEFICS (Laurençon et al., 2023), this would have been an unfair comparison with BLIP-2 style models because the former adds more layers to the original resampler architecture. The other option would be to provide the visual embeddings and the text embeddings to the probe, but this defeats the purpose of the probing classifier as probe since it would have to perform the necessary multimodal fusion internally; thus making any comparisons uninterpretable. Consequently, our study does not encompass the entirety of available models adopting resamplers, and the findings may not be fully representative of the broader V&L model landscape.

We also recognize the limitation in our exploration of spatial understanding as an emergent ability in V&L models. The question of whether spatial understanding materializes as a natural consequence of model scale remains unanswered in our study. A more in-depth investigation controlling the pretraining dataset, the size of the models as well, and the training hyperparameters is required in order to truly understand the capacity of these models to develop fine-grained and disentangled representations that facilitate spatial understanding.

Acknowledgements

We would like to thank the reviewers for their valuable feedback during the ARR process. Additionally, we would like to thank Malvina Nikandrou, and Ioannis Konstas for their suggestions with regards to the experimental setup. This work was supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA. PMLR.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $ &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *ArXiv*.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023a. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023b. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. 2023. Deep learning tuning playbook. Version 1.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2022. Perceiver io: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*.

Ivana Kajic and Aida Nematzadeh. 2022. Probing representations of numbers in vision and language models. In *SVRHM 2022 Workshop@ NeurIPS*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.

Adam Dahlgren Lindström, Johanna Björklund, Suna Bensch, and Frank Drewes. 2020. Probing multimodal embeddings for linguistic properties: the visual-semantic case. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

LMSYS ORG. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer.

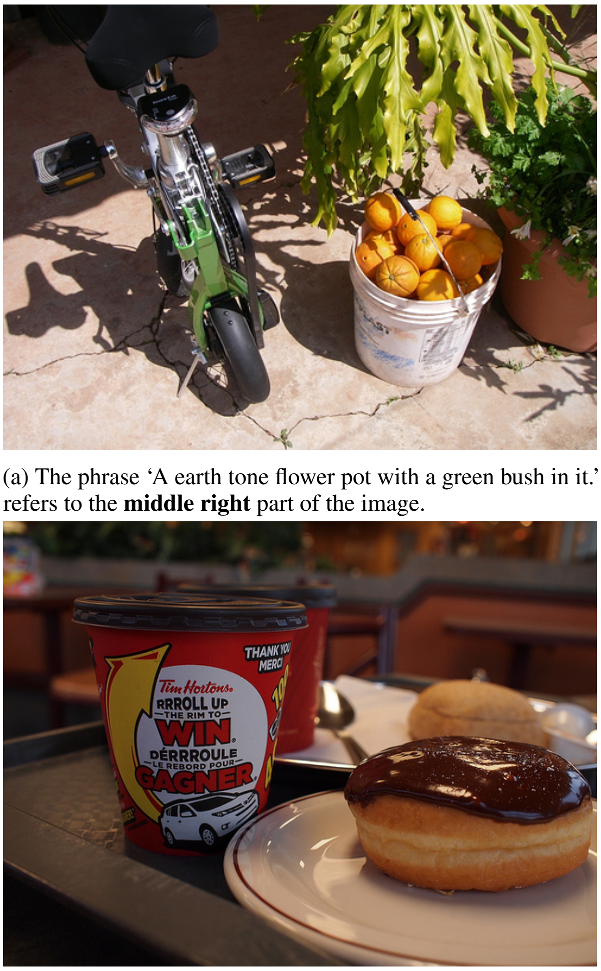
Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Region Cell Matching

The purpose of Region Cell Matching (RCM) is to evaluate a model's capacity to perform visual grounding in an implicit manner, i.e., the model does not need to provide a specific region within an image, but it has to substantiate if a given description belongs to a certain region within an image. To make the task even easier, this region is not arbitrary, it corresponds to one of the cells of a 3×3 grid on top of the image. Each of these cells is mapped to a natural language description, for example, top left, middle, top right, etc. Figure 3 illustrates one positive and one negative example from the dataset.



(a) The phrase 'A earth tone flower pot with a green bush in it.' refers to the **middle right** part of the image.

(b) The phrase 'A tan and brown donut with a thick coating of chocolate on top.' refers to the **middle** part of the image.

Figure 3: Illustration of positive (a) and negative (b) examples from the RCM task.

To create the dataset, we started from RefCOCOg examples and assigned each bounding box to one of the cells within the grid by matching its center to the closest cell. To prevent overpopulating the dataset with examples where the bounding box is centered, we downsampled the dataset so that the distribution of the cells is balanced. With this process, we created a subset of N positive examples that are evenly distributed between the 9 cells. In order to prevent biases related to the distribution of the cells we additionally created N negative examples as follows: For each grid cell i with N_i positive examples we selected $N_i/8$ from every other cell j as negative examples. We repeat the steps for train, validation, and test sets resulting in 46k, 3k, and 5.5k samples, respectively.

Human Performance Apart from fine-tuning ViLT, we established a human baseline by estimating the performance of humans in the task. We developed a Gradio interface (Abid et al., 2019) where participants received the input image, the region description as well as the assigned cell and they were asked to provide a binary response to the question 'Does the phrase match the location in the image?'. In order to imitate the training and evaluation setting in our experiments, we did not provide any additional information (e.g., there was no visible grid on top of the image as this would have trivialized the task) to the participants, with the exception of a few introductory examples before actually completing the task.

Since a region may overlap with multiple grid cells, we also gave participants the option to provide up to 4 grid cells ranked in terms of priority. Additionally, participants may refrain from answering the question if the phrase is factually incorrect (e.g., the phrase 'A dog with a frisbee' is factually incorrect if there is no dog within the image). We decided to include this option to avoid any potential confusion and introduce unnecessary noise to the annotation.

We recruited a total of five participants who were informed about the study and the use of their data. Each participant annotated 100 examples from the test set (50 positive / 50 negative). To estimate a human baseline, we removed the instances where each annotator assigned either multiple cells or labeled an instance as factually incorrect. Finally, we measured the annotator agreement with the Fleiss' kappa coefficient (Fleiss, 1971): $k = 80.98$.

B Implementation Details

In our experiments we used BLIP2's (Li et al., 2023b) Q-Former from the first pretraining stage which is pretrained using contrastive, image-text matching, and masked language modeling losses.

Task	Hyperparameters		Q-Former		IBLIP Q-Former	
	Name	Value	❄	🔥	❄	🔥
RefCOCOg	lr	[1e-5, 5e-4]	4.85e-4	1.03e-4	2.55e-4	1.08e-4
	epochs	{20, 30, 40}	40	20	40	40
	GIoU scale	{1, 2x, $x \in \{1, \ldots, 10\}$}	6	20	16	20
	L1 scale	{1, 2x, $x \in \{1, \ldots, 10\}$}	20	18	18	8
VSR	lr	[1e-5, 5e-4]	3.92e-4	4.59e-4	1.03e-4	2.49e-5
	epochs	{3, 5, 10, 15, 20}	5	10	15	20
RCM	lr	[1e-5, 5e-4]	1.94e-5	4.74e-4	4.34e-4	3.10e-5
	epochs	{50, 100, 150}	100	150	150	50

Table 4: Hyperparameters used during bayesian optimization. Additionally, we performed early stopping for RCM with a patience of 10 epochs.

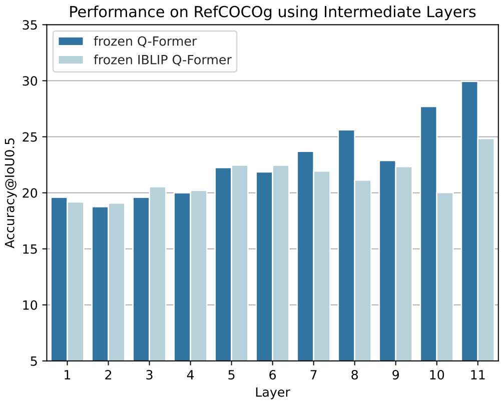


Figure 4: Performance of Q-Former on RefCOCOg per intermediate layer.

In this stage the Q-Former is trained as a standalone component, i.e, there is no language modeling loss from an LLM. For InstructBLIP (Dai et al., 2023), we used the Q-Former that is trained to prompt the Vicuna-7B model (LMSYS ORG, 2023).

For all experiments we used AdamW optimizer with weight decay of 0.01 and 10% warmup. We used a fixed batch size of 128 and tuned exclusively the learning rate and the number of steps following (Godbole et al., 2023). For RefCOCOg we also tuned the scale of GIoU and L1 loss. Table 4 shows the hyperparameters that were tuned, their minimum and maximum values, and the best configuration for frozen and unfrozen resamplers. All training logs regarding the main experiments as well as the experiments using intermediate representations are available here.

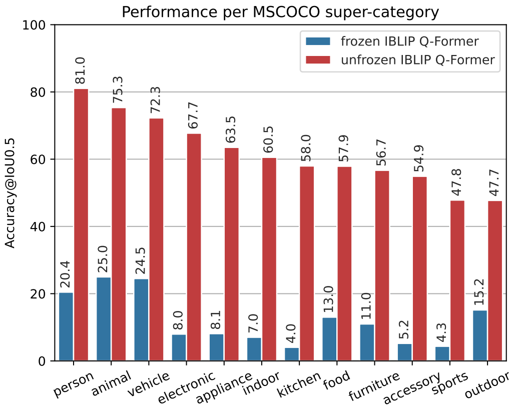


Figure 5: Performance of InstructBLIP Q-Former on RefCOCOg per MSCOCO super-category.

	Q-Former		IBLIP Q-Former	
	❄	🔥	❄	🔥
# examples (train)	0.63	0.42	0.33	0.42
Area (test)	0.84	0.45	0.66	0.45
Distance (test)	-0.51	-0.42	-0.63	-0.42

Table 5: Kendall correlation coefficient between performance of resamplers and 1) # training examples, 2) bounding box area of test examples, and 3) distance between the center of the bounding box and the center of an image of test examples. Numbers illustrate p-values greater than 0.05.

C Additional Results

For completeness, Figure 4 shows the results on RefCOCOg after obtaining the representations of the queries from intermediate layers. We observe a similar pattern as in Figure 2a, where there is a clear boost when obtaining the representations from deeper layers from the BLIP2's Q-Former but in the case of the InstructBLIP we observe fluctuations in the performance.

RefCOCOg performance analysis per object category In order to better understand variations in performance between the different object categories, we used the distribution of 1) # training examples, 2) bounding box area of test examples, and 3) distance between the center of the bounding box and the center of an image of test examples. Table 5 shows the Kendall correlation coefficient between the performance on different super-categories and the three conditions.

Relationship between performance of probe and the visual LLM With regards to the relation between probing and the performance of the visual LLM, we prompted InstructBLIP on VSR and RCM with the prompts reported in the original paper and ranked the logits for positive / negative answers. The performance of the InstructBLIP model is 61% on VSR and 51% on RCM. While this is a performance increase in the case of probing on VSR, it shows that even the full stack of the MLMM is unable to robustly retrieve spatial information from the compressed visual sequence. In the case of RCM we observe a notable drop which we assume is due to the lack of any similar tasks during the instruction-tuning phase.