Data Efficient Masked Language Modeling for Vision and Language

Yonatan Bitton◇   **Gabriel Stanovsky**◇   **Michael Elhadad**♠   **Roy Schwartz**◇

◇School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

♠ Department of Computer Science, Ben Gurion University, Israel

{yonatanbitton,gabis,roys}@cs.huji.ac.il   elhadad@cs.bgu.ac.il

Abstract

Masked language modeling (MLM) is one of the key sub-tasks in vision-language pretraining. In the cross-modal setting, tokens in the sentence are masked at random, and the model predicts the masked tokens given the image and the text. In this paper, we observe several key disadvantages of MLM in this setting. First, as captions tend to be short, in a third of the sentences no token is sampled. Second, the majority of masked tokens are stop-words and punctuation, leading to under-utilization of the image. We investigate a range of alternative masking strategies specific to the cross-modal setting that address these shortcomings, aiming for better fusion of text and image in the learned representation. When pre-training the LXMERT model, our alternative masking strategies consistently improve over the original masking strategy on three downstream tasks, especially in low resource settings. Further, our pre-training approach substantially outperforms the baseline model on a prompt-based probing task designed to elicit image objects. These results and our analysis indicate that our method allows for better utilization of the training data.[1]

1 Introduction

Pre-trained vision-language (VLP) models such as ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019) and UNITER (Chen et al., 2020) have recently improved the state-of-the-art across various vision and language benchmarks. One of the primary pre-training objectives of VLP is masked language modeling (MLM). Motivated by the single-modal MLM task, most models perform as introduced in BERT (Devlin et al., 2019) for text-only data, randomly masking tokens with a probability of 15% (Shin et al., 2021).
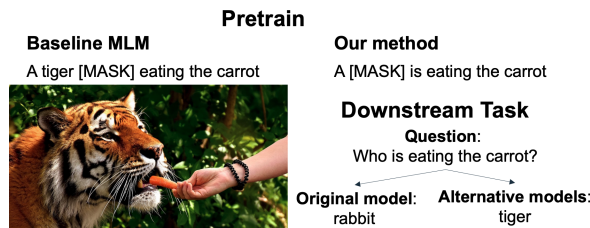

Figure 1: Illustration of our approach. The baseline MLM masks a random token with 15% probability, where ≈50% of the masked tokens are stop-words or punctuation. Our method masks words that require the image in order to be predicted (e.g., physical objects). Our pre-train masking strategy consistently improves over the baseline strategy in two evaluation setups.

The main difference in the cross-modal setting[2] is that the model takes into account both the textual context and the image, and the latter can help it resolve ambiguities. For example, in Figure 1, given the masked sentence "A [MASK] is eating the carrot", without the image, the model might predict *rabbit*, since it is correlated with *carrot*. But the image reveals that the answer is *tiger*.

In this work, we find that the MLM pre-training method is sub-optimal for VLP, as it does not make efficient use of the training data. This manifests in two major shortcomings, common to many popular pre-train datasets (Lin et al., 2014; Krishna et al., 2017; Sharma et al., 2018; Ordonez et al., 2011). First, we observe that image captions, which form the textual part of these corpora, tend to be much shorter than the documents in BERT's pre-train data. As a result, uniformly masking tokens at 15% probability results in many cases where no token is being masked (e.g., about one third in LXMERT).

Second, we note that 45%–50% of the masked tokens are stop-words or punctuation. While this seems a common phenomena also in text-only

---

[1]Our code, pre-trained, and fine-tuned models are published at https://github.com/yonatanbitton/data_efficient_masked_language_modeling_for_vision_and_language.

[2]This task is often referred to as "cross-modality MLM", or "MLM conditioned on image regions" (Chen et al., 2020), to emphasize the difference from the text-only MLM task.

3013

datasets, we show that this causes the image to be under-used in MLM pre-training for VLP. Evidently, for the popular LXMERT model, we find that the MLM validation accuracy on stop-words and punctuation is almost perfect even when omitting the image.

To address these limitations, we propose alternative strategies aiming to mask words that require the image (e.g., physical objects). We pre-train the LXMERT model with these strategies and demonstrate their benefits in two evaluation setups. First, on three VLP downstream tasks (GQA, Hudson and Manning, 2019; VQA, Goyal et al., 2017; NLVR2, Suhr et al., 2019), our masking strategies consistently improve over the traditional MLM, especially in low resource settings. Second, we experiment with prompt based object detection (Radford et al., 2021), a probing task designed to elicit image objects by presenting the pre-trained models with prompts such as "A photo of [MASK]" and compare their top predictions with image objects. Our results show that our strategy substantially improves over the baseline sampling approach, even when trained over only a third of its epochs and half of its training data.

In our analysis, we introduce a new metric (Δ *image loss*) to estimate the necessity of the image for a masked word during MLM. We extract the Δ *image loss* value for each token in LXMERT validation pre-train data. We then present a hierarchy of semantic classes ranked by this metric, and find that the frequently masked tokens in our strategies indeed increase the image necessity.

Our main contributions are: (1) We show that the current MLM pre-training method is sub-optimal for VLP, and it does not make efficient use of pre-train data. (2) We propose alternative masking strategies, and show that models trained with these strategies outperform the baseline strategy in two evaluation setups, especially in low resource settings. (3) We introduce the Δ *image loss* metric, which aims to explain the relation between a masked token and the image; we publicly release the computed values of this metric for the LXMERT validation set; this data may be used in future work to devise improved masking strategies.

2 Limitations of MLM Approaches for Vision and Language

In this section, we present the limitations of the MLM approach to vision and language tasks. We start by reviewing the way MLM is currently applied in cross-modal models, and analyzing the pre-train datasets used by most models. We observe the following two major limitations in the current approach: (1) no token is masked in roughly a third of the sentences; (2) a substantial part of the masked tokens are stop-words or punctuation, which can be predicted based on textual context alone, and do not require the image.

2.1 Background

Multiple studies have been proposed to modify the MLM objective in text-only domains (Joshi et al., 2020; Sun et al., 2019; Clark et al., 2020; Levine et al., 2021). However, less research has been dedicated to the implications of MLM in vision and language tasks.

Shin et al. (2021) recently reviewed how the transformer architecture (Vaswani et al., 2017) has been incorporated into vision-language cross-modal tasks. They show that most VLP models perform MLM in the same way as introduced in BERT (Devlin et al., 2019) for text-only data, randomly masking tokens with 15% probability. Further, virtually all models are pre-trained on a handful of pre-training cross-modal datasets, including Conceptual Captions (CC; Sharma et al., 2018); SBU captions (Ordonez et al., 2011) and the LXMERT pre-train dataset, which is a combination of COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), VQA (Goyal et al., 2017), VG-QA (Zhu et al., 2016), and GQA (Hudson and Manning, 2019).

Importantly, all these datasets consist of <sentence, image> pairs, where the sentence is usually a caption describing the image or, in VQA, an image-related question.

2.2 Limitations

In many cases, no token is masked. Image captions tend to be shorter than the documents in BERT pre-train data, such as Wikipedia articles. BERT input sequence length is 512 tokens, while in VLP datasets the sequence length is ≈20 tokens. For this reason, when masking 15% of the tokens in the VLP models, there are cases where *no token* is masked. For example, in LXMERT we find that in 36% of the sentences, no token is masked.

Many masked words are stop-words and punctuation. We observe that over 45-50% of tokens masked by either LXMERT, CC, and SBU are stop-

words or punctuation marks.[3] We now describe an experiment that shows that this distribution causes the image to be under-utilized during MLM pre-training.

We follow the approach of amnesic probing (Elazar et al., 2021). The intuition is that if the image is being used for cross-modal MLM, then the removal of the image should negatively influence the ability of the model to solve the task. If the removal of the image has little or no influence on the ability to solve cross-modal MLM, then the image is not a contributing factor in this task.

We consider the published pre-trained LXMERT model.[4] We evaluate it at inference time with the MLM task twice: with and without the image,[5] using different masking strategies. We use the LXMERT pre-train validation data (≈214K sentences). To estimate the image necessity for a masked token during MLM, we introduce the Δ *image loss* metric, which is the difference in validation loss of the model prediction with and without the image. For example, in Figure 2, the loss *without the image* for predicting "motorcycle" is 3.96, and the loss with the image is 0.25, the Δ *image loss* is 3.71. In addition, we report the *Accuracy@5* metric, which is whether the label is among the top 5 most confident predictions of the model. We compare three masking strategies, keeping a 15% probability to mask a token: (1) Baseline MLM masking strategy, where a token is masked uniformly at 15% probability; (2) masking only stop-words and punctuation; and (3) masking only content words, which is the complementary group of stop words and punctuation.

Results are presented in Table 1. We observe that the model validation accuracy on stop-words and punctuation is almost perfect (96%) even without the image. On the other hand, in the case of content words, accuracy is much lower without the image, and adding it increases accuracy by roughly 20%.

3 Alternative Masking Strategies

To overcome the limitations presented in the previous section, we introduce several alternative masking strategies for cross-modal MLM. The proposed strategies use several semantic classes, which are introduced in Section 3.1, and then used in Section 3.2.

3.1 Semantic Classes

Objects*, *Attributes*, and *Relationships   We use the definitions of *objects*, *attributes*, and *relationships* as described in Visual Genome (Krishna et al., 2017). *Objects* represent physical entities in the image (e.g., a tiger, or a carrot). *Attributes* are properties of objects, such as colors or physical state (e.g., upright). Finally, *relationships* connect between two objects. These can be actions (e.g., a tiger is *eating* a carrot), spatial relations (e.g., the tiger is *behind* the carrot), etc.

In order to mask the tokens that belong to those semantic classes, we first need to identify them in a given sentence. Some datasets (e.g., GQA) include scene-graph annotations of these classes for each image. We use the annotations as ground-truth and develop heuristics to identify them automatically. For example, an *Object* can be reliably annotated by identifying nouns which are also in the Visual Genome objects list. This simple heuristic achieves an accuracy of ≈90% and recall of ≈97% for ientifying objects on the LXMERT pre-train dataset. We elaborate on these heuristics in Appendix A.1.

Concreteness   We hypothesize the image contributes more when predicting concrete concepts (e.g., tiger) compared to abstract concepts (e.g., hunger). To that end, we use a dataset of lexical concreteness presented in (Brysbaert et al., 2014). This dataset provides concreteness scores (on a scale of 1-5) for over 91% of the lemmas in LXMERT pre-training dataset.

3.2 Proposed Strategies

We consider the following masking strategies:

- *Baseline MLM*: the original masking strategy as defined in the LXMERT paper, 15% random token masking.

- *Objects*: Randomly mask one object word.[6]

- *Content words*: Mask exactly one word in each sentence. Instead of almost 50–50 partition between masking stop-words and content words, increase the probability to mask content word to 80%.

---

[3]We used nltk and gensim stop words lists.

[4]https://github.com/airsplay/lxmert

[5]Without the image, we block access to the image and use the model as a single-stream model, without the co-attention layers from the image to the text. The model receives only the text and needs to complete the masked tokens.

[6]In > 97.2% of the sentences there is at least one object. In other cases, we mask a word at random.

Sentence	A person performs a stunt jump on a [MASK].
Masked token	motorcycle
Top 5 predictions	motorcycle, bike, ramp, bicycle, cycle
Top 5 predictions w/o image	building, wall, beach, field, street
Loss	0.25
Loss w/o image	3.96
Δ *image loss*	3.71

Figure 2: An example from the extracted Δ *image loss* data. The masked word is *motorcycle*. Model predictions ("Top 5 predictions") are better correlated with the image when it is given, and the loss is 0.25. Without the image, the predictions ("Top 5 predictions w/o image") are tokens that do not appear in the image, and the loss is much higher (3.96). The Δ *image loss* is the gap: 3.71.

Masking strategy	With Image		Without Image		Image Necessity	
Metric	image loss (exp)	Accuracy @ 5	image loss (exp)	Accuracy @ 5	Δ *image loss* (exp)	Accuracy @ 5
Baseline MLM	3.2	89%	8.9	78%	5.7	10%
Stop-words & punctuation, 15%	1.5	98%	2.9	96%	1.4	2%
Content words, 15%	9.4	76%	38.7	56%	29.3	20%

Table 1: Performance of the LXMERT model on the MLM task, when different words are masked, with and without the image. Accuracy on stop-words and punctuation is almost perfect even when no image is present. However, for content words, the image does contribute to increased accuracy.

- *Top concrete*: Mask one of the top concrete words in the sentence, weighted by their order.[7]

- *Stop-words & punctuation*: as baseline, mask only stop-words & punctuation, keeping a 15% probability of masking.

- *Random 1 word*: An ablation of masking a single random word.

Tokenization: The words in the sentences are tokenized using BERT tokenizer. For strategies requiring word-level masking (Objects, Content words, Top concrete, Baseline MLM, Random 1 word), we mask all of the corresponding word-pieces (e.g., "A tiger is eat #ing" is masked as "A tiger is [MASK] [MASK]").

4 Experiments

To evaluate the value of our proposed strategies, we conduct experiments by pre-training models with different masking strategies and evaluate them on two evaluation setups. We describe the experimental setups below.

4.1 Downstream Tasks

Experimental setup We pre-train the LXMERT architecture with the proposed masking strategies, experimenting with increasing amounts of pre-training data (10%, 20%, 50%, 100%), training for 7 epochs.[8] All other hyper-parameters are the same as the original implementation. We only modify the MLM objective, fine-tuning on three downstream tasks (VQA, GQA, NLVR2). For VQA and GQA, we report the mean of two experiments with different random seeds. The NLVR2 dataset is smaller ($\approx$10% of GQA), so we report three experiments with different random seeds. Following common practice (Tan and Bansal, 2019), we test GQA on the *test-dev* split; NLVR2 on the public test set *test-P*; and VQA on the *minival* split. See corresponding papers for more details.

---

[7]Of the three words with the highest concreteness value in the sentence, mask the most concrete word with 55% probability, the second most concrete with 30% probability, and the third most with 15% probability.

[8]While the published LXMERT model was pre-trained for 20 epochs, we pre-train for 7 epochs because we conduct multiple pre-train experiments, and prefer to spend our budget on more experiments than a few very expensive ones.

Published LXMERT Objects	bathroom, beach, city, kitchen, woman
	motorcycle, bathroom, parade, man, crowd
Ground truth objects	glasses, gang, motorcycle, shirt, man, parade, ...

Figure 3: Example of top 5 predictions for the prompt based object detection task, for the prompt "A photo of a [MASK]". Green underline indicate that the model predicted an object that appear in the ground truth objects (obtained from the scene graph). The model trained with *Objects* masking strategy is more responsive to the image content compared to the baseline model.

Results Figure 4 presents our downstream tasks results.[9] For brevity, we focus on the *Objects* masking strategy, though the trend is similar for the other alternative strategies. We observe that our alternative masking strategies consistently outperform the *Baseline MLM* strategy, especially in low resource settings. Pre-training with the *Objects* strategy yields gains of 0.72–0.86% on VQA and GQA, and 4% on NLVR2 with 10% of the pre-train data; 0.64–0.95% gains on VQA and GQA, and 1.35% on NLVR2 with 20%; 0.5–1.02% gains on VQA and GQA, and 1.6% in NLVR2 with 50%. With 100%, the improvement is minor in GQA, VQA, but still noticeable (1.08%) on NLVR2 (The Content words strategy achieves 0.49 gain on GQA with 100%). [10]

Ablation studies The gains observed when using our proposed strategies can result from both changes we made to address the limitations of standard MLM presented in Section 2: masking a single word in each sentence (rather than not masking any word in some cases) and deciding which word to mask (rather than randomly masking tokens). To isolate the contributing factors, we design additional experiments. We pre-train with 10% and 20% of the data with the *random 1 word* strategy, and present the mean accuracy on the VQA and

GQA in Figure 5. We see that this strategy outperforms the *Baseline MLM* strategy, but underperforms *Objects*. In addition, in Appendix B we show experiments of varying masking probabilities rather than the baseline's 15%, with and without multiple masked tokens per sentence, and allowing sentences without any masked token. Out of all tested settings, masking a single word achieves the best downstream results. We conclude that the benefit of our proposed strategies comes from both choosing a single word to mask, and masking tokens that are more important.

For completeness, we experiment with the *stopwords & punctuation* strategy with 10% and 20% of the data on VQA and GQA. As expected, this strategy under-performs the *Baseline MLM*; by 1.4% when pre-training with 10% of the data, and 3.37% with 20% the data.

4.2 Prompt Based Object Detection

To further examine the value of our proposed masking strategies, we examine in what way the pretrained models trained with different strategies differ. To do so, we use prompts, and study whether a model trained for only completing *Objects* (for example) will be more responsive to the image contents compared to the baseline model.

For example, given the image in Figure 1, we can query the model using the prompt "A photo of a [MASK]", and count how many of the objects ("tiger", "carrot") are in its top k predictions. We compare our alternative pre-trained models, pre-trained on 50% of the data, with the original pre-trained LXMERT model. We evaluate them on 2193 images from the LXMERT *minival* split, which the model did not observe during pretraining. Given a (prompt, image) pair, we intersect each model's top k predictions with the ground-truth objects list obtained from the image ground truth scene-graph, available for these images. We use several prompts: "A photo of a [MASK]" (inspired by CLIP (Radford et al., 2021)), "A [MASK] in the photo", and "A [MASK]". We present a precision for different values of k in Figure 6.

Our models achieve improved precision score over published LXMERT, despite training over only a third of its epochs and half of its training data. The precision metric is simply the number of correct predictions (intersection of predictions with ground-truth objects), divided by the number of predictions. For example, when con-

---

[9]Results tables presented in Appendix B.3.
[10]Preliminary experiments show that increasing the number of epochs leads to smaller gains, which emphasizes the benefits of our method in low resource settings.
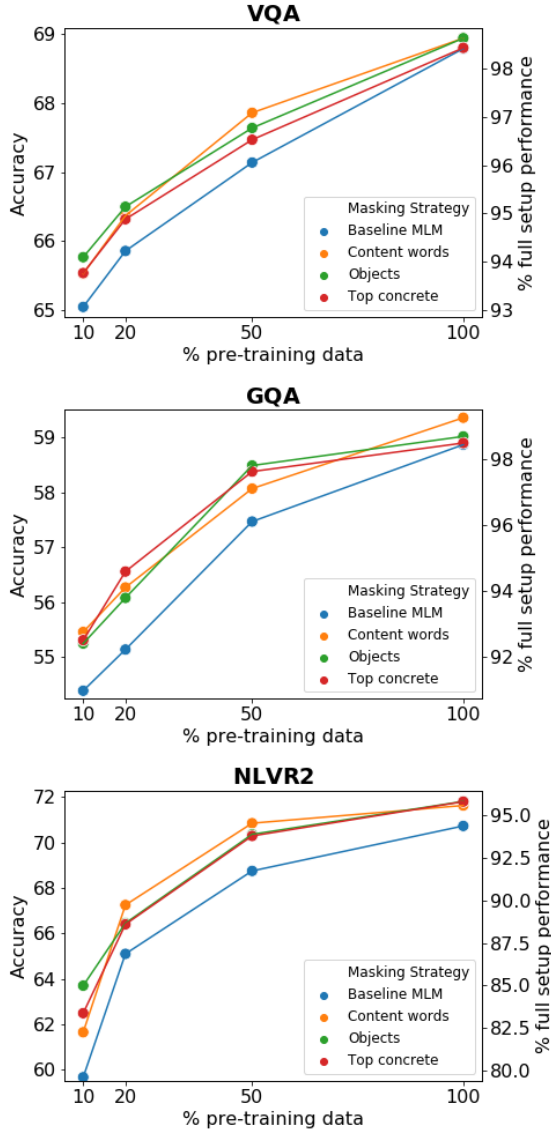
Figure 4: VQA, GQA and NLVR2 downstream tasks results for models with different masking strategies and increasing amounts of pre-train data. The left Y axis describes the accuracy, the right Y axis describes the percentage of the full setup performance (trained with 20 epochs and 100% of the pre-train data). Our alternative masking strategies consistently improve over the Baseline MLM masking strategy, especially in low resource settings.

sidering five top predictions (k=5), the published LXMERT achieves 10% precision, compared to 18% precision for the model trained with *Content words* masking strategy. When k=10, the improvement is 11% → 16%, etc. Additional results and ROC curve are available in Section B.3 in the Appendix. Our results indicate that our proposed models are more responsive to the image compared to the model trained with the Baseline MLM strategy.
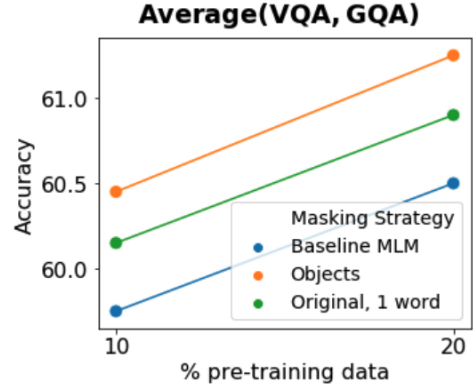

Figure 5: Ablation results for randomly masking a single word. The plot shows the average results for GQA and VQA. A model that masks a single word outperforms one with the original strategy of randomly masking 15% of the tokens, but under-performs a model that masks a single *object* word. We conclude that the gain of our proposed strategies comes from both masking a single word, and selecting tokens that are more important.

An example comparing the Baseline MLM model and model trained with *Objects* masking strategy is presented in Figure 3. Four of the top five predictions of the model trained with *Objects* masking strategy appear in the list of ground-truth objects, while the model trained with Baseline MLM strategy predicts only one of the ground-truth objects.

5 Analysis and Discussion

5.1 Hierarchy of Masked Semantic Classes

We have shown that our strategies improve results over the Baseline MLM. In this section, we aim to understand if the tokens we mask make the model actively rely on the image. For this purpose, we extract the image necessity for a masked token using the Δ *image loss* metric (see Section 2.2) for every token. We use the original LXMERT pre-trained model and validation data. For each sentence, we iterate over each token, mask and predict it with and without the image. An example from the extracted Δ *image loss* data is presented in Figure 2.[11] Following, Figure 7 presents a hierarchy of the different semantic classes described in Section 3.1, ranked by their Δ *image loss*.[12]

We draw several observations based on that plot. First, we note that objects that appear in both text and the scene graph (dubbed grounded objects, e.g.,

---

[11] We publish this extracted data for future work.

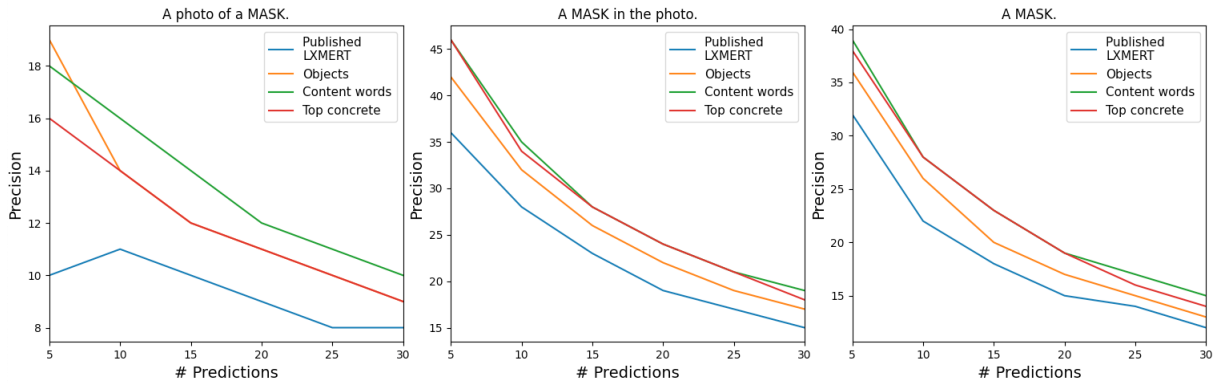[12] The groups are not mutually exclusive.

Figure 6: Precision/recall curve for prompt-base object detection task. Our models substantially improve over the published LXMERT, despite training over only a third of its epochs and half of its training data.

"tiger") are more important than non-grounded objects. Our intuition is that grounded concepts have higher Δ *image loss* compared to non-grounded concepts, as the model benefits from masking the latter. For example, consider the sentence "Is there a *tiger* in the image?", for an image without any tiger (i.e., *tiger* is not grounded). In this case, the model would not have the ability to differentiate the true word (*tiger*) from any other object in the vocabulary that is also not in the image.

In addition, we observe that the objects semantic class is the most important one. We see a connection between the hierarchy and downstream performance obtained by our different strategies. *Stop-words & punctuation* are ranked the lowest, and indeed pre-training with the *Stop-words & punctuation* strategy achieves the lowest results. The strategies of *Objects* and *Top concrete* are ranked high, and indeed they achieve improved results compared to the Baseline MLM.

5.2 MLM Performance across Word Classes

Many works (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020) assume that a VLP model should include an MLM component that is capable of predicting *every* masked token, including objects, properties, but also stop words and punctuation. Does a model that uses our *Objects* strategy, and masks only objects, learn to complete words from other classes? If not, can such a pre-training strategy be effective?

To examine this questions, we extend the experiment described in Section 2 to additional masking strategies, comparing between the different models pre-trained on 50% of the data. Results are presented in Table 2. We see that the model trained with the *Baseline MLM* masking strategy is able to

complete masked words from different classes (performance are above 70% for all cases). However, the model trained with *Objects* masking strategy indeed learned to complete only objects. Nonetheless, its downstream performance is in fact higher than the *Baseline MLM* model. We conclude that a model does not necessarily need to be able to complete all semantic classes, and some classes are more beneficial than others. For example, the *Objects* model's performance is quite low on both completing stop-words (4%), which is considered an easy task, and on attributes (22%).

A possible explanation for these findings might be that the model is evaluated mostly on retrieving objects, and had we tested it on other classes, its performance would have substantially decreased. To test this hypothesis, we inspect the same model's performance on questions with answers from different semantic types. To do so, we experiment with the GQA dataset, which includes partitioning of the answers into different semantic types, including *Objects*, *Relations* (subject or object of a described relation, e.g., "what is the girl wearing?"), and *Attributes* (the properties or position of an object).

The results for the semantic type partition are presented in Table 3. Comparing between the models trained with *Objects* and *Baseline MLM* masking strategies, the *Objects* masking strategy achieves improved performance in *Relationships* and *Attributes*, although it never masked these kinds of tokens, and its MLM performance on these classes is considerably lower. It seems that masking only objects might assist the models to learn additional semantic classes.
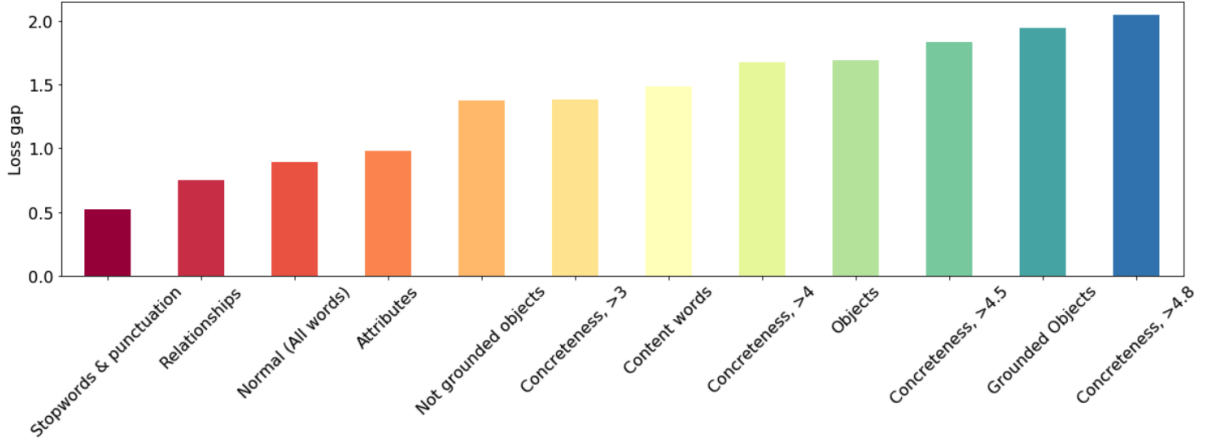
3019

Figure 7: Hierarchy of semantic classes and its importance by the Δ *image loss* metric (Loss without image - Loss with image).

Model Masking Strategy	Baseline MLM	Objects	Content words	Top concrete
Baseline MLM	87%	27%	70%	36%
Stop-words & punctuation, 15%	98%	4%	80%	13%
Content words, 15%	74%	57%	62%	62%
Objects	76%	85%	82%	83%
Attributes	70%	22%	59%	50%
Relationships	89%	15%	75%	25%

Table 2: MLM Validation Accuracy@5 for different pre-training strategies, tested on different masking strategies. Interestingly, the model trained with *Objects* strategy achieves low performance on all semantic classes except objects, but still achieves improved results compared to the model trained with Baseline MLM strategy.

Question semantic type	# Questions	Masking Strategy Baseline MLM	Objects
Objects	778	86.89	87.79
Attributes	5,186	63.17	63.96
Relations	5,308	49.72	50.47

Table 3: GQA semantic types partition performance. The model trained with *Objects* masking strategy achieves improved performance compared to the baseline model on *Relationships* and *Attributes*, although it never masked these kind of tokens.

6 Related Work

6.1 Vision Language Pre-training (VLP)

Recently, many VLP models have been proposed (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020). The pre-training objectives in many cases are: (1) Masked language modeling (MLM), where a model predicts masked tokens given the sentence and the image. (2) Masked region modeling (MRM), where the model predicts masked visual object features, and (3) Sentence-image matching, where the model predicts whether the sentence belongs to the image. Some models also add the visual question answering objective during the pre-training phase (Tan and Bansal, 2019; Li et al., 2021). Previous works have found that the MLM objective is an important pre-training task affecting the quality of the learned representations (Chen et al., 2020; Huang et al., 2020; Hendricks et al., 2021). However, the MRM objective was not always found to be important (Su et al., 2020; Hendricks et al., 2021), and the same for sentence-image prediction (Hendricks et al., 2021; Li et al., 2019). For this reason, we focus on the MLM objective.

6.2 Alternative MLM objectives in vision and language

Concurrently with our work, Zellers et al. (2021) presented an approach for pre-training over YouTube videos. They suggested a strategy of corrupting highly visual words in the masked lan-

guage modeling task, observing that vanilla BERT-style often masks ungrounded words like "umm" or "yeah". We share the same motivation to mask highly visual words.

6.3 Challenges in VQA generalization

Visual understanding Language and vision tasks inherently demand deep understanding of both the text and the image. However, many works show that models can succeed on VQA *datasets* using strong language priors, and by relying on superficial cues, and there are still challenges to overcome for tasks with more compositional structure (Jabri et al., 2016; Zhang et al., 2016; Goyal et al., 2017; Agarwal et al., 2020; Bitton et al., 2021; Dancette et al., 2021). Balanced datasets such as VQA 2.0 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019) have been presented to address these challenges. Novel models with richer visual representations (Zhang et al., 2021) were also presented, and some works tried to encourage the model to look at the "correct" image regions (Liu et al., 2021; Yang et al., 2020).

Bias Yang et al. (2021) and Hendricks et al. (2018) have shown that attention-based vision-language models suffer from bias that misleads the attention module to focus on spurious correlations in training data, and leads to poor generalization. Some examples are presented in Appendix B.4, Figure 9. To mitigate the language priors bias, it may be beneficial to increase the focus on the image during pre-training.

7 Conclusions

We have shown that the current MLM pre-training method is sub-optimal for visual language pre-training, as this process tends to focus on stop words and punctuation, and in many cases does not mask any word in the sentence. We proposed alternative masking strategies that better utilize the image during pre-training, for example, focusing on physical objects. We found improved results in two evaluation setups, especially in low resource settings. We introduced the Δ *image loss* metric, which aims to explain the relation between a masked token and the image. Our analysis includes a hierarchy that describes the necessity of the image for different semantic classes. We publicly release the extracted data with this metric on the LXMERT pre-train validation data. Future work can use this information to devise new masking strategies, and

progress towards VLP models that better leverage the visual aspect of the cross-modal tasks.

Acknowledgements

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9687–9695. IEEE.

Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. 2021. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of GQA. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 94–105, Online. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. *arXiv preprint arXiv:2104.03149*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. PMI-masking: Principled masking of correlated spans. In *Proc. of ICLR*.

Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. 2021. Semvlp: Vision-language pre-training by aligning semantics at multiple levels. *arXiv preprint arXiv:2103.07829*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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 *European conference on computer vision*, pages 740–755. Springer.

Yibing Liu, Yangyang Guo, Jianhua Yin, Xuemeng Song, Weifeng Liu, and Liqiang Nie. 2021. Answer questions with right image regions: A visual attention regularization approach. *arXiv preprint arXiv:2102.01916*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151.

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. *arXiv preprint arXiv:2103.00020*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Andrew Shin, Masato Ishii, and Takuya Narihira. 2021. Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision. *arXiv preprint arXiv:2103.04037*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jianwei Yang, Jiayuan Mao, Jiajun Wu, Devi Parikh, David D Cox, Joshua B Tenenbaum, and Chuang Gan. 2020. Object-centric diagnosis of visual reasoning. *arXiv preprint arXiv:2012.11587*.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *arXiv preprint arXiv:2106.02636*.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5014–5022. IEEE Computer Society.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

A Appendix

Reproducibility The experiments have been performed with the LXMERT model (Tan and Bansal, 2019) with the public implementation.[13] The experiments were performed with NVIDIA RTX2080 GPUs.

Pre-training data	10%	20%	50%	100%
# Epochs	7	7	7	7
Batch size	64	64	100	256
# GPUs	1	1	3	4
Runtime	2 days	3 days	3 days	3 days

Table 4: Pre-training experiments configurations.

A.1 Detection of Objects, Attributes and Relationships

Using the annotated scene-graph as ground truth A simple way to detect *objects*, *attributes*, and *relationships* in captions, is to obtain it, given that the image has scene-graph annotation from Visual-Genome or GQA. In LXMERT pre-training data, 83% of the sentences have scene-graph annotations for their corresponding image. For example, given the sentence, image pair: "The rabbit is eating the orange carrot", and an image, the ground truth by the scene-graph will include *Objects*: rabbit, carrot; *Attributes*: orange; and *Relationships*: eating. When obtained from the scene-graph, we call it "Grounded" (Grounded objects, grounded attributes, and grounded relationships).

Predicting *objects*, *attributes*, and *relationships* in each caption: For more general and scalable method when scene-graph is not available, we can use matching heuristics. We use the Part-of-speech tagging (POS), and we aggregate lists of Objects, Attribute and Relationships from Visual Genome dataset annotations.[14] Those are our heuristics:[15]

- *Objects* are words with POS = "NOUN" and in Visual Genome objects list.

- *Attributes* are words with POS = "ADJ" and in Visual Genome attributes list.

[13] https://github.com/airsplay/lxmert
[14] http://visualgenome.org/api/v0/api_home.html
[15] Our full code, including code to detect the semantic type tokens will be published

Epoch	Baseline MLM	Content words	Objects	Top Concrete
1	1.70	3.07	3.23	3.26
2	1.46	2.11	2.28	2.29
3	1.40	1.97	2.14	2.15
4	1.36	1.88	2.04	2.05
5	1.33	1.81	1.96	1.98
6	1.30	1.75	1.90	1.91
7	1.27	1.71	1.84	1.86
8	1.25			
9	1.27			
10	1.23			
11	1.21			
12	1.19			
13	1.17			
14	1.16			
15	1.14			
16	1.12			
17	1.11			
18	1.09			

Table 5: Training loss for models trained in different masking strategies. The training loss for the original is obtained from the original model repository. Because we focus on tokens that are more difficult for the model to complete, the training loss is higher.

	# items	Accuracy	Recall
Objects	7,484,940	89.89	97.39
Attributes	3,240,096	92.91	79.91
Relationships	3,195,345	86.42	96.88

Table 6: Detection performance of *Objects*, *Attributes*, and *Relationships*.

- *Relationships* are words with POS = "ADP" or "VERB", and in Visual Genome relationships list.

Those simple rules are our predictions for detecting *Objects*, *Attributes*, and *Relationships* in a sentence.

Validation of the *objects attributes* and *relationships* task: We can now evaluate the predicted *objects*, *attributes* and *relationships* with the ground-truth obtained from the scene-graph. The grounding method (matching between the caption and the scene-graph) we use is simple: exact match between the word in the scene-graph and the caption. Using a more complex grounding algorithm will not change our predictions, but it can only improve our results (For example, if the caption has "women" that was predicted as *Object*, and the scene-graph has "woman", it is currently counted as "False-Positive" because it's not exact match). Results are presented at Table 6.

A.2 Concrete and Abstract definitions

The concreteness annotation dataset (Brysbaert et al., 2014) is annotated by 20-30 annotators. The rating scale is 1-5, when 1 is abstract, and 5 is concrete. This is how they define concrete: "A concrete word comes with a higher rating and refers to something that exists in reality ; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do. The easiest way to explain a word is by pointing to it or by demonstrating it."

This is how they define abstract: "An abstract word comes with a lower rating and refers to something you cannot experience directly through your senses or actions. Its meaning depends on language. The easiest way to explain it is by using other words".

B Additional Experiments

B.1 How good is current pre-training?

We want to asses contribution of the current LXMERT pre-training. We conduct fine-tune experiments with LXMERT without pre-tain. Results are presented at Table 7. We see that pre-training adds ≈6.5 in GQA, ≈4.8 in VQA, and ≈23.8 in NLVR2.

Dataset	GQA	VQA	NLVR2
No pre-train	53.24	65.10	51.07
Pre-training all data Reported LXMERT GitHub results	59.80	69.90	74.95

Table 7: Downstream task performance for limited pre-training methods.

B.2 How to change the 15% masking amount?

In Section 2 we discussed that 15% with short captions (≈6.86) causes that with third of the cases no token is masked, in another third 1 token is masked, and in the last third, multiple tokens are masked.

We isolate those factors by conducting 3 experiments:

- Not allowing 0 masked (if 0 tokens were masked, sampling 1 token to mask).

- Not allowing multiple masked (if multiple tokens were masked, sample 1 token from them to mask)

- Masking only 1 word.

	GQA	VQA	NLVR2
Baseline MLM	54.4	65.06	58.55
Don't allow 0 masked	54.98	65.4	59.45
Don't allow multiple masked	54.46	65	58.82
Mask 1 word	55.07	65.26	61.25

Table 8: Changing 15% masking amount. Masking 1 word achieves the higher downstream tasks results.

Results are presented at Table 8.

We can see that not allowing multiple masked tokens helps a bit. Not allowing 0 masked tokens helps more. And masking 1 word is the better overall strategy.

B.3 Full results tables

B.4 Examples

| | **% of pre-train data** | | | |
Masking Strategy	**10**	**20**	**50**	**100**
Baseline MLM	$65.05_{\pm0.02}$	$65.86_{\pm0.06}$	$67.14_{\pm0.2}$	$68.79_{\pm0.02}$
Content words	$65.53_{\pm0.04}$	$66.37_{\pm0.04}$	$67.86_{\pm0.08}$	$68.94_{\pm0.05}$
Objects	$65.77_{\pm0.05}$	$66.5_{\pm0.04}$	$67.64_{\pm0.08}$	$68.94_{\pm0.06}$
Top concrete	$65.54_{\pm0.21}$	$66.32_{\pm0.02}$	$67.47_{\pm0.1}$	$68.8_{\pm0.03}$

Table 9: Full VQA 2.0 results, mean$_{\pm std}$

| | **% of pre-train data** | | | |
Masking Strategy	**10**	**20**	**50**	**100**
Baseline MLM	$54.39_{\pm0.01}$	$55.14_{\pm0.02}$	$57.47_{\pm0.13}$	$58.87_{\pm0.04}$
Content words	$55.46_{\pm0.04}$	$56.27_{\pm0.33}$	$58.07_{\pm0.09}$	$59.36_{\pm0.08}$
Objects	$55.25_{\pm0.21}$	$56.08_{\pm0.10}$	$58.49_{\pm0.01}$	$59.02_{\pm0.03}$
Top Concrete	$55.31_{\pm0.12}$	$56.56_{\pm0.35}$	$58.38_{\pm0.25}$	$58.9_{\pm0.04}$

Table 10: Full GQA results, mean$_{\pm std}$

| | **% of pre-train data** | | | |
Masking Strategy	**10**	**20**	**50**	**100**
Baseline MLM	$59.67_{\pm1.04}$	$65.1_{\pm1.13}$	$68.75_{\pm0.53}$	$70.73_{\pm0.65}$
Content words	$61.65_{\pm0.95}$	$67.25_{\pm0.48}$	$70.85_{\pm0.06}$	$71.63_{\pm0.44}$
Objects	$63.7_{\pm0.14}$	$66.45_{\pm1.2}$	$70.36_{\pm0.91}$	$71.81_{\pm0.51}$
Top Concrete	$62.49_{\pm0.72}$	$66.4_{\pm0.56}$	$70.29_{\pm0.22}$	$71.8_{\pm0.1}$

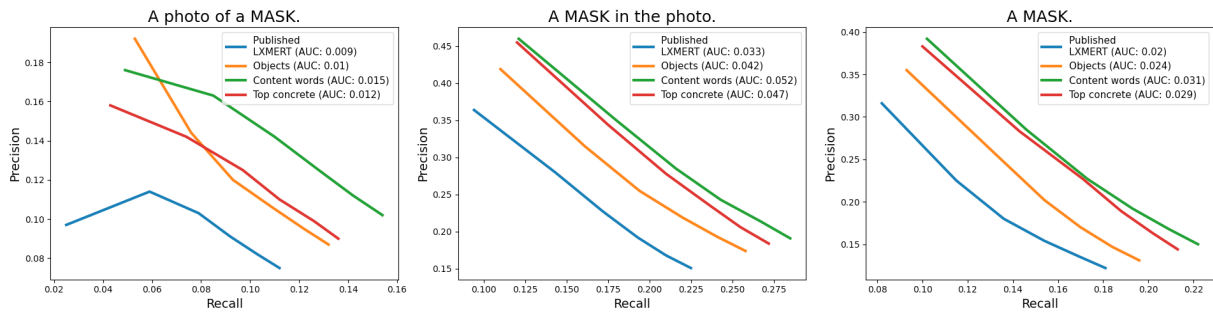Table 11: Full NLVR2 results, mean mean$_{\pm std}$

Figure 8: Precision-recall curve for prompt-base object detection task. Our models achieve improved results over published LXMERT, although trained with a half of the pre-train data and a third of the epochs.

Question:
where is the laptop?

Model answer: table
Ground truth: bed

Question:
What are the people sitting on?

Model answer: bench
Ground truth: ski-lift

Question:
What is the umbrella used for?

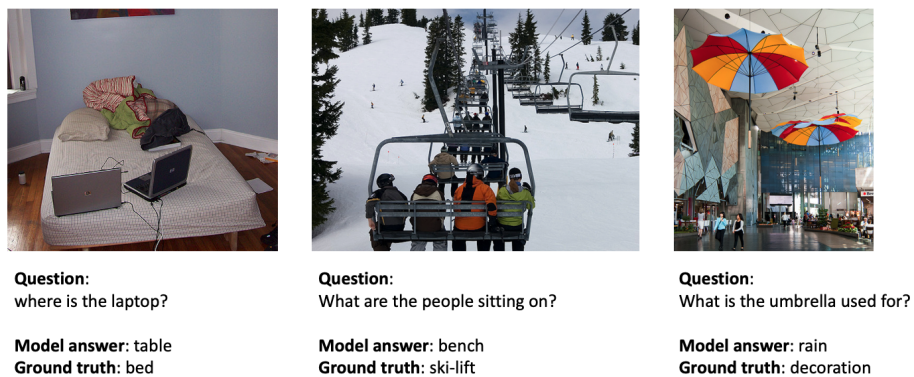**Model answer:** rain
Ground truth: decoration

Figure 9: LXMERT mistakes observed on examples from GQA and VQA. The tendency of VLP models is to predict something that is correlated with the text, or common answers. In many cases, the prediction is not an item that even appears in the image.

Published LXMERT Objects	bathroom, kitchen, bedroom, beach, city
	bathroom, restroom, sink, toilet, mirror
Ground truth objects	tile, toilet, wash cloth, tub, sink, mirror, ...

Published LXMERT Objects	beach, field, bathroom, woman, man
	beach, field, baseball, woman, game
Ground truth objects	bat, shirt, catcher, glove, lot, distance, ...

Figure 10: Additional examples of top 5 predictions for the prompt based object detection task, for the prompt "A photo of a [MASK]". Green underline indicate that the model predicted an object that appear in the ground truth objects (obtained from the scene graph).