Curriculum Masking in Vision-Language Pretraining to Maximize Cross Modal Interaction

Kraig Yuheng Tou
Georgia Tech
ytou3@gatech.edu

Zijun Sun
University of Bologna
zijun.sun@studio.unibo.it

Abstract

Many leading methods in Vision and language (V+L) pretraining utilize masked language modeling (MLM) as a standard pretraining component, with the expectation that reconstruction of masked text tokens would necessitate reference to corresponding image context via cross/self attention and thus promote representation fusion. However, we observe that the minimization of MLM loss in earlier training stages can depend disproportionately on local text signals, leading to poor training efficiency and inconsistency with the goal of representation fusion. The extent of this lack of cross modal interaction depends strongly *which* token(s) are masked. To address this issue, we propose a curriculum masking scheme as a replacement for random masking. Tokens are selected to be masked at a frequency proportional to the expected level of cross modal interaction necessary to reconstruct them. This is achieved using a parallel mask selection agent that measures the cross modal flow of information and treats it as a reward to be maximized. By additionally masking contiguous spans that include key objects and their relations, we also achieve better relational understanding, which has been shown to be lacking in many SOTA models. Our experiments on a wide range of V+L tasks show that we trail closely behind state-of-the-art methods despite pretraining on 300x to 1000x less data and we also achieve either top or runner-up performance on tasks from the ARO benchmark which tests compositional relationships. Finally, we demonstrate the potential of our method to scale to larger pretraining data.

1 Introduction

In the short time since large scale pretraining was first introduced to the multimodal setting (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2019), performance on leading benchmarks has risen by up to ten percentage points (Yu et al., 2022; Wang et al., 2022; Alayrac et al., 2022), driven in no small part by the trend towards larger, often custom curated image-text pair datasets (Yuan et al., 2021). However, the topic of pretraining efficiency has not received the attention it deserves. Masked language modeling is widely used as a pretraining component (Li et al., 2021; Wang et al., 2022; Kwon et al., 2022), with the expectation that reconstruction of masked text tokens would necessitate reference to corresponding image context via cross/self attention and thus promote representation fusion. The issue is that resulting cross modal interaction is highly dependent on the token that is masked. From an intuitive perspective, some tokens—such as prepositions and stop words (1)—can be reconstructed entirely disregarding image context, while others (2) can be narrowed down to a few possibilities from surrounding text alone. As can be seen in figure 1, when we replace paired images with random ones and calculate cosine similarities between the fused representations of masked tokens for original text-image pairs and their random-image variants, this swapped context similarity (SCS) score remains relatively static for many epochs at the start even though MLM loss rapidly decreases. It is not until a third of the way through training that image context noticeably affects learned representations. The large variance in SCS score at the start also reflects on how mask selection affects cross modal interaction. Clearly, alternatives to random masking warrant exploration. In the interest of pretraining efficiency, (1) could be masked less often compared to tokens requiring targeted attention to image context to reconstruct. (2) would ideally be masked less early on, until the model learns to venture beyond local textual context.

While several adaptive masking strategies have been proposed (, levine2020pmi, yang2022learning) they apply to the unimodal case and cannot be directly adapted to maximize cross modal interaction. Another pressing issue

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3672–3688

is that many advanced models fail to demonstrate relational understanding; many were recently shown to achieve at-or-below-chance performance at discerning between captions like "The horse is eating the grass" and "The grass is eating the horse" (Yuksekgonul et al., 2022), despite their rich pretraining data. Their behavior was likened to "a bag of words model", suggesting the lack of nuanced cross modal information flow.

A particular property of transformer architectures can be exploited to measure cross modal interaction and help address these issues– the attention between text and visual tokens, based on which we formulate an interaction score. Since we hope to maximize this score over mask selection rather than model parameters, we cannot set it up as an optimization problem baked into the loss function etc. We thus treat the problem as a reinforcement learning one, with mask configuration as the action space and the interaction score as the reward, and use a parallel model with shared parameters to select masks while the main model performs multimodal MLM. This curriculum masking strategy also includes masking contiguous spans and learned factorization order, making it a generalized form of language modeling, of which masked and autoregressive language modeling are special cases. We find that this strategy results in not only more data-efficient pretraining, but also a natural solution to the issue of relational understanding, as reconstruction of a masked spans that include key objects and their relations forces relational learning. Our contributions are as follows:

1. We propose a measure for cross modal information flow R_t based on value-weighted attention and a curriculum masking strategy whereby tokens/spans to be masked as well as factorization order are selected at a frequency proportional to their expected $\mathcal{R}_i$. Selection is performed by a parallel reinforcement learning network with partially shared parameters that treats masking and order selection as stochastic actions and $\mathcal{R}_i$ as the reward. There is no delay as the parallel network performs masking/order selection for batch $t+1$ while the main model learns from batch t. Our masking strategy is also compatible with most multimodal transformer architectures and can serve as a complement to concurrent works in multimoddal pretraining.

2. We achieve state-of-the-art performance on VGR and VGA tasks of the Attribution, Relation, and Order (ARO) benchmark (Yuksekgonul et al., 2022) and runner-up performance on its caption selection tasks.

3. We achieve performance that trails slightly behind SOTA models on a range of multimodal understanding, generation, and zero-shot tasks despite pretraining on significantly less data (300x to 1000x less). Our method also shows better performance in pretrain-data-equated scenarios and exhibits a trend line that demonstrates scaling potential.

4. Our method using RL can directly maximize cross model interaction and do so over the entire course of training. Non-RL methods may be less tedious but fall short in at least one of these two aspects, as discussed in section 5.3.

2 Related Work

2.1 Dynamic and curriculum based mask selection

Several works have explored adaptive, non-random masking in the text-only scenario, using selection strategies such as pointwise mutual information (Levine et al., 2020) and POS-tag weighting (Yang et al., 2022). On the masked image modeling frontier, learned masking of image patches that correspond to semantic entities via adversarially choosing them (Shi et al., 2022) or semantic part learning (Li et al., 2022a) has also been recently proposed. These varied approaches all tackle the issue of models reconstructing masks solely by latching onto nearby tokens that are part of the same semantic constituent (e.g. reconstructing [igen] in "eigenvector" without utilizing the rest of the sentence), by attempting to mask the said semantic constituent as a whole. They are more adaptive than brute approaches like masking noun phrases etc. but are not alone sufficient for the multimodal scenario where masking of tokens that require cross modal fusion to reconstruct is desired, as two correlated spans of tokens that both form semantic entities can require a completely different level of reference to the other modality to reconstruct.

There is also a line of work termed "concept based curriculum masking" (Lee et al., 2022) that starts by masking simple concepts like "car" before expanding to more sophisticated concepts that

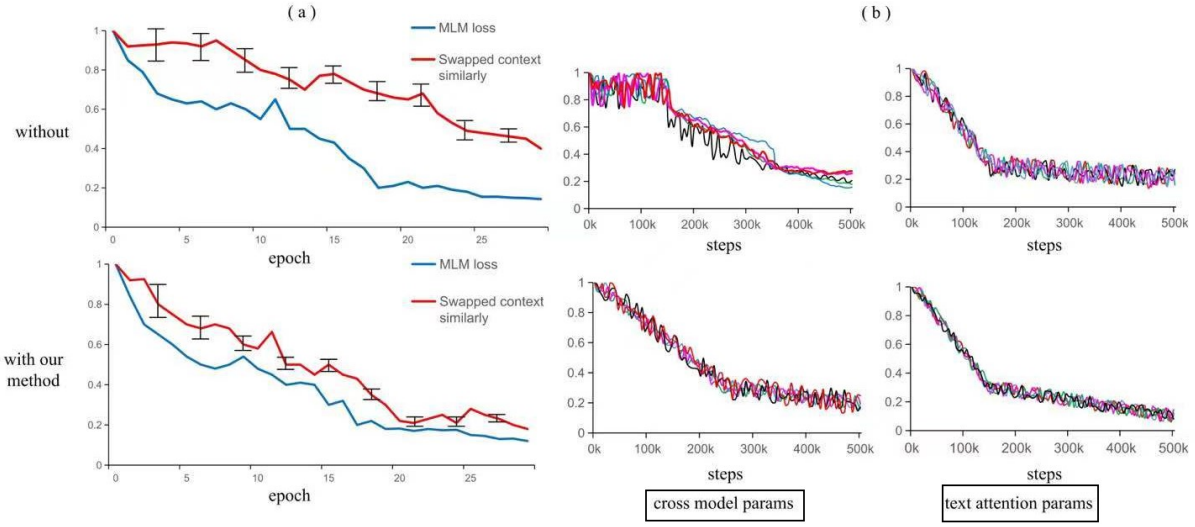


Figure 1: Left: How fused representations change when images are swapped for random ones (a). The y axis shows normalized cosine similarities between original representations and swapped-image variants. Right: Parameter distances to their final values over training (b). Top 6 principal components are shown. Further details in appendix G

are proximal on a knowledge graph such as "self-driving car". Again, the level of sophistication of a concept does not directly correspond to the level of cross modal fusion required for reconstruction.

2.2 Using a side model as the basis of a curriculum

AtnnMask (Kakogeorgiou et al., 2022) forms a teacher model using exponential moving average weights and masks image patches strongly attended to by its CLS token. The thought process that a side model with a similar level of learning is a good indicator of appropriate difficulty can be traced back to earlier works in curriculum learning (Sachan and Xing, 2016). Different from these works, our proposed method features a curriculum that is centered around cross modal interaction maximization, albeit also using a side model.

2.3 Cross Modal Interaction

To our knowledge, there have been no works that directly maximize cross modal interaction in masked language or image modeling. The authors of MLIM (Arici et al., 2021) state the goal of maximizing cross modal information flow, but do not measure and optimize for it. Instead, they alternate heavy image masking and heavy text masking (80% mask probability) to limit the possibility of reconstruction using local-modal context. Measuring cross modal interaction uing gradient-based approaches has been explored (Liang et al., 2022), though only for interpretability. Converting gradient based approaches to a mask selection strategy is not intractable, but not straightforward either.

2.4 Language modeling with factorization orders different from MLM or left-to-right

Language modeling with factorization orders other than left-to-right has been explored in the unimodal case by UniLM-v2 (Bao et al., 2020) and XLnet (Yang et al., 2019), but they use a random factorization order rather than a learned one. The former also uses masked spans. In the multimodal case, SimVLM (Wang et al., 2021) uses prefix LM which has characteristics of both MLM and autoregressive LM.

3 Proposed Method

3.1 Multimodal pretraining background

Given a set of n modalities $M = M_1, M_2, ..., M_n$, each with corresponding feature space $\mathcal{F}_i$, a function $f : \mathcal{F}_1 \times \mathcal{F}_2 \times ... \times \mathcal{F}_n \rightarrow \mathcal{R}$ that maps them to a common representation space $\mathcal{R}$ is trained via self-supervised tasks. One common task is masked language modeling (MLM) (Li et al., 2021; Singh et al., 2022; Lu et al., 2019; Alayrac et al., 2022). Denoting text as the first modality, M_1, text tokens $t_i \in M_1$ are randomly masked with fixed probability p, most often chosen to be 15%; then, masked tokens are reconstructed from non-masked tokens and paired images. The training objective is to minimize

$$\mathcal{L}_{MMLM} = -\mathbb{E}_{\mathbf{m}\sim D}\left[\log P(\mathbf{x}_{mask}|\mathbf{x}_{\backslash mask}, \mathbf{m}_2, \mathbf{m}_3, ..., \mathbf{m}_n)\right] \quad (1)$$

Contrastive learning (Jia et al., 2021) can be used to align image/text pairs before fusion as in ALBEF (Li et al., 2021) and masked image modeling (MIM) (Wang et al., 2022; Dou et al., 2022) can also be included as a pretraining task. Autoregressive language modeling (ALM) may also be included for generation capabilities as in COCA (Yu et al., 2022). The overall loss function to be minimized thus comprises a set of individual self-supervised tasks: $\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}i(f)$.

3.2 Generalized formulation

To mask important tokens more often, we use dynamic p_i's in place of a fixed p: $t_i \rightarrow$ "[MASK]" if $u_i(0,1) < p_i$, for $i = 1, 2, ..., n$. Furthermore, we mask contiguous spans to allow for prediction of concepts captured by phrases. Using s_i to denote span-masking probability, the full masking procedure can be expressed as

$$\hat{t_i} = \begin{cases} \text{"[MASK]"}, & \text{w/ probability } p_i \\ t_i, & \text{w/ probability } (1-p_i) \\ \text{"[MASK]"}, & \text{w/ probability } (1-p_i)\times s_i \\ & \text{if } \hat{t_{i-1}} = \text{"[MASK]"} \end{cases} \quad (2)$$

Finally, we predict masked tokens/spans according to different factorization orders q to learn their interrelationships and encourage relational understanding of image concepts. With ordered sequence $q = \langle x^i_{mask}\rangle_{i\in|\text{mask}|}$, the objective becomes

$$\mathcal{L}_{MMLM} = -\mathbb{E}_{\mathbf{m}\sim D}\left[\sum_{i\in q} \log P(\mathbf{x}_i|\mathbf{x}_{\backslash q}, \mathbf{x}_{<i}, \mathbf{m}_2, \mathbf{m}_3, ..., \mathbf{m}_n)\right] \quad (3)$$

and Autoregressive language modeling would be a special case where q is the entire text-input ordered from left to right. We direct readers to appendix I for more details on generalized LM.

3.3 From vanilla pretraining to cross-modal interaction maximizing pretraining

We measure cross modal interaction and treat it as a reward signal to be maximized over the range of pretraining parameters, which we reformulate below as action space probabilities.

Cross modal information flow In calculating the cross modal flow of information for MLM, text is the destination modality (query side) and image is the source modality (key side). The measure is computed for a chosen destination token, over all source tokens, using value weighted attention. Letting W_q, W_k, W_v respectively denote query, key, and value matrices in either cross or self attention, and h_i, t source and destination tokens, value weighted attention (Kobayashi et al., 2020) can be calculated as attention weight, $\alpha_i = \text{Softmax}\left(\frac{W_Q t\ (W_K h_i)^T}{\sqrt{d_k}}\right)$, times the norm of the value transformed source vector $\|W_v h_i\|$. The average score of the top 8 "attended to" source tokens is then taken, after normalization by mean interaction score:

$$R_t = \sum 0.125 \cdot Top8\left(\frac{\alpha_i \cdot \|W_v h_i\|}{\sum_i(\alpha_i \cdot \|W_v h_i\|)/n}\right) \quad (4)$$

(shown visually in appendix D) The calculation for R_t is presented as though there were a single attention head and a single layer for better readability. In practice, R_t is calculated for each head in each layer and averaged. Normalization of the interaction scores is performed layerwise. **Remark:** Alternative choices of R_t are discussed in appendix A.

p,s,q as action space probabilities On each forward pass, cross modal information flow is measured, yielding an interactive environment that provides feedback on the masking configuration and factorization order used, the selection of which can be viewed as a markov decision process $MDP(\mathcal{S}, \mathcal{A}, \mathcal{G}, \gamma, \mathcal{T})$ and solved with reinforcement learning (RL). States $\mathcal{S}$ are given by the multimodal input, while rewards $\mathcal{G} : \mathcal{S}_T \times \mathcal{A} \rightarrow \mathbb{R}^1 = \sum_{t\in mask} R_t + \sum_t \frac{R_t}{n}$ at the end of each episode are determined by the sum of information flows for each masked token/span plus the average over the entire input. Each action $a \in \mathcal{A}$ is a choice of masking configuration and factorization order q, which we represent as a non-independent composite action made up of individual token/span selections until a masking quota is reached, followed by a selection of q. Mathematically, $\pi_\theta(a|s) = \prod_0^L \left(a^l|a^{<l}, s\right)$ where $a^L = q \in \mathfrak{S}_q$ and $a^{<L} = t_i, t_j ...$ A table of the symbols used is given in appendix D, along with pseudocode in appendix E

Implementation wise, the masking network outputs action probabilities **p** and **s** of length n, the

maximum sequence length. A token position is sampled from **p** to be masked and subsequent tokens have probability s_i of being added to the span, as in eq 2. After each token/span selection, <m> </m> tags are appended around the masked position for the next forward pass, until 50% of tokens are masked, at which point factorization order is selected from the action distribution **q**. An illustrated layout is presented in figure 2. Invalid actions due to variable sequence length and already-masked positions are filtered out and probabilities are renormalized over remaining actions. While the main model learns from a batch, the agent selects masking configuration/factorization order for the next batch. An analysis of speed is provided in section 5.

Policy Learning We learn a parameterized policy $\pi_\theta : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ using PPO (Schulman et al., 2017). A vanilla advantage function $A_t^\pi(s_t, a_t) = Q_t^\pi(s, a) - V_t^\pi$ is used, with standard definitions for $V_t^\pi = \mathbb{E}_{a\sim\pi}\left[\sum_t^T \gamma G(s_t, a_t)\right]$ and $Q_t^\pi(s, a) = G(s_t, a) + \gamma\mathbb{E}_{s_{t+1}\sim\mathcal{T}}\left[V_{t+1}^\pi(s_{t+1})\right]$. It is desirable to choose reward maximizing masking configurations/factorization orders more often, but even low reward settings must be chosen at a certain minimum probability in order to learn the full pre-training data distribution. Thus, we add entropy regularization $\beta H_\theta(s_t) = -\sum_a \pi_\theta(a|s_t)log(\pi_\theta(a|s_t))$ with entropy coefficient β to gradient updates to achieve a more stochastic policy. Since a is a composite, non-independent action, we use an unbiased estimator $\tilde{H}_\theta(s_t) = \sum_{A_i\in A}\sum_{a\in A_i} p_\theta(a|a_{i-1})log(p_\theta(a|a_{i-1}))$ (Zhang et al., 2018). (Pseudocode provided in appendices).

3.4 Architecture

Curriculum masking is compatible with many V+L transformer architectures. More details on compatibility are provided in appendix H. The illustrated layout in figure 2 depicts our method mounted onto a generic dual stream architecture in order to more clearly depict the respective roles of, and interactions between, the main model and agent. For our experiments below, we use a 12 layer multiway transfomer architecture (Bao et al., 2022).

4 Experiments

4.1 Pre-training and evaluation tasks

We pretrain on 4m image-text pairs taken from Conceptual Captions 3m (Sharma et al., 2018) (2.95m pairs), Visual Genome (Krishna et al., 2017) (100k images, 770k text), SBU (Ordonez et al., 2011) (860k pairs), and COCO Captions (Lin et al., 2014) (115k images, 560k text) for 30 epochs as a base model for fair comparison with other models trained on a 4m dataset; and also add Conceptual Captions 12m(Changpinyo et al., 2021) to the pre-training dataset, after removing the 63k pairs overlapping with CC3m, to test the scaling potential of our model. Pretraining, finetuning, and task details are provided in appendices C and B.

Classic multimodal understanding, generation, and retrieval tasks The classic understanding tasks we evaluate on are Visual Question and Answering (VQA-va) (Goyal et al., 2017), Natural Language for Visual Reasoning (NLVR) (Suhr et al., 2018), and Visual Entailment (VE) (Xie et al., 2019). For image retrieval (IR), text retrieval (TR), and caption generation we evaluate on COCO (Lin et al., 2014) and FLickr30k (Plummer et al., 2015).

Attribution, Relation, and Order (ARO) benchmark This recently proposed benchmark (Yuksekgonul et al., 2022) tests relational, attributive, and order understanding using four tasks: Visual Genome Relation (VGR), Visual Genome Attribution (VGA), COCO order (Co) and Flickr order (Fo). Many previously tested state-of-the-art models displayed near or below chance level of performance. VGR involves picking between correct and reversed prepositional/verb orders concerning two objects in a given image. An illustrative example provided by the authors is "The horse is eating the grass" vs. "The grass is eating the horse".

Zero shot tasks We evaluate zero-shot transfer capabilities on two downstream tasks, image/text retrieval and caption generation. The out-of-domain split from Nocaps (Agrawal et al., 2019) is used as the benchmark for the latter and Flickr30k (Plummer et al., 2015) is used for the former.

4.2 Discussion of Results

We achieve SOTA performance on relational and attributive understanding, as shown in Table 2, and runner up performance on COCO/Flickr or-

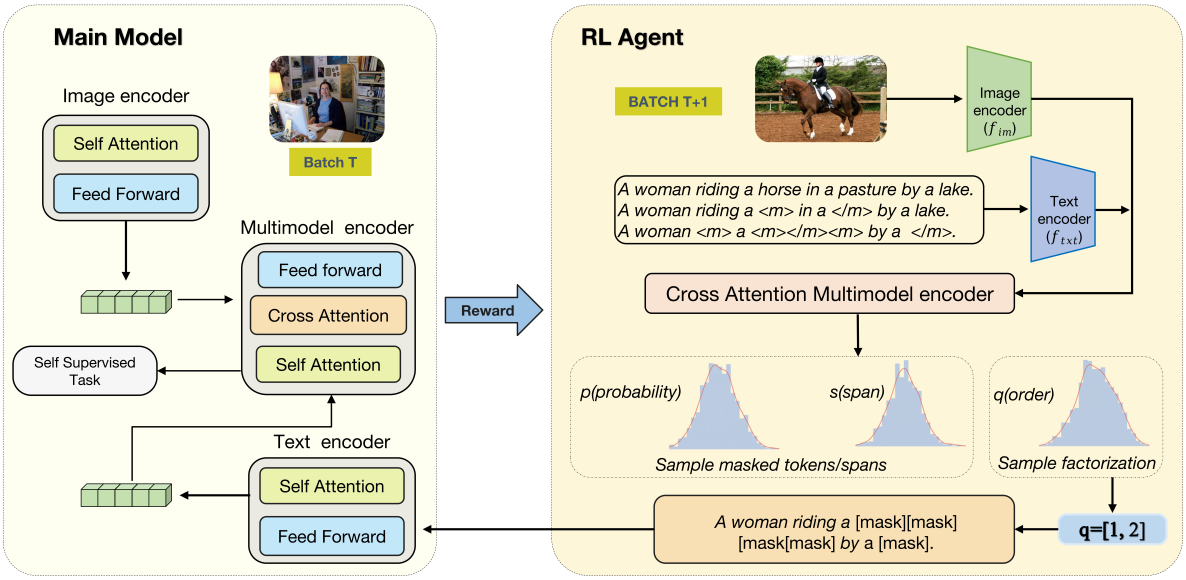


Figure 2: Illustrated layout: The agent depicted on the right selects masking configuration and factorization order for batch t+1 while the main model learns from batch t. P, S, Q are stochastic actions selected by the agent. **p** is a distribution over tokens, so softmax is used. **s** is a probability of span masking a token, conditioned on its previous token having been masked, so sigmoid is used..

der understanding. It is important to point out that NegClip, the leading method on the latter two tasks, was fine-tuned on hard negatives consisting of swapped linguistic structures, making it highly tailored for those tasks. Our results on VGR confirms our hypothesis that masking key spans and using different factorization orders forces the learning of interrelationships, something vanilla MLM and autoregressive LM fails to do.

Comparison against SOTA methods on VQA, NLVR, SNLI-VE, and COCA-Caption generation is reported in Table 1. Our proposed method closes the gap with the two leading models, COCA (Yu et al., 2022) and BEiT-3 (small) (Wang et al., 2022), despite the fact that they were trained on 1000x and 500x as much pretraining data, respectively, not to mention COCA's 10x larger model size. We even outperform SimVLM-large (1.8m pretrain size) on multiple tasks. The substantial increases in performance from our 14m model also demonstrates the scaling potential of our proposed method. Image and text retrieval results are reported in Table 3. Our proposed method falls short of the leading model, Florence (Yuan et al., 2021), which was pretrained on 500x the amount of data, but it outperforms all other SOTA models trained on a similar amount of data except for MaskVLM (Kwon et al., 2022) and MAP(Ji et al., 2023) on some setups. Their performance does not detract from the potential of our approach as they are fully compatible with our generalized method. Zero-shot performance is reported in Table 4. We achieve runner up performance on zero-shot generation, behind SimVLM-large (Wang et al., 2021) and trail within 2% of models pretrained on significantly more data, even though pre-training data size has been observed to be even more critical for zero shot tasks.

4.3 Scaling and Ablations

To demonstrate scaling potential and the competitiveness of our proposed method under equated pretraining conditions, we compare a leading model, BEiT-3 (base version) (Wang et al., 2022), with and without our proposed method mounted on. Using pretrain data consisting of 1m, 2m, 4m, and 14m image/text pairs sampled from our 14m dataset, we evaluate on VQA test-std and report our results in figure 4 (right). Mounting our method on demonstrably improves performance at each tested data scale.

To study whether each newly introduced feature stacks additional benefit, we perform combinatorial ablations on VQA test-dev as well as Visual Genome Relations. As evident from the results in Figure 4(left), masking contiguous spans plays a critical role in understanding object relation. Additional ablation studies, analysis, and details are

Method	#pretrain images	VQA		NLVR		VE		COCA-Captions			
		test-dev	test-std	dev	test-p	val	test	B@4	M	C	S
FLAVA(Singh et al., 2022)	68m	72.8	-	-	-	-	79.0	-	-	-	-
SimVLM-base(Wang et al., 2021)	1.8b	77.87	78.14	81.72	81.77	84.20	84.15	39.0	32.9	134.8	24.0
ALBEF(Li et al., 2021)	4m	74.54	74.70	80.24	80.50	80.14	80.30	-	-	-	-
OSCAR(Li et al., 2020)	4m	73.61	73.82	79.12	80.37	-	-	41.7	30.6	140.0	24.5
Codebook(Duan et al., 2022)	4m	74.86	74.97	80.50	80.84	80.47	80.40	-	-	-	-
UNITER(Chen et al., 2020)	4m	73.82	74.02	79.12	79.98	79.39	79.38	-	-	-	-
MaskVLM(Kwon et al., 2022)	4m	75.45	75.40	81.58	81.98	80.37	80.67	-	-	-	-
SimVLM-large (Wang et al., 2021)	1.8b	79.32	79.56	84.13	84.84	85.68	85.62	40.3	33.4	142.6	24.7
COCA(Yu et al., 2022)	4.8b	82.3	82.3	87.0	87.1	86.1	87.0	40.9	33.9	143.6	24.7
Flamingo (Alayrac et al., 2022)	2b	82.0	82.1	-	-	-	-	-	-	-	-
BEiT-3-base (Wang et al., 2022)	2b	81.16	81.05	86.91	87.06	-	-	40.2	31.8	143.8	24.6
VinVL(Zhang et al., 2021)	5.65m	76.56	76.60	82.67	83.98	-	-	41.0	31.1	140.9	25.2
Florence (Yuan et al., 2021)	900m	80.16	80.36	-	-	-	-	-	-	-	-
MAP (Ji et al., 2023)	4m	78.03	-	83.30	83.48	81.40	81.39	-	-	-	-
MPlug2 (Xu et al., 2023)	17m	81.11	81.13	-	-	-	-	-	-	-	-
C-mask	4m	80.27	80.32	84.14	83.76	85.09	85.01	39.8	33.1	136.7	24.2
C-mask	14m	81.66	81.45	85.77	85.81	85.30	85.27	40.2	33.3	142.4	24.5

Table 1: Evaluation on classic multimodal understanding and generation tasks. Models highlighted in yellow pretrain on web-scale data

Method	VGR	VGA	COCO-order-PRC	Flickr-order-PRC
FLAVA(Singh et al., 2022)	0.25	0.73	0.004	0.13
BLIP(Li et al., 2022b)	0.59	0.88	0.32	0.37
XVLM(Zeng et al., 2021)	0.73	0.87	0.36	0.48
CLIP(Radford et al., 2021)	0.59	0.63	0.46	0.60
NegCLIP(Yuksekgonul et al., 2022)	0.81	0.71	0.91	0.86
BEiT-3-base (Wang et al., 2022)	0.71	0.88	0.47	0.56
C-mask(14m)	0.83	0.90	0.88	0.81

Table 2: Evaluation on new Attribute, Relations, and Orders Benchmark

presented in appendix F.

5 Analysis and Discussion

5.1 On pretraining efficiency resulting from curriculum masking

We use HiRes-Cam (Draelos and Carin, 2020) to visualize how tokens refer to different image regions over the course of training. From the examples shown in figure 3 (top), it is clear that curriculum masking encourages the model to zoom in on semantically relevant image context earlier. Learned factorization orders also has this effect compared to left-to-right order, as demonstrated in figure 3 (bottom-left). Further evidence of increased pretraining efficiency can be found in figure 1 (a), which shows the average cosine similarities for representations produced from text-image pairs and their random-image variants.A high swapped similarity score indicates that image context is not utilized much, as texts produce similar fused representations even when paired with random images. Additional validation can be seen from the trajectories of squared distances between cross-attention parameters and their final values shown in figure 1 (b). With our method, there is no inertia in parameter space at the start of training associated with delayed learning of image context utilization

Parallel agent effect on speed and memory As mentioned, the agent learns from batch t while the parallel agent selects masks from batch t+1, so there is no delay in training the main model. Moreover, the agent shares the main model's parameters for the first six layers. To assess if latency mismatch poses an issue, we compare the training speeds in steps per second with and without curriculum masking. 5

The difference per second was found to be minimal, though the cumulative effect over the entire training process would depend on factors like dataset and batch size. As for additional memory overhead, curriculum masking resulted in 1.27x peak memory usage. This number would likely go down on setups using larger main models.

Agent training A comparison of rewards 3.3) obtained with and without our method are shown

Method	#pretrain	COCO						Flickr					
		IR			TR			IR			TR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER(Chen et al., 2020)	4m	52.9	79.9	88.0	65.7	88.6	93.8	75.6	94.1	96.8	87.3	98.0	99.8
OSCAR(Li et al., 2020)	4m	54.0	80.8	88.5	70.0	91.1	95.5	-	-	-	-	-	-
ALBEF(Li et al., 2021)	4m	56.8	81.5	89.2	73.1	91.4	96.0	82.8	96.7	98.4	94.3	99.4	99.8
Codebook(Duan et al., 2022)	4m	58.7	82.8	89.7	75.3	92.6	96.6	83.3	96.1	97.8	95.1	99.4	99.9
ALIGN(Jia et al., 2021)	1.8b	59.9	83.3	89.8	77.0	93.5	96.9	84.9	97.4	98.6	95.3	99.8	100.0
FLAVA(Singh et al., 2022)	68m	38.4	67.5	-	42.7	76.8	-	65.2	89.4	-	67.7	94.0	-
MaskVLM (Kwon et al., 2022)	4m	60.1	83.6	90.4	76.3	93.8	96.8	84.5	96.7	98.2	95.6	99.4	99.9
Florence (Yuan et al., 2021)	900m	63.2	85.7	-	81.8	95.2	-	87.9	98.1	-	97.2	99.9	-
BEiT-3-base (Wang et al., 2022)	2b	63.2	84.3	90.8	80.9	95.1	97.3	87.3	97.9	99.1	97.1	99.6	99.9
MAP(Ji et al., 2023)	4m	60.9	86.2	93.1	79.3	94.8	97.6	83.8	97.2	98.7	94.9	99.5	99.8
C-mask	4m	60.9	83.6	90.2	78.5	94.0	95.2	83.6	95.8	98.6	94.6	99.0	99.1
C-mask	14m	61.7	84.7	90.2	78.9	95.3	96.9	84.2	96.9	99.1	95.1	99.4	99.9

Table 3: Evaluation on Image Retrieval (IR) and Text Retrieval (TR). Models trained on significantly more data are highlighted in yellow

Method	NoCaps OOD	Flickr					
		IR			TR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER(Chen et al., 2020)	-	66.2	88.4	92.9	80.7	95.7	98.0
ALBEF(Li et al., 2021)	-	68.2	88.6	93.0	84.9	97.2	99.0
⋄ SimVLM-base(Wang et al., 2021)	82.5	-	-	-	-	-	-
⋄ SimVLM-large(Wang et al., 2021)	96.3	-	-	-	-	-	-
FLAVA(Singh et al., 2022)	-	65.2	89.4	-	67.7	94.0	-
OSCAR(Li et al., 2020)	80.3	-	-	-	-	-	-
VinVL(Zhang et al., 2021)	83.8	-	-	-	-	-	-
⋄ CLIP (Jia et al., 2021)	-	68.7	90.6	95.2	88.0	98.7	99.4
⋄ ALIGN(Li et al., 2021)	-	75.7	93.8	96.8	88.6	98.7	99.7
MaskVLM (Kwon et al., 2022)	-	75.0	92.5	95.8	87.0	97.9	99.3
⋄Florence (Yuan et al., 2021)	-	76.7	93.6	-	90.9	99.1	-
⋄BEiT-3-base (Wang et al., 2022)	-	77.6	93.4	96.2	90.9	99.0	99.0
C-mask (4m)	87.2	74.2	91.8	95.1	86.2	96.3	99.1
C-mask (14m)	94.5	74.7	92.1	95.3	87.9	97.9	99.3

Table 4: Zero shot performance on retrieval (Flickr) and caption generation (NoCaps out of domain split). Models highlighted in green were finetuned on COCO-splits for NoCaps and models with a ⋄ preceding them were trained on significantly more data.

in figure 3 (bottom-right). In our experiments, the non-stationarity of the environment has not resulted in failure to converge. State and action spaces as well as transition probabilities remain the same; only the reward function changes as the main model learns. The ratios of cross model interactions also likely stay similar as their magnitudes change. Finally, the reward function changes at a very gradual pace.

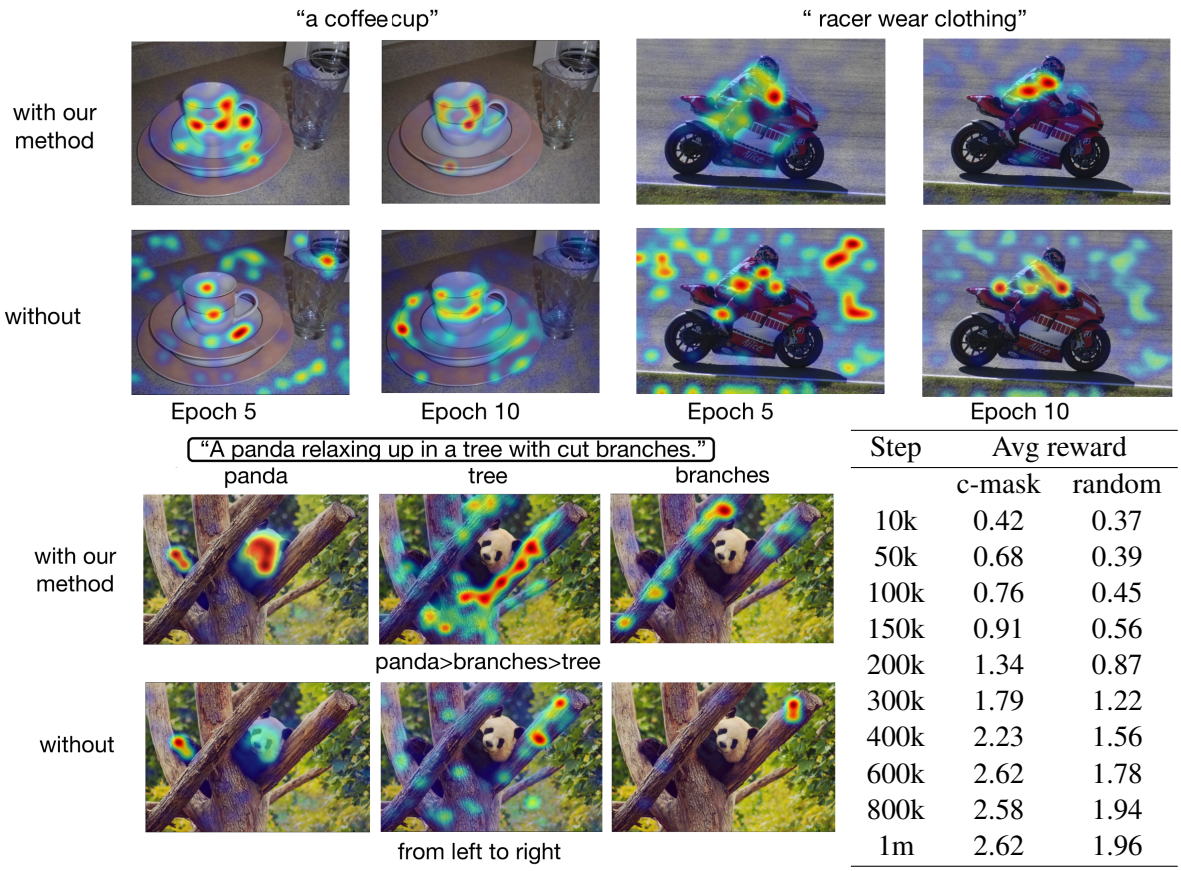
5.2 Choice of reward

Theoretically grounded measures for cross modal interaction (Liang et al., 2022) have been developed, but they are designed for interpretability/visualization and cannot be readily adapted for mask selection. Therefore, we use a value weighted attention based score as a proxy. According to their framework, a multimodal function f can be decomposed into unimodal subfunctions and cross modal interaction: $f(x_1, x_2) = g_1(x_1) + g_2(x_2) + g_{1,2}(x1, x2)$

The cross modal term, $g_{1,2}(x1, x2)$ (CM), can be isolated by taking second-order gradients of f to zero out unimodal terms, resulting in an interaction-per-pixel intensity map. Through a series of experiments in Appendix A, we show that our reward score is highly correlated with CM intensity scores in image regions pertinent to queried text tokens.

5.3 RL versus other approaches

Our method using RL is the only approach out of several possible ones that can directly maximize cross model interaction and do so over the entire course of training. Curriculum setups can be divided into static and dynamic variations, and the latter can be further subdivided into data based and model based approaches. With a static curriculum of preset masks for each data point, the change in



Step	Avg reward	
	c-mask	random
10k	0.42	0.37
50k	0.68	0.39
100k	0.76	0.45
150k	0.91	0.56
200k	1.34	0.87
300k	1.79	1.22
400k	2.23	1.56
600k	2.62	1.78
800k	2.58	1.94
1m	2.62	1.96

Figure 3: High-res cam showing the image areas attended to at early stages in training (epochs 5 and 10) (top) and for different factorization orders(bottom-left). The bottom images of the panda were taken from epoch 20. Bottom-right shows average rewards using c-mask compared to random masking

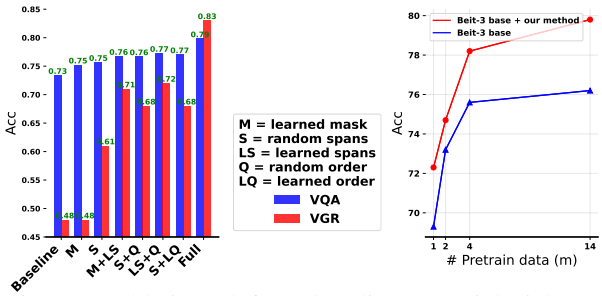


Figure 4: Ablations (left) and scaling potential (right)

attention patterns as the model learns is not factored in. Data based approaches (Levine et al., 2020), leveraging measures of statistical association like mutual information, do offer dynamic mask selection, but they require a certain level of alignment between image and text feature space, which only arises midway through training. For this reason they are not readily transferable to the multimodal scenario despite their success in text only pretraining. Finally, model based approaches leverage artifacts like model weights and/or a side model, based on the notion that a model with a similar level of learning is a good indicator of appropriate difficulty (Sachan and Xing, 2016). Limitations of previously discussed methods do not apply. Our approach using a parallel model with shared weights as an RL agent falls into this category. An alternative approach in this category would be a student-teacher like in AttnMask (Kakogeorgiou et al., 2022). This method, however, only maximizes cross modal interaction indirectly.

method	training speed (steps/sec)	peak memory multiplier
with c-mask	6.1	1.27x
without c-mask	6.3	1x

Table 5: Effect on speed and memory

6 Conclusion

In this paper, we present a curriculum masking strategy for V+L pretraining. Masks and factorization orders are selected by a parallel agent that aims to maximize cross modal interaction. Better pretraining efficiency and relational understanding are achieved, as demonstrated on various downstream tasks and experiments, at a reasonable speed and memory cost. The proposed method is also able to complement concurrent methods in V+L pretraining.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

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*, 35:23716–23736.

Tarik Arici, Mehmet Saygin Seyfioglu, Tal Neiman, Yi Xu, Son Train, Trishul Chilimbi, Belinda Zeng, and Ismail Tutar. 2021. Mlim: Vision-and-language model pre-training with masked language and image modeling. *arXiv preprint arXiv:2109.12178*.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

Rachel Lea Draelos and Lawrence Carin. 2020. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv e-prints*, pages arXiv–2011.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*.

Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2022. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. 2023. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23262–23271.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. 2022. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318. Springer.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

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*, 123:32–73.

Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto.

2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*.

Mingyu Lee, Jun-Hyung Park, Junho Kim, Kang-Min Kim, and SangKeun Lee. 2022. Efficient pre-training of masked language model via concept-based curriculum masking. *arXiv preprint arXiv:2212.07617*.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2020. Pmi-masking: Principled masking of correlated spans. *arXiv preprint arXiv:2010.01825*.

Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. 2022a. Semmae: Semantic-guided masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

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*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. 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.

Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2022. Multiviz: An analysis benchmark for visualizing and understanding multimodal models. *arXiv preprint arXiv:2207.00056*.

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.

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

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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.

Mrinmaya Sachan and Eric Xing. 2016. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 453–463.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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 ACL*.

Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. 2022. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pages 20026–20040. PMLR.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*.

Dongjie Yang, Zhuosheng Zhang, and Hai Zhao. 2022. Learning better masking for better language model pre-training. *arXiv preprint arXiv:2208.10806*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

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? *arXiv e-prints*, pages arXiv–2210.

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

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Yiming Zhang, Quan Ho Vuong, Kenny Song, Xiao-Yue Gong, and Keith W Ross. 2018. Efficient entropy for policy gradient with multidimensional action space. *arXiv preprint arXiv:1806.00589*.

A Choice of Reward

Here we present empirical support for our choice of R_i. Theoretically grounded measures for cross modal interaction have been developed (Liang et al., 2022), but they are designed for interpretability/visualization and cannot be readily adapted for mask selection. Therefore, we use a proxy in their place. According to their framework, a multimodal function f can be decomposed into unimodal subfunctions and cross modal interaction:

$$f(x_1, x_2) = g_1(x_1) + g_2(x_2) + g_{1,2}(x1, x2)$$

The cross modal term, $g_{1,2}(x1, x2)$ (CM), can be isolated by taking second-order gradients of f to zero out unimodal terms. This results in an interaction-per-pixel intensity map. It can be used to visualize important regions, but our use case requires a composite score. Thus we adopt a R_i based on value weighted attention, which measures how much an image token's representation is fused into a text token's. We show it to be highly correlated with CM intensity scores in image regions pertinent to queried text tokens in Appendix C. Pertinent regions are found using an object detector. We calculate R_i by averaging the value weighted attentions for the top 8 "attended to" image patches as this signifies concentrated attention on key image regions. Using the average over all patches would reward blindly attending to random image content or dispersed attention around relevant objects. We found 8 to work best, but this can be selected as a hyperparameter as well.

Experiments for choice of reward Reward R_i is designed to encourage cross model interaction. As discussed, there exist theoretically grounded measures for cross modal interaction, but they cannot be readily adapted for our purposes. We thus use our value weighted attention score as a proxy. The theoretically grounded measure based on second order gradients (Liang et al., 2022) requires first taking a gradient of f with respect to the text token we are interested in finding cross modal interaction for. A second order gradient is then taken with respect to all image pixels. The resulting output is thus of the same dimension as image pixels. Values show the amount a pixel interacts with the queried text token. This output adds tremendous value for visualization, but its per-pixel format is not a score we can use as a reward to maximize over. Furthermore, it is not relevant-object-discriminative. If a function of the per-pixel interaction map like its sum were used, it would reward interaction with a large number of irrelevant pixels. Thus, we opt for value weighted attention and specifically the value weighted attention of the top 8 image to-

Top 8 value weighted attention	1	2	3	4	5	6	7
0.91	0.64	0.73	0.69	0.58	0.84	0.82	0.82

Table 6: Correlations of different candidate reward scores with W-IOU scores

Hyperparameter	VQA	VE	NLVR	COCO	No-caps
Batch size	264	264	128	128	264
Warm up steps	1k	1k	1k	500	1k
train steps	100k	50k	50k	50k	10k
Start LR	5e-4	5e-3	5e-4	8e-4	8e-4
Min LR	0	0	5e-7	8e-7	0
Grad Clip	1.0				
WD	0.05				
Scheduler	One-Cycle				
Optim	AdamW				
Optim-epsilon	1e-8				
Optim-betas	(0.9, 0.999)				

Table 7: Fine tuning hyperparameters

kens, hypothesizing that highly concentrated value weighted attention would signify interaction with a relevant object. Our experiment detailed below validates this hypothesis.

We point out that our top-8 value weighted attention score is a composite score over the entire image, thus we cannot make comparisons to relevant objects found using an object detector. The second order gradient interaction score is, however, given pixel-by-pixel and can thus be compared in such a manner, meaning that we can first use it, along with an object detector, to compute interaction with relevant image objects and then calculate the correlation between our proposed reward and this discriminative-interaction.

We thus select 200,000 image-text pairs randomly and compute the second order gradient interaction scores for a random word from epoch 0 to 30. Next, we locate objects corresponding to the queried tokens using a pretrained object detector, DETR, for our 600k data points. Weighted intersection over union (w-IOU) scores are then computed to find the overlap between detected objects and high-interaction pixels. Next, we find the correlation between our top-8 value weighted attention reward scores and w-IOU scores, which indicate not just cross modal interaction, but cross moddal interaction with relevant regions. Since samples are taken from epoch 0 to 30, the ability to gauge relevant cross modal interaction is assessed for various stages of training. We find a 0.91 correlation for our proposed score, a much higher number than other candidate scores. Other measures we test include aggregating second order gradient score by mean (1), top 100 pixels (2), top 500 pixels (3), top 200 pixels (4), top 4 value weighted attention (5), top 15 value weighted attention (6), and top 50 value weighted attention (7). Correlations are presented in table 6:

In summary, our our reward serves as a suitable proxy for cross modal interaction with relevant image regions.

B Fine tuning and task details

Details for fine tuning on downstream tasks are listed in table 7.

For NLVR, we deal with (image-1, image-2, text) triplets by using a [sep] token between images. An MLP layer follows to output binary predictions of whether the image-pair is described by the text.

In-domain splits from Nocaps are not used as they contain data from COCO, which we used in pretraining. Similarly, COCO results are not reported for zero-shot retrieval in order to truly assess zero shot transfer.

VGA tests the ability to correctly assign attributes to objects, e.g. "The crouched cat and the open door" vs. "The open cat and the crouched door". Co and Fo require discerning the correct ordering of a caption for a given image from permuted orderings.

Batch Size	LR	Min LR	Warmup	Optim	epsilon	betas	Dropout	clipping	wd
6144	5e-4	0	10k	AdamW	1e-6	(0.9, 0.98)	No	1.5	0.05

Table 8: Pretraining hyperparameters

C Pretraining details

Details for pretraining are listed in table 8.

Weights are initialized from unimodal stagewise pretrained checkpoints (ImageNet-22k followed by English Wikipedia/BookCorpus/OpenWebText). Image-text contrastive loss (ITC) is added for retrieval tasks and a decoder is head used for generation tasks. The parallel agent uses parameters from the first 6 layers of the main model and 4 additional layers.

D Illustrated reward calculation and explanation of symbols

Presented in figure 5 top is a visual illustration of the reward calculation, and on the botom is a table of symbols used.

E Pseudocode

Pseudocode for training PPO in parallel to the main model is provided below in algorithm 1. We use a beta of 1e-1 for entropy regularization. There is no replay buffer/mini batch size as we perform updates on the whole batch at once. An LR of 1e-4 is used and 0.2 is selected as the clipping parameter. Since episdoes are short, there is no horizon parameter either.

F Additional ablations

Additional ablations are presented in table 9.

Details The tested configurations are: learned mask selection only, random masked spans only, learned mask + masked span selection, random masked spans + random factorization order, learned masked spans + random factorization order, random masked spans + learned factorization order, learned mask selection + masked span selection + factorization order (full model), and no new features (vanilla MLM baseline).

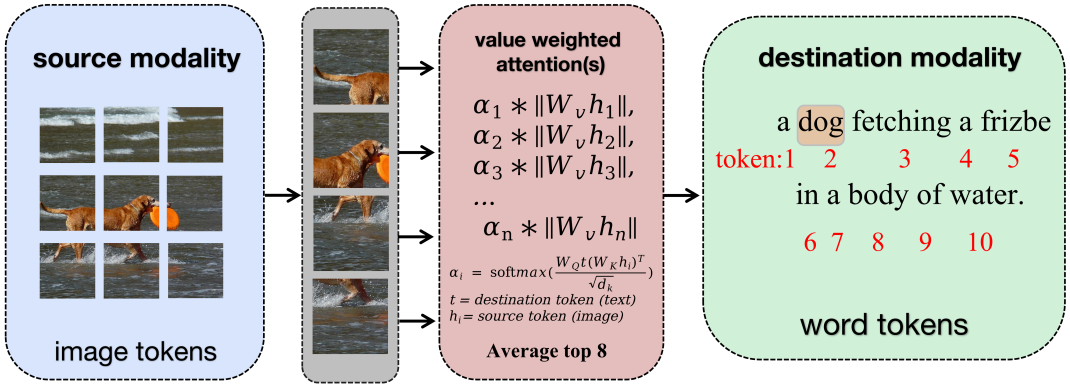
Analysis Even random spans boosts performance on VGR considerably. Learned masks alone, on the other hand, do not have this effect. Learned factorization order boosts performance over all tested tasks and boosts VGR performance in greater proportion. Random factorization order also brings a slight boost over vanilla MLM.

G Setup for experiments shown in figure 1

Experiments whose output is shown in figure 1 are designed to test the limitations of random masking in V+L pretraining. The model architecture used is as follows: A two stream encoder consisting of separate vision and text encoders are used, followed by a fusion block. The vision encoder is initiated from pretrained weights of a ViT-base and the text encoder is initiated from the first six layers of BERT. The fusion encoder consists of the last six layers of bert, with added cross attention to each layer to allow for attention to the output of the vision encoder. The architecture is similar to that used in ALBEF (Li et al., 2021). Pretraining datasets and hyper-parameters are same as those used for our main experiments. We use a different architecture from our main experiments for these "motivation experiments" because those used in motivation experiments are more widely used and thus better reflect the limitations that random masking bring in practice. Furthermore, the two stream architecture lends itself better to interpretability, as attention to image context happens in an isolated cross attention layer. If attention to image and text context share attention weights, this advantage is not present.

H Class of compatible functions f

Our proposed method is compatible with both single and dual stream architectures. Expressing the general form of f as $f(m_1, m_2, ..., m_n) = g(h_1(m_1), h_2(m_2), ..., h_n(m_n))$, where $h_i : \mathcal{M}_i \rightarrow \mathcal{H}_i$ are modality-specific encoders that maps features to a modality-specific representation space and $g :\rightarrow \mathcal{R}$ is a fusion function that combines them into the common representation space R, the only constraint is that g includes a transformer. h_i may be modality specific transformers or simply the identity function $\mathcal{I}$ in the single stream case. They may also perform preprocessing



Symbol/eq	Def.
R_i	Cross modal interaction for text token t
$\mathcal{S}$	State space, i.e. the multimodal input
$\mathcal{G}$	Rewards, i.e. mean cross modal interaction
$\mathcal{A}$	Action space, i.e. mask configuration and factorization order
a	action within action space (composite action)
γ	Discount rate
$\mathcal{T}$	Transition probs for states
π	Policy learnt by RL agent
q	Factorization Order
$a^L = q$	Last action in composite action is the selection of factorization order
$a^{<q} = t_1, t_2 ...$	Actions before last are selection of tokens to mask
V_t^π	State value estimate following policy π
Q_t^π	Action value estimate following policy π
A_t^π	Advantage function following policy π
H_θ	Entropy regularization term

Figure 5: Top: Illustration of the calculation for cross modal interaction score' Bottom: Symbols and definitions

steps like <img> </img> tag appending, as in recent breakthroughs Kosmos1 (Huang et al., 2023) and Palm-e (Driess et al., 2023), which concatenate tagged modalities before feeding them into a transformer. On that note, g must include a transformer, but can also include other modules like concatenation or contrastive mapping.In the dual stream case, self attention is used for intra-modality attention and cross attention is used for inter-modality attention. In the single stream case, self attention is used to attend to both inter and intra modality contexts.

I Generalized language modeling

Masked language modeling comes with a conditional independence assumption. From equation 1, it can be seen that the prediction of one masked token does not depend on other masked tokens; only unmasked tokens are used as context. This assumption is lifted in the case of autoregressive language modeling in which previously predicted tokens (to the left) are used as context. They can in fact be viewed as masks that were predicted at previous factorization steps. If the conditional independence assumption were removed from MLM, every single token were masked, and prediction occurred from left to right, we would have autoregressive language modeling, hence the statement that autoregressive language modeling is a special case of generalized language modeling. Furthermore, ARLM assumes one token is predicted at each factorization step. Generalized language modeling, in contrast, allows for predicting multiple tokens in parallel at each step. (MLM also features this, as there is only one "step") We thus allow for prediction of contiguous spans to maximize relational learning and cross model interaction. In summary, generalized language modeling allows for masking of any number of tokens, prediction of any subset of them at each step, and prediction in any factorization order. There are other terms to describe this such as partially autoregressive language modeling (Yang et al., 2019) and permutation language modeling (Bao et al., 2020). Implementation-wise, we prevent tokens from attending to tokens that come

Algorithm 1 Training masking agent in parallel with main model

Initialize parameters for main model θ
Share first 6 layers of θ with agent and initialize remaining three layers θ_{agent}
Initialize value function parameters ϕ
Set clip parameter ϵ and entropy regularization coefficient β

procedure ROLLOUT PHASE(batch)
    Use current θ_{agent} to compute multi-discrete action probs
    Obtain mask config and factorization order
    Send to main model and get reward
end procedure

procedure LEARNING PHASE(batch)
    Compute advantages estimates $\hat{A}$
    Compute ratio of new and old probabilities $r(\theta) = \frac{P_\theta(a|s)}{P_{\text{old}}(a|s)}$
    Compute clipped surrogate objective
    $L(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s',d)\in\mathcal{B}} \min\left(r(\theta)\hat{A}, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)\hat{A}\right) - \beta H(P_\theta)$
    Update agent's policy (last 3 layers) by ascending the stochastic gradient with respect to θ_{agent}: $\theta_{agent} \leftarrow \theta_{agent} + \alpha\nabla_{\theta_{agent}} L(\theta)$
    Update value function by descending the stochastic gradient with respect to ϕ: $\phi \leftarrow \phi - \alpha\nabla_\phi V_\phi(s)$
end procedure

for (batch t, batch t-1) $\leftarrow$ 1 to n-batches **do**
    Rollout-phase (batch t), Learning-phase (batch t-1)
    Train main model (batch t-1)
end for

} in parallel

later in the factorization sequence using "attention masks" (not to be confused with masked tokens) and train with teacher forcing. This is akin to how in ARLM tokens can only attend to context to their left. Pseudo masks are used as in UniLm-v2 (Bao et al., 2020) to allow for teacher forcing during training. Pseudo masks with the same positional embeddings as the original tokens are appended to the sequence, alongside original tokens. The reason is that, if regular [masks] were used, teacher forcing would not be possible. Pseudo masks are of course prevented from attending to original tokens to prevent information leakage. We direct readers to figures 4 and 5 of the UniLM-v2 paper, which provides excellent graphical layout. An image is truly worth a thousand words in this case.

Configuration	NLVR test-p	VE-test	Flickr-IR-R@1	Flickr-TR-R@1
baseline	78.57	78.83	80.1	89.8
m	79.23	79.01	80.0	90.1
s	79.26	79.73	80.2	90.7
m+ls	79.32	80.45	81.0	91.2
s+q	80.46	81.02	82.3	91.6
ls+q	81.29	81.56	83.1	92.2
s + lq	81.26	82.33	83.1	92.7
full model	83.76	85.01	83.6	94.6

Table 9: Further ablations