# DIFFERENCE-MASKING: Choosing What to Mask in Continued Pretraining

**Alex Wilf**[*,1], **Syeda Nahida Akter**[*,1], **Leena Mathur**[1], **Paul Pu Liang**[2], **Sheryl Mathew**[3], **Mengrou Shou**[3], **Eric Nyberg**[1], **Louis-Philippe Morency**[1]
[1] Language Technologies Institute, Carnegie Mellon University
[2] Machine Learning Department, Carnegie Mellon University
[3] Computer Science Department, Carnegie Mellon University
{awilf,sakter,lmathur,pliang,sherylm,mshou,morency}@cs.cmu.edu

## Abstract

Self-supervised learning (SSL) and the objective of masking-and-predicting, in particular, have led to promising SSL performance on a variety of downstream tasks. However, while most approaches randomly mask tokens, there is strong intuition from the field of education that deciding *what to mask* can substantially improve learning outcomes. We introduce DIFFERENCE-MASKING, an approach that automatically chooses what to mask during continued pretraining by considering what makes an unlabelled target domain *different* from the pretraining domain. Empirically, we find that DIFFERENCE-MASKING outperforms baselines on continued pretraining settings across four diverse language and multimodal video tasks. The cross-task applicability of DIFFERENCE-MASKING supports the effectiveness of our framework for SSL pretraining in language, vision, and other domains.

## 1 Introduction

Self-supervised learning (SSL) strategies have recently been applied to adapt pretrained models to specific downstream tasks by continuing to pretrain models on in-domain unlabelled data from tasks (Dery et al., 2023; Gururangan et al., 2020). This *continued pretraining* setting can be useful to test different SSL strategies such as masking.

Inspired by the distributional hypothesis in the language domain (Harris, 1954), masking is an SSL objective in which a portion of the data is hidden and the model attempts to reconstruct the hidden portion from the surrounding context. Masking has enabled breakthrough performance on tasks in a variety of domains, such as language, vision, and speech (Devlin et al., 2019; Li et al., 2021; Hsu et al., 2021; Ericsson et al., 2022), motivating interest in reseearching how masking strategies can influence representation learning in SSL.

While prior work has studied how the *amount masked* influences model learning (He et al., 2022),

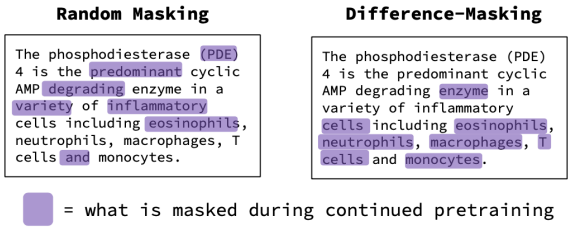



Figure 1: DIFFERENCE-MASKING automatically selects *what to mask* based on what makes the task domain *different* from the pretraining domain. In this sample from a chemistry relation extraction task (ChemProt), random masking masks more domain-irrelevant words (left) compared to our DIFFERENCE-MASKING approach, which masks domain-relevant words (right).

most masking approaches choose which parts of the data to mask randomly. Although it is understudied in SSL, deciding *what to mask* is a critical component in human education. Educators designing "fill-in-the-blank" assessments for students must decide what content to mask in order to effectively assess student understanding of a domain (Pajares and Miller, 1997; Bjork and Linn, 2006; Bae and Lee, 2018). For example, in a real-world "fill-in-the-blank" chemistry test, a teacher might choose to mask out domain-specific words ("density", "silicon") to assess student learning, instead of masking domain-irrelevant words ("example", "process").

We propose DIFFERENCE-MASKING, a novel approach for automatically selecting *what to mask* during continued pretraining for improved performance on downstream tasks. Our approach first identifies "seeds" that describe what is unique or different about an unlabeled target domain and then uses these seeds to choose what tokens to mask in continued pretraining.

In experiments spanning diverse language and multimodal video tasks (ACL-ARC, ChemProt, TVQA, Social-IQ), we find that DIFFERENCE-MASKING outperforms strong baselines, validating our hypothesis that *masking based on*

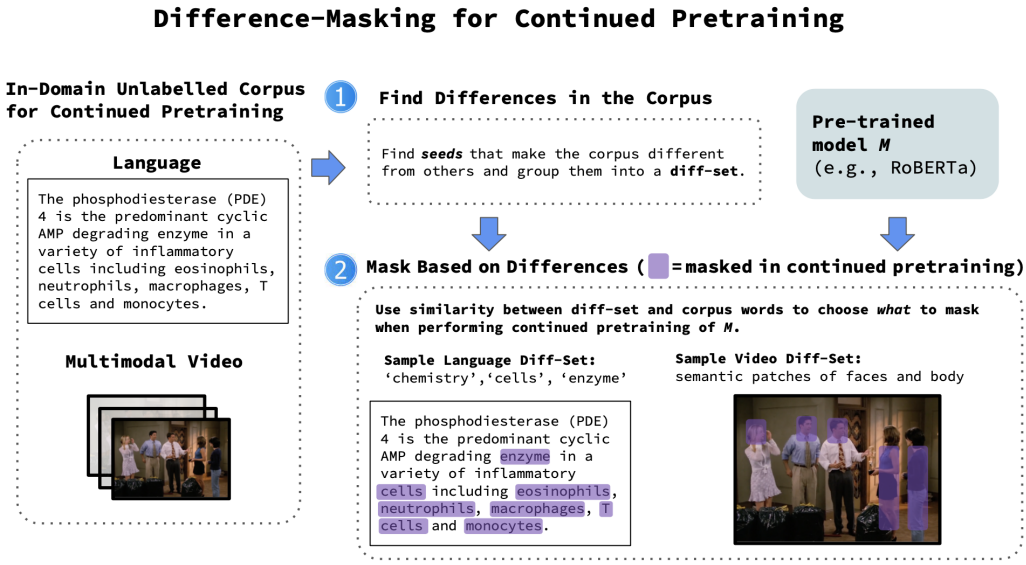



Figure 2: DIFFERENCE-MASKING: an approach to masking during continued pretraining that first selects *seed topics* relating to the downstream task, then masks tokens based on their similarity to those seed topics.

*what is unique* about a domain or task provides stronger representation learning transfer than the alternative. We provide intuition to explain the strong performance of DIFFERENCE-MASKING, along with extensive analyses and ablations to better understand the performance of our method. Code is available at https://github.com/abwilf/Difference-Masking.

## 2 Related Work

Prior masking strategies in NLP have considered task-specific approaches but have not generalized these methods to different tasks in continued pretraining. Prior masking strategies in vision have been primarily based on signals from the modeling process and *do not consider what makes domains unique*. In this section, we present an overview of masking strategies in NLP, vision, and SSL.

### 2.1 Masking in NLP

Masking relies on the distributional hypothesis, which posits that the meaning of a word can be inferred from its context (Harris, 1954). Masking in NLP has functioned as an effective SSL strategy when training models such as BERT (Devlin et al., 2019) and XL-Net (Yang et al., 2019). While most prior masking approaches in NLP approaches have been random masking, some approaches have considered non-random masking. For example, Salient Span Masking (SSM) (Guu et al., 2020) is the closest to our work. This approach uses a named entity recognition model to mask out named entities for the task of open-domain QA. Similarly, Studying Strategically (Ye et al., 2020) learns to find the answer within the context and mask it out. However, these approaches are not flexibly extended beyond open-domain QA, for example towards the chemistry (ChemProt) and citation intent (ACL-ARC) tasks we include in our experiments (Kringelum et al., 2016; Jurgens et al., 2018). Our DIFFERENCE-MASKING approach can be flexibly applied across different domains and tasks.

Another relevant work, (Arefyev et al., 2021), uses the weights of a naive-bayes classifier to determine what to mask. However, this setting differs from ours, in that our setting does not have access to labels. Finally, AANG (Dery et al., 2023) combines masking strategies such as BERT-style (Devlin et al., 2019) and XL-Net-style (Yang et al., 2019) masking and meta-learns weights for each of the different strategies. However, this work does not learn a masking strategy that is non-randomized and decided by the downstream task.

### 2.2 Masking in Vision

Prior work in vision has used the attention of the model during SSL training to determine what to mask. MST (Li et al., 2021) uses attention maps to determine "non-essential regions" to mask, while AttnMask (Kakogeorgiou et al., 2022) does the opposite by masking the most attended-to regions. SemMAE (Li et al., 2022) uses attentions to guide

the masking process by creating "semantic parts". iBot (Zhou et al., 2022) learns an online tokenizer during SSL which modifies how masks are selected, and ADIOS (Shi et al., 2022) learns an adversary which attempts to mask those regions that are most difficult for the model to predict. Although these prior approaches learn non-random masking strategies, they do not use domain-specific data to guide their masking strategy.

### 2.3 Masking in SSL

Empirical and theoretical studies of SSL have demonstrated that masking can be seen as an instance of contrastive learning (Tsai et al., 2021; Shi et al., 2022). Prior work has shown that contrastive learning has the potential to extract task-relevant information (Oord et al., 2018; Bachman et al., 2019; Zhang et al., 2016b) and discard task-irrelevant information (Tsai et al., 2021), under certain critical assumptions where the views or modalities differ only by the task-relevant information (Tian et al., 2020). Our paper contributes deeper empirical contributions to parallel these SSL studies by demonstrating how DIFFERENCE-MASKING can automatically identify task-relevant information to mask during continued pretraining, in order to improve downstream task performance.

## 3 DIFFERENCE-MASKING

This section describes the motivation and implementation of DIFFERENCE-MASKING.

### 3.1 Notation

We are given a model $M$ which has been pretrained on large amounts of multi-domain data, drawn from domain distribution $D_*$ (e.g., a model such as RoBERTa pretrained on a large multi-domain corpus). We are given downstream target task data drawn from domain distribution $D_T$, but we are not given task labels $y$.

We denote $D_{T/*}$ as a masking over $D_T$ that conceals the information that makes the domain $D_T$ different from $D_*$. For example, in a corpus related to understanding chemistry, $D_{T/*}$ would mask chemistry concepts which are likely to be in $D_T$ and unlikely to be in $D_*$. We term $D_{T*}$ as a masking over $D_T$ that masks concepts likely to be in both domains. The relationships among these variables are visualized in **Figure 3**.

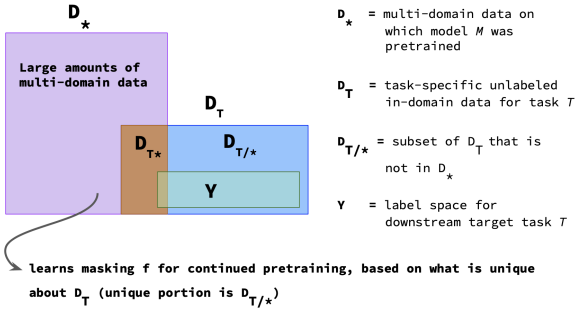



Figure 3: DIFFERENCE-MASKING tests the hypothesis that models pretrained on multi-domain data $D_*$ and trained through continued pretraining on $D_T$ will perform better on eventual finetuning if the masking strategy prioritizes information that is *unique* to $D_T$ (information in $D_{T/*}$), than if the masking strategy relies on randomly-selected masks.

### 3.2 Motivation

We build on the intuition from (Gururangan et al., 2020), who study continued pretraining and find that in-domain data can be beneficial for **task-adaptive pretraining** to adapt models to tasks defined on data from specialized domains (e.g. chemistry). DIFFERENCE-MASKING is motivated by the intuition that continued pretraining can benefit from a **task-adaptive masking strategy** as well: a masking strategy that prioritizes tokens that are more related to the downstream task's domain than tokens related to many domains. This assumption can be expressed through mutual information:

$$I(D_{T/*}; y) > I(D_{T*}; y) \quad (1)$$

We assume that $D_{T/*}$ shares more mutual information with the task label than $D_{T*}$ does.

### 3.3 Objective and Approach

The goal of DIFFERENCE-MASKING is to learn representations during continued pretraining that capture $D_{T/*}$. Most prior works mask tokens randomly during pretraining, which may choose masks in $D_{T*}$ that are not task-adaptive.

Formally, the objective of DIFFERENCE-MASKING is to create a function $f$ that masks the portions of $D_T$ that make it unique, recovering $D_{T/*}$, as denoted by the following expression:

$$\max_f I(f(D_T); D_{T/*}) \quad (2)$$

In practice, we evaluate our method using the following objective: minimizing the finetuning loss after continued pretraining ($CPT$) with masking

strategy $f$. CPT takes as input the pretrained model $M$, unlabelled in-domain data $D_T$, and a masking function $f$. CPT trains $M$ on $D_T$ using $f$ and then outputs an updated model. Finetuning ($FT$) takes this updated model and task data $D_{FT}$ as input, finetunes the model, and outputs the loss $L_{FT}$. The objective of DIFFERENCE-MASKING is represented by the following expression:

$$\min_f L_{\text{FT}}(\text{CPT}(D_T, M, f), D_{FT}) \quad (3)$$

To systematize this intuition, DIFFERENCE-MASKING proceeds in two steps:

1. **Finding Differences**: at a high level, we create a modified TF-IDF (Jones, 1972) topic model to determine which words are most commonly found in the in-domain data that are *not* commonly found in other domains. We term these words **seeds**, and the group of seeds is referred to as our **diff-set**.

2. **Masking Based on Differences**: we mask tokens based on their similarity to our diff-set.

### 3.4 Finding Differences: TF-ICF

To find what makes a domain or task unique, we use a modified TF-IDF topic model. The standard TF-IDF determines the ratio of how frequently a word appears in a *document* compared to how frequently the word appears in *other documents in a corpus*. In our case, because we are attempting to find words that make a *corpus* different from other *corpora*, the score of a word is highest when it appears frequently in our corpus (which we term $D_{CPT}$) and infrequently in any other corpus (which we term $D_*$). We denote our approach as **TF-ICF** for term-frequency, inverse-*corpus*-frequency, expressed by the following equation:

$$\text{score}_{\text{word}} = \frac{\text{freq}(\text{word}, D_{\text{CPT}})}{\text{freq}(\text{word}, D_*)} \quad (4)$$

Our diff-set is comprised of the top $K$ scoring words, per TC-ICF scoring.

To effectively capture word frequencies in the general distribution of the English Language ($D_*$), we use unigram counts derived from the Google Web Trillion Word Corpus (Brants and Franz, 2006; Norvig, 2009).

### 3.5 Masking Based on Differences

Intuitively, we aim to mask tokens based on how similar they are to the words in our diff-set. Because each token in question is a single vector and the diff-set contains $K$ vectors, it is not immediately clear how to determine a similarity score. Our intuition is that, for a set of $K$ seeds which may describe multiple different concepts defining what makes a domain or task unique, we would like to mask tokens if they relate closely to *any* of those concepts. For this reason, we determine the token's similarity based on its similarity to its nearest-neighbor seed.

Formally, we refer to the word embeddings for the words in the diff-set as $E_{diff}$ and the embedding for a given token $t_i$ in a sequence of $N$ tokens is referred to as $E_{t_i}$. The likelihood $P(M_{t_i})$ that a token $t_i$ should be masked is, then, proportional to the cosine similarity of that token's embedding and its nearest-neighbor of the diff-set embeddings:

$$P(M_{t_i}) = \frac{\max_k \cos(E_{t_i}, E_{\text{diff}_k})}{\sum_{j=1}^{N} \max_k \cos(E_{t_i}, E_{\text{diff}_k})} \quad (5)$$

where the denominator is a normalization over the length of the sequence, to ensure that the probability distribution sums to 1. We empirically validate our intuition for the nearest-neighbor strategy as opposed to a similarity function using an aggregate vector such as the centroid of the diff-set vectors in Section 5.3.

## 4 Experimental Settings

Experiments are performed to allow each model to learn as long as needed during Continued Pre-Training, only stopping when validation error increased (aka early-stopping). So all models, including the random-masking baseline, have pre-trained as much as they need to before overfitting. More experimental details can be found in Appendix D.

### 4.1 Language Experiments

#### 4.1.1 Datasets and Tasks

We consider the task-adaptive continued pretraining setting (TAPT) for language tasks as in (Gururangan et al., 2020), performing auxiliary learning on a pretrained model. TAPT is a common SSL setting for two reasons: (1) it represents a computationally-feasible way to test the effectiveness of self-supervised representation learning methods, and (2) it is realistic to modern approaches which rely heavily on pretrained models (Dery et al., 2023).

We conduct experiments with the **ChemProt** task (Kringelum et al., 2016), a relation classification task that uses chemistry documents. ChemProt

is a low-resource classification task with a large amount of in-domain unlabelled data, making it a realistic settings in which SSL is helpful in continued pretraining. The primary metric used in prior ChemProt work is accuracy.

We also conduct experiments with the **ACL-ARC** task (Jurgens et al., 2018), a citation intent task based on the ACL Anthology Reference Corpus (Bird et al., 2008). We use the same ACL-ARC dataset as prior works (Gururangan et al., 2020; Dery et al., 2023).

#### 4.1.2 Modeling

For our experiments, we reproduce the multitask setting from AANG (Dery et al., 2023) and use a pretrained RoBERTa$_{base}$ to learn separate classification heads for each auxiliary objective (in our case, the auxiliary objectives are the primary finetuning task and the SSL masking task). This is in slight contrast to the original setting from (Gururangan et al., 2020), which contained a pretraining followed by finetuning step.

Most masking approaches mask individual tokens with a random masking ratio, but we found it to be more effective to mask tokens grouped by the word they correspond to, regardless of the subword-tokenization. Our intuition is that for specialized domains, such as the domain in the ChemProt context, words such as "phosphates" would be tokenized into "phos" and "-phates", either of which is easy to predict given the other, but which does not correspond to learning a general understanding of the specialized domain. We provide further discussion of this design decision in Appendix A.

### 4.2 Multimodal Video Experiments

#### 4.2.1 Datasets and Tasks

We consider the same task-adaptive setting for two multimodal video understanding tasks that have an emphasis on social interactions. We chose these tasks because of the intuition in Gururangan et al. (2020) that TAPT will be most effective when task data is a narrowly-defined subset of the broader domain. Since pretrained multimodal models such as MERLOT (Zellers et al., 2022) are trained on a broad domain of 20 million videos that do not all contain social interactions, we hypothesize that these two tasks will serve as effective benchmarks for understanding DIFFERENCE-MASKING's capabilities as a task-adaptive masking strategy.

**TVQA** TVQA (Lei et al., 2018) is a dataset containing 21,792 videos from 6 American television shows and questions and answers related to the videos. Each question is paired with 5 answer choices, one correct and 4 incorrect, and corresponding video, audio, and subtitles.

**Social-IQ** Social-IQ (Zadeh et al., 2019) is a dataset containing 1250 videos of social situations and questions and answers pertaining to the videos. Each question has corresponding video, audio, and subtitles. Each question has 3 incorrect and 4 correct answers, resulting in 12 samples for each question (with 1 correct option and 3 incorrect options for each sample).

#### 4.2.2 Modeling

The TAPT experiments were conducted using the MERLOT-Reserve (base) model (Zellers et al., 2022). MERLOT-Reserve is a large multimodal transformer pretrained on YT-Temporal-1B, a dataset of 20 million Youtube videos. MERLOT-Reserve learns multimodal representations of videos, given audio, subtitle text, and video frames by masking and predicting segments of either audio or text given the corresponding video frames. We reproduce MERLOT-Reserve's original training on TVQA: we decompose samples in Social-IQ and TVQA from the form (Question, All Answers, Video Information) into a list of 3-tuples: (Question, Candidate Answer, Video Information). MERLOT scores each candidate answer independently, given the question and video, and is trained with loss that encourages the model to minimize estimated likelihood of incorrect answers and maximize likelihood of correct answers. MERLOT's training hyperparameters are in Appendices B.4 and D.1.2 of their paper (Zellers et al., 2022).

From video frames, we mask image patches into 16x16 patches as determined by MERLOT-Reserve's backbone image transformer ViT (Dosovitskiy et al., 2021). Similar to our language task, where we mask tokens corresponding to the same word, we found that SSL training improved when we masked image patches semantically-grouped together based on object type. Details about this experimental design decision are in Appendix B.

### 4.3 Baselines

**Random Masking** Most masking approaches choose tokens to mask with uniform random probability. Formally, the likelihood $P(M_{t_i})$ that a token

| | | Language | | Multimodal | |
|---|---|---|---|---|---|
| | | **ACL-ARC** | **ChemProt** | **TVQA** | **Social-IQ** |
| (1) | Random Masking | 66.30 | 82.82 | 73.75 | 69.05 |
| (2) | AttnMask | 65.57 | 82.11 | 81.57 | 69.61 |
| (3) | Selective Masking* | 69.06 | 82.94 | - | - |
| (4) | EntityBERT* | - | 82.04 | - | - |
| (5) | DIFFERENCE-MASKING | **74.04** | **83.94** | **81.73** | **71.37** |

Table 1: We find that DIFFERENCE-MASKING outperforms both the widely-used Random Masking approach and AttnMask approach in both the language and multimodal experimental settings. DIFFERENCE-MASKING also outperforms the Selective Masking* and EntityBERT* baselines in language domain. (*) We note that Selective Masking and Entity Masking baselines are only designed for language, and EntityBERT can only be implemented with a domain-specific pretrained model. In this case, we use BioBERT to select entities in ChemProt.

$t_i$ in a sequence of length $N$ will be masked is

$$P(M_{t_i}) = \frac{1}{N} \tag{6}$$

**AttnMask** AttnMask (Kakogeorgiou et al., 2022) is a *domain-agnostic* approach in which the likelihood of masking a given token is proportional to how attended-to that token is by the [CLS] token, averaged across the different heads of the transformer. Formally, we define a function $f_{att}$ which takes in model $M$, sequence of tokens $t$, and target token index $i$ and outputs how attended-to token $t_i$ is. The probability that token $t_i$ will be masked is

$$P(M_{t_i}) \propto f_{att}(M, t, i) \tag{7}$$

**Selective Masking** Selective masking (Gu et al., 2020) chooses tokens to mask based on whether adding each token will improve downstream task accuracy. Notably, this approach *uses downstream task labels to guide the choice of mask* in continued pretraining, whereas DIFFERENCE-MASKING *is entirely self-supervised*. The probability that token $t_i$ will be masked in Selective Masking is proportional to the difference between the downstream task performance when using the full sequence $t_{[:]}$ versus using only the sequence up to and including the token $t_i$. We utilize the same corpus for both $D_D$ and $D_T$ to ensure a fair comparison with our proposed approach in the TAPT setting, and report results on the Language experiments because the method has not been tested yet for different vision tokenization strategies.

$$P(M_{t_i}) \propto P(y \mid t_{[:]}) - P(y \mid t_{[:i]}) \tag{8}$$

**EntityBERT** EntityBERT (Lin et al., 2021) presents an approach to masking tokens based on whether they are part of "entities", as defined by a domain-specific entity-tagger. The original paper, tested on the clinical domain, continually pretrains the PubMedBERT model. However, because the ChemProt domain is different, we implement this baseline by using the BioBERT (Lee et al., 2019) model that more closely aligns with our downstream task. To identify the entities, we use the BioBERT model fine-tuned in NER task with BC5CDR-chemicals (Li et al., 2016) and BC4CHEMD (Krallinger et al., 2015) corpus. We report results for the EntityBERT approach on the ChemProt task, only, because EntityBERT can only be implemented with a domain-specific pretrained model. It is infeasible to reproduce domain-specific tagging for ACL-ARC domains without unfairly representing the method. The probability that a token $t_i$ is masked is related to whether it is part of a recognized entity or not.

## 5 Results and Analysis

We find that DIFFERENCE-MASKING outperforms both the Random Masking and AttnMask baselines in both the language and multimodal TAPT settings. Moreover, our approach outperforms our other two baselines (Selective Masking and Entity Masking) specifically in language domain. The results are presented in **Table 1**. We investigate three questions to analyze the performance of DIFFERENCE-MASKING:

1. Why does DIFFERENCE-MASKING outperform the baselines? How does the choice of *what is masked* impact downstream task performance? (Section 5.1 and 5.2)

2. How does our choice of nearest-neighbor approach in DIFFERENCE-MASKING impact performance? (Section 5.3)

3. Why does DIFFERENCE-MASKING perform better relative to baselines on some tasks than on others? (Section 5.4)

### 5.1 What is masked?

We investigate why our method outperforms the baselines by looking at which words are masked during DIFFERENCE-MASKING. We focus this section's analysis on our language tasks and further analyze our video tasks in Section 5.4.

Surprisingly, in the ChemProt task we find that some of the words selected for masking by DIFFERENCE-MASKING are *the same words as the labels for the downstream task*. ChemProt is a relation extraction task, where labels include "inhibitor", "antagonist", and "activation". As shown in **Figure 4(a)**, the word most often masked by DIFFERENCE-MASKING is "activity", followed by "inhibited", "inhibitor", and, some words later, "antagonist". We find that DIFFERENCE-MASKING's automatic process for finding seeds in **unlabelled** domain data is reconstructing the labels of the downstream task without using any label information. *This highly-promising phenomena suggests that* DIFFERENCE-MASKING *is capable of masking in such a way that the task performed during SSL is similar to the downstream task.*

We find that the most frequently masked words in the ACL-ARC task had an interesting grounding in human intuition as well: the most frequently masked words closely-aligned with the ACL paper submission tracks describing the high-level topics for papers. For example, some of the most frequently masked words were "learning", "information", "translation", "semantic", and "lexical". These words closely correspond to the ACL submission tracks "Machine Learning for NLP", "Information Extraction", "Machine Translation", and "Semantics: Lexical". A full visualization of the most frequently masked words is presented in **Figure 4(b)**. Since submission tracks for ACL can be seen as a set of topics that span the space of ACL papers, this provides further evidence for our hypothesis that *masked words will align closely with what makes each domain unique*.

### 5.2 Why does DIFFERENCE-MASKING outperform baselines on language tasks?

We hypothesize that our approach outperforms the Selective Masking and EntityBERT approaches because both have limitations that restrict their ability to generalize to our downstream task domains. In the original work, Selective Masking uses in-domain data $D_D$ that is three orders of magnitude larger than the task data $D_T$. However, in our TAPT setting we only have access to the lower-resource unlabelled data from $D_T$. Our experiments suggest that Selective Masking fails to generalize well to a smaller continued pretraining dataset.

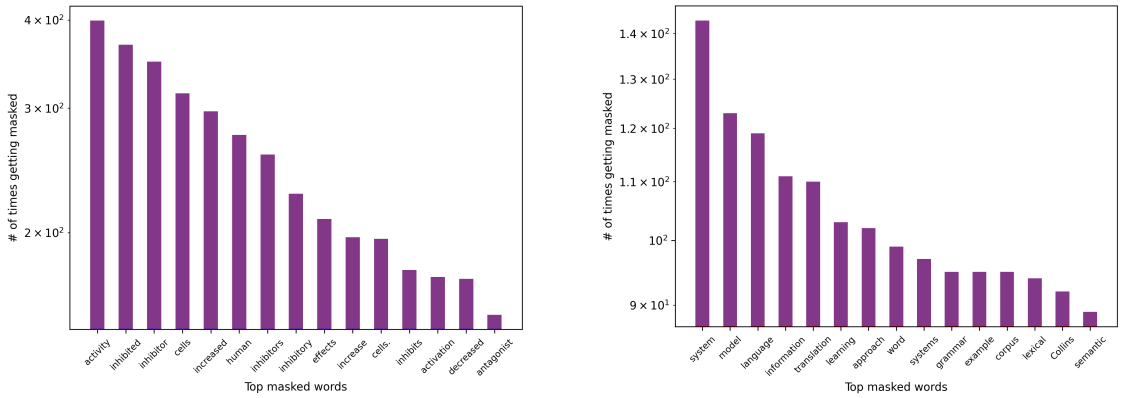
Similarly, the EntityBERT masking strategy was also trained on a much larger dataset (4.6M sentences as opposed to 5k), and struggles to generalize to a different domain. The EntityBERT approach uses a domain-specific model for tagging which, while well suited to the original clinical domain, does not seem to generalize well to the academic chemistry domain. In initial experiments, we found that the EntityBERT approach with the original model performed slightly worse than the Random Masking baseline. We implemented these experiments using BioBERT for tagging, a model which is specifically designed for scientific biomedical texts, but it still did not perform as well as DIFFERENCE-MASKING.

Although these masking strategies work well in their original settings, they have difficulty generalizing to the challenging TAPT setting across domains. This is a strength of our method, which is able to use a relatively small continued pretraining dataset and is domain-agnostic. We analyze the differences between the other baselines: Random and Attn-Mask, and DIFFERENCE-MASKING in detail in section 5.4 below.

### 5.3 How does the nearest-neighbor seed selection contribute to performance?

We hypothesize that, by selecting the nearest-neighbor seed to each token, *our algorithm becomes robust to poor seed selections* made by the TF-ICF algorithm described in Section 3.5.

We validate this hypothesis by comparing TF-ICF seed word rankings to the seeds that were most commonly "chosen" (the closest to a word that was masked). For example, the word "charniak" was ranked first in the TF-ICF scores, which led to its selection as a seed word. "Charniak" is a highly-specific concept that is not the type of seed we would expect would perform well at articulating the unique aspects of the space of ACL papers. However, relatively few words that were masked were closest to "charniak" as a seed, making it only the 11/20th most "chosen" seed. Figure 5 depicts




(a) ChemProt (b) ACL-ARC

Figure 4: The most frequently masked words chosen by the DIFFERENCE-MASKING algorithm across the ChemProt and ACL-ARC tasks. We find that for the ChemProt dataset, the masks we find automatically through unlabelled data partially recover the end task labels.

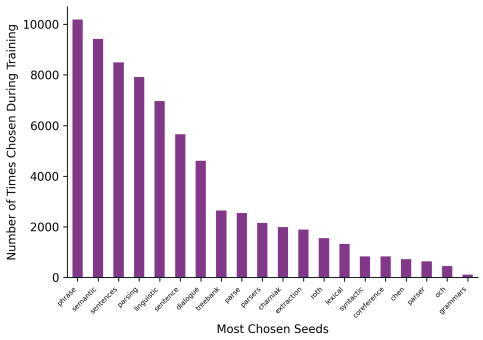



Figure 5: For the ACL-ARC task, we display how often each chosen seed word was the word which caused a token to be masked during training, by virtue of being the most similar seed to the given token.

the most masked words for the ACL-ARC task. Their ordering aligns well with the uniqueness of the ACL-ARC task space, *qualitatively supporting our hypothesis that nearest-neighbor reduces the effect of poor seed selection*.

We further investigate this hypothesis by evaluating a different strategy for token-scoring: the centroid. Instead of scoring a token based on its similarity with the *most similar seed*, we score each token based on its similarity with the *centroid of the seed word embeddings*. While poor seed choices would directly affect the centroid through the mean operation, they may affect the nearest-neighbor strategy less because the poor seed may not be chosen as the closest seed for masking. We re-conducted our TAPT experiments with this centroid strategy and report our results in Table 2. We find that the nearest-neighbor strategy does, in fact, outperform the centroid strategy, especially on the ACL-ARC task, further validating our hypothesis.

| | **ACL-ARC** | **ChemProt** |
|---|---|---|
| Centroid | 69.02 | 83.66 |
| Nearest-Neighbor | **74.04** | **83.94** |

Table 2: Ablating DIFFERENCE-MASKING's seed-scoring function based on nearest-neighbor and replacing it with one based on similarity with the seed embeddings' centroids leads to a performance degradation. This provides evidence for our hypothesis that the nearest-neighbor scoring function helps make DIFFERENCE-MASKING robust to poor seed selections.

This result leads us to hypothesize that *some of the effectiveness of the nearest neighbor strategy is due to the variance of the cluster of seed embeddings*. With a very low variance, we would expect that the space spanned by embeddings would be too restrictive, because tokens would be masked very similarly. We investigate this hypothesis by evaluating DIFFERENCE-MASKING across four values of $K$, ranging from 5 to 20 in increments of 5. We find that, for both tasks, the variance of seed embeddings correlates strongly with the results. Results are shown in **Table 3**. Because we see clear correlation values, we hypothesize that future work may find it fruitful to consider optimizing DIFFERENCE-MASKING's seed selection further, based on whether individual datasets benefit from higher or lower variance seed clusters.

| ACL-ARC | ChemProt |
|---|---|
| -0.9576 | 0.8463 |

Table 3: We find high Pearson's Correlation Coefficient results between the variance of the seed embeddings and the performance of the model.

### 5.4 Why does DIFFERENCE-MASKING perform better relative to baselines on Social-IQ than on TVQA?

We were particularly curious why, in the TAPT settings, DIFFERENCE-MASKING outperformed baselines by a larger margin on Social-IQ than on TVQA. We hypothesize that this result suggests that *the information contained in the visual representations of people is less relevant to TVQA than to Social-IQ*.

Qualitatively, we test this hypothesis by analyzing both the Social-IQ and TVQA questions and answers. We find that 55.6% of the TVQA questions begin with "what", and that many of these questions correspond to abstract concepts or objects, instead of people. For example, "What was on the back of the pills when the patient asked for cough pills?" is a TVQA question that does not relate to visual representations of people. As (Lei et al., 2018) mention, other questions are designed to rely on multiple modalities to infer answers. For example the question "What was Castle right about when Beckett is speaking?" relies on the visual modality to localize when Beckett is speaking, and the rest of the inference relies on text. These types of questions could make continued pretraining on the visual domain, alone, less relevant.

In contrast to TVQA, all questions in Social-IQ are focused on understanding social interaction; therefore, we posit that masking out visual representations of people (faces and bodies) should be beneficial for learning this downstream task, even if the questions rely on multimodal data. For example, some questions in Social-IQ include "What is the man in the black T shirt shirt thinking?" and "What is the overall tone of the conversation?" These are examples of questions in which a more nuanced understanding of the visual modality would prove helpful to the Social-IQ task performance.

Empirically, we validate this hypothesis by analyzing how often DIFFERENCE-MASKING masks tokens from within person bounding boxes for each of the tasks. If our hypothesis that visual representations of people are less important to TVQA than to Social-IQ holds, then we would expect our algorithm to mask fewer tokens from person bounding boxes in TVQA videos than in Social-IQ videos.

| Method | TVQA | Social-IQ |
|---|---|---|
| Random | .17 | .15 |
| AttnMask | .38 | .19 |
| DIFFERENCE-MASKING | .40 | .90 |

Table 4: For each method, we analyze how often tokens are chosen to be masked from within bounding boxes over people as opposed to objects.

In **Table 4**, we confirm this expectation. We find that the best setting of DIFFERENCE-MASKING masks substantially fewer visual tokens corresponding to people than to other objects in TVQA (.40) as opposed to Social-IQ (.90). In Social-IQ, where the performance difference over the closest baseline is more pronounced (1.76%), the best setting of DIFFERENCE-MASKING draws 90% of its masked visual tokens from representations of people.

## 6 Conclusion

In this paper we introduce DIFFERENCE-MASKING, a method for identifying what makes a corpus unique and using this information to guide a strategy that chooses *what to mask* during SSL continued pretraining. We find that our method outperforms strong baselines across diverse language and multimodal video understanding tasks. We provide a detailed discussion of *what is masked* in DIFFERENCE-MASKING and why our method performs well on various tasks. The cross-task applicability of DIFFERENCE-MASKING supports the effectiveness of our framework for SSL pretraining in language, vision, and other domains.

## 7 Limitations

As described in Section 3, DIFFERENCE-MASKING is based on the intuition that it is more beneficial to mask based on what is unique ($D_{T/*}$) about a downstream task's domain. However, it is challenging to find what makes a domain unique; therefore, our method is an approximation of $D_{T/*}$. We believe future work may find it fruitful to investigate additional methods for approximating $D_{T/*}$. In Section 5, we provided intuition, empirical results, and analysis to understand why our method outperformed attention

masking baselines by a larger margin on Social-IQ than on TVQA. A broader investigation of why DIFFERENCE-MASKING during pretraining is beneficial by a larger margin to some downstream tasks would be helpful to the community and represents a fruitful research direction.

## 8 Ethics Statement

We believe strongly that self-supervised learning is a promising direction for the machine learning community. This does not discount, in any way, the salient arguments made about the social and enviromental risks of large models (Bender et al., 2021; Strubell et al., 2019). We believe that works such as ours, which study SSL in a resource-constrained context, both increase access to those with limited compute resources and conform to a more environmentally-sustainable way of doing research.

## 9 Acknowledgements

This material is based upon work partially supported by BMW, National Science Foundation awards 1722822, 1750439, DGE2140739, and National Institutes of Health awards R01MH125740, R01MH132225, R01MH096951 and R21MH130767. In addition, this work was supported by the TPU Research Cloud (TRC) program, which made running the MERLOT experiments accessible to the authors. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred.

## References

Nikolay Arefyev, Dmitrii Kharchev, and Artem Shelmanov. 2021. Nb-mlm: Efficient domain adaptation of masked language models for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9114–9124.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

Minryoung Bae and Byungmin Lee. 2018. Effects of text length and question type on test-takers' performance on fill-in-the-blank items in korean csat. *English Teaching*, 73(4):149–174.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the sixth international conference on language resources and evaluation (LREC'08)*.

Robert A Bjork and Marcia C Linn. 2006. The science of learning and the learning of science. *Aps Observer*, 19(3).

Thorsten Brants and Alex Franz. 2006. All our n-gram are belong to you. https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html. Accessed: 2023-05-22.

Lucio M. Dery, Paul Michel, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. 2023. AANG : Automating auxiliary learning. In *The Eleventh International Conference on Learning Representations*.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. 2022. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62.

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020. Train no evil: Selective masking for task-guided pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6966–6974, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey,

and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

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 *Computer Vision – ECCV 2022*, pages 300–318. Springer Nature Switzerland.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

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

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016. Baw068.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. 2021. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.

Peter Norvig. 2009. Natural language corpus data. *Beautiful data*, pages 219–242.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Frank Pajares and M David Miller. 1997. Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *The journal of experimental education*, 65(3):213–228.

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.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*.

Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. 2021. Do different tracking tasks require different appearance

models? *Advances in Neural Information Processing Systems*, 34:726–738.

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.

Qinyuan Ye, Belinda Z Li, Sinong Wang, Benjamin Bolte, Hao Ma, Wen-tau Yih, Xiang Ren, and Madian Khabsa. 2020. Studying strategically: Learning to mask for closed-book qa. *arXiv preprint arXiv:2012.15856*.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016a. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Richard Zhang, Phillip Isola, and Alexei A Efros. 2016b. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2022. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*.

## A Masking Language Tokens

In Section 4.1.2 we describe the motivation for using a word-level instead of token-level masking strategy. Empirically, we find that this improved performance substantially, as shown in the results in Table 5.

| | ACL-ARC | ChemProt |
|---|---|---|
| Token | 0.6501 | 0.8224 |
| Word | 0.7404 | 0.8394 |

Table 5: We validate our hypothesis that masking tokens using DIFFERENCE-MASKING at the word-level is more effective than masking at the token-level.

## B Masking Video Tokens

Following the intuition from language, we hypothesize that masking and predicting small patches of an image may be testing *local* capabilities (e.g. determining what an eye looks like from the rest of the face) rather than *global* capabilities (e.g. determining what a person's face looks like from the rest of the scene, including other people's faces).

Accordingly, instead of masking low-level image patches, we mask groups of patches corresponding to a higher level semantic entity: bounding boxes over objects in the image. We see this approach as a visual analogue for masking at the word-level instead of the token-level in our language experiments. We found that $K = 1$ performed much better than other values, where the selected seed word was "person". We considered two possible bounding boxes associated with people: bounding boxes over faces and bodies. We evaluated both options and found that considering entire bounding boxes over people's bodies (including their faces) performed the best. These results are shown in Table 6.

| Masking Strategy | TVQA | SiQ |
|---|---|---|
| Random Masking | 73.75 | 69.05 |
| DIFFERENCE-MASKING (Face) | 81.51 | 69.13 |
| DIFFERENCE-MASKING (Body) | **81.73** | **71.37** |

Table 6: Results of DIFFERENCE-MASKING on multimodal video understanding benchmarks TVQA and Social IQ. DIFFERENCE-MASKING leads to an improvement of 8% and 2% accuracy; metrics are 5- and 4-class accuracy, respectively.

We extracted body detection coordinates using UniTrack (Wang et al., 2021) and face detection coordinates using MTCNN (Zhang et al., 2016a).

## C Performance by Varying Number of Seed Words

We conducted an analysis on the effective number of seed words for calculating similarity and observed the impact on performance for two tasks. The results, as shown in Figure 6, indicate that using 20 seed words yields the optimal performance for both tasks.

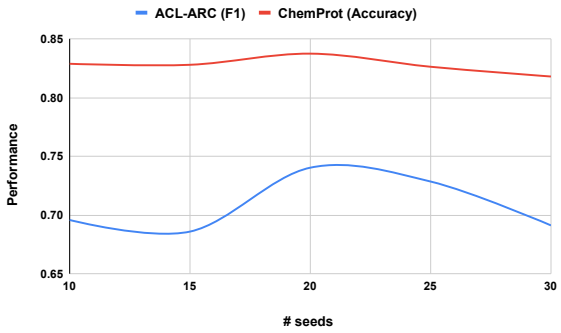



Figure 6: Number of seed words vs performance on DIFFERENCE-MASKING. We get the best performance for both tasks using 20 seed words.

## D Detailed Experimental Setup

In this section, we provide an overview of the experimental conditions utilized in our study. To ensure fair comparisons with our baselines, we maintain a consistent set of hyperparameters for both continuous pretraining and fine-tuning. For language tasks, we largely adhere to the hyperparameters employed in (Gururangan et al., 2020). Throughout our experiments, we maintain a masking ratio of 25% in both language and multimodal settings. We adopt a static masking strategy, replacing masked tokens with random values.

| Hyperparameters | CPT | | FT | |
|---|---|---|---|---|
| | Language | Multimodal | Language | Multimodal |
| *learning_rate* | 0.0001 | 0.000005 | 1.00E-06 | 5.00E-06 |
| *num_train_epochs* | 150 | 20 | 10 | 20 |
| *eval_every_n_epochs* | 30 | 1 | 1 | 1 |
| *patience* | 20 | 5 | 3 | 5 |

Table 7: List of hyperparameters used in both continuous pretraining (CPT) and finetuning (FT).