Look Ahead or Look Around?
A Theoretical Comparison Between Autoregressive and Masked Pretraining

Qi Zhang [* 1] **Tianqi Du** [* 1] **Haotian Huang** [2] **Yifei Wang** [3] **Yisen Wang** [1 4]

Abstract

In recent years, the rise of generative self-supervised learning (SSL) paradigms has exhibited impressive performance across visual, language, and multi-modal domains. While the varied designs of generative SSL objectives lead to distinct properties in downstream tasks, a theoretical understanding of these differences remains largely unexplored. In this paper, we establish the first theoretical comparisons between two leading generative SSL paradigms: autoregressive SSL and masked SSL. Through establishing theoretical frameworks, we elucidate the strengths and limitations of autoregressive and masked SSL within the primary evaluation tasks of classification and content generation. Our findings demonstrate that in classification tasks, the flexibility of targeted tokens in masked SSL fosters more inter-sample connections compared to the fixed position of target tokens in autoregressive SSL, which yields superior clustering performance. In content generation tasks, the misalignment between the flexible lengths of test samples and the fixed length of unmasked texts in masked SSL (vs. flexible lengths of conditional texts in autoregressive SSL) hinders its generation performance. To leverage each other's strengths and mitigate weaknesses, we propose diversity-enhanced autoregressive and variable-length masked objectives, which substantially improve the classification performance of autoregressive SSL and the generation performance of masked SSL. Code is available at https://github.com/PKU-ML/LookAheadLookAround.

---
[*]Equal contribution [1]National Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, China [2]Sun Yat-Sen University, China [3]MIT CSAIL, USA [4]Institute for Artificial Intelligence, Peking University, China. Correspondence to: Yisen Wang <yisen.wang@pku.edu.cn>.
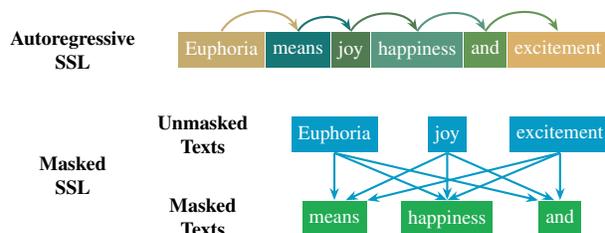
Figure 1. Illustration of two primary paradigms in generative SSL: autoregressive SSL and masked SSL.

1. Introduction

Recently, self-supervised learning (SSL) paradigms have emerged to be promising in a wide variety of domains (He et al., 2022; Radford et al., 2018; Devlin et al., 2019; Radford et al., 2021). Generative SSL, characterized by its ability to reconstruct target tokens from conditional tokens within the same sequence, stands out for its distinct advantages in fine-tuning (He et al., 2022), zero-shot learning (Brown et al., 2020; Radford et al., 2021), and many other downstream tasks. In general, generative SSL can be divided into two categories: autoregressive SSL (represented by GPT (Radford et al., 2018)) and masked SSL (represented by BERT (Devlin et al., 2019)). As shown in Figure 1, the key difference between them lies in the choice of target tokens. Taking the language models as an example, autoregressive SSL predicts the next word in a sentence based on preceding words, while masked SSL predicts randomly masked words in a sentence based on bidirectional words. As a result, these two generative SSL models demonstrate significant differences in many aspects. For example, on two primary evaluation tasks of pretrained models, i.e., classification and content generation tasks, masked SSL tends to perform better in classification tasks, whereas autoregressive SSL shows superior performance in content generation tasks (Table 1). Despite empirical evidence supporting their differing efficacies, the underlying reasons explaining the distinct properties of autoregressive and masked SSL remain unclear.

In this paper, we propose the first theoretical comparison between autoregressive and masked SSL, with a focus on classification and content generation tasks as widely con-

Table 1. The autoregressive and masked SSL models with similar scales of parameters show distinct properties in classification (evaluated by GLUE score (Wang et al., 2018a)) and content generation tasks (evaluated by perplexity on WikiText-2 (Merity, 2016)). More details can be found in Appendix B.1.

Pretraining	GLUE ($\uparrow$)	Perplexity ($\downarrow$)
Autoregressive	72.8	**22.8**
Masked	**80.5**	34.1

sidered in the literature (Radford et al., 2018; Devlin et al., 2019; Dong et al., 2019). To analyze the classification performance of autoregressive and masked SSL uniformly, we establish the unified connections between generative SSL objectives and matrix decomposition objectives. Based on these connections, we establish a theoretical downstream guarantee for generative SSL and describe the downstream classification performance by the properties of generative SSL co-occurrence matrices. From this perspective, we conduct both empirical and theoretical comparisons of autoregressive and masked SSL. Our analysis demonstrates that randomly selected target tokens in masked SSL could lead to more connections between different semantically similar samples, leading to superior downstream classification performance compared with autoregressive SSL. To bridge this gap, for autoregressive SSL we propose a diversity-enhanced autoregressive objective, which improves its classification performance in real-world datasets.

On content generation tasks (Dong et al., 2019), we first compare the generation ability of masked and autoregressive SSL on different sample lengths. We observe that the masked SSL exhibits particularly inferior performance when generating short texts. Intuitively, it is attributed to a misalignment between the pretraining and downstream objectives. For instance, in a question-answering scenario, the length of questions is variable while the length of masked texts in the masked SSL objective is fixed. As a result, the model cannot find the specific part of the information that is used to infer the masked texts during the pretraining process and may output a random guess. To formalize this intuition, we theoretically compare autoregressive and masked SSL generation abilities in the sample with different lengths. We prove that the performance gap is decided by the degree of the misalignment between pretraining objectives and downstream evaluation tasks. Building on this analysis, for masked SSL we introduce variable-length masked objectives, which exhibit significant improvements in its content generation performance.

We summarize our contributions as follows:

- We establish the first unified theoretical framework to compare the downstream classification performance

of autoregressive and masked SSL, showcasing the unique advantage of masked SSL in enhancing sample connectivity.

- We build a theoretical comparison in generation abilities between autoregressive and masked SSL, which reveals that masked SSL's underperformance is attributed to the misalignment between the pretraining and downstream objectives.

- To further verify the theoretical insights, we propose the diversity-enhanced autoregressive and variable-length masked objectives, which substantially improve the classification performance of autoregressive SSL and the generation performance of masked SSL.

2. Related Work

Self-supervised Learning. Traditional deep learning relies on label information to cluster semantically similar samples in the feature space (Krizhevsky et al., 2012). However, due to the expensive costs of labeled data, the performance of deep learning is constrained. As a result, self-supervised (SSL) paradigms are proposed to extract meaningful representations from unlabeled data. Among various self-supervised objectives (e.g., generative SSL objectives (Devlin et al., 2019; He et al., 2022), contrastive SSL objectives (Chen et al., 2020b; Giorgi et al., 2021) and context-based SSL objectives (Larsson et al., 2016; Gidaris et al., 2018)), generative paradigms have achieved impressive performance in vision, language, and multi-modal domains.

Generative SSL. Generative SSL objectives can be broadly categorized into two main types: autoregressive SSL (predicting the next token of the prefix information) and masked SSL (predicting random tokens with the bidirectional information). In language models, the autoregressive representative model GPT (Radford et al., 2018) and its variants (Radford et al., 2019; Brown et al., 2020) obtain promising performance in content generation, long-text comprehension, and many other downstream tasks. Concurrently, the masked SSL representative language model BERT (Devlin et al., 2019) and its variants (Lan et al., 2020; Liu et al., 2019) also show unique advantages and achieve remarkable performance in downstream tasks including sequence labeling, text classification, etc. Besides the language tasks, the generative SSL objectives have also shown great potential in visual representation learning (He et al., 2022; Xie et al., 2022) and multi-modal representation learning (Bachmann et al., 2022).

Theoretical Understanding of SSL. Despite the impressive empirical performance exhibited by various self-supervised objectives in diverse downstream tasks, the theoretical understanding of the mechanisms behind them (especially for the generative SSL objectives) is still under-explored. In

the language domain, previous studies mainly demonstrated the advantages of different pretraining objectives through empirical comparisons. For example, Wang et al. (2022a) shows that the autoregressive SSL models can achieve superior performance in zero-shot tasks, while the masked SSL models perform better in fine-tuning tasks. Besides, Dong et al. (2019) observes that the autoregressive SSL is more proficient in generation tasks, while the masked SSL outperforms in classification tasks. In the vision domain, researchers mainly focus on the theoretical analysis of contrastive objectives instead of generative SSL objectives. For example, they have established the theoretical connections between the contrastive objectives and the downstream performance (Saunshi et al., 2019; HaoChen et al., 2021) or analyzed them from a mutual information perspective (Oord et al., 2018; Tian et al., 2020). Different from prior studies, our work mainly focuses on uncovering the mechanism behind the generative SSL objectives by theoretically comparing the two major generative SSL objectives.

3. Mathematical Formulation

We begin by introducing the basic notations of autoregressive and masked SSL. Without loss of generality, we take the language models as an example, and the analysis can be easily extended to visual and other domains. In general, both autoregressive and masked SSL comprise two stages: self-supervised pretraining and downstream evaluation.

3.1. Pretraining Objectives

Taking an unsupervised language dataset $\mathcal{D} = \{x_i\}_{i=1}^N$ as an example, each sample $x_i = (x_{i,1}, \cdots, x_{i,s})$ consists of s tokens. The objective of autoregressive SSL is to predict the subsequent word based on prefix texts, i.e., minimizing the following Negative Log-Likelihood (NLL) Loss:

$$\mathcal{L}_{ar}(\Theta) = -\mathbb{E}_{x_i} \sum_k \log P(x_{i,k}|x_{i,1}, \cdots, x_{i,k-1}; \Theta),$$

where Θ are the weights of the neural network.

During the masked SSL pretraining process, for each sample $x_i = (x_{i,1}, \cdots, x_{i,s})$, we draw a random mask $m \in \{0,1\}^s$ (drawing 0 with probability ρ_m, i.e., the mask ratio) and apply the mask to generate the unmasked text x_i^1 and masked text x_i^2:

$$x_i^1 = x_i[m] \in \mathbb{R}^{s(1-\rho_m)}, x_i^2 = x_i[1-m] \in \mathbb{R}^{s\rho_m}.$$

The objective of masked SSL is to predict the masked tokens from unmasked texts, i.e.,

$$\mathcal{L}_m(\Theta) = -\mathbb{E}_{(x_i^1, x_i^2)} \sum_k \log P(x_{i,k}^2|x_{i,1}^1, \cdots, x_{i,s(1-\rho_m)}^1; \Theta),$$

where $x_{i,k}^1, x_{i,k}^2$ respectively denote the k-th token of the unmasked text x_i^1 and masked text x_i^2.

In practice, both the representative autoregressive SSL model GPT (Radford et al., 2018) and the representative masked SSL model BERT (Devlin et al., 2019) use the multi-layer Transformer to generate the predicted distribution and apply the Cross-Entropy (CE) loss. Let $f(X_i) \in \mathbb{R}^t$ be the output of the network and $W \in \mathbb{R}^t \times \mathbb{R}^{N_D}$ be the token embedding matrix that transforms the output features to the predicted distribution of N_D different tokens. Then the predicted probability in autoregressive and masked SSL can be uniformly reformulated as:

$$P(X_i^+|X_i; \Theta) = \frac{\exp((Wf(X_i))^\top \mathbb{1}_{X_i^+})}{\sum_{X_i^-} \exp((Wf(X_i))^\top \mathbb{1}_{X_i^-})},$$

where X_i are the conditional tokens, X_i^+ are the target tokens and X_i^- are the independent tokens. To be specific, taking the sentence in Figure 1 as an example ("Euphoria means joy, happiness and excitement"), for autoregressive SSL, X_i represents the prefix tokens (e.g., [Euphoria, means, joy, happiness, and]) and X_i^+ is the subsequent token ([excitement]). For masked SSL, X_i represents the unmasked tokens (e.g., [Euphoria, joy, excitement]) while X_i^+ represents one of the masked tokens (e.g., [happiness]). Besides, we use a one-hot vector $\mathbb{1}_{X_i^+}$ to indicate the target tokens (the X_i^+-th position of $\mathbb{1}_{X_i^+}$ is set to 1). The corresponding objective is:

$$\mathcal{L}_{CE}(f, W) = -\mathbb{E}_{(X_i, X_i^+)} \log \frac{\exp((Wf(X_i))^\top \mathbb{1}_{X_i^+})}{\sum_{X_i^-} \exp((Wf(X_i))^\top \mathbb{1}_{X_i^-})},$$

We note that the main objective of generative SSL loss is to maximize the probability of target words and minimize the independent word probability. Following the same spirits, we consider the following spectral loss (HaoChen et al., 2021) to simplify our theoretical analysis:

$$\begin{aligned} \mathcal{L}_{SL}(f, W) = & -\mathbb{E}_{(X_i, X_i^+)}(Wf(X_i))^\top \mathbb{1}_{X_i^+} \\ & + \mathbb{E}_{X_i, X_i^-}((Wf(X_i))^\top \mathbb{1}_{X_i^-})^2. \end{aligned} \quad (1)$$

3.2. Revisiting Objectives from a Matrix Perspective

When theoretically comparing the generalization performance of autoregressive and masked SSL pretraining, it's challenging to analyze the overall optimal solutions for their objectives as both of them are usually seen as instance-level tasks, which hinders us from demonstrating the downstream performance of pretrained models. Inspired by the previous work (HaoChen et al., 2021) which addressed analogous challenges in contrastive learning, we consider reformulating the spectral loss and analyzing it from a matrix decomposition perspective.

We start by introducing the co-occurrence matrix A, which formulates the joint distribution $P_M(X_i, X_i^+)$ of the conditional texts and target texts as a matrix, i.e.,

$$A_{X_i, X_i^+} = P_M(X_i, X_i^+) \geq 0.$$

And we denote $P_C(X_i) = \sum_{X_i^+} P_M(X_i, X_i^+)$, $P_G(X_i^+) = \sum_{X_i} P_M(X_i, X_i^+)$ as the marginal distributions of conditional and target texts. Based on these definitions, we establish a crucial connection: the spectral loss is equivalent to an asymmetric matrix decomposition objective on the co-occurrence matrix.

Theorem 3.1. *Let $\bar{A}$ be the normalized co-occurrence matrix, i.e., $\bar{A}_{X_i,X_i^+} = \frac{A_{X_i,X_i^+}}{\sqrt{P_C(X_i)P_G(X_i^+)}}$. Then we obtain*

$$\mathcal{L}_{SL}(f, W) = \|\bar{A} - FW'^{\top}\|^2 + const.$$

where the X_i-th row of F and the X_i^+-th row of W' respectively represents encoded features and token embeddings, i.e., $F_{X_i} = \sqrt{P_C(X_i)}f(X_i)^{\top}$, $W'_{X_i^+} = \sqrt{P_G(X_i^+)}W_{X_i^+}$.

3.3. Downstream Tasks

Classification tasks and content generation tasks are the most commonly adopted evaluation protocols for pretrained models (Radford et al., 2018; Brown et al., 2020; Dong et al., 2019; Devlin et al., 2019; Yang et al., 2019; Chen et al., 2020a; He et al., 2022). Therefore, in this paper, we mainly consider the classification tasks and content generation tasks for generality.

Linear classification. For the classification tasks, we assume that the labels of $\mathcal{D}$ can be accessed, i.e., $\mathcal{D} = \{(x_i, y(x_i))\}_{i=1}^N$, where $y(x_i)$ is the ground-truth label of x_i. For each sample x_i, we first encode it using the pretrained model f and then apply a linear classifier g following that to generate the predicted distribution of different labels. For ease of theoretical analysis, we follow the settings of linear evaluation (Saunshi et al., 2019), i.e., the pretrained encoder f is frozen during the downstream evaluation process. Then we evaluate the prediction error

$$\mathcal{E}(f) = \mathbb{E}_{(x_i, y(x_i))}P(\hat{y}_i(x_i) \neq y(x_i)),$$

where $\hat{y}_i(x_i) = \arg\max(g(f(x_i)))$.

Content Generation. For the content generation tasks, we consider an unsupervised text dataset $\mathcal{D}_u = \{x_i\}_{i=1}^M$ that consists of M samples, we still assume that each x consists of s tokens, i.e., $x_i = (x_{i,1}, \cdots, x_{i,s})$. The downstream objective is to evaluate the following likelihood with the pretrained model weights:

$$\mathcal{L}_{gen}(\Theta) = -\mathbb{E}_{x_i} \sum_k \log P(x_{i,k+1}|x_{i,1}, \cdots, x_{i,k}; \Theta),$$
(2)

The exponential of this downstream objective is also called as perplexity, which is an important metric to evaluate the language model.

4. A Theoretical Comparison between Autoregressive and Masked SSL

4.1. Generalization on Linear Classification

In this section, we compare the generalization performance between autoregressive and masked SSL on linear classification tasks. To achieve this, we first establish a unified downstream classification guarantee of generative SSL and then compare autoregressive and masked SSL based on that.

4.1.1. DOWNSTREAM CLASSIFICATION GUARANTEES OF GENERATIVE SSL

Theorem 3.1 establishes the mathematical equivalence between the generative SSL objectives and the asymmetric matrix decomposition objectives. Leveraging the Eckart-Young Theorem (Eckart & Young, 1936), we can explicitly characterize the optimal solutions of the asymmetric decomposition objectives, which allows us to characterize the ideal features learned with generative SSL objectives. Substituting these pretrained features into the downstream classification objective yields the following theorem, which establishes the downstream guarantees for generative SSL.

Theorem 4.1. *We define the labeling error α as the probability that the conditional texts and the target text have different labels, i.e.,*

$$\alpha = \mathbb{E}_{(X_i, X_i^+)}\mathbb{1}[y(X_i) \neq y(X_i^+)],$$

where $y(\cdot)$ denotes the ground-truth label. Let $f^{\star}$ be the optimal solutions of $\mathcal{L}_{SL}(f, W)$, then we obtain

$$\mathbb{E}_{x_i}(\hat{y}(x_i) \neq y(x_i)) \leq c_1 \sum_{j=t+1}^{N_D} \sigma_j^4 + c_2\alpha$$

where σ_j^2 is the j-th largest singular value of the normalized co-occurrence matrix and c_1, c_2 are the constants.

As shown in Theorem 4.1, the downstream performance is mainly decided by two critical factors: the labeling error and the singular values of the co-occurrence matrix. In this paper, we follow the assumptions used in (Saunshi et al., 2019; Wang et al., 2022b) that the labeling error between different tokens of the same samples is negligible. Consequently, our primary focus in this paper is on elucidating how the diverse objectives of autoregressive and masked SSL impact the singular values of the co-occurrence matrix.

A common way to understand the singular values is from a graph perspective. More precisely, the co-occurrence matrix A can be viewed as the adjacent matrix of a bipartite graph $\mathcal{G}$, where the nodes represent the conditional and target texts, and the edge weights denote the joint probability between them. The spectral graph theory (Chung, 1997) states that the smaller singular values of the adjacent matrices
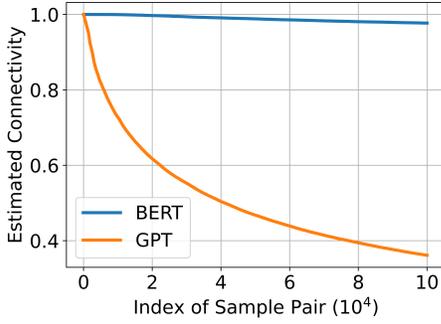
Figure 2. Comparisons on estimated connectivity of the co-occurrence matrices of GPT and BERT. Details in Appendix B.2

correspond to stronger connectivity in the graphs, indicating properties such as fewer disconnected sub-graphs and shorter diameters. Consequently, Theorem 4.1 suggests that achieving superior downstream classification performance requires enhanced connectivity in the co-occurrence matrix.

This perspective provides insight into the differences between masked and autoregressive SSL models in downstream classification tasks. Intuitively, the multiple positions of target tokens in the masked SSL objectives have the potential to generate more connections than the single position in the autoregressive SSL objectives. In the following, we conduct both empirical and theoretical comparisons of the connectivity in autoregressive and masked co-occurrence matrices.

4.1.2. COMPARING AUTOREGRESSIVE AND MASKED SSL

Based on the perspective above, we first empirically investigate the connectivity of autoregressive and masked co-occurrence matrices in the real-world dataset PILE(Gao et al., 2020). As the co-occurrence matrices of the real-world datasets are not accessed, we calculate the average feature similarity between different texts as a surrogate metric for the sample connectivity. To ensure a fair comparison, we pretrain the same scale network with autoregressive and masked SSL objectives respectively. More details can be in Appendix. As shown in Figure 2, the estimated connectivity of the masked SSL co-occurrence matrices is significantly stronger than that of autoregressive SSL.

Besides, to theoretically compare the connectivity of autoregressive and masked SSL, we construct a toy model and calculate the singular values of the respective co-occurrence matrices.

Toy model. For the dataset $\mathcal{D}_{sim} = \{(x_i, y(x_i))\}$, we denote that there exist r classes in the dataset, i.e., $y(x_i) \in \{1, \cdots, r\}$. For the k-th token in the text x_i, we assume that $x_{i,k}$ are uniformly selected from the set $\{k \cdot T \cdot r + y_i, k \cdot T \cdot r + y_i \cdot 2, \cdots, k \cdot T \cdot r + y_i \cdot T\}$ with T elements. Then we can explicitly calculate the singular values of the

co-occurrence matrix.

Theorem 4.2. *For the normalized co-occurrence matrix $\bar{A}, \bar{A}'$ of autoregressive and masked SSL on the toy model dataset $\mathcal{D}_{sim}$, when the length of masked texts $s\rho_m > 1$, we obtain*

$$\begin{cases} \sigma_j = \sigma'_j = 1, & j \leq r, \\ \sigma_j = 1 > \sqrt{\dfrac{s(1-\rho_m)}{(s\rho_m)(s-1)}} = \sigma'_j, & r < j \leq rs, \\ \sigma_j = \sigma'_j = 0, & j > rs, \end{cases}$$

where σ_j, σ'_j are the j-th largest singular values of $\bar{A}, \bar{A}', r$ is the number of class, s is the length of a sample and ρ_m is the mask ratio.

According to Theorem 4.2, the singular values of the masked SSL co-occurrence matrix are smaller than that of the autoregressive SSL, which verifies the intuition that the masked SSL can foster more connections by the multiple-lace predictions. Combined with Theorem 4.1, the masked SSL objective has a superior downstream guarantee. Additionally, it is noteworthy that the singular values of the masked SSL co-occurrence matrix decrease with a larger mask ratio. This observation implies that an aggressive mask ratio can effectively cluster more samples in the feature space, which is consistent with previous empirical findings in generative SSL (He et al., 2022; Devlin et al., 2019).

In summary, both the empirical and theoretical analysis verify that multiple-place word predictions can bring more connections compared to the next word prediction, which suggests that the autoregressive SSL objective should introduce more diverse predictions to strengthen the connectivity between different texts. Naturally, we propose the following diversity-enhanced autoregressive SSL objective:

$$\mathcal{L}_{dar,t}(\Theta) = -\mathbb{E}_{x_i} \sum_k \log P(x_{i,[k+1,k+t]}|x_{i,1}, \cdots, x_{i,k}; \Theta),$$

$$(3)$$

where $x_{i,[k+1,k+t]}$ is a token randomly selected from $\{x_{i,k+1}, \cdots, x_{i,k+t}\}$, $t \geq 1$. This objective lets the conditional sequence $x_{i,1}, \cdots, x_{i,k+1}$ additionally predict $t-1$ more subsequent words, which introduces more diverse prediction targets and helps bring more connections. We will verify the effectiveness of this objective in Section 5.

4.2. Generalization on Content Generation

Besides the classification tasks we analyzed above, another important evaluation task for pretrained models is the content generation ability (Radford et al., 2018; Devlin et al., 2019). Previous works have empirically shown that the autoregressive SSL models outperform masked SSL models in downstream content generation tasks (Dong et al., 2019). To further understand the advantages of autoregressive SSL
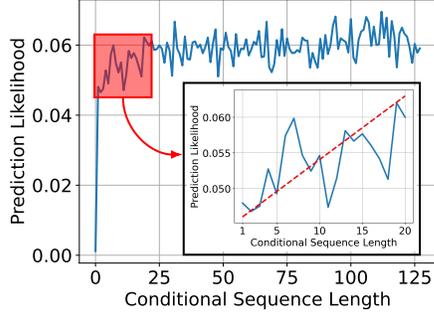
Figure 3. Comparison between different conditional sequence lengths in generation evaluation of masked SSL. Shorter conditional sequences suffer from low prediction likelihood.

models in content generation tasks, we first observe the generation ability of masked SSL models with different types of texts.

As shown in Figure 3, we observe that the masked SSL model performs particularly unsatisfactory when generating from short texts. Intuitively, it can be attributed to a misalignment between the pretraining objective and the downstream evaluation tasks. To be specific, when predicting the masked words in the pretraining process, the lengths of inputs are fixed (e.g., 15% of the texts). Consequently, the masked SSL model may struggle to accurately infer the complete texts due to limited information in the downstream generation tasks. In the subsequent analysis, we aim to theoretically substantiate this intuition. For the ease of our theoretical analysis, we consider using the linear attention as the pretrained encoder to compare the generation abilities of autoregressive and masked SSL models.

Linear Attention. The general form of linear attention is given by:

$$f(x_i) = x_i W^Q (x_i W^K)^T x_i W^V = QK^T V,$$

where W^Q, W^K, W^V are projections.

By comparing the evaluation error on the downstream generation tasks, we obtain the following theorem.

Theorem 4.3. *Let f_{mask} be the model pretrained by masked SSL. We establish the upper bound of masked SSL in content generation tasks:*

$$\mathcal{L}_{gen}(f_{mask}) \leq \frac{\sum_k \left(\frac{w_k^2}{(k-1)^6} + w_k \|W\|_2^2 \eta \right)}{2s(1 - \rho_m)} + \delta_{mask} + 1,$$

where s is the length of a sequence, ρ_m is the mask ratio, $w_k = ((s(1 - \rho_m))^3 - (k-1)^3)$ is the misalignment degree of input lengths, $\eta = \max \|x_{i,a} W^Q (x_{i,b} W^K)^\top x_{i,c} W^V - x_{i,\alpha} W^Q (x_{i,\beta} W^K)^\top x_{i,\gamma} W^V\|$ represents the difference between pretrained model outputs of different positions, and $\delta_{mask} = \max(-(W f(X_i))^\top \mathbb{1}_{X_i^+} + ((W f(X_i))^\top \mathbb{1}_{X_i^-})^2)$ represents the pretraining error of masked SSL.

For the autoregressive SSL, we obtain

$$\mathcal{L}_{gen}(f_{ar}) \leq \delta_{ar}.$$

Comparing two upper bounds, the gap between masked and autoregressive SSL is:

$$\frac{\sum_k \left(\frac{w_k^2}{(k-1)^6} + w_k \|W\|_2^2 \eta \right)}{2s(1 - \rho_m)} + 1 + \delta_{mask} - \delta_{ar}. \quad (4)$$

Consequently, autoregressive SSL obtains a smaller error when the pretraining errors are negligible.

As stated in Theorem 4.3, besides the canonical pretraining error, the autoregressive SSL models have superior performance guarantees in downstream content generation tasks compared to masked SSL models, and the performance gap is mainly decided by two crucial factors: the alignment of input lengths (w_k) and the consistency of different positions (η). In the following, we respectively discuss these two factors.

Alignment of input lengths. We note that when the lengths of test samples are close to the length of unmasked texts in the pretraining objective, the performance of masked SSL models can approach autoregressive SSL models, which proves the length misalignment between unmasked texts in pretraining and test examples in downstream is a crucial reason for the inferior performance of masked SSL models.

Consistency of different positions. The term η in Equation (4) evaluates the difference in model output across various positions in a sequence. When the output distribution is more consistent, the performance of the masked SSL model is better. Consequently, it is advantageous to encourage the model to generate predictions based on the entire sentence rather than focusing on specific tokens. Besides, the consistency of different positions offers some other potential benefits. For example, by uniformly considering the semantic information across different positions in a sentence, the model can avoid generating shortcuts and overfit solutions.

In summary, the theoretical bounds in Theorem 4.3 provide two principled guidelines to improve the content generation performance of masked SSL: (1) we should try to mitigate the length misalignment between pretraining and downstream examples, and (2) the model predictions should be encouraged to be consistent with different tokens in the same sentences.

Inspired by the theoretical analysis, predicting the masked texts with different lengths of unmasked texts is a straightforward solution to mitigate the misalignment and improve the generation performance of masked SSL. Consequently, we propose the following variable-length masked objective:

$$\mathcal{L}_{vm}(\Theta)$$
$$= -\mathbb{E}_{\rho_m} \mathbb{E}_{(x_i^1, x_i^2)} \sum_k \log P(x_{i,k}^2 | x_{i,1}^1, \cdots, x_{i,s(1-\rho_m)}^1; \Theta).$$
(5)

In this objective, the mask ratio ρ_m is not fixed. Instead, it is randomly sampled from a range. The variable mask ratio has the potential to mitigate the length misalignment between pretraining and downstream examples, improving the content generation performance. We will verify its effectiveness in Section 5.

4.3. Discussion

In summary, this section introduces the first theoretical comparison between autoregressive and masked SSL. As autoregressive and masked SSL exhibit contrasting advantages in downstream classification and content generation tasks (Radford et al., 2018; Devlin et al., 2019), the two most crucial evaluation tasks for pretrained models seem to be contradictory. However, by revealing the limitations of autoregressive SSL in classification tasks (fixed position of target tokens) and limitations of masked SSL in content generation tasks (fixed length of unmasked texts), we note that the limitations of these two paradigms are not contradictory to each other. Instead, they deliver a consistent message: we should encourage diversity (in inputs, prediction objectives, etc) to the generative SSL objectives to improve the generalization performance. As a verification of our analysis, we will show the performance of our proposed objectives across both language and vision tasks in the following sections.

5. Experiments

In this section, we will verify the effectiveness of our new proposed objectives presented in Equation (3) and Equation (5). We will conduct experiments on both vision and language tasks to demonstrate the generality of our methods.

5.1. Diversity-enhanced Autoregressive Objective Improves Classification Ability

In this part, we will consider the diversity-enhanced autoregressive objective as proposed in Equation (3). The diversity-enhanced autoregressive objective is to predict the next t tokens of conditional texts. Inspired by (Wang et al., 2018b), which predicts several tokens simultaneously, we perform group-autoregressive modeling on the sequence to effectively realize this objective. Specifically, the sample sequence $(x_{i,1}, \ldots, x_{i,s})$ is divided into several consecutive groups $G_{i,1}, \ldots, G_{i,l}$ in order, with each group containing any number of tokens. The prediction is performed group by group, i.e., within each group, the tokens are predicted in parallel, while across group, the predictions are sequential. In this way, given $G_{i,1}, \ldots, G_{i,s}$, the model is

to predict all tokens from $G_{i,s+1}$. Therefore, by setting $|G_{i,2}| = \cdots = |G_{i,l-1}| = t$ and summing $|G_{i,1}|$ from 1 to t, the semi-autoregressive modeling task

$$\mathcal{L}_{semi,t}(\Theta)$$
$$= -\mathbb{E}_{x_i} \sum_{|G_1|} \sum_s \log P(G_{i,s+1} | G_{i,1}, \ldots, G_{i,s}; \Theta)$$
(6)

is equivalent to Equation (3). In our experiments, we set $t = 2$ and uniformly sampling $|G_{i,1}|$ from $\{1, 2, \ldots, t\}$ instead. Since there exist multiple prediction targets for one condition sequence, it is difficult for vanilla Transformer (Vaswani et al., 2017) to model. Therefore, for language tasks, we will leverage and extend the two-stream attention module proposed in XLNet (Yang et al., 2019) by designing causal masks for semi-autoregressive modeling, where the two-stream attention can be used to model more various dependency relationship between tokens compared to vanilla Transformer. More details on the realization will be illustrated in Appendix B.3. Similarly, for vision tasks, we use the two-stream attention version of ViT (Dosovitskiy et al., 2021), which is also adopted in the previous work (Hua et al., 2022).

Language Tasks. The model is pretrained on the Pile dataset (Gao et al., 2020), which contains content from 22 diverse sources. We use a decoder-only Transformer with 16 layers and hidden size of 768. We train the model for 100K steps with batch size of 8192. The other pretraining procedure follows the protocol proposed by Cramming (Geiping & Goldstein, 2023). The model is then finetuned with 5 epochs and evaluated on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018a), which is a collection of several language understanding tasks.

Table 2 presents the GLUE test results. These results reveal that the model trained with the diversity-enhanced autoregressive objective consistently exhibits improvement in classification tasks. Our proposed objective has a gain of 1.9% on the average score and 3.3% on the hardest task CoLA. These improvements support the effectiveness of our theoretical findings regarding the classification ability of autoregressive models.

Vision Tasks. The model is pretrained on ImageNet-1K with ViT-Base and ImageNet-100 with ViT-Small (Deng et al., 2009). The training epoch is 200 with a warm-up of 10 epochs. The batch size is set to 4096 following previous works (He et al., 2022; Hua et al., 2022; Xie et al., 2022). We set the base learning rate as 5e-4 and use Adam optimizer (Kingma & Ba, 2015). After the pretraining is finished, we perform downstream tasks under two protocols: For linear probing, we train a linear classifier on the frozen pretrained encoder for 90 epochs; For non-linear fine-tuning, we train both the pretrained encoder and the linear classifier with the

Table 2. GLUE test set results of autoregressive objective and diversity-enhanced autoregressive objective scored with 5 epochs fine-tuning on each test set. We report F1 scores for QQP and MRPC, Pearson correlation coefficient for STS-B, Matthews correlation coefficient for CoLA and accuracy scores for the other tasks.

Objective	MNLI	SST-2	STSB	RTE	QNLI	QQP	MRPC	CoLA	**Avg**
Autoregressive	78.2/79.3	**90.3**	82.1	54.3	84.7	85.6	81.9	30.9	74.1
Diversity-enhanced Autoregressive	**80.5/80.6**	89.9	**83.7**	**56.4**	**87.2**	**86.3**	**85.1**	**34.2**	**76.0**

Table 3. Experiment results of autoregressive objective and diversity-enhanced autoregressive objective on image classification tasks. ViT-S(mall) is trained on ImageNet-100 and ViT-B(ase) is trained on ImageNet-1K. LP ACC. refers to linear probing accuracy (%). FT Acc. refers to fine-tuning accuracy (%).

Arch.	Objective	LP Acc.	FT Acc.
ViT-S	Autoregressive	33.1	81.1
	Diversity-enhanced	**36.2**	**83.2**
ViT-B	Autoregressive	56.2	82.5
	Diversity-enhanced	**59.4**	**82.9**

soft target cross entropy loss (Peterson et al., 2019) for 100 epochs. The batch size is set to 4096.

The experiment results are presented in Table 3. The diversity-enhanced autoregressive objective improves linear probing accuracy by ∼3%. It also improves fine-tuning accuracy on ViT-S by 2.1% and on ViT-B by 0.4%. These positive outcomes affirm the effectiveness of our proposed objective, emphasizing our theoretical insight that it is beneficial to have more connections in generative SSL.

We also conduct various extensive experiments on vision tasks including few-shot learning as Dubois et al. (2023), multi-epoch training as Li et al. (2023) and transfer learning as Kong et al. (2023).

Few-shot learning. We conduct experiments with the diversity-enhanced autoregressive objective. In the stage of the downstream tasks, we only use 1% of the label. Finetuning and linear probing results are shown in Table 4. As shown in the above table, the diversity-enhanced autoregressive objective improves both the few-shot linear and finetuning accuracy, which further verifies the effectiveness of our proposed objectives across different tasks.

Table 4. Few-shot evaluation accuracy (%) on ImageNet-1K with ViT Base. Only 10% of the training data is given during training.

Objective	LP Acc.	FT Acc.
Autoregressive	41.2	70.2
Diversity-enchanced	**45.3**	**71.1**

Multi-epoch training. We conduct experiments with the

diversity-enhanced autoregressive objective. We respectively pretrain the models with 100 epochs and 200 epochs. Finetuning and linear probing results are shown in Table 5. As shown in the above table, the diversity-enhanced autoregressive models with different training epochs consistently achieve significant improvement in classification tasks, which also verifies the effectiveness of our proposed objectives.

Table 5. Test accuracy (%) with multi-epoch training on ImageNet-1K with ViT Base.

Epochs	Objective	LP Acc.	FT Acc.
100	Autoregressive	48.2	81.4
	Diversity-enchanced	**52.6**	**82.1**
200	Autoregressive	56.2	82.5
	Diversity-enchanced	**59.4**	**82.9**

Transfer learning. We conduct experiments with the diversity-enhanced autoregressive objective. We evaluate the transfer learning performance pretrained on ImageNet-1K on 9 downstream datasets as Zhao et al. (2022) and Ericsson et al. (2021), which are FGVC Aircraft, Caltech-101, Stanford Cars, CIFAR10, CIFAR100, DTD, Oxford 102 Flowers, Food-101 and Oxford-IIIT Pets. For linear evaluation, multinomial logistic regression is fit on the extracted features. Results are shown in Table 6. As shown in the table, the diversity-enhanced autoregressive models also show superior performance in transfer learning, which further verifies the generalization performance of our models.

Table 6. Transfer learning average accuracy (%) on 9 downstream datasets with ViT Base.

Objective	Average accuracy
Autoregressive	76.2
Diversity-enchanced	**78.9**

The experiments across three tasks and multiple datasets show that the modified objectives consistently improve the performance of downstream tasks. This proves the superiority of our objective and verifies the effectiveness of our proposed theory.

Table 7. Perplexity of models trained with masked objective and variable-length masked objective on the Pile dataset and C4 dataset. Smaller perplexity indicates better generation ability. $[0.15, 0.3]$ means uniformly sampling from $[0.15, 0.3]$ during training.

Objective (Mask ratio)	Pile ($\downarrow$)	C4 ($\downarrow$)
Masked (0.15)	59.6	71.2
Masked (0.3)	50.2	63.8
Variable-length ($[0.15, 0.3]$)	**45.1**	**59.4**

5.2. Variable-length Masked Objective Improves Generation Ability

In this part, we will consider the variable-length MIM objective as proposed in Equation (5). Recall that the variable-length MIM objective leverages variable mask ratio to mitigate the length misalignment problem in masked models. We will consider different sampling strategies for the mask ratio in language tasks and vision tasks.

Language Tasks. Similar to the experiments on diversity-enhanced autoregressive Objective, the model is pretrained on the Pile dataset (Gao et al., 2020). We use a decoder-only Transformer (without causal mask) with 16 layers and hidden size of 768. The pretraining procedure follows the protocol proposed by Cramming (Geiping & Goldstein, 2023). Regarding the mask ratio in the masked model, we explore three options: (1) retaining the original BERT value of 0.15, (2) opting for a larger ratio of 0.3, and (3) uniformly sampling from the range $[0.15, 0.3]$, aligning with our proposed variable-length objective. We adopt perplexity as the evaluation metric, which is the exponential of Equation (2). We will evaluate the model on two datasets: (1) the tail of the Pile dataset, which is never exhibited to the model during the pretraining process; (2) a subset of C4 dataset (Raffel et al., 2020), which is a colossal, cleaned version of Common Crawl's web crawl corpus.

Table 7 demonstrates that the model trained with the variable-length objective exhibits significantly lower perplexity compared to the masked models trained with a fixed mask ratio on both datasets. This suggests that the variable-length objective can improve the generation ability of the model. These enhancements support our theoretical insights into the generation ability of masked models.

Vision Tasks. We select MAE (He et al., 2022) as the baseline model, with pretraining hyper-parameters consistent with those in the diversity-enhanced experiments. After the pretraining is finished, we perform image reconstruction tasks. The image is first randomly masked out with 25% portion. Then the model predicts the masked part token by token in an autoregressive way. Reconstruction quality is evaluated using two metrics between the reconstructed

Table 8. Reconstruction ability of models trained with masked objective and variable-length masked objective on ImageNet dataset. There are two evaluation metrics: L2 loss and Perceptual loss. $[0.5, 0.75]$ means uniformly sampling from $[0.5, 0.75]$ during training.

Objective (Mask ratio)	L2 ($\downarrow$)	Perceptual ($\downarrow$)
Masked (0.75)	0.127	0.201
Masked (0.5)	0.115	0.192
Variable-length ($[0.5, 0.75]$)	**0.072**	**0.136**

and original images: L2 loss at the pixel level and LPIPS loss (Zhang et al., 2018) at the high-level feature level. The average loss is calculated over the ImageNet validation set.

Table 8 reveals that the model trained with the variable-length objective experiences significantly lower (approximately 30%) reconstruction loss compared to the masked models trained with a fixed mask ratio. These enhancements underscore our theoretical insight that mitigating misalignment of unmasked lengths between pretraining and test examples is crucial for generation ability.

6. Conclusion

In this paper, we propose the first theoretical comparison between two primary generative self-supervised (SSL) paradigms: autoregressive SSL and masked SSL. In particular, we establish the theoretical guarantees for autoregressive SSL and masked SSL in the most common evaluation tasks, i.e., classification and content generation. Through empirical and theoretical analyses, we delineate the advantages of masked SSL in classification tasks and the advantages of autoregressive SSL in content generation tasks. Building upon the insights from our theoretical analysis, we formulate principled guidelines for generative SSL and introduce two improved generative SSL objectives. Empirically, we verify that our proposed objectives significantly improve the performance of generative SSL models accoss both visual and language datasets.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *ICML*, 2020a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, 2020b.

Chung, F. R. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Dubois, Y., Hashimoto, T., and Liang, P. Evaluating self-supervised learning via risk decomposition. In *ICML*, 2023.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1936.

Ericsson, L., Gouk, H., and Hospedales, T. M. How well do self-supervised models transfer? In *CVPR*, 2021.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, 2020.

Geiping, J. and Goldstein, T. Cramming: Training a language model on a single gpu in one day. In *ICML*, 2023.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

Giorgi, J., Nitski, O., Wang, B., and Bader, G. Declutr: Deep contrastive learning for unsupervised textual representations. In *ACL*, 2021.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Hua, T., Tian, Y., Ren, S., Raptis, M., Zhao, H., and Sigal, L. Self-supervision through random segments with autoregressive coding (randsac). In *ICLR*, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kong, L., Ma, M. Q., Chen, G., Xing, E. P., Chi, Y., Morency, L.-P., and Zhang, K. Understanding masked autoencoders via hierarchical latent variable models. In *CVPR*, 2023.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*, 2020.

Larsson, G., Maire, M., and Shakhnarovich, G. Learning representations for automatic colorization. In *ECCV*, 2016.

Li, S., Wu, D., Wu, F., Zang, Z., and Li, S. Z. Architecture-agnostic masked image modeling—from vit back to cnn. In *ICML*, 2023.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, 2019.

Merity, S. The wikitext long term dependency language modeling dataset. *Salesforce Metamind*, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv*, 2018.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In *ICCV*, 2019.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning. In *NeurIPS*, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP*, 2018a.

Wang, C., Zhang, J., and Chen, H. Semi-autoregressive neural machine translation. In *EMNLP*, 2018b.

Wang, C., Li, M., and Smola, A. J. Language models with transformers. *ArXiv*, 2019.

Wang, T., Roberts, A., Hesslow, D., Le Scao, T., Chung, H. W., Beltagy, I., Launay, J., and Raffel, C. What language model architecture and pretraining objective works best for zero-shot generalization? In *ICML*, 2022a.

Wang, Y., Zhang, Q., Wang, Y., Yang, J., and Lin, Z. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *ICLR*, 2022b.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.

Zhang, Q., Wang, Y., and Wang, Y. On the generalization of multi-modal contrastive learning. In *ICML*, 2023a.

Zhang, Q., Wang, Y., and Wang, Y. Identifiable contrastive learning with automatic feature importance discovery. In *NeurIPS*, 2023b.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Zhao, X., Du, T., Wang, Y., Yao, J., and Huang, W. Arcl: Enhancing contrastive learning with augmentation-robust representations. In *ICLR*, 2022.

A. Proofs

A.1. Proof of Theorem 3.1

Proof. Expanding the decomposition object and we obtain,

$$
\begin{aligned}
\|\bar{A} - FW'^{\top}\|^2 &= \sum_{X_i, X_i^+} \left(\bar{A}_{X_i, X_i^+} - F_{X_i}(W'_{X_i^+})^{\top} \right)^2 \\
&= \sum_{X_i, X_i^+} \left(\frac{P_M(X_i, X_i^+)}{\sqrt{P_C(X_i)P_G(X_i^+)}} - \sqrt{P_C(X_i)}f(X_i)^{\top}\sqrt{P_G(X_i^+)}(W_{X_i^+})^{\top} \right)^2 \\
&= \sum_{X_i, X_i^+} \left(\frac{P_M(X_i, X_i^+)^2}{P_C(X_i)P_G(X_i^+)} + P_C(X_i)P_G(X_i^+)\left(f(X_i)^{\top}(W_{X_i^+})^{\top} \right)^2 - 2P_M(X_i, X_i^+)f(X_i)^{\top}(W_{X_i^+})^{\top} \right) \\
&= \sum_{X_i, X_i^+} \left(\frac{P_M(X_i, X_i^+)^2}{P_C(X_i)P_G(X_i^+)} \right) - 2\mathbb{E}_{(X_i, X_i^+)}f(X_i)^{\top}(W_{X_i^+})^{\top} + \mathbb{E}_{X_i, X_i^-}\left(f(X_i)^{\top}(W_{X_i^-})^{\top} \right)^2 \\
&= \sum_{X_i, X_i^+} \left(\frac{P_M(X_i, X_i^+)^2}{P_C(X_i)P_G(X_i^+)} \right) - \mathbb{E}_{(X_i, X_i^+)}(Wf(X_i))^{\top}\mathbb{1}_{X_i^+} + \mathbb{E}_{X_i, X_i^-}((Wf(X_i))^{\top}\mathbb{1}_{X_i^-})^2 \\
&= \mathcal{L}_{SL}(f, W) + \text{const.}
\end{aligned}
$$

$\square$

A.2. Proofs of Theorem 4.1

We first introduce a lemma that will be used in the following proofs.

Lemma A.1 (Lemma 3.1 in (HaoChen et al., 2021))**.** *For two learned embedding matrices F, $\widetilde{F}$, a diagonal matrix D and an invertible matrix R, if $F = D\widetilde{F}R$, they have the equal linear probing error, i.e.,*

$$
\mathcal{E}(F) = \mathcal{E}(\widetilde{F}).
$$

Proof. The proofs mainly follow the theoretical framework proposed in (Zhang et al., 2023a). According to Eckart-Young Theorem (Eckart & Young, 1936), the optimal solution $F^{\star}, (W')^{\star}$ of the decomposition objective $\mathcal{L}_{MF}(F, W') = \|\bar{A} - FW'^{\top}\|^2$ satisfy:

$$
F^{\star}(W'^{\star})^{\top} = U^t \operatorname{diag}(\sigma_1, ..., \sigma_t)(V^t)^{\top},
$$

where we denote $\bar{A} = U\Sigma V^{\top}$ as the singular value decomposition of $\bar{A}$, $(\sigma_1, ..., \sigma_t)$ are the t-largest singular values of $\bar{A}$, the t-th column of $U^t \in \mathbb{R}^{N \times t}$ contains the corresponding eigenvectors of the t-th largest singular values and $V^t \in \mathbb{R}^{N_D \times t}$ is a unitary matrix. Then we respectively obtain the optimal solutions $F^{\star}$ and $(W')^{\star}$:

$$
\begin{aligned}
F^{\star} &= U^t DR, \\
(W')^{\star} &= V^t \operatorname{diag}(\sigma_1, ..., \sigma_t)D^{-1}R,
\end{aligned}
$$

where $R \in \mathbb{R}^{t \times t}$ is a unitary matrix and D is an invertible diagonal matrix. Then we define a symmetric spectral loss:

$$
\mathcal{L}_{SCL}(f) = -\mathbb{E}_{(X_i, X_i')}f(X_i)^{\top}f(X_i') + \mathbb{E}_{X_i, (X_i')^-}(f(X_i)^{\top}f((X_i')^-))^2, \tag{7}
$$

where $(X_i, X_i') \sim P_T, P_T(X_i, X_i') = \mathbb{E}_{X_i^+}P_M(X_i|X_i^+)P_M(X_i'|X_i^+)$, and $X_i, X_i' \sim P_C(X_i)$. Following the proof of theorem 3.1, the symmetric spectral loss is also equivalent to a matrix decomposition loss, i.e., $\mathcal{L}_{SCL}(f) = \|\tilde{P}_T - FF^{\top}\|^2 +$

$const$, where $(\tilde{P}_T)_{(X_i, X_i')} = \frac{P_T(X_i, X_i')}{\sqrt{\mathcal{P}_C(X_i)\mathcal{P}_C(X_i')}}$ and $(F)_{X_i} = \frac{f(X_i)^\top}{\sqrt{P_C(X_i)}}$. Then we consider the objective $\|\tilde{P}_T - FF^\top\|^2$. Similar to the asymmetric decomposition objective, the optimal solution can be represented as:

$$(F^\star)' = U_T^t D_T R_T,$$

where $U_T^t \in \mathbb{R}^{N \times t}$ contains t corresponding eigenvectors of t largest singular values of $\tilde{P}_T$, $D_T \in \mathbb{R}^{t \times t}$ is an invertible diagonal matrix and $R_T \in \mathbb{R}^{t \times t}$ is a unitary matrix. In the next step, we analyze the relationship between $\bar{A}$ and $\tilde{P}_T$. Considering the (X_i, X_i')-th element of $\bar{A}\bar{A}^\top$, we have

$$
\begin{aligned}
(\bar{A}\bar{A}^\top)_{X_i, X_i'} &= \sum_{x_i^+} (\bar{A})_{X_i, X_i^+} (\bar{A})_{X_i', X_i^+} \\
&= \sum_{X_i^+} \frac{\mathcal{P}_M(X_i, X_i^+)\mathcal{P}_M(X_i', X_i^+)}{\mathcal{P}_G(X_i^+)\sqrt{\mathcal{P}_C(X_i)\mathcal{P}_C(X_i')}} \\
&= \frac{1}{\sqrt{\mathcal{P}_C(X_i)\mathcal{P}_C(X_i')}} \sum_{X_i^+} \mathcal{P}_G(X_i^+)\mathcal{P}_M(X_i|X_i^+)\mathcal{P}_M(X_i'|X_i^+) \quad \left(\mathcal{P}_M(X_i, X_i^+) = \mathcal{P}_M(X_i|X_i^+)\mathcal{P}_G(X_i^+)\right) \\
&= \frac{\mathbb{E}_{X_i^+}\mathcal{P}_M(X_i|X_i^+)\mathcal{P}_M(X_i'|X_i^+)}{\sqrt{\mathcal{P}_C(X_i)\mathcal{P}_C(X_i')}} \\
&= (\tilde{P}_T)_{X_i, X_i'}.
\end{aligned}
$$

We know that $\tilde{P}_T = \bar{A}\bar{A}^\top$, so $\tilde{P}_T$ and $\bar{A}$ share the same eigenvectors, i.e., $U^t = U_T^t$. As D, D_2, R, D_T, R_T are invertible matrices and the product of the invertible matrices is still invertible, we obtain

$$F^\star = (F^\star)'T,$$

where $T = (D_T)^{-1}(R_T)^{-1}DR$ is an invertible matrix. With Lemma A.1, we obtain

$$\mathcal{E}(f^\star) = \mathcal{E}(f_{SCL}^\star).$$

$\square$

Then we introduce another lemma:

Lemma A.2 (Theorem 5.1 in (Zhang et al., 2023b)). *We denote $\tilde{\alpha}$ as the probability that the conditional texts have different labels, i.e., $\tilde{\alpha} = \mathbb{E}_{(X_i, X_i') \sim P_T} \mathbb{1}[y(X_i) \neq y(X_i')]$. Let $f_{SCL}^\star$ be the optimal solutions of the symmetric spectral objective Then, we have the downstream classification guarantee:*

$$\mathcal{E}(f_{SCL}^\star) \leq c_1 \sum_{i=t+1}^{N_D} \tilde{\sigma}_i^2 + c_2 \cdot \tilde{\alpha}, \tag{8}$$

where $\tilde{\sigma}_i$ is the i-th largest eigenvalue of $\tilde{P}_T$, and c_1, c_2 are constants.

And we continue the proofs.

Proof. Combined with Lemma A.2, we obtain

$$\mathcal{E}(f^\star) = \mathcal{E}(f_{SCL}^\star) \leq c_1 \sum_{i=t+1}^{N_D} \tilde{\sigma}_i^2 + c_2 \cdot \tilde{\alpha}.$$

As $\tilde{P}_T = \bar{A}\bar{A}^\top$, we obtain $\tilde{\sigma} = \sigma^2$. And for the labeling error, we have

$$
\begin{aligned}
\tilde{\alpha} &= \sum_{X_i, X_i'} \mathcal{P}_T(X_i, X_i') \mathbb{1}[y(X_i) \neq y(X_i')] \\
&= \sum_{X_i, X_i'} \mathbb{E}_{X_i^+} \left[\mathcal{P}_M(X_i | X_i^+) \mathcal{P}_M(X_i | X_i^+) \mathbb{1}[y(X_i) \neq y(X_i')] \right] \\
&\leq \sum_{X_i, X_i'} \mathbb{E}_{X_i^+} \left[\mathcal{P}_M(X_i | X_i^+) \mathcal{P}_M(X_i | X_i^+) (\mathbb{1}[y(X_i) \neq y(X_i^+)] + \mathbb{1}[y(X_i') \neq y(X_i^+)]) \right] \\
&= 2\mathbb{E}_{X_i^+}[\mathcal{P}_M(X_i | X_i^+) \mathbb{1}[y(X_i) \neq y(X_i^+)]] \\
&= 2\mathbb{E}_{X_i, X_i^+} \mathbb{1}[y(X_i) \neq y(X_i^+)] \\
&= 2\alpha.
\end{aligned}
$$

So we obtain

$$
\mathcal{E}(f^\star) \leq c_1 \sum_{i=t+1}^{N_D} \sigma_i^4 + c_2' \cdot \alpha, \tag{9}
$$

which finishes the proofs of Theorem 4.1. $\qquad\square$

A.3. Proofs of Theorem 4.2

We first introduce a lemma that will be used in the following proofs.

Lemma A.3 (Theorem 4.2 in (Zhang et al., 2023a)). *For a block matrix:*

$$
P = \begin{pmatrix}
p_a & \cdots & p_a & p_b & \cdots & p_b & \cdots & p_b & \cdots & p_b \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
p_a & \cdots & p_a & p_b & \cdots & p_b & \cdots & p_b & \cdots & p_b \\
p_b & \cdots & p_b & p_a & \cdots & p_a & \cdots & p_b & \cdots & p_b \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
p_b & \cdots & p_b & p_a & \cdots & p_a & \cdots & p_b & \cdots & p_b \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
p_b & \cdots & \cdots & \cdots & \cdots & p_b & \cdots & p_a & \cdots & p_a \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
p_b & \cdots & \cdots & \cdots & \cdots & p_b & \cdots & p_a & \cdots & p_a
\end{pmatrix},
$$

each row and each columns has $s_a \cdot s_b$ elements. Among them, s_a elements are p_a and else are p_b. Then the singular values of it are:

$$
\begin{aligned}
\sigma_1 &= s_a \cdot p_a + (s_b - 1) \cdot s_a \cdot p_b, \\
\sigma_2 &= \cdots = \sigma_{s_b} = s_a \cdot |p_b - p_a|, \\
\sigma_{s_a+1} &= \cdots = \sigma_{s_a \cdot s_b} = 0.
\end{aligned}
$$

Proof. For the autoregressive objective, the co-occurrence matrix size is $\left(\frac{T - T^s}{1 - T} r \times STr \right)$.

As the tokens are uniformly selected, we know that $A_{X, X^+} = \frac{1}{rsT^{i+1}}$ (when $i \neq 0$), where i is the length of the condition texts X. Consequently, the elements in the normalized $\bar{A}$ satisfy $\bar{A}_{X, X^+} = \frac{1}{\sqrt{T^{i+1}}}$.

As the samples of different classes are disconnected, we only need to consider the singular values of the sub-matrix of intra-class samples. We note that each sub-matrix has s non-zero diagonal blocks. The i-th block size is $(T^i \times T^i)$ and the

elements in the same block are equal to $\frac{1}{\sqrt{T^{i+1}}}$, i.e., the sub-matrix is:

$$
\begin{bmatrix}
\frac{1}{\sqrt{T^2}} & \cdots & \frac{1}{\sqrt{T^2}} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{1}{\sqrt{T^2}} & \cdots & \frac{1}{\sqrt{T^2}} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & \cdots & 0 & \cdots & \cdots & \cdots & 0 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 0 & \cdots & \cdots & \cdots & 0 & \cdots & 0 \\
0 & \cdots & \cdots & \cdots & \cdots & 0 & \frac{1}{\sqrt{T^{s+1}}} & \cdots & \frac{1}{\sqrt{T^{s+1}}} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & 0 & \frac{1}{\sqrt{T^{s+1}}} & \cdots & \frac{1}{\sqrt{T^{s+1}}}
\end{bmatrix}.
$$

Then we consider the matrix $\bar{A}^\top \bar{A}$, it is still a block matrix. It has s non-zero diagonal blocks and the size of each block is $T \times T$ The elements are $\frac{1}{T}$, i.e., the sub-matrix is

$$
\begin{bmatrix}
\frac{1}{T} & \cdots & \frac{1}{T} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{1}{T} & \cdots & \frac{1}{T} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & \cdots & 0 & \cdots & \cdots & \cdots & 0 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 0 & \cdots & \cdots & \cdots & 0 & \cdots & 0 \\
0 & \cdots & \cdots & \cdots & \cdots & 0 & \frac{1}{T} & \cdots & \frac{1}{T} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & 0 & \frac{1}{T} & \cdots & \frac{1}{T}
\end{bmatrix}.
$$

With Lemma A.3, the eigenvalues of the sub-matrix are

$$
\sigma_1 = 1,
$$
$$
\sigma_2 = \cdots = \sigma_s = T \cdot \frac{1}{T} = 1,
$$
$$
\sigma_s = \cdots = \sigma_{sT \cdot sT} = 0.
$$

Combing different the samples of different classes, the eigenvalues of $\bar{A}^\top \bar{A}$ are

$$
\sigma_1 = \cdots = \sigma_r = 1,
$$
$$
\sigma_{r+1} = \cdots = \sigma_{r \cdot s} = T \cdot \frac{1}{T} = 1,
$$
$$
\sigma_{r \cdot s+1} = \cdots = \sigma_{sT \cdot sT} = 0.
$$

With the definition of singular values, the singular values of $\bar{A}$ are

$$
\begin{cases}
\sigma_j = 1, & j \leq rs, \\
\sigma_j = 0, & \text{else},
\end{cases}
$$

For masked models, the co-occurrence matrix size is $(\mathcal{C}_s^{(1-\rho_m)s} T^{(1-\rho_m)s} r \times STr)$. When the positions of conditional texts and target tokens are overlapped, the joint probability is 0. So each row in the co-occurrence matrix has $\mathcal{C}_{s-1}^{s_m}$ zeros, where $s_m = s(1 - \rho_m)$. And the else elements is $\frac{1}{\mathcal{C}_{s-1}^{s_m}(T)^{s_m} sTr}$. By normalizing the co-occurrence matrix, the non-zero elements become $\bar{A}'_{X,X+} = \frac{\sqrt{ST\mathcal{C}_s^{s_m}(T)^{s_m}}}{\mathcal{C}_{s-1}^{s_m}}$.

15

Similar to autoregressive SSL, we calculate the intra-class sub-matrix of $\bar{A}'^{\top}\bar{A}'$ and obtain:

$$
\begin{pmatrix}
\frac{1}{(s-s_m)T} & \cdots & \frac{1}{(s-s_m)T} & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{1}{(s-s_m)T} & \cdots & \frac{1}{(s-s_m)T} & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} \\
\frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \frac{1}{(s-s_m)T} & \cdots & \frac{1}{(s-s_m)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \frac{1}{(s-s_m)T} & \cdots & \frac{1}{(s-s_m)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \cdots & \cdots & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{1}{(s-s_m)T} & \cdots & \frac{1}{(s-s_m)T} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \cdots & \cdots & \cdots & \frac{(s-s_m-1)}{(s-s_m)(s-1)T} & \cdots & \frac{1}{(s-s_m)T} & \cdots & \frac{1}{(s-s_m)T}
\end{pmatrix},
$$

With Lemma A.3, the eigenvalues of the sub-matrix are

$$\sigma_1 = 1,$$

$$\sigma_2 = \cdots = \sigma_s = T \cdot \left(\frac{1}{(s-s_m)T} - \frac{s-1-s_m}{(s-s_m)(s-1)T} \right) = \frac{s_m}{(s-s_m)(s-1)}$$

$$\sigma_s = \cdots = \sigma_{sT \cdot sT} = 0.$$

Combing different the samples of different classes, the eigenvalues of $\bar{A}'^{\top}\bar{A}'$ are

$$\sigma_1 = \cdots = \sigma_r = 1,$$

$$\sigma_{r+1} = \cdots = \sigma_{r \cdot s} = \frac{s_m}{(s-s_m)(s-1)},$$

$$\sigma_{r \cdot s+1} = \cdots = \sigma_{sT \cdot sT} = 0.$$

When $s_m < (s-1)$, we obtain $\frac{s_m}{(s-s_m)(s-1)} < 1$. With the definition of singular values, the singular values of $\bar{A}'$ are

$$
\begin{cases}
\sigma_j = 1, & j \le r, \\
\sigma_j = \sqrt{\frac{s_m}{(s-s_m)(s-1)}} < 1, & r < j \le rs, \\
\sigma_j = 0, & \text{else.}
\end{cases}
$$

$\square$

A.4. Proofs of Theorem 4.3

Proof. Given a sample x_i, let $X^i_{<k}$ be the first $k-1$ tokens of x_i, $(X^i_k)^+$ be the k-th token of x_i and $(X^i_k)^-$ is any independent token. Expanding $-(Wf_{\text{mask}}(X^i_{<k}))^{\top}\mathbb{1}_{(X^i_k)^+}$ and we have

$$
\begin{aligned}
-(Wf_{mask}(X^i_{<k}))^{\top}\mathbb{1}_{(X^i_k)^+} &= \frac{1}{2}\left\|Wf_{mask}(X^i_{<k}) - \mathbb{1}_{(X^i_k)^+}\right\|^2_2 - \frac{1}{2}\left\|Wf_{mask}(X^i_{<k})\right\|^2_2 - \frac{1}{2}\left\|\mathbb{1}_{(X^i_k)^+}\right\|^2_2 \\
&= \frac{1}{2}\left\|Wf_{mask}(X^i_{<k}) - \mathbb{1}_{(X^i_k)^+}\right\|^2_2 - 1.
\end{aligned}
\tag{10}
$$

Let the mask m_k satisfy that $\{x_{i,1}, \cdots, x_{i,k-1}\}$ are all included in $x_i[m_k]$ and $x_{i,k} \in x_i[1-m_k]$. We denote $X^i_{m_k} = x_i[m_k]$ and have the following equation:

$$
\frac{1}{2}\left\|Wf_{mask}(X^i_{<k}) - \mathbb{1}_{(X^i_k)^+}\right\|^2_2 \le \frac{1}{2}\left\|Wf_{mask}(X^i_{m_k}) - Wf_{mask}(X^i_{<k})\right\|^2_2 + \frac{1}{2}\left\|Wf_{mask}(X^i_{m_k}) - \mathbb{1}_{(X^i_k)^+}\right\|^2_2, \tag{11}
$$

The upper bound of $\frac{1}{2}\left\|Wf_{\text{mask}}(X^i_{m_k}) - \mathbb{1}_{(X^i_k)^+}\right\|^2_2$ is:

$$
\frac{1}{2}\left\|Wf_{mask}(X^i_{m_k}) - \mathbb{1}_{(X^i_k)^+}\right\|^2_2 = 1 - (Wf_{mask}(X^i_{m_k}))^{\top}\mathbb{1}_{(X^i_k)^+} \le 1 + \delta_{mask}. \tag{12}
$$

Next, we consider the term $\frac{1}{2} \left\| W f_{mask}(X^i_{m_k}) - W f_{mask}(X^i_{<k}) \right\|_2^2 = \frac{1}{2} \left\| W(f_{mask}(X^i_{m_k}) - f_{mask}(X^i_{<k})) \right\|_2^2$. We denote $X^i_{m,>k} = X^i_{m_k} - X^i_{<k}$ and $g(x_{i,u}, x_{i,v}, x_{i,w}) = x_{i,u} W^Q (x_{i,v} W^K)^\top x_{i,w} W^V$. Expanding $f_{\text{mask}}(X^i_{m_k})$ and $f_{\text{mask}}(X^i_{<k})$ separately and we obtain:

$$
\begin{aligned}
f_{mask}(X^i_{m_k}) &= X^i_{m_k} W^Q (X^i_{m_k} W^K)^\top X^i_{m_k} W^V \\
&= \sum_{x_{i,u} \in X^i_{m_k}} \sum_{x_{i,v} \in X^i_{m_k}} \sum_{x_{i,w} \in X^i_{m_k}} g(x_{i,u}, x_{i,v}, x_{i,w}) \\
&= \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{m_k}} \sum_{x_{i,w} \in X^i_{m_k}} g(x_{i,u}, x_{i,v}, x_{i,w}) + \sum_{x_{i,u} \in X^i_{m,>k}} \sum_{x_{i,v} \in X^i_{m_k}} \sum_{x_{i,w} \in X^i_{m_k}} g(x_{i,u}, x_{i,v}, x_{i,w}) \\
&= \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{<k}} \sum_{x_{i,w} \in X^i_{<k}} g(x_{i,u}, x_{i,v}, x_{i,w}) + \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{m,>k}} \sum_{x_{i,w} \in X^i_{<k}} g(x_{i,u}, x_{i,v}, x_{i,w}) \\
&+ \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{<k}} \sum_{x_{i,w} \in X^i_{m,>k}} g(x_{i,u}, x_{i,v}, x_{i,w}) + \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{m,>k}} \sum_{x_{i,w} \in X^i_{m,>k}} g(x_{i,u}, x_{i,v}, x_{i,w}) \\
&+ \sum_{x_{i,u} \in X^i_{m,>k}} \sum_{x_{i,v} \in X^i_{m_k}} \sum_{x_{i,w} \in X^i_{m_k}} g(x_{i,u}, x_{i,v}, x_{i,w})
\end{aligned}
\tag{13}
$$

with

$$
f_{mask}(X^i_{<k}) = \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{<k}} \sum_{x_{i,w} \in X^i_{<k}} g(x_{i,u}, x_{i,v}, x_{i,w}).
\tag{14}
$$

For any $(x_{i,u}, x_{i,v}, x_{i,w})$ and $(x_{i,a}, x_{i,b}, x_{i,c})$, we define:

$$
\varepsilon^{u,v,w}_{a,b,c} = g(x_{i,u}, x_{i,v}, x_{i,w}) - g(x_{i,a}, x_{i,b}, x_{i,c}).
$$

We can obtain:

$$
\begin{aligned}
\sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{m,>k}} \sum_{x_{i,w} \in X^i_{<k}} g(x_{i,u}, x_{i,v}, x_{i,w}) &= \left| X^i_{<k} \right| \left| X^i_{m,>k} \right| \left| X^i_{<k} \right| g(x_{i,a}, x_{i,b}, x_{i,c}) \\
&+ \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{m,>k}} \sum_{x_{i,w} \in X^i_{<k}} \varepsilon^{u,v,w}_{a,b,c},
\end{aligned}
$$

$$
\begin{aligned}
\sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{<k}} \sum_{x_{i,w} \in X^i_{m,>k}} g(x_{i,u}, x_{i,v}, x_{i,w}) &= \left| X^i_{<k} \right| \left| X^i_{<k} \right| \left| X^i_{m,>k} \right| g(x_{i,a}, x_{i,b}, x_{i,c}) \\
&+ \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{<k}} \sum_{x_{i,w} \in X^i_{m,>k}} \varepsilon^{u,v,w}_{a,b,c},
\end{aligned}
$$

$$
\begin{aligned}
\sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{m,>k}} \sum_{x_{i,w} \in X^i_{m,>k}} g(x_{i,u}, x_{i,v}, x_{i,w}) &= \left| X^i_{<k} \right| \left| X^i_{m,>k} \right| \left| X^i_{m,>k} \right| g(x_{i,a}, x_{i,b}, x_{i,c}) \\
&+ \sum_{x_{i,u} \in X^i_{<k}} \sum_{x_{i,v} \in X^i_{m,>k}} \sum_{x_{i,w} \in X^i_{m,>k}} \varepsilon^{u,v,w}_{a,b,c},
\end{aligned}
$$

$$
\begin{aligned}
\sum_{x_{i,u} \in X^i_{m,>k}} \sum_{x_{i,v} \in X^i_{m_k}} \sum_{x_{i,w} \in X^i_{m_k}} g(x_{i,u}, x_{i,v}, x_{i,w}) &= \left| X^i_{m,>k} \right| \left| X^i_{m_k} \right| \left| X^i_{m_k} \right| g(x_{i,a}, x_{i,b}, x_{i,c}) \\
&+ \sum_{x_{i,u} \in X^i_{m,>k}} \sum_{x_{i,v} \in X^i_{m_k}} \sum_{x_{i,w} \in X^i_{m_k}} \varepsilon^{u,v,w}_{a,b,c},
\end{aligned}
$$

where $\{x_{i,a}, x_{i,b}, x_{i,c}\} \in X^i_{<k}, \left|X^i_{<k}\right| = k - 1, \left|X^i_{m_k}\right| = s(1 - \rho_m), \left|X^i_{m,>k}\right| = s(1 - \rho_m) - k + 1$. It is easy to to calculate the summation:

$$\left|X^i_{<k}\right|\left|X^i_{m,>k}\right|\left|X^i_{<k}\right| + \left|X^i_{<k}\right|\left|X^i_{<k}\right|\left|X^i_{m,>k}\right| + \left|X^i_{<k}\right|\left|X^i_{m,>k}\right|\left|X^i_{m,>k}\right| + \left|X^i_{m,>k}\right|\left|X^i_{m_k}\right|\left|X^i_{m_k}\right|$$
$$= (s(1 - \rho_m))^3 - (k - 1)^3.$$

Let

$$\varepsilon_{a,b,c} = \sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{m,>k}} \sum_{x_{i,w}\in X^i_{<k}} \varepsilon^{u,v,w}_{a,b,c} + \sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{<k}} \sum_{x_{i,w}\in X^i_{m,>k}} \varepsilon^{u,v,w}_{a,b,c}$$
$$+ \sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{m,>k}} \sum_{x_{i,w}\in X^i_{m,>k}} \varepsilon^{u,v,w}_{a,b,c} + \sum_{x_{i,u}\in X^i_{m,>k}} \sum_{x_{i,v}\in X^i_{m_k}} \sum_{x_{i,w}\in X^i_{m_k}} \varepsilon^{u,v,w}_{a,b,c}$$

we have

$$\sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{m,>k}} \sum_{x_{i,w}\in X^i_{<k}} g(x_{i,u}, x_{i,v}, x_{i,w}) + \sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{<k}} \sum_{x_{i,w}\in X^i_{m,>k}} g(x_{i,u}, x_{i,v}, x_{i,w})$$
$$+ \sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{m,>k}} \sum_{x_{i,w}\in X^i_{m,>k}} g(x_{i,u}, x_{i,v}, x_{i,w}) + \sum_{x_{i,u}\in X^i_{m,>k}} \sum_{x_{i,v}\in X^i_{m_k}} \sum_{x_{i,w}\in X^i_{m_k}} g(x_{i,u}, x_{i,v}, x_{i,w}) \tag{15}$$
$$= ((s(1 - \rho_m))^3 - (k - 1)^3) g(x_{i,a}, x_{i,b}, x_{i,c}) + \varepsilon_{a,b,c},$$

with

$$\|\varepsilon_{a,b,c}\|_2^2 \leq ((s(1 - \rho_m))^3 - (k - 1)^3)\eta. \tag{16}$$

Since equation 15 holds for any $x_{i,a}, x_{i,b}, x_{i,c} \in X^i_{<k}$, taking the average over $(x_{i,a}, x_{i,b}, x_{i,c})$ yields

$$\sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{m,>k}} \sum_{x_{i,w}\in X^i_{<k}} g(x_{i,u}, x_{i,v}, x_{i,w}) + \sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{<k}} \sum_{x_{i,w}\in X^i_{m,>k}} g(x_{i,u}, x_{i,v}, x_{i,w})$$
$$+ \sum_{x_{i,u}\in X^i_{<k}} \sum_{x_{i,v}\in X^i_{m,>k}} \sum_{x_{i,w}\in X^i_{m,>k}} g(x_{i,u}, x_{i,v}, x_{i,w}) + \sum_{x_{i,u}\in X^i_{m,>k}} \sum_{x_{i,v}\in X^i_{m_k}} \sum_{x_{i,w}\in X^i_{m_k}} g(x_{i,u}, x_{i,v}, x_{i,w}) \tag{17}$$
$$= \left(\frac{(s(1 - \rho_m))^3}{(k - 1)^3} - 1\right) f_{mask}(X^i_{<k}) + \frac{1}{(k - 1)^3} \sum_{x_{i,a}\in X^i_{<k}} \sum_{x_{i,b}\in X^i_{<k}} \sum_{x_{i,c}\in X^i_{<k}} \varepsilon_{a,b,c}.$$

Let $\varepsilon_k = \frac{1}{(k-1)^3} \sum_{x_{i,a}\in X^i_{<k}} \sum_{x_{i,b}\in X^i_{<k}} \sum_{x_{i,c}\in X^i_{<k}} \varepsilon_{a,b,c}$. Using equation 16 we have

$$\|\varepsilon_k\|_2^2 \leq ((s(1 - \rho_m))^3 - (k - 1)^3)\eta. \tag{18}$$

Combining equation 17 and equation 13 we have

$$f_{mask}(X^i_{m_k}) = \frac{(s(1 - \rho_m))^3}{(k - 1)^3} f_{mask}(X^i_{<k}) + \varepsilon_k,$$

which implies that

$$f_{mask}(X^i_{m_k}) - f_{mask}(X^i_{<k}) = \left(\frac{(s(1 - \rho_m))^3}{(k - 1)^3} - 1\right) f_{mask}(X^i_{m_k}) - \varepsilon_k. \tag{19}$$

The upper bound of $\frac{1}{2}\left\|Wf_{mask}(X^i_{m_k}) - Wf_{mask}(X^i_{<k})\right\|^2_2$ is given by

$$
\begin{aligned}
\frac{1}{2}\left\|Wf_{mask}(X^i_{m_k}) - Wf_{mask}(X^i_{<k})\right\|^2_2 &= \frac{1}{2}\left\|\left(\frac{(s(1-\rho_m))^3}{(k-1)^3} - 1\right)Wf_{mask}(X^i_{m_k}) - W\varepsilon_k\right\|^2_2 \\
&\leq \frac{1}{2}\left\|\left(\frac{(s(1-\rho_m))^3}{(k-1)^3} - 1\right)Wf_{mask}(X^i_{m_k})\right\|^2_2 + \frac{1}{2}\left\|W\varepsilon_k\right\|^2_2 \\
&\leq \frac{1}{2}\left(\frac{(s(1-\rho_m))^3 - (k-1)^3}{(k-1)^3}\right)^2 + \frac{1}{2}((s(1-\rho_m))^3 - (k-1)^3)\left\|W\right\|^2_2\eta \\
&= \frac{w^2_k}{2(k-1)^6} + \frac{w_k}{2}\left\|W\right\|^2_2\eta.
\end{aligned}
$$

This result, combined with equation 10 to equation 12, imply that

$$
-(Wf_{\text{mask}}(X^i_{<k}))^\top \mathbb{1}_{(X^i_k)^+} \leq \frac{w^2_k}{2(k-1)^6} + \frac{w_k}{2}\left\|W\right\|^2_2\eta + \delta_{mask} \tag{20}
$$

Meanwhile, the upper bound of $((Wf_{mask}(X^i_{<k}))^\top \mathbb{1}_{(X^i_k)^-})^2$ is given by:

$$
((Wf_{mask}(X^i_{<k}))^\top \mathbb{1}_{(X^i_k)^-})^2 \leq \left\|Wf_{mask}(X^i_{<k})\right\|^2_2 \left\|\mathbb{1}_{(X^i_k)^-}\right\|^2_2 = 1. \tag{21}
$$

The two upper bounds imply the following result:

$$
\begin{aligned}
\mathcal{L}_{gen}(f_{mask}) &= -\mathbb{E}_{(X^i_{<k},(X^i_k)^+)}(Wf_{\text{mask}}(X^i_{<k}))^\top \mathbb{1}_{(X^i_k)^+} + \mathbb{E}_{X^i_{<k},(X^i_k)^-}((Wf_{\text{mask}}(X^i_{<k}))^\top \mathbb{1}_{(X^i_k)^-})^2 \\
&= -\mathbb{E}_{x_i}\frac{\sum_k (Wf_{\text{mask}}(X^i_{<k}))^\top \mathbb{1}_{(X^i_k)^+}}{s(1-\rho_m)} + \mathbb{E}_{X^i_{<k},(X^i_k)^-}((Wf_{\text{mask}}(X^i_{<k}))^\top \mathbb{1}_{(X^i_k)^-})^2 \\
&\leq \frac{\sum_k\left(\frac{w^2_k}{(k-1)^6} + w_k\left\|W\right\|^2_2\eta\right)}{2s(1-\rho_m)} + \delta_{mask} + 1.
\end{aligned} \tag{22}
$$

which finishes the proof. $\square$

B. Experimental Details

B.1. Model details in Table 1

In the comparison of GLUE score, the autoregressive model is chosen to be GPT-2 medium (345M parameters) (Radford et al., 2019) and the masked model is chosen to be BERT large (340M parameters) (Devlin et al., 2019). In the comparison of perplexity, the autoregressive model is chosen to be GPT-2 medium (345M parameters) and the masked model is chosen to be BERT-Large-CAS (395M parameters) (Wang et al., 2019).

B.2. More details in Figure 2

We employ GPT and BERT models that have been trained on a subset of the Pile dataset (Gao et al., 2020) using the Cramming protocol (Geiping & Goldstein, 2023). Following pretraining, we randomly sample several examples from the remaining Pile dataset and obtain output features using the pretrained models. We choose 8000 features for each model and calculate their similarity by computing the inner product for every pair of features. Subsequently, we sort these similarities and select the top 10^5 pairs to form Figure 2.

B.3. More details on Two-Stream Attention Transformer.

In our objective of Equation 6, we predict several tokens given one conditional sequence. It seems that we can modify the causal mask to suit the dependency relationship between tokens. But in fact, we are not capable of modeling the objective if we use a single stream. This is because of two requirements that are contradictory in a standard decoder-only

Transformer architecture, which have been discussed in XLNet (Yang et al., 2019): (1) Suppose the output of the network in position z_t is parameterized by $g_\theta(x_{\mathbf{z}<t}, z_t)$. In order to predict the token x_{z_t}, the network output in the z_t position in one layer $g_\theta(x_{\mathbf{z}<t}, z_t)$ should only use the position z_t and not the content x_{z_t}, (2) to predict the other tokens x_{z_j} with $j > t$, $g_\theta(x_{\mathbf{z}<t}, z_t)$ should also encode the content x_{z_t} to provide full contextual information. Therefore, it is not enough with only one stream. Since the two-stream attention proposed by XLNet facing the same problem works well, we build upon the two-stream attention mechanism proposed by XLNet and extend it by excavating important usage of the causal masks.

In the two-stream attention Transformer, the content stream h_θ encodes the contextual information, and the query stream g_θ predicts the targets with the help of the content stream. The two stream share the same weights of the attention block but differ in the causal mask. Intuitively, the causal masks in the two streams enable us to easily formulate various dependency relationship between tokens. The main difference for the two causal masks is that the content stream should ensure x_{z_t} can be attended to itself while the query stream is the contrary. The two-stream attention allows us to arbitrarily change the sequential order without confronting the inconsistency as mentioned above. For each self-attention layer $l = 1, \ldots, L$, the two streams of representations are schematically updated with a shared set of parameters. In the l-th layer, the outputs of a self-attention head $\mathbf{A}_h^{(l)}$ and $\mathbf{A}_g^{(l)}$ in the two-stream are computed in the form of:

$$
\begin{aligned}
\mathbf{Q}_h &= \mathbf{H}^{(l-1)}\mathbf{W}_Q^l, \ \mathbf{K}_h = \mathbf{H}^{(l-1)}\mathbf{W}_K^l, \ \mathbf{V}_h = \mathbf{H}^{(l-1)}\mathbf{W}_V^l \\
\mathbf{Q}_g &= \mathbf{G}^{(l-1)}\mathbf{W}_Q^l, \ \mathbf{K}_g = \mathbf{H}^{(l-1)}\mathbf{W}_K^l, \ \mathbf{V}_g = \mathbf{H}^{(l-1)}\mathbf{W}_V^l \\
\mathbf{A}_h^{(l)} &= \mathrm{softmax}(\frac{\mathbf{Q}_h\mathbf{K}_h^\top}{\sqrt{d_k}} + \mathbf{M}_h)\mathbf{V}_h, \mathbf{A}_g^{(l)} = \mathrm{softmax}(\frac{\mathbf{Q}_g\mathbf{K}_g^\top}{\sqrt{d_k}} + \mathbf{M}_g)\mathbf{V}_g,
\end{aligned}
\tag{23}
$$

where $\mathbf{H}^{(l-1)}$ and $\mathbf{G}^{(l-1)}$, the representations before the $(l-1)$-th layer of the content stream and the query stream, are linearly projected to queries, keys and values using shared trainable parameter matrices $\mathbf{W}_Q^l, \mathbf{W}_K^l, \mathbf{W}_V^l$ respectively, and d_k is the dimension of the representations. $\mathbf{M}_h, \mathbf{M}_g \in \mathbb{R}^{s \times s}$ are the causal masks in the content stream and the query stream, respectively, which are the key factors in our framework. The value of each element $(\mathbf{M})_{ij}$ in either causal mask can only be 0 or $-\infty$, where $(\mathbf{M})_{ij} = 0$ means the $j-$th token **can** be attended to the $i-$th token and $(\mathbf{M})_{ij} = -\infty$ means the $j-$th token **cannot** be attended to the $i-$th token. This indicates that the causal mask has a one-to-one correspondence with the dependency relationship between tokens. Therefore, **we only need to alter the causal masks M_h and M_g to adapt to different choices of the group setting, which eliminate the need of using multiple architectures.** We will show how to construct the corresponding attention masks.

Note that the value of each element $(\mathbf{M})_{ij}$ in either causal mask can only be 0 or $-\infty$, where $(\mathbf{M})_{ij} = 0$ means the $j-$th token **can** be attended to the $i-$th token and $(\mathbf{M})_{ij} = -\infty$ means the $j-$th token **cannot** be attended to the $i-$th token. This implicates that by finding out the dependency between each token we can determine the causal mask. We provide the causal masks corresponding to Equation (6): Let $f(t)$ denote the index of the group to which x_t belongs. In this situation, x_i will be dependent on x_j if $f(i) > f(j)$. Besides, tokens in the same group is dependent on each other in the content stream to provide contextual information. The causal masks for the two streams should be

$$
(\mathbf{M}_h)_{ij} = \begin{cases} 0 & f(i) \geq f(j) > 0 \\ -\infty & f(j) > f(i) > 0 \end{cases}, \ (\mathbf{M}_g)_{ij} = \begin{cases} 0 & f(i) > f(j) > 0 \\ -\infty & f(j) \geq f(i) > 0 \end{cases}
\tag{24}
$$

C. Ablation Study

We conduct more ablation experiments on the choice of t in the vision domain. We vary t in $[1, 2, 3, 4]$ and investigate the mixture of $t = 1$ and $t = 2$. We use ImageNet as the dataset. The results are shown in the Table 9. The results indicate that there exists a sweet point between $t = 2$ and $t = 3$. This means that t should not be too large. The mixture choice of t does not bring much improvement as well.

Table 9. Ablation study on the choice of t in the diversity-enhanced autoregressive objective.

t	Linear probing accuracy	Finetuning accuracy
1 (original autoregressive objective)	56.2	82.5
2 (used in the previous experiments)	59.4	82.9
3	59.6	82.7
4	58.8	82.5
mixture of 1 and 2	58.1	82.5