How Transformers Learn Diverse Attention Correlations in Masked Vision Pretraining

Yu Huang*† UPenn
Zixin Wen*‡ CMU
Yuejie Chi§ CMU
Yingbin Liang¶ OSU

March 2024; Revised May 2024

Abstract

Masked reconstruction, which predicts randomly masked patches from unmasked ones, has emerged as an important approach in self-supervised pretraining. However, the theoretical understanding of masked pretraining is rather limited, especially for the foundational architecture of transformers. In this paper, to the best of our knowledge, we provide the first end-to-end theoretical guarantee of learning one-layer transformers in masked reconstruction self-supervised pretraining. On the conceptual side, we posit a mechanism of how transformers trained with masked vision pretraining objectives produce empirically observed **local and diverse** attention patterns, on data distributions with spatial structures that highlight *feature-position correlations*. On the technical side, our end-to-end characterization of training dynamics in softmax-attention models simultaneously accounts for input and position embeddings, which is developed based on a careful analysis tracking the interplay between feature-wise and position-wise attention correlations.

Contents

1 Introduction **2**

2 Problem Setup **5**
2.1 Masked Vision Pretraining . . . 5
2.2 Data Distribution . . . 5
2.3 Masked Vision Pretraining with Transformers . . . 6

3 Attention Patterns and Feature-Position Correlations **7**

4 Statements of Main Results **8**

5 Proof Sketch **10**

6 Conclusion **11**

A Additional Related Work **16**

B Experiments **16**

*The first two authors contributed equally.
†Department of Statistics and Data Science, Wharton School, University of Pennsylvania. yuh42@wharton.upenn.edu
‡Machine Learning Department, Carnegie Mellon University. zixinw@andrew.cmu.edu
§Department of Electrical and Computer Engineering, Carnegie Mellon University. yuejiec@andrew.cmu.edu
¶Department of Electrical and Computer Engineering, The Ohio State University. liang.889@osu.edu

C Overview of the Proof Techniques **17**
C.1 GD Dynamics of Attention Correlations . . . 18
C.2 Phase I: Decoupling the Global FP Correlations . . . 19
C.3 Phase II: Growth of Target Local FP Correlation . . . 19
C.4 Learning Processes in Other Scenarios . . . 20

D Preliminaries **20**
D.1 Gradient Computations . . . 21
D.2 High-probability Event . . . 28
D.3 Properties of Loss Function . . . 29

E Overall Induction Hypotheses and Proof Plan **31**
E.1 Positive Information Gap . . . 31
E.2 Negative Information Gap . . . 31
E.3 Proof Outline . . . 32

F Analysis for the Local Area with Positive Information Gap **32**
F.1 Phase I, Stage 1 . . . 32
F.2 Phase I, Stage 2 . . . 38
F.3 Phase II, Stage 1 . . . 42
F.4 Phase II, Stage 2 . . . 45

G Analysis for Local Areas with Negative Information Gap **48**

H Analysis for the Global area **50**

I Proof of Main Theorems **51**
I.1 Proof of Induction Hypotheses . . . 51
I.2 Proof of Theorem 4.1 and Theorem 4.2 with Positive Information Gap . . . 52
I.3 Proof of Theorem 4.1 and Theorem 4.2 with Negative Information Gap . . . 53

1 Introduction

Self-supervised learning has been a leading approach to pretrain neural networks for downstream applications since the introduction of BERT [10] and GPT [35] in natural language processing (NLP). On the other hand, in vision, self-supervised learning was initially focused more on *discriminative* methods, which include contrastive learning [20, 7] and non-contrastive learning methods [16, 7, 5, 62]. Inspired by masked language models in NLP, together with the crucial progress made by [11] in successfully implementing vision transformers (ViTs), *generative* approaches, such as masked reconstruction-based methods, have become popular in self-supervised vision pretraining, especially since the emergence of masked autoencoders (MAE) [19] and SimMIM [60].

In masked vision pretraining, neural networks are instructed to reconstruct some or all patches of an image given a masked version, aiming to learn certain abstract semantics of visual contents when trained to fill in the missing patches. In practice, this approach not only proves to be very successful with remarkable finetuning performance but also shows intriguing phenomena that differ significantly from the behaviors observed in other self-supervised learning methods. The seminal work of [19] showed that MAE can conduct visual reasoning when filling in masked patches even with very high masking rates, suggesting that masked image modeling learns some complex relationships between visual objects. Some critical observations from recent research [55, 32, 59] have suggested that the ViTs trained via masked reconstruction objectives display **diverse attention patterns**: different query patches pay attention to distinct local areas. This is in sharp contrast to the discriminative self-supervised learning approaches, whose attention maps mostly capture the most significant global pattern regardless of where the query patches are, leading to a phenomenon known as "attention collapse", as shown in Figure 1. Given these empirical observations, it naturally prompts the question: from a *theoretical* standpoint, how do ViTs learn these varied attention patterns during masked pretraining?

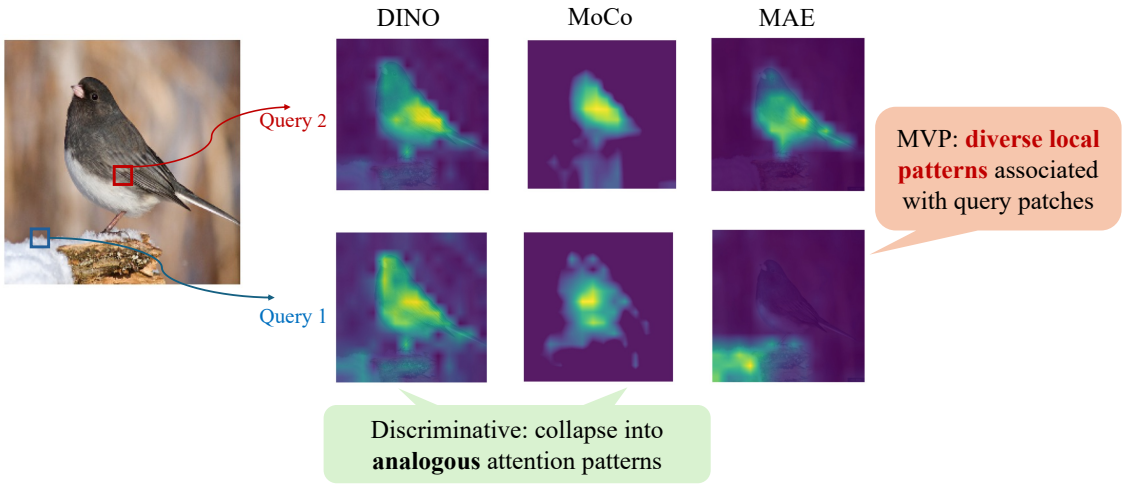

Figure 1: Visualization of attention maps in the last layer of ViT for query patches from two different spatial locations, trained by the generative self-supervised learning approach of masked reconstruction (MAE) and two discriminative self-supervised learning methods (DINO [5] and MoCo [9]).

Despite the substantial empirical effort dedicated to investigating masked vision pretraining, our theoretical understanding of it is still nascent. Most existing theories for self-supervised pretraining focused on discriminative methods [2, 8, 37, 18, 57, 43, 52, 58], such as contrastive learning. Among very few attempts towards masked image modeling, [4] studied the patch-based attention via an integral kernel perspective; [63] analyzed MAE through an augmentation graph framework, which connects MAE to contrastive learning. [30] characterized the optimization process of MAE with shallow convolutional neural networks (CNNs). However, **transformers**, as the prevailing architecture in current deep learning practice, were not addressed in the aforementioned theoretical studies of vision pretraining and, more broadly, self-supervised pretraining, leaving a significant gap in the literature.

Building on the mind-blowing empirical advances and recognizing the lack of theoretical understanding of *masked reconstruction-based methods with transformers* in self-supervised learning, we are motivated to ask:

Our Research Questions

Can we theoretically characterize what solutions transformers converge to in masked vision pretraining? How do transformers learn **diverse local attention patterns** rather than object-focused global attention in such pretraining?

Contributions. In this paper, we take a step towards answering the above questions by studying the training process of one-layer transformers trained by gradient descent (GD) for masked reconstruction objectives. We focus on spatially structured data distributions, where each image is sampled from different clusters characterized by different patch-wise associations. Under such setting:

1. We provide the global convergence guarantee of the masked reconstruction loss and characterize how attention is distributed at convergence to demonstrate the non-collapsed diverse local attention patterns obtained by masked pretraining , which to our knowledge is the first end-to-end theory of learning transformers for masked-reconstruction type self-supervised methods.
2. We analyze the training dynamics of *attention correlations* (see Definition 3.1) and prove that ViTs manage to capture desirable diverse local patterns in masked pretraining by learning the target **feature-position attention correlations** for all visual features irrespective of their significance i.e., whether they are global or local features. This marks the first result of learning the softmax attention model that jointly considers input and position encodings.
3. We design a novel empirical metric, termed as the **attention diversity metric**, to probe vision transformers trained by different methods. Our new observations (see Figure 2) provide further evidence for the diverse

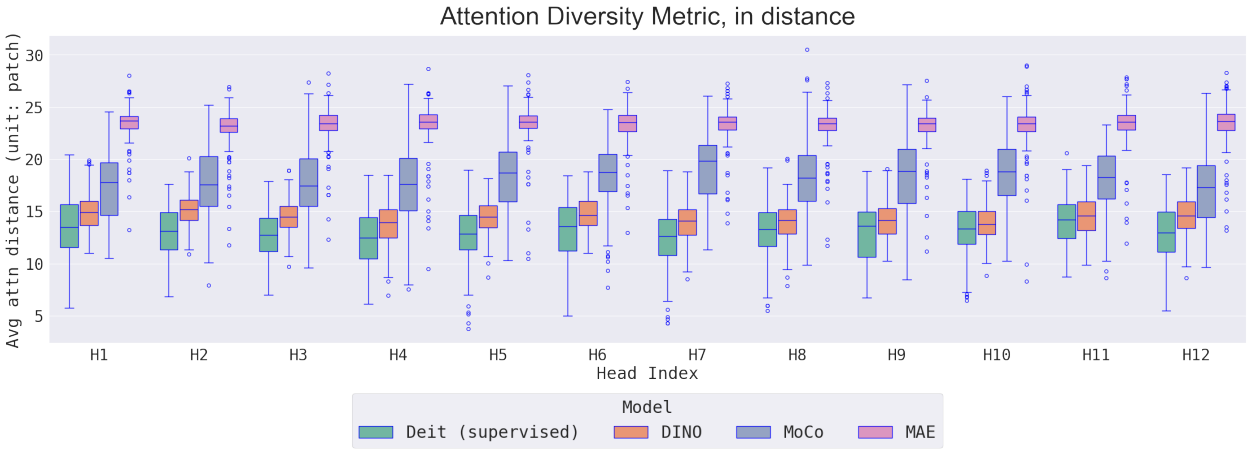

Figure 2: **Attention Diversity Metric:** We examined the last layer of ViTs trained by masked reconstruction (MAE), discriminative self-suprvised learning (MoCo, DINO), and supervised learning (DeiT). Lower values of this metric signify focused attention on a similar area across different patches, reflecting a global pattern of focus. Conversely, higher values suggest that attention is dispersed, focusing on different, localized areas. The results show that the MAE model excels in capturing diverse local patterns (see Appendix B for more details).

local patterns revealed through masked image modeling.

Comparisons with prior works. [1] A few works [22, 30] have studied topics related to ours. Here we summarize the differences between our work and theirs regarding settings and analysis.

- [22] pioneered the characterization of the training dynamics for transformers in supervised learning. They provided the first convergence result for one-layer softmax-attention transformers trained on a simple visual data distribution, in which the patch association is **fixed** (see Definition 2.1 in their paper). Their assumptions require learning only the position-position correlations, which is rather limited. In Section 3, after formally defining the feature-position correlations, we will highlight the inadequacy of such a setting to fully demonstrate masked pretraining's capability to learn locality in detail. In this work, we draw inspiration from their data assumptions and generalize them to allow **varying** spatial structures of patches (see Definition 2.1). Due to this generalization, we need to analyze the learning processes of *different* spatial correlations among visual features simultaneously, which poses key challenges in our overall analysis.
- [30] analyzed the feature learning process of MAE with CNNs architectures, without any patch-level positional structure in the network. The main implication of their theoretical result is that the trained *CNN* can provably identify all global patterns in masked pretraining, which does not capture the findings in recent self-attention investigations [32, 55] and Figure 1 that masked pretraining with *transformers* can learn diverse local patterns and avoid collapsing solutions. This suggests that the masked reconstruction-based method's success under *transformers* may fall beyond what *CNN*-based analysis could reveal, emphasizing the importance of studying *attention* patterns from a theoretical point of view. This is the focus of this paper.

Notation. We introduce notations to be used throughout the paper. For any two functions $h(x)$ and $g(x)$, we use $h(x) = \Omega(g(x))$ $\big($resp. $h(x) = O(g(x))\big)$ to denote that there exist some universal constants $C_1 > 0$ and a_1, s.t. $|h(x)| \geq C_1|g(x)|$ $\big($resp. $|h(x)| \leq C_1|g(x)|\big)$ for all $x \geq a_1$; Furthermore, $h(x) = \Theta(g(x))$ indicates $h(x) = \Omega(g(x))$ and $h(x) = O(g(x))$ hold simultaneously. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function, and let $[N] := \{1, 2, \ldots, N\}$. We use $\widetilde{O}$, $\widetilde{\Omega}$, and $\widetilde{\Theta}$ to further hide logarithmic factors in the respective notations. We use $\mathrm{poly}(P)$ and $\mathrm{polylog}(P)$ to represent large constant-degree polynomials of P and $\log(P)$, respectively.

[1] More detailed discussions for related work can be found in Appendix A.

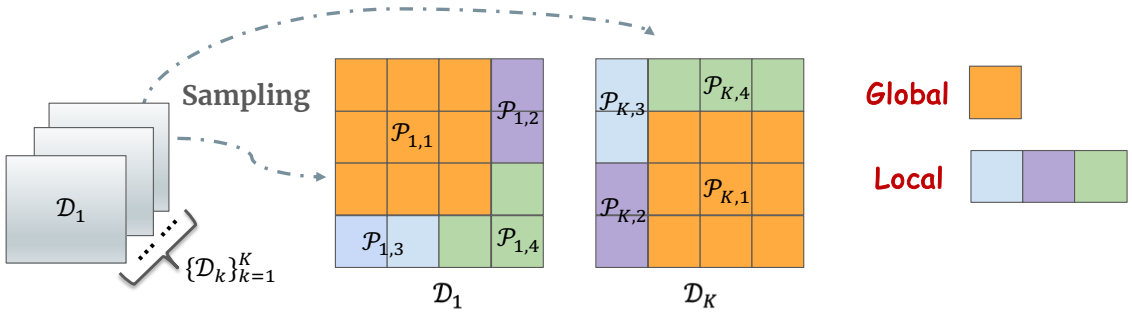

Figure 3: Illustration of our data distribution (see Definition 2.1). Each cluster $\mathcal{D}_k$ is segmented into distinct areas $\mathcal{P}_{k,j}$, with squares in the same color representing the same area $\mathcal{P}_{k,j}$. The global area $\mathcal{P}_{k,1}$ (depicted in orange) contains a larger count of patches compared to any other local areas. It is important to note that while we use spatially contiguous partitions for clarity in this illustration, our data model is also applicable to non-contiguous cases.

2 Problem Setup

In this section, we present our problem formulations for studying the training process of transformers during masked vision pretraining. We begin with some background information, followed by a description of our dataset settings. We then detail the masked pretraining strategy and the specific transformer architecture considered in this paper.

2.1 Masked Vision Pretraining

We follow the masked reconstruction frameworks in [19, 60]. Each data sample $X \in \mathbb{R}^{d\times P}$ has the form $X = (X_{\mathbf{p}})_{\mathbf{p}\in\mathcal{P}}$, which has $|\mathcal{P}| = P$ patches, and each patch $X_{\mathbf{p}} \in \mathbb{R}^d$. Given a collection of images $\{X_i\}_{i\in[n]}$, we select a masking set $\mathcal{M}_i \subset [P]$ for each image X_i, and mask these patches to a uniform value $\mathsf{M} \in \mathbb{R}^d$. The resulting masked images $\{\mathsf{M}(X_i)\}_{i\in[n]}$ are given by

$$\mathsf{M}(X_i)_{\mathbf{p}} = \begin{cases} [X_i]_{\mathbf{p}} & \mathbf{p} \in \mathcal{U}_i \\ \mathsf{M} & \mathbf{p} \in \mathcal{M}_i \end{cases}, \qquad i \in [n], \tag{2.1}$$

where $\mathcal{U}_i = \mathcal{P} \setminus \mathcal{M}_i$ is the index set of unmasked patches. Let $F : X \mapsto \widehat{X}$ be an architecture that outputs a reconstructed image $\widehat{X} \in \mathbb{R}^{d\times P}$ for any given input $X \in \mathbb{R}^{d\times P}$. The pretraining objective is then defined as the mean-squared reconstruction loss over a series of subsets $\mathcal{P}_i' \subset \mathcal{P}$ of the image as follows:

$$\mathcal{L}(F) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \sum_{\mathbf{p}\in\mathcal{P}_i'} \left\| [X_i]_{\mathbf{p}} - [F(\mathsf{M}(X_i))]_{\mathbf{p}} \right\|_2^2. \tag{2.2}$$

[19] chose the subset $\mathcal{P}_i'$ as the set of masked patches $\mathcal{M}_i$, whereas [60] aimed to reconstruct the full image $\mathcal{P}_i' = \mathcal{P}$. We do not explore the trade-offs between these two approaches in our study. As shall be seen momentarily, our theoretical framework is based on F being a simplified version of vision transformers [11] which utilizes the attention mechanism [49].

2.2 Data Distribution

We assume the data samples $X \in \mathbb{R}^{d\times P}$ are drawn independently based on some data distribution $\mathcal{D}$. To capture the *feature-position (FP) correlation* in the learning problem, we consider the following setup for vision data. We assume that the data distribution consists of many different clusters, where each cluster captures a distinct spatial pattern, and hence is defined by a different partition of patches with a different set of visual features. We define the data distribution $\mathcal{D}$ formally as follows. An intuitive illustration of data generation is given in Figure 3.

Definition 2.1 (Data distribution $\mathcal{D}$)**.** The data distribution $\mathcal{D}$ has $K = O(1)$ different clusters $\{\mathcal{D}_k\}_{k=1}^K$. For every cluster $\mathcal{D}_k, k \in [K]$, there is a corresponding partition of $\mathcal{P}$ into N_k disjoint subsets $\mathcal{P} = \bigcup_{j=1}^{N_k} \mathcal{P}_{k,j}$ which we call **areas**. For each sample $X = (X_{\mathbf{p}})_{\mathbf{p}\in\mathcal{P}}$, its sampling process is as follows:

- We draw $\mathcal{D}_k$ uniformly at random from all clusters and draw a sample X from $\mathcal{D}_k$.
- Given $k \in [K]$, for any $j \in [N_k]$, all patches $X_{\mathbf{p}}$ in the area $\mathcal{P}_{k,j}$ are given the same content $X_{\mathbf{p}} = v_{k,j} z_j(X)$, where $v_{k,j} \in \mathbb{R}^d$ is the *visual* feature and $z_j(X)$ is the latent variable. We assume $\bigcup_{k=1}^K \bigcup_{j=1}^{N_k} \{v_{k,j}\}$ are orthogonal to each other with unit norm.
- Given $k \in [K]$, for any $j \in [N_k]$, $z_j(X) \in [L, U]$, where $0 \le L < U$ are on the order of $\Theta(1)$.[2]

Global and local features in an image. Image data naturally contains two types of features: the global features and the local features. For instance, in an image of an object, global features can capture the shape and texture of the object, such as the fur color of an animal, whereas local features describe specific details of local areas, such as the texture of leaves in the background. Recent empirical studies on self-supervised pretraining with transformers [32, 55] and observations in Figure 1 collectively show that masked pretraining exhibits the capacity to avoid attention collapse concentrating towards those global shapes by identifying diverse local attention patterns. Consequently, unraveling their mechanisms necessitates a thorough examination of data characteristics that embody both global and local features. In this paper, we characterize these two types of features by the following assumption on the data.

Assumption 2.2 (Global feature vs local feature)**.** Let $\mathcal{D}_k$ with $k \in [K]$ be a cluster from $\mathcal{D}$. We let $\mathcal{P}_{k,1}$ be the **global area** of cluster $\mathcal{D}_k$, and all the other areas $\mathcal{P}_{k,j}, j \in [N_k] \setminus \{1\}$ be the **local areas**. Since each area corresponds to an assigned feature, we also call them the global and local features, respectively. Moreover, we assume:

- Global area: given $k \in [K]$, we set $C_{k,1} = |\mathcal{P}_{k,1}| = \Theta(P^{\kappa_c})$ with $\kappa_c \in [0.5005, 1]$, where $C_{k,1}$ is the number of patches in the global area $\mathcal{P}_{k,1}$.
- Local area: given $k \in [K]$, we choose $C_{k,j} = \Theta(P^{\kappa_s})$ with $\kappa_s \in [0.001, 0.5]$ for $j > 1$, where $C_{k,j}$ denotes the number of patches in the local area $\mathcal{P}_{k,j}$.

The rationale for defining the global feature in this manner stems from observing that patches representing global features ($C_{k,1}$) typically occur more frequently than those representing local features ($C_{k,j}$, for $j > 1$), since global features capture the primary visual information of an image, offering a dominant view, while local features focus on subtler details within the image.

2.3 Masked Vision Pretraining with Transformers

Transformer architecture. A transformer block [49, 11] consists of a self-attention layer followed by an MLP layer. The self-attention layer has multiple heads, each of which consists of the following components: a query matrix W^Q, a key matrix W^K, and a value matrix W^V. Given an input X, the output of one head in the self-attention layer can be described by the following mapping:

$$G(X; W^Q, W^K, W^V) = \text{softmax}\left((W^Q X)^\top W^K X\right) \cdot (W^V X)^\top, \tag{2.3}$$

where the softmax$(\cdot)$ function is applied row-wisely and for a vector input $z \in \mathbb{R}^P$, the i-th entry of softmax(z) is given by $\frac{\exp(z_i)}{\sum_{s=1}^P \exp(z_s)}$.

To simplify the theoretical analysis, we consolidate the product of query and key matrices $(W^Q)^\top W^K$ into one weight matrix denoted as Q. Furthermore, we set W^V to be the identity matrix and fixed during the training. These simplifications are often taken in recent theoretical works [22, 21, 64] in order to allow tractable analysis. With these simplifications in place, (2.3) can be rewritten as

$$G(X; Q) = \text{softmax}\left(X^\top Q X\right) \cdot X^\top. \tag{2.4}$$

[2]The distribution of $z_j(X)$ can be arbitrary within the above support set.

Input tokens in transformers are indistinguishable without explicit spatial information. Therefore, positional encodings should be added to the input embeddings to retain this crucial positional context as in practices [11, 19]. Our assumptions regarding the positional encodings are as follows:

Assumption 2.3 (Positional encoding)**.** We assume fixed positional encodings, which is consistent with the implementation in MAE [19]: $E = (e_{\mathbf{p}})_{\mathbf{p}\in\mathcal{P}} \in \mathbb{R}^{d\times P}$ where positional embedding vectors $e_{\mathbf{p}}$ are orthogonal to each other and to all the features $v_{k,j}$, and are of unit-norm.

We now include positional embeddings in (2.4) and introduce the network architecture for masked reconstruction used in this study.

Definition 2.4 (Transformer network for masked pretraining)**.** We assume that our vision transformer $F(X;Q)$ consists of a single-head self-attention layer with an attention weight matrix $Q \in \mathbb{R}^{d\times d}$. For an input image $X \sim \mathcal{D}$, we add positional encoding by letting $\widetilde{X} = X + E$. The attention score from patch $X_{\mathbf{p}}$ to patch $X_{\mathbf{q}}$ is denoted by

$$\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}(X;Q) \coloneqq \frac{e^{\widetilde{X}_{\mathbf{p}}^\top Q \widetilde{X}_{\mathbf{q}}}}{\sum_{\mathbf{r}\in\mathcal{P}} e^{\widetilde{X}_{\mathbf{p}}^\top Q \widetilde{X}_{\mathbf{r}}}}, \quad \text{for } \mathbf{p},\mathbf{q} \in \mathcal{P}. \tag{2.5}$$

Then the output of the transformer is given by

$$[F(X;Q)]_{\mathbf{p}} = \sum_{\mathbf{q}\in\mathcal{P}} \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}(X;Q) \cdot X_{\mathbf{q}}, \quad \text{for } \mathbf{p} \in \mathcal{P}. \tag{2.6}$$

Then we formally define the masking operation and the objective for our masked pretraining task.

Definition 2.5 (Random masking)**.** Let $\mathsf{M}(X) \to \mathbb{R}^{d\times P}$ denote the random masking operation, which randomly selects (without replacement) a subset of patches $\mathcal{M}$ in X with a masking ratio $\gamma = \Theta(1) \in (0,1)$ and masks them to be $\mathsf{M} \coloneqq \mathbf{0} \in \mathbb{R}^d$. The masked samples obey (2.1).

Masked reconstruction objective. To train the transformer model $F(\mathsf{M}(X);Q)$, we minimize the squared reconstruction error (2.2) only on masked patches, where $\mathsf{M}(X)$ follows Definition 2.5. The training objective thus can be written as

$$\mathcal{L}(Q) \coloneqq \frac{1}{2}\mathbb{E}\left[\sum\nolimits_{\mathbf{p}\in\mathcal{P}} \mathbb{1}\{\mathbf{p}\in\mathcal{M}\}\left\|[F(\mathsf{M}(X);Q)]_{\mathbf{p}} - X_{\mathbf{p}}\right\|^2\right], \tag{2.7}$$

where the expectation is with respect to both the data distribution and the masking. Note that our objective remains highly nonconvex with the model defined in Definition 2.4.

Training algorithm. The learning objective in (2.7) is minimized via GD with learning rate $\eta > 0$. At $t = 0$, we initialize $Q^{(0)} \coloneqq \mathbf{0}_{d\times d}$ as the zero matrix. The parameter is updated as follows:

$$Q^{(t+1)} = Q^{(t)} - \eta\nabla_Q \mathcal{L}(Q^{(t)}).$$

In the following, any variable with a superscript $^{(t)}$ represents that variable at the t-th step of training. Note that the initialization of $Q^{(0)}$ results in any query patch uniformly attending to all patches.

3 Attention Patterns and Feature-Position Correlations

To show the significance of the data distribution design and understand the nature of the masked reconstruction task, in this section, we will provide some preliminary implications of the spatial structures in Definition 2.1. Intuitively, for a given cluster $\mathcal{D}_k$, to reconstruct a missing patch $\mathbf{p} \in \mathcal{P}_{k,j} \cap \mathcal{M}$, the attention head should exploit all *unmasked* patches in the *target* area $\mathcal{P}_{k,j}$ to find the same visual feature $v_{k,j}$ to fill in the blank, which emphasizes the *locality* for $\mathbf{p}$ in different areas. This approach remains effective even at high masking rates, provided that a small number of patches in $\mathcal{P}_{k,j}$ remain unmasked. We will elaborate on this point

by describing the *area attentions* and illustrating the intuition about how they can be learned via *attention correlations* (Definition 3.1).

Area attention. We first define a new notation for a cleaner presentation. For $X \sim \mathcal{D}$ and $\mathbf{p} \in \mathcal{P}$, we write the attention of patch $X_{\mathbf{p}}$ to a subset $\mathcal{A} \subset \mathcal{P}$ of patches by

$$\widetilde{\mathbf{Attn}}_{\mathbf{p}\to\mathcal{A}}(X;Q) \coloneqq \textstyle\sum_{\mathbf{q}\in\mathcal{A}} \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}(X;Q).$$

Let us explain why the above notion of area attention matters in understanding how attention works in masked reconstruction. Suppose we have a sample X picked from $\mathcal{D}_k$, and the patch $X_{\mathbf{p}}$ with $\mathbf{p} \in \mathcal{P}_{k,j}$ is masked. Then the prediction of $X_{\mathbf{p}}$ given masked input $\mathsf{M}(X)$ can be written as

$$\begin{aligned}[F(\mathsf{M}(X);Q)]_{\mathbf{p}} &= \textstyle\sum_{\mathbf{q}\in\mathcal{P}} \mathsf{M}(X)_{\mathbf{q}} \cdot \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}(\mathsf{M}(X);Q) \\ &= \textstyle\sum_{i\in[N_k]} z_i(X) v_{k,i} \cdot \widetilde{\mathbf{Attn}}_{\mathbf{p}\to\mathcal{U}\cap\mathcal{P}_{k,i}}(\mathsf{M}(X);Q) \text{ (since } \mathsf{M}(X)_{\mathbf{q}} = \mathbf{0} \text{ if } \mathbf{q} \in \mathcal{M}).\end{aligned}$$

To reconstruct the original patch $X_{\mathbf{p}}$, the transformer should not only focus on the correct area $\mathcal{P}_{k,j}$, but must also prioritize attention to the *unmasked* patches within this area. This specificity is denoted by the area attention $\widetilde{\mathbf{Attn}}_{\mathbf{p}\to\mathcal{U}\cap\mathcal{P}_{k,j}}$ over $\mathcal{U} \cap \mathcal{P}_{k,j}$, a requirement imposed by masking operations. To further explain how transformers perform such prioritization, we introduce the following quantities, which capture the major insights of our analysis that differentiate from those in [22].

Definition 3.1. (Attention correlations) Let $\mathbf{p} \in \mathcal{P}$, and we define attention correlations as:
1. Feature-Position (FP) Correlation: $\Phi_{\mathbf{p}\to v_{k,m}} \coloneqq e_{\mathbf{p}}^\top Q v_{k,m}$, for $k \in [K]$ and $m \in [N_k]$;

2. Position-Position (PP) Correlation: $\Upsilon_{\mathbf{p}\to\mathbf{q}} \coloneqq e_{\mathbf{p}}^\top Q e_{\mathbf{q}}$, $\forall \mathbf{q} \in \mathcal{P}$.

Due to our (zero) initialization of $Q^{(0)}$, we have $\Phi^{(0)}_{\mathbf{p}\to v_{k,m}} = \Upsilon^{(0)}_{\mathbf{p}\to\mathbf{q}} = 0$.

These two types of attention correlations, FP correlation $\Phi_{\mathbf{p}\to v_{k,m}}$ and PP correlation $\Phi_{\mathbf{p}\to\mathbf{q}}$, act as the exponent terms within the softmax calculations for attention scores. Given $\mathbf{p} \in \mathcal{P}_{k,j}$ is masked, the (unnormalized) attention $\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}$ directed towards an *unmasked* patch $\mathbf{q}$ is influenced jointly by these correlations. Hence, the described attention pattern can emerge from either a substantial FP correlation $\Phi_{\mathbf{p}\to v_{k,j}}$ or a significant PP correlation $\Phi_{\mathbf{p}\to\mathbf{q}}$ for $\mathbf{q}$ in the same area as $\mathbf{p}$. However, in our setting, the latter mechanism—learning via PP correlation—fails to produce desired attention patterns, as explained below on a notable gap that prior work [22] did not address.

Can pure positional attention explain the transformer's ability to learn locality? [22] theoretically analyzed how ViTs can identify spatially localized patterns by minimizing the supervised cross-entropy loss with GD. Their analysis focused on a spatially structured dataset equivalent to our settings when $K = 1$ without distinguishing the global and local features. Hence, the association between patch $\mathbf{p}$ and $\mathbf{q}$ is fixed since there is only a single cluster. Based on such invariant patch-wise relations, they showed that ViTs can achieve optimal attention patterns by solely learning the "patch association", where $e_{\mathbf{p}}^\top Q e_{\mathbf{q}}$ is large for $\mathbf{p}, \mathbf{q}$ coming from the same area, corresponding to large PP correlation $\Gamma_{\mathbf{p}\to\mathbf{q}}$ in our settings. However, such an assumption of invariant patch associations is often unrealistic for vision datasets in practice, for instance, a cube-shaped building typically requires a different attention pattern from a bird inside the woods. Therefore, when multiple patterns appear in the data distribution (e.g., in our settings with $K > 1$ clusters), relying solely on PP correlations is insufficient, and is often undesirable due to variations of PP correlations across clusters. This highlights the necessity of examining FP correlations for a deeper understanding of the local representation power of transformers.

4 Statements of Main Results

In this section, we present the main theorems of this paper. Our results are structured into two parts: *i).* analysis of convergence (Theorem 4.1), which includes the global convergence guarantee of the masked

reconstruction loss and characterization of the attention pattern at the end of training to demonstrate the diverse locality; *ii).* learning dynamics of attention correlations (Theorem 4.2), which shows how transformers capture target FP correlations while downplaying PP correlations as discussed in Section 3.

Information gap and a technical condition. Based on our data model in Section 2.2, we introduce a notion of *information gap* to quantify the difference of significance between global and local areas (cf. Assumption 2.2). Denoted as Δ, the information gap is defined as follows:

$$\Delta \coloneqq (1-\kappa_s) - 2(1-\kappa_c). \tag{4.1}$$

Our study focuses on the regime where Δ is not too close to zero, i.e., $|\Delta| = \Omega(1)$, which allows for cleaner induction arguments. This condition could be potentially relaxed via more involved analysis.

Notations for theorem and proof presentations. Based on the crucial role of those unmasked patches for the reconstruction task, we further define the *unmasked area attention* as follows:

$$\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}(\mathsf{M}(X);Q) \coloneqq \widetilde{\mathbf{Attn}}_{\mathbf{p}\to\mathcal{U}\cap\mathcal{P}_{k,m}}(\mathsf{M}(X);Q).$$

We abbreviate $\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}(\mathsf{M}(X);Q^{(t)})$ $(\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}(\mathsf{M}(X);Q^{(t)}))$ as $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}$ ($\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}}$). Moreover, to properly evaluate the reconstruction performance, we further introduce the following notion of the reconstruction loss with respect to a specific patch $\mathbf{p}\in\mathcal{P}$:

$$\mathcal{L}_{\mathbf{p}}(Q) = \frac{1}{2}\mathbb{E}\left[\mathbb{1}\{\mathbf{p}\in\mathcal{M}\}\left\|[F(\mathsf{M}(X);Q)]_{\mathbf{p}} - X_{\mathbf{p}}\right\|^2\right]. \tag{4.2}$$

Now we present our first main result regarding the convergence of masked pretraining.

Theorem 4.1 (Training convergence)**.** *Suppose the information gap $\Delta\in[-0.5,-\Omega(1)]\cup[\Omega(1),1]$. For any $0<\epsilon<1$, suppose* $\mathrm{polylog}(P)\gg\log(\frac{1}{\epsilon})$*. We train the transformer in Def. 2.4 by GD to minimize reconstruction loss in* (2.7) *with $\eta\ll\mathrm{poly}(P)$. Then for each patch $\mathbf{p}\in\mathcal{P}$, we have*

1. *Loss converges: $\mathcal{L}_{\mathbf{p}}(Q^{(T^\star)})-\mathcal{L}^\star_{\mathbf{p}}\le\epsilon$ in $T^\star = O\Big(\frac{1}{\eta}\log(P)P^{\max\{2(\frac{U}{L}-1),1\}(1-\kappa_s)}+\frac{1}{\eta\epsilon}\log\left(\frac{P}{\epsilon}\right)\Big)$ iterations, where $\mathcal{L}^\star_{\mathbf{p}}$ is the global minimum of the patch-level reconstruction loss in* (4.2)*.*
2. **Area-wide** *pattern of attention: given cluster $k\in[K]$, and $\mathbf{p}\in\mathcal{P}_{k,j}$ for some $j\in[N_k]$, if $X_{\mathbf{p}}$ is masked, then the one-layer transformer nearly "pays all attention" to all unmasked patches in the same area $\mathcal{P}_{k,j}$ as $\mathbf{p}$, i.e.,*
$$\left(1-\mathbf{Attn}^{(T^\star)}_{\mathbf{p}\to\mathcal{P}_{k,j}}\right)^2\le O(\epsilon).$$

Theorem 4.1 indicates that, at the time of convergence, for any masked query patch $X_{\mathbf{p}}$ in the k-th cluster, the transformer exhibits an *area-wide* pattern of attention, concentrating on those unmasked patches within the area that $\mathbf{p}$ lies in, as demonstrated in Section 3. The location of the patch determines such area-wide attention and can be achieved no matter if $\mathbf{p}$ belongs to the global or local areas, which jointly highlight the **diverse local patterns** for masked vision pretraining.

Next, we detail the training phases of attention correlations in the following theorem, which explicitly confirms that the model **learns** target FP correlations while **ignoring** PP correlations to achieve the desirable area-wide attention patterns as suggested in Section 3. The intuitive explanations are provided in Section 5.

Theorem 4.2 (Learning Feature-Position correlations)**.** *Following the same assumptions in Theorem 4.1, for $\mathbf{p}\in\mathcal{P}$, given $k\in[K]$, if $\mathbf{p}\in\mathcal{P}_{k,j}$ for some $j\in[N_k]$, we have*

For positive information gap $\Delta\in[\Omega(1),1]$:

a. *Global areas ($j=1$) learn FP correlation in* **one-phase:** $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ *monotonically increases to $O(\log(P)/\eta)$ throughout the training, with all other attention correlations remaining close to 0.*

b. *Local areas ($j > 1$) learn FP correlation in* **two-phase***: In phase one, FP correlation $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ between the local area and the global area feature quickly decreases to $-\Theta(\log(P))$ whereas all other attention correlations stay close to zero; In phase two, FP correlation $\Phi^{(t)}_{\mathbf{p}\to v_{k,j}}$ for the target local area starts to grow until convergence with all other attention correlations nearly unchanged.*

For negative information gap $\Delta \in [-0.5, -\Omega(1)]$:

c. *All areas learn FP correlation through* **one-phase:** *$\Phi^{(t)}_{\mathbf{p}\to v_{k,j}}$ monotonically increases to $O(\log(P)/\eta)$ throughout the training, with all other attention correlations remain close to 0.*

The training dynamics are different depending on whether Δ is positive or negative, and further vary for positive Δ depending on whether $X_{\mathbf{p}}$ is situated in global or local areas. Typically, the target FP correlations are learned directly in a single phase. However, for a positive information gap Δ, when patch $\mathbf{p}$ is located in a local area, the learning process contains an additional decoupling phase, to reduce the FP correlation with the non-target global features. Such behavior of the learning dynamics will be further explained in Section 5 through our main proof ideas.

5 Proof Sketch

The framework for constructing attention correlations, as outlined in Section 3, offers insights into how transformers may achieve specific localities during masked pretraining. Within this framework, we start this section with a warm-up discussion to further clarify why target FP correlations, rather than PP correlations, should be the focus of learning. We then offer an overview of our proof, which analyzes the dynamics of these attention correlations.

Warm-up intuition for attention correlations. Let us start by exploring the role of attention correlations in determining the area attention scores. Given a masked input $\mathsf{M}(X)$ with $\mathbf{p} \in \mathcal{P}_{k,j} \cap \mathcal{M}$, it holds that

$$\widetilde{\mathbf{Attn}}_{\mathbf{p}\to\mathcal{P}_{k,j}}(\mathsf{M}(X); Q) \propto \sum_{\mathbf{q}\in\mathcal{U}\cap\mathcal{P}_{k,j}} e^{\Phi_{\mathbf{p}\to v_{k,j}} + \Upsilon_{\mathbf{p}\to\mathbf{q}}} + \sum_{\mathbf{q}\in\mathcal{M}\cap\mathcal{P}_{k,j}} e^{\Upsilon_{\mathbf{p}\to\mathbf{q}}}.$$

The first term on the RHS is proportional to our unmasked area attention $\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,j}}$, which is balanced by the relative magnitudes between Feature-Position and Position-Position correlations. Upon initial observation, large FP correlation $\Phi_{\mathbf{p}\to v_{k,j}}$ or PP correlation $\Upsilon_{\mathbf{p}\to\mathbf{q}}$ for patch $\mathbf{q} \in \mathcal{P}_{k,j}$ may both contribute to the attention towards the area $\mathcal{U} \cap \mathcal{P}_{k,j}$. However, two key issues can arise from the PP correlation, which jointly lead to the suppression of PP correlation for reconstruction: *i).* such a mechanism inadvertently directs attention to the *masked* patches, which is not desirable; *ii).* such position association could be vulnerable to the variation across different clusters, i.e., $\mathbf{p}, \mathbf{q} \in \mathcal{P}_{k,j}$ does not necessarily hold for all $k \in [K]$. Consequently, as illustrated in Figure 4, to obtain a perfect reconstruction of the masked patch, learning a relatively large FP correlation $\Phi_{\mathbf{p}\to v_{k,j}}$ from the patches to the correct features is required.

Main idea of the proof. The main idea of our analysis is to track the dynamics of those attention correlations. We first provide the following lemma of GD updates for $\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}$ and $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}$.

Lemma 5.1 (Gradient of attention correlations, informal)**.** *Given $\mathbf{p}, \mathbf{q} \in \mathcal{P}$, let $\alpha^{(t)}_{\mathbf{p}\to v_{k,m}} = \frac{1}{\eta}\big(\Phi^{(t+1)}_{\mathbf{p}\to v_{k,m}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,m}}\big)$ with $k \in [K], m \in [N_k]$, and $\beta^{(t)}_{\mathbf{p}\to\mathbf{q}} = \frac{1}{\eta}\big(\Upsilon^{(t+1)}_{\mathbf{p}\to\mathbf{q}} - \Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}\big)$. We use $a_{k,\mathbf{p}}$ to indicate that the index of the area that patch $\mathbf{p}$ is located in the cluster $\mathcal{D}_k$.*

a. *For the same area, $\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}} \approx \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\right)^2$;*

b. *If $a_{k,\mathbf{p}} \neq 1$, for the global area, $\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} \approx -\big(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\big)^2$;*

c. *For other area $m \notin \{a_{k,\mathbf{p}}\} \cup \{1\}$, $|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}} + \mathbb{1}\{n\neq 1\}|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{N_k}\Big)$;*

d. *$\beta^{(t)}_{\mathbf{p}\to\mathbf{q}} = \sum_{k\in[K]} \beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$, where $|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{|\alpha^{(t)}_{k,\mathbf{p}\to a_{k,\mathbf{q}}}|}{C_{k,a_{k,\mathbf{q}}}}\Big)$.*

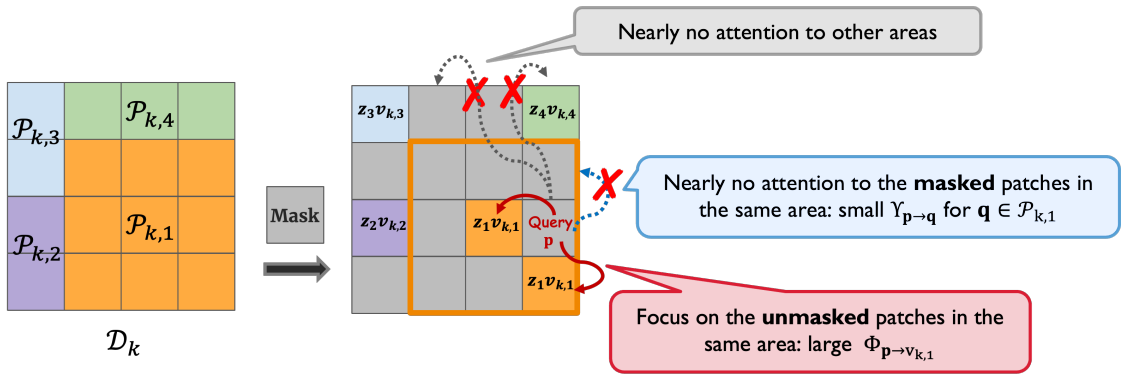

Figure 4: The mechanism of how the masked patch attends to other patches via attention correlations.

Here, $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ can be interpreted as the projected gradient for PP correlation $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}$ on the k-th cluster. Thus, Lemma 5.1.d. directly implies that the overall increment of PP correlation $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}$ should be negligible compared to the corresponding FP correlations $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{q}}}}$ on certain cluster $\mathcal{D}_k$ since $C_{k,a_{k,\mathbf{q}}} \gg 1$ and $K = \Theta(1)$, which implies that **all PP correlations are small.**

Moreover, from Lemma 5.1, it is observed that the target FP correlation $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ exhibits a monotonically increasing trend, while FP-correlations for feature $v_{k,1}$ will monotonically decrease if $a_{k,\mathbf{p}} \neq 1$. Moreover, FP gradients of the patch $\mathbf{p}$ to other unrelated local features are significantly smaller and become negligible since $N_k \gg 1$. Based on these observations, we now explain how the trend of FP-correlations determines the different learning behaviors varied across different locations $a_{k,\mathbf{p}}$ and information gap Δ.

- **For local areas with $a_{k,p} > 1$:** at the beginning of training, $\mathbf{attn}^{(0)}_{\mathbf{p}\to\mathbf{q}} = \frac{1}{P}$ for any $\mathbf{q} \in \mathcal{P}$ due to zero initialization. However, with high probability, the count of unmasked patches in the global area $\mathcal{P}_{k,1}$ significantly exceeds that in other areas.

  1. $\Delta = (1-\kappa_s) - 2(1-\kappa_c) \geq \Omega(1)$: $|\alpha^{(0)}_{\mathbf{p}\to v_{k,1}}| \approx \left(\mathbf{Attn}^{(0)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)^2 = \Omega(\frac{1}{P^{2(1-\kappa_c)}}) \gg \Theta(\frac{1}{P^{1-\kappa_s}}) = \mathbf{Attn}^{(0)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}} \approx \alpha^{(0)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$. Thus, $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ enjoys a much larger decreasing rate initially, which defines the *decoupling phase* of non-target global FP correlations. The significant decrease in $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ leads to a reduction in $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}$, which in turn makes a drop in $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|$. As the gradient decreases below $\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,pb}}}$, $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ within the target local area begins to have a larger gradient, which triggers the *growth of target FP correlation* phase, leading to continuous growth until convergence.
  2. $\Delta \leq -\Omega(1)$: the above relationship reverses, and $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ starts with a fast initial increase, which moves directly into the *growth* phase, eliminating the need to decouple $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$.

- **For global areas with $a_{k,\mathbf{p}} = 1$:** the global area is exactly the target area that FP correlations should be learned, and thus the second term in Lemma 5.1 will not exist in this case. As a result, the training process also skips the initial decoupling phase and starts with the *growth* phase for $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$. Such an effect is independent of the value of Δ.

6 Conclusion

In this work, we study the training process of masked vision pretraining with a one-layer softmax-based transformer. Our key contribution lies in showing that transformers trained with masked reconstruction objectives exhibit local and diverse patterns by learning correlations between patch features and patch positions. To our knowledge, our work is the first in analyzing softmax self-attention on both patch and position embedding simultaneously. Our proof techniques leveraged ideas from phase decomposition based on the interplay between feature-position and position-wise correlations, which do not need to disentangle

patches and positional encodings as in prior works. We anticipate that our theory can be useful for future studies of the spatial-or-temporal structures of state-of-the-art transformers and can promote theoretical studies relevant to deep learning practice.

References

[1] K. Ahn, X. Cheng, H. Daneshmand, and S. Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

[2] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[3] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021.

[4] S. Cao, P. Xu, and D. A. Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] S. Chen and Y. Li. Provably learning a multi-head attention layer. *arXiv preprint arXiv:2402.04084*, 2024.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] T. Chen, C. Luo, and L. Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

[9] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] B. L. Edelman, S. Goel, S. Kakade, and C. Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

[13] A. Ghiasi, H. Kazemi, E. Borgnia, S. Reich, M. Shu, M. Goldblum, A. G. Wilson, and T. Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022.

[14] A. Giannou, S. Rajput, J.-y. Sohn, K. Lee, J. D. Lee, and D. Papailiopoulos. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.

[15] E. Greene and J. A. Wellner. Exponential bounds for the hypergeometric distribution. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 23(3):1911, 2017.

[16] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[17] M. Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

[18] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011, 2021.

[19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[21] Y. Huang, Y. Cheng, and Y. Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

[22] S. Jelassi, M. Sander, and Y. Li. Vision transformers provably learn spatial structure. In *Advances in Neural Information Processing Systems*, volume 35, pages 37822–37836, 2022.

[23] J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 309–323, 2021.

[24] H. Li, M. Wang, S. Liu, and P.-Y. Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *arXiv preprint arXiv:2302.06015*, 2023.

[25] H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen. Training nonlinear transformers for efficient in-context learning: A theoretical learning and generalization analysis, 2024.

[26] H. Li, M. Wang, T. Ma, S. Liu, Z. Zhang, and P.-Y. Chen. What improves the generalization of graph transformer? a theoretical dive into self-attention and positional encoding. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*, 2023.

[27] B. Liu, D. J. Hsu, P. Ravikumar, and A. Risteski. Masked prediction: A parameter identifiability view. In *Advances in Neural Information Processing Systems*, pages 21241–21254, 2022.

[28] L. Melas-Kyriazi. Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet. *arXiv preprint arXiv:2105.02723*, 2021.

[29] E. Nichani, A. Damian, and J. D. Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

[30] J. Pan, P. Zhou, and Y. Shuicheng. Towards understanding why mask reconstruction pretraining helps in downstream tasks. In *The Eleventh International Conference on Learning Representations*, 2022.

[31] N. Park and S. Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[32] N. Park, W. Kim, B. Heo, T. Kim, and S. Yun. What do self-supervised vision transformers learn? In *The Eleventh International Conference on Learning Representations*, 2023.

[33] S. Paul and P.-Y. Chen. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022.

[34] A. Pokle, J. Tian, Y. Li, and A. Risteski. Contrasting the landscape of contrastive and non-contrastive learning. *arXiv preprint arXiv:2203.15702*, 2022.

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

[36] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

[37] J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[39] C. Sanford, D. Hsu, and M. Telgarsky. Transformers, parallel computation, and logarithmic depth. *arXiv preprint arXiv:2402.09268*, 2024.

[40] C. Sanford, D. J. Hsu, and M. Telgarsky. Representational strengths and limitations of transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] D. A. Tarzanagh, Y. Li, C. Thrampoulidis, and S. Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

[42] D. A. Tarzanagh, Y. Li, X. Zhang, and S. Oymak. Max-margin token selection in attention mechanism. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[43] Y. Tian, X. Chen, and S. Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

[44] Y. Tian, Y. Wang, B. Chen, and S. Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arXiv:2305.16380*, 2023.

[45] Y. Tian, Y. Wang, Z. Zhang, B. Chen, and S. S. Du. JoMA: Demystifying multilayer transformers via joint dynamics of MLP and attention. In *The Twelfth International Conference on Learning Representations*, 2024.

[46] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[47] A. Trockman and J. Z. Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

[48] B. Vasudeva, P. Deora, and C. Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *arXiv preprint arXiv:2402.05738*, 2024.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[50] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[51] J. Vuckovic, A. Baratin, and R. T. d. Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.

[52] X. Wang, X. Chen, S. S. Du, and Y. Tian. Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint arXiv:2110.04947*, 2021.

[53] C. Wei, Y. Chen, and T. Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022.

[54] C. Wei, S. M. Xie, and T. Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *Advances in Neural Information Processing Systems*, 34:16158–16170, 2021.

[55] Y. Wei, H. Hu, Z. Xie, Z. Zhang, Y. Cao, J. Bao, D. Chen, and B. Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

[56] G. Weiss, Y. Goldberg, and E. Yahav. Thinking like transformers. In *International Conference on Machine Learning*, pages 11080–11090. PMLR, 2021.

[57] Z. Wen and Y. Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.

[58] Z. Wen and Y. Li. The mechanism of prediction head in non-contrastive self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 24794–24809, 2022.

[59] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, and Y. Cao. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14475–14485, 2023.

[60] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[61] C. Yun, S. Bhojanapalli, A. S. Rawat, S. J. Reddi, and S. Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

[62] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[63] Q. Zhang, Y. Wang, and Y. Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022.

[64] R. Zhang, S. Frei, and P. L. Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

[65] Y. Zhang, F. Zhang, Z. Yang, and Z. Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023.

Appendix

A Additional Related Work

Empirical studies of transformers in vision. A number of works have aimed to understand the transformers in vision from different perspectives: comparison with CNNs [36, 13, 31], robustness [3, 33], and role of positional embeddings [28, 47]. Recent studies [59, 55, 32] have delved into ViTs with self-supervision to uncover the mechanisms at play, particularly through visualization and analysis of metrics related to self-attention. [59] compared the masked image modeling (MIM) method with supervised models, revealing MIM's capacity to enhance diversity and locality across all ViT layers, w which significantly boosts performance on tasks with weak semantics following fine-tuning. Building on MIM's advantages, [55] further proposed a simple feature distillation method that incorporates locality into various self-supervised methods, leading to an overall improvement in the finetuning performance. [32] conducted a detailed comparison between masked image modeling (MIM) and contrastive learning. They demonstrated that contrastive learning will make the self-attentions collapse into homogeneity for all query patches due to the nature of discriminative learning, while MIM leads to a diverse self-attention map since it focuses on local patterns.

Theory of self-supervised learning. A major line of theoretical studies falls into one of the most successful self-supervised learning approaches, contrastive learning [57, 37, 8, 2], and its variant non-contrastive self-supervised learning [58, 34, 52]. Some other works study the mask prediction approach [23, 54, 27], which is the focus of this paper. [23] provided statistical downstream guarantees for reconstructing missing patches. [54] studied the benefits of head and prompt tuning with masked pretraining under a Hidden Markov Model framework. [27] provided a parameter identifiability view to understand the benefit of masked prediction tasks, which linked the masked reconstruction tasks to the informativeness of the representation via identifiability techniques from tensor decomposition.

Theory of transformers and attention models. Prior work has studied the theoretical properties of transformers from various aspects: representational power [61, 12, 51, 53, 39], internal mechanism [41, 56], limitations [17, 40], and PAC learning [6]. Recently, there has been a growing body of research studying in-context learning with transformers due to the remarkable emergent in-context ability of large language models [65, 50, 14, 1, 64, 21, 29, 25]. Regarding the training dynamics of attention-based models, [24] studied the training process of shallow ViTs in a classification task. Subsequent research expanded on this by exploring the graph transformer with positional encoding [26] and in-context learning performance of transformers with nonlinear self-attention and nonlinear MLP [25]. However, all of these analyses rely crucially on stringent assumptions on the initialization of transformers and hardly generalize to our setting. [44] mathematically described how the attention map evolves trained by SGD for one-layer transformer but did not provide any convergence guarantee, and the follow-up work [45] considered a generalized case with multiple layers. [42, 48] investigated the implicit bias for self-attention models trained with GD. Furthermore, [21] proved the in-context convergence of a one-layer softmax transformer trained via GD and illustrated the attention dynamics throughout the training process. More recently, [29] studied GD dynamics on a simplified two-layer attention-only transformer and proved that it can encode the causal structure in the first attention layer. However, none of the previous studies analyzed the training of transformers under self-supervised learning, which is the focus of this paper.

B Experiments

Previous studies on the attention mechanisms of ViT-based pre-training approaches have mainly utilized a metric known as the attention distance [11]. Such a metric quantifies the average spatial distance between the query and key tokens, weighted by their self-attention coefficients. The general interpretation is that larger attention distances indicate global understanding, and smaller values suggest a focus on local features.

However, such a metric does not adequately determine if the self-attention mechanism is identifying a unique global pattern. A high attention distance could result from different patches focusing on varied distant areas, which does not necessarily imply that global information is being effectively synthesized. To address this limitation, we introduce a novel and revised version of average attention distance, called attention diversity metric, which is designed to assess whether various patches are concentrating on a similar region, thereby directly capturing global information.

Attention diversity metric, in distance. This metric is computed for self-attention with a single head of the specific layer. For a given image divided into $P \times P$ patches, the process unfolds as follows: for each patch, it is employed as the query patch to calculate the attention weights towards all P^2 patches, and those with the top-n attention weights are selected. Subsequently, the coordinates (e.g. (i, j) with $i, j \in [P]$) of these top-n patches are concatenated in sequence to form a $2 \times n$-dimensional vector. The final step computes the average distance between all these $2n$-dimensional vectors, i.e., $P^2 \times P^2$ vector pairs.

Setup. In this work, we compare the performance of ViT-B/16 encoder pre-trained on ImageNet-1K [38] among the following four models: masked reconstruction model (MAE), contrastive learning model (MoCo v3 [9]), other self-supervised model (DINO [5]), and supervised model (DeiT [46]). We focus on 12 different attention heads in the last layer of ViT-B on different pre-trained models. The box plot visualizes the distribution of the top-10 averaged attention focus across 152 example images, as similarly done in [11].

Implications. The experiment results based on our new metric are provided in Figure 2. Lower values of the attention diversity metric signify a focused attention on a coherent area across different patches, reflecting a global pattern of focus. On the other hand, higher values suggest that attention is dispersed, focusing on different, localized areas. It can be seen that the masked pretraining model is particularly effective in learning more diverse attention patterns, setting it apart from other models that prioritize a uniform global information with less attention diversity. This aligns with and provides further evidence for the findings in [32].

C Overview of the Proof Techniques

In this section, we explain our key proof techniques in analyzing the masked pretraining of transformers. We focus on the reconstruction of a specific patch $X_{\mathbf{p}}$ for $\mathbf{p} \in \mathcal{P}$. We aim to elucidate the training phases through which the model learns FP correlations related to the area associated with $\mathbf{p}$ across different clusters $k \in [K]$.

Our characterization of training phases differentiates between whether $X_{\mathbf{p}}$ is located in the global or local areas and further varies based on whether Δ is positive or negative. Specifically, for $\Delta \in [\Omega(1), 1]$, we observe distinct learning dynamics for FP correlations between local and global areas:

- Local area attends to FP correlation in two-phase: given $k \in [K]$, if $a_{k,\mathbf{p}} \neq 1$, then
  1. $\Phi^{(t)}_{\mathbf{p} \to v_{k,1}}$ first quickly decreases whereas all other $\Phi^{(t)}_{\mathbf{p} \to v_{k,m}}$ with $m \neq 1$ and $\Upsilon^{(t)}_{\mathbf{p} \to \mathbf{q}}$ do not change much;
  2. after some point, the increase of $\Phi^{(t)}_{\mathbf{p} \to v_{k,a_{k,\mathbf{p}}}}$ takes dominance. Such $\Phi^{(t)}_{\mathbf{p} \to v_{k,a_{k,\mathbf{p}}}}$ will keep growing until convergence with all other FP and PP attention correlations nearly unchanged.
- Global areas learn FP correlation in one-phase: given $k \in [K]$, if $a_{k,\mathbf{p}} = 1$, the update of $\Phi^{(t)}_{\mathbf{p} \to v_{k,1}}$ will dominate throughout the training, whereas all other $\Phi^{(t)}_{\mathbf{p} \to v_{k,m}}$ with $m \neq 1$ and learned PP correlations remain close to 0.

For $\Delta \in [-0.5, -\Omega(1)]$, the behaviors of learning FP correlations are uniform for all areas. Namely, all areas learn FP correlation through one-phase: given $k \in [K]$, throughout the training, the increase of $\Phi^{(t)}_{\mathbf{p} \to v_{k,a_{k,\mathbf{p}}}}$ dominates, whereas all other $\Phi^{(t)}_{\mathbf{p} \to v_{k,m}}$ with $m \neq a_{k,\mathbf{p}}$ and PP correlations $\Upsilon^{(t)}_{\mathbf{p} \to \mathbf{q}}$ remain close to 0.

For clarity, this section will mainly focus on the learning of *local* feature correlations with a positive information gap $\Delta \geq \Omega(1)$ in Appendices C.2 and C.3, which exhibits a two-phase process. The other scenarios will be discussed briefly in Appendix C.4.

C.1 GD Dynamics of Attention Correlations

Based on the crucial roles that attention correlations play in determining the reconstruction loss, the main idea of our analysis is to track the dynamics of those attention correlations. We first provide the following GD updates of $\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}$ and $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}$ (see Appendix D.1.1 for formal statements).

Lemma C.1 (FP correlations, informal)**.** *Given $k \in [K]$, for $\mathbf{p} \in \mathcal{P}$, denote $n = a_{k,\mathbf{p}}$, let $\alpha^{(t)}_{\mathbf{p}\to v_{k,m}} = \frac{1}{\eta}\big(\Phi^{(t+1)}_{\mathbf{p}\to v_{k,m}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,m}}\big)$ for $m \in [N_k]$, and suppose $X_{\mathbf{p}}$ is masked. Then*

1. *for the same area, $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \approx \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2$;*
2. *if $k \in \mathcal{B}_{\mathbf{p}}$, for the global area,*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} \approx -\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \cdot \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right) + \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\right);$$

3. *for other area $m \notin \{n\} \cup \{1\}$,*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,m}} \approx \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\left(\mathbb{1}\{n \neq 1\}\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)^2 - \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right).$$

From Lemma C.1, it is observed that for $\mathbf{p} \in \mathcal{P}_{k,n}$, the feature correlation $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}$ exhibits a monotonically increasing trend over time because $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \geq 0$. Furthermore, if $n > 1$, i.e., $\mathcal{P}_{k,n}$ is the local area, $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ will monotonically decrease.

Lemma C.2 (PP attention correlations, informal)**.** *Given $\mathbf{p}, \mathbf{q} \in \mathcal{P}$, let $\beta^{(t)}_{\mathbf{p}\to\mathbf{q}} = \frac{1}{\eta}\big(\Upsilon^{(t+1)}_{\mathbf{p}\to\mathbf{q}} - \Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}\big)$, and suppose $X_{\mathbf{p}}$ is masked. Then $\beta^{(t)}_{\mathbf{p}\to\mathbf{q}} = \sum_{k\in[N]}\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$, where $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ satisfies*

1. *if $a_{k,\mathbf{p}} = a_{k,\mathbf{q}} = n$, $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} \approx \mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}}\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2$;*
2. *if $k \in \mathcal{B}_{\mathbf{p}} \cap \mathcal{C}_{\mathbf{q}}$, where $a_{k,\mathbf{p}} = n > 1$ and $a_{k,\mathbf{q}} = 1$:*

$$\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} \approx -\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \cdot \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right) + \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\right);$$

3. *if $a_{k,\mathbf{q}} = m \notin \{n\} \cup \{1\}$, where $a_{k,\mathbf{p}} = n$,*

$$\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} \approx \mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \cdot \left(\mathbb{1}\{n \neq 1\}\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)^2 - \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right).$$

Based on the above gradient update for $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}$, we further introduce the following auxiliary quantity $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$, which can be interpreted as the PP attention correlation "projected" on the k-th cluster $\mathcal{D}_k$, and will be useful in the later proof.

$$\Upsilon^{(t+1)}_{k,\mathbf{p}\to\mathbf{q}} := \Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} + \eta\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}, \quad \text{with } \Upsilon^{(0)}_{k,\mathbf{p}\to\mathbf{q}} = 0. \tag{C.1}$$

We can directly verify that $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}} = \sum_{k\in[K]}\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$.

The key observation by comparing Lemma C.1 and C.2 is that the gradient of projected PP attention $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ is smaller than the corresponding FP gradient $\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{q}}}}$ in magnitude since $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \approx \frac{\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{q}}}}}{(1-\gamma)C_{k,a_{k,\mathbf{q}}}}$. We will show that the interplay between the increase of $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}$ and the decrease of $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ determines the learning behaviors for the local patch $\mathbf{p} \in \mathcal{P}_{k,n}$ with $n > 1$, and which effect will happen first depends on the initial attention, which is also determined by the value of information gap Δ.

C.2 Phase I: Decoupling the Global FP Correlations

We now explain how the attention correlations evolve at the initial phase of the training to decouple the correlations of the non-target global features when $\mathbf{p}$ is located in the local area for the k-th cluster. This phase can be further divided into the following two stages.

Stage 1. At the beginning of training, $\Phi^{(0)}_{\mathbf{p}\to v_{k,m}} = \Upsilon^{(0)}_{k,\mathbf{p}\to\mathbf{q}} = 0$, and hence $\mathbf{attn}^{(0)}_{\mathbf{p}\to\mathbf{q}} = \frac{1}{P}$ for any $\mathbf{q}\in\mathcal{P}$, which implies that the transformer equally attends to each patch. However, with high probability, the number of unmasked global features in the global area $\mathcal{P}_{k,1}$ is much larger than others. Hence, $\mathbf{Attn}^{(0)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = \frac{|\mathcal{U}\cap\mathcal{P}_{k,1}|}{P} \geq \Omega(\frac{1}{P^{1-\kappa_c}}) \gg \Theta(\frac{1}{P^{1-\kappa_s}}) = \mathbf{Attn}^{(0)}_{\mathbf{p}\to\mathcal{P}_{k,m}}$ for $m>1$. Therefore, by Lemma C.1 and C.2, we immediately obtain

- $\alpha^{(0)}_{\mathbf{p}\to v_{k,1}} = -\Theta\Big(\frac{1}{P^{2(1-\kappa_c)}}\Big)$, whereas $\alpha^{(0)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}} = \Theta\Big(\frac{1}{P^{(1-\kappa_s)}}\Big)$;
- all other FP correlation gradients $\alpha^{(0)}_{\mathbf{p}\to v_{k,m}}$ with $m\neq 1, a_{k,\mathbf{p}}$ are small;
- all projected PP correlation gradients $\beta^{(0)}_{k,\mathbf{p}\to\mathbf{q}}$ are small.

Since $\Delta = (1-\kappa_s) - 2(1-\kappa_c) \geq \Omega(1)$, it can be seen that $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ enjoys a much larger decreasing rate initially. This captures the decoupling process of the feature correlations with the global feature $v_{k,1}$ in the global area for $\mathbf{p}$. It can be shown that such an effect will dominate over a certain period that defines stage 1 of phase I. At the end of this stage, we will have $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \leq -\Omega\left(\log(P)\right)$, whereas all FP attention correlation $\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}$ with $m>1$ and all projected PP correlations $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ stay close to 0 (see Appendix F.1).

During stage 1, the significant decrease of the global FP correlation $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ leads to a reduction in the attention score $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}$. Meanwhile, attention scores $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}$ (where $m>1$) for other patches remain consistent, reflecting a uniform distribution over unmasked patches within each area. By the end of stage 1, $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}$ drops to a certain level, resulting in a decrease in $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|$ as it approaches $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}$, which indicates that stage 2 begins.

Stage 2. Soon as stage 2 begins, the dominant effect switches as $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|$ reaches the same order of magnitude as $\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$. The following result shows that $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ must update during stage 2.

Lemma C.3 (Switching of dominant effects (See Appendix F.2))**.** *Under the same conditions as Theorem 4.1, for $\mathbf{p}\in\mathcal{P}$, there exists $\widetilde{T}_1$, such that at iteration $t=\widetilde{T}_1+1$, we have*

a. $\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}} \geq \Omega\left(\log(P)\right)$*, and* $\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,1}} = -\Theta(\log(P))$*;*

b. all other FP correlations $\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}$ with $m\neq 1, a_{k,\mathbf{p}}$ are small;

c. all projected PP correlations $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ are small.

Intuition of the transition. Once $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ decreases to $-\frac{\Delta}{2L}\log(P)$, we observe that $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|$ is approximately equal to $\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$. After this point, reducing $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ further is more challenging compared to the increase in $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$. To illustrate, a minimal decrease of $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ by an amount of $\frac{0.001}{L}\log(P)$ will yield $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \leq O(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{0.002}})$. Such a discrepancy triggers the switch of the dominant effect.

C.3 Phase II: Growth of Target Local FP Correlation

Moving beyond phase I, FP correlation $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ within the target local area $\mathbf{p}$ already enjoys a larger gradient $\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ than other $\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}$ with $m\neq a_{k,\mathbf{p}}$ and all projected PP correlations $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$. We can show that the growth of $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ will continue to dominate until the end of training by recognizing the following two stages.

Rapid growth stage. At the beginning of phase II, $\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ is mainly driven by $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}$ since $1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}$ remains at the constant order. Therefore, the growth of $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ naturally results in a boost in $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}$, thereby promoting an increase in its own gradient $\alpha^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$, which defines the rapid growth stage. On the other hand, we can prove that the following gap holds for FP and projected PP correlation gradients (see Appendix F.3):

- all other FP correlation gradients $\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}$ with $m \neq a_{k,\mathbf{p}}$ are small;
- all projected PP correlation gradients $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ are small.

Convergence stage. After the rapid growth stage, the desired local pattern with a high target feature-position correlation $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ is learned. In this last stage, it is demonstrated that the above conditions for non-target FP and projected PP correlations remain valid, while the growth of $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ starts to decelerate as $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ reaches $\Theta(\log(P))$, resulting in $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \approx \Omega(1)$, which leads to convergence (see Appendix F.4).

C.4 Learning Processes in Other Scenarios

In this section, we talk about the learning process in other settings, including learning FP correlations for the local area when the information gap is negative, learning FP correlations for the global area, and failure to learn PP correlations.

What is the role of positive information gap? As described in stage 1 of phase 1 in Appendix C.2, the decoupling effect happens at the beginning of the training because $\alpha^{(0)}_{\mathbf{p}\to v_{k,1}} \gg \alpha^{(0)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ attributed to $\Delta \geq \Omega(1)$. However, in cases where $\Delta \leq -\Omega(1)$, this relationship reverses, with $\alpha^{(0)}_{\mathbf{p}\to v_{k,1}}$ becoming significantly smaller than $\alpha^{(0)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$. Similarly, other FP gradients $\alpha^{(0)}_{\mathbf{p}\to v_{k,m}}$ with $m \neq 1, a_{k,\mathbf{p}}$ and all the projected gradients of PP correlation $\beta^{(0)}_{\mathbf{p}\to\mathbf{q}}$ are small in magnitude. Consequently, $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ starts with a larger gradient, eliminating the need to decouple FP correlations for the global area. As a result, training skips the initial phase, and moves directly into Phase II, during which $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ continues to increase until it converges (see Appendix G).

Learning FP correlations for the global area. When the patch $X_{\mathbf{p}}$ is located in the global area of cluster k, i.e., $a_{k,\mathbf{p}} = 1$, the attention score $\mathbf{Attn}^{(0)}_{\mathbf{p}\to\mathcal{P}_{k,1}}$ directed towards the target area $\mathcal{P}_{k,1}$ is initially higher compared to other attention scores due to the presence of a significant number of unmasked patches in the global area. This leads to an initially larger gradient $\alpha^{(0)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$. Such an effect is independent of the value of Δ. As a result, the training process skips the initial phase, which is typically necessary for the cases where $a_{k,\mathbf{p}} > 1$ with a positive information gap, and moves directly into Phase II (see Appendix H).

All PP correlations are small. Integrating the analysis from all previous discussions, we establish that for every cluster $k \in [K]$, regardless of its association with $\mathcal{C}_{\mathbf{p}}$ (global area) or $\mathcal{B}_{\mathbf{p}}$ (local area), and for any patch $X_{\mathbf{q}}$ with $\mathbf{q} \in \mathcal{P}$, the projected PP correlation $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ remains nearly zero in comparison to the significant changes observed in the FP correlation, because the gradient $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ is relatively negligible. Therefore, the overall PP correlation $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}} = \sum_{k=1}^{K} \Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ also stays close to zero, given that the number of clusters $K = \Theta(1)$.

D Preliminaries

In this section, we will introduce warm-up gradient computations and probabilistic lemmas that establish essential properties of the data and the loss function, which are pivotal for the technical proofs in the upcoming

sections. Throughout the appendix, we assume $N_k = N$ and $C_{k,n} = C_n$ for all $k \in [K]$ for simplicity. We will also omit the explicit dependence on X for $z_n(X)$. We use $k_X \in [K]$ to denote the cluster index that a given image X is drawn from.

D.1 Gradient Computations

We first calculate the gradient with respect to Q. We omit the superscript '(t)' and write $\mathcal{L}(Q)$ as $\mathcal{L}$ here for simplicity.

Lemma D.1. *The gradient of the loss function with respect to Q is given by*

$$\frac{\partial \mathcal{L}}{\partial Q} = -\mathbb{E}\left[\sum_{\mathbf{p}\in\mathcal{M}}\sum_{\mathbf{q}} \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} \mathsf{M}(X)_{\mathbf{q}}^\top (X_{\mathbf{p}} - [F(\mathsf{M}(X);Q)]_{\mathbf{p}}) \cdot \right.$$
$$\left. \widetilde{\mathsf{M}}(X)_{\mathbf{p}} \left(\widetilde{\mathsf{M}}(X)_{\mathbf{q}} - \sum_{\mathbf{r}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} \widetilde{\mathsf{M}}(X)_{\mathbf{r}}\right)^\top\right].$$

Proof. We begin with the chain rule and obtain

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial Q} &= \mathbb{E}[\sum_{\mathbf{p}\in\mathcal{M}} \frac{\partial [F(\mathsf{M}(X);Q)]_{\mathbf{p}}}{\partial Q}([F(\mathsf{M}(X);Q)]_{\mathbf{p}} - X_{\mathbf{p}})] \\ &= \mathbb{E}[\sum_{\mathbf{p}\in\mathcal{M}}\sum_{\mathbf{q}} \frac{\partial \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}}{\partial Q} \mathsf{M}(X)_{\mathbf{q}}^\top ([F(\mathsf{M}(X);Q)]_{\mathbf{p}} - X_{\mathbf{p}})]. \end{aligned} \tag{D.1}$$

We focus on the gradient for each attention score:

$$\begin{aligned}\frac{\partial \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}}{\partial Q} &= \sum_{\mathbf{r}} \frac{\exp\left(\widetilde{\mathsf{M}}(X)_{\mathbf{p}}^\top Q(\widetilde{\mathsf{M}}(X)_{\mathbf{r}} + \widetilde{\mathsf{M}}(X)_{\mathbf{q}})\right)}{\left(\sum_{\mathbf{r}} \exp(\widetilde{\mathsf{M}}(X)_{\mathbf{p}}^\top Q \widetilde{\mathsf{M}}(X)_{\mathbf{r}})\right)^2} \widetilde{\mathsf{M}}(X)_{\mathbf{p}} (\widetilde{\mathsf{M}}(X)_{\mathbf{q}} - \widetilde{\mathsf{M}}(X)_{\mathbf{r}})^\top \\ &= \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} \sum_{\mathbf{r}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} \widetilde{\mathsf{M}}(X)_{\mathbf{p}} (\widetilde{\mathsf{M}}(X)_{\mathbf{q}} - \widetilde{\mathsf{M}}(X)_{\mathbf{r}})^\top \\ &= \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} \widetilde{\mathsf{M}}(X)_{\mathbf{p}} \cdot \left[\widetilde{\mathsf{M}}(X)_{\mathbf{q}} - \sum_{\mathbf{r}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} \widetilde{\mathsf{M}}(X)_{\mathbf{r}}\right]^\top. \end{aligned}$$

Substituting the above equation into (D.1), we complete the proof. □

Recall that the quantities $\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}$ and $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}$ are defined in Definition 3.1. These quantities are associated with the attention weights for each token, and they play a crucial role in our analysis of learning dynamics. We will restate their definitions here for clarity.

Definition D.2. (Attention correlations) Given $\mathbf{p}, \mathbf{q} \in \mathcal{P}$, for $t \geq 0$, we define two types of attention correlations as follows:

1. Feature Attention Correlation: $\Phi^{(t)}_{\mathbf{p}\to v_{k,m}} \coloneqq e_{\mathbf{p}}^\top Q^{(t)} v_{k,m}$ for $k \in [K]$ and $m \in [N]$;
2. Positional Attention Correlation: $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}} \coloneqq e_{\mathbf{p}}^\top Q^{(t)} e_{\mathbf{q}}$.

By our initialization, we have $\Phi^{(0)}_{\mathbf{p}\to v_{k,m}} = \Upsilon^{(0)}_{\mathbf{p}\to\mathbf{q}} = 0$.

Next, we will apply the expression in Lemma D.1 to compute the gradient dynamics of these attention correlations.

D.1.1 Formal Statements and Proof of Lemma C.1 and C.2

We first introduce some notations. Given $\mathbf{r} \in \mathcal{U}$, for $\mathbf{p} \in \mathcal{P}$, $k \in [K]$ and $n \in [N]$ define the following quantities:

$$
\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &\coloneqq \mathsf{M}(X)_{\mathbf{r}}^{\top}(X_{\mathbf{p}} - [F(\mathsf{M}(X); Q)]_{\mathbf{p}}) \\
I_{\mathbf{r}}^{\mathbf{p},k,n} &\coloneqq \left(\widetilde{\mathsf{M}}(X)_{\mathbf{r}} - \sum_{\mathbf{w}\in\mathcal{P}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} \widetilde{\mathsf{M}}(X)_{\mathbf{w}}\right)^{\top} v_{k,n} \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &\coloneqq \left(\widetilde{\mathsf{M}}(X)_{\mathbf{r}} - \sum_{\mathbf{w}\in\mathcal{P}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} \widetilde{\mathsf{M}}(X)_{\mathbf{w}}\right)^{\top} e_{\mathbf{q}}
\end{aligned}
$$

Lemma D.3 (Formal statement of Lemma C.1)**.** *Given $k \in [K]$, for $\mathbf{p} \in \mathcal{P}$, denote $n = a_{k,\mathbf{p}}$, let $\alpha_{\mathbf{p}\to v_{k,m}}^{(t)} = \frac{1}{\eta}\left(\Phi_{\mathbf{p}\to v_{k,m}}^{(t+1)} - \Phi_{\mathbf{p}\to v_{k,m}}^{(t)}\right)$ for $m \in [N_k]$, then*

a. for $m = n$,

$$
\begin{aligned}
\alpha_{\mathbf{p}\to v_{k,n}}^{(t)} = \mathbb{E}\Bigg[&\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}^{(t)} \cdot \\
&\left(z_n^3 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}^{(t)}\right)^2 + \sum_{a\neq n} z_a^2 z_n \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a}}^{(t)}\right)^2\right)\Bigg];
\end{aligned}
$$

b. for $m \neq n$,

$$
\begin{aligned}
\alpha_{\mathbf{p}\to v_{k,m}}^{(t)} = \mathbb{E}\Bigg[&\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}^{(t)} \cdot \Bigg(\sum_{a\neq m,n} z_a^2 z_m \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a}}^{(t)}\right)^2 - \\
&\left(z_m z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}^{(t)}\right) \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}^{(t)} + z_m^3 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}^{(t)}\right) \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}^{(t)}\right)\Bigg)\Bigg].
\end{aligned}
$$

Proof. From Lemma D.1, we have

$$
\begin{aligned}
\alpha_{\mathbf{p}\to v_{k,m}}^{(t)} &= e_{\mathbf{p}}^{\top}(-\frac{\partial \mathcal{L}}{\partial Q}) v_{k,m} \\
&= \mathbb{E}[\mathbb{1}\{\mathbf{p} \in \mathcal{M}\} \sum_{\mathbf{r}\in\mathcal{U}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} J_{\mathbf{r}}^{\mathbf{p}} \cdot I_{\mathbf{r}}^{\mathbf{p},k,m}] \\
&= \mathbb{E}[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \sum_{\mathbf{r}\in\mathcal{U}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} J_{\mathbf{r}}^{\mathbf{p}} \cdot I_{\mathbf{r}}^{\mathbf{p},k,m}]
\end{aligned}
$$

where the last equality holds since when $k_X \neq k$, $I_{\mathbf{r}}^{\mathbf{p},k,m} = 0$ due to orthogonality. Thus, in the following, we only need to consider the case $k_X = k$.

Case 1: $m = n$**.**

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,n}$, since $v_{k,n'} \perp v_{k,n}$ for $n' \neq n$, and $v_{k,n} \perp \{e_{\mathbf{q}}\}_{\mathbf{q}\in\mathcal{P}}$ we have

$$
\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_n v_{k,n}^{\top} \left(z_n v_{k,n} - \sum_{\mathbf{q}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} z_n v_{k,n}\right) \\
&= z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \\
I_{\mathbf{r}}^{\mathbf{p},k,n} &= (z_n v_{k,n} - \sum_{\mathbf{q}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} z_n v_{k,n})^{\top} v_{k,n} = J_{\mathbf{r}}^{\mathbf{p}}/z_n
\end{aligned}
$$

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,n'}$ with $n' \neq n$

$$
\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_{n'} v_{k,n'}^{\top} \left(z_n v_{k,n} - \sum_{\mathbf{q} \in \mathcal{U} \cap \mathcal{P}_{k,n'}} \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} z_{n'} v_{k,n'} \right) \\
&= -z_{n'}^2 \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n'}} \\
I_{\mathbf{r}}^{\mathbf{p},k,n} &= \left(z_{n'} v_{k,n'} - \sum_{\mathbf{q} \in \mathcal{U} \cap \mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} z_n v_{k,n} \right)^{\top} v_{k,n} \\
&= -z_n \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}}
\end{aligned}
$$

Putting it together, then we obtain:

$$
\begin{aligned}
e_{\mathbf{p}}^{\top} (-\frac{\partial L}{\partial Q}) v_{k,n} = \mathbb{E} \Big[&\mathbb{1}\{\{\mathbf{p} \in \mathcal{M}, k_X = k\}\} \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}}^{(t)} \cdot \\
&\left(z_n^3 \left(1 - \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}}^{(t)} \right)^2 + \sum_{a \neq n} z_a^2 z_n \left(\mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,a}}^{(t)} \right)^2 \right) \Big]
\end{aligned}
$$

Case 2: $m \neq n$**.** Similarly

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,n}$

$$
\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_n v_{k,n}^{\top} \left(z_n v_{k,n} - \sum_{\mathbf{q} \in \mathcal{U} \cap \mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} z_n v_{k,n} \right) \\
&= z_n^2 (1 - \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}}) \\
I_{\mathbf{r}}^{\mathbf{p},k,m} &= \left(z_n v_{k,n} - \sum_{\mathbf{q} \in \mathcal{U} \cap \mathcal{P}_{k,m}} \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} z_m v_{k,m} \right)^{\top} v_{k,m} \\
&= -z_m \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,m}}
\end{aligned}
$$

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,m}$

$$
\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_m v_{k,m}^{\top} \left(z_n v_{k,n} - \sum_{\mathbf{q} \in \mathcal{U} \cap \mathcal{P}_{k,m}} \mathbf{attn}_{\mathbf{p} \to \mathbf{q}}^{(t)} z_m v_{k,m} \right) \\
&= -z_m^2 \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,m}} \\
I_{\mathbf{r}}^{\mathbf{p},k,n} &= \left(z_m v_{k,m} - \sum_{\mathbf{q} \in \mathcal{U} \cap \mathcal{P}_{k,m}} \mathbf{attn}_{\mathbf{p} \to \mathbf{q}}^{(t)} z_m v_{k,m} \right)^{\top} v_{k,m} \\
&= z_n (1 - \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,m}})
\end{aligned}
$$

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,a}$, $a \neq n, m$

$$
J_{\mathbf{r}}^{\mathbf{p}} = z_a v_{k,a}^{\top} \left(z_n v_{k,n} - \sum_{\mathbf{q} \in \mathcal{U} \cap \mathcal{P}_{k,a}} \mathbf{attn}_{\mathbf{p} \to \mathbf{q}}^{(t)} z_a v_{k,a} \right)
$$

$$
\begin{aligned}
&= -z_a^2 \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a}} \\
I_{\mathbf{r}}^{\mathbf{p},k,n} &= \left(z_a v_{k,a} - \sum_{\mathbf{q}\in\mathcal{U}\cap\mathcal{P}_{k,m}} \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}^{(t)} z_m v_{k,m} \right)^\top v_{k,m} \\
&= -z_m \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}
\end{aligned}
$$

Putting them together, then we complete the proof. □

Lemma D.4 (Formal statement of Lemma C.2)**.** *Given* $\mathbf{p},\mathbf{q}\in\mathcal{P}$*, let* $\beta_{\mathbf{p}\to\mathbf{q}}^{(t)} = \frac{1}{\eta}\big(\Upsilon_{\mathbf{p}\to\mathbf{q})}^{(t+1)} - \Upsilon_{\mathbf{p}\to\mathbf{q})}^{(t)}\big)$*, then*

$$
\beta_{\mathbf{p}\to\mathbf{q}}^{(t)} = \sum_{k\in[K]} \beta_{k,\mathbf{p}\to\mathbf{q}}^{(t)}, \qquad \textit{where } \beta_{k,\mathbf{p}\to\mathbf{q}}^{(t)} \textit{ satisfies}
$$

a. if $a_{k,\mathbf{p}} = a_{k,\mathbf{q}} = n$*,*

$$
\begin{aligned}
\beta_{k,\mathbf{p}\to\mathbf{q}}^{(t)} = \mathbb{E}\Bigg[&\mathbb{1}\{\mathbf{p}\in\mathcal{M}, k_X = k\}\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}^{(t)} \cdot \Bigg(\sum_{a\neq n} z_a^2 \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a}}^{(t)}\right)^2 + \\
&z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}^{(t)}\right)\left(\mathbb{1}\{\mathbf{q}\in\mathcal{U}\} - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}^{(t)}\right)\Bigg)\Bigg];
\end{aligned}
$$

b. for $a_{k,\mathbf{p}} = n \neq m = a_{k,\mathbf{q}}$*,*

$$
\begin{aligned}
\beta_{k,\mathbf{p}\to\mathbf{q}}^{(t)} = \mathbb{E}\Bigg[&\mathbb{1}\{\mathbf{p}\in\mathcal{M},, k_X = k\}\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}^{(t)} \cdot \Bigg(\sum_{a\neq n} z_a^2 \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a}}^{(t)}\right)^2 - \\
&\left(z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}^{(t)}\right)\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}^{(t)} + \mathbb{1}\{\mathbf{q}\in\mathcal{U}\} z_m^2 \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}^{(t)}\right)\Bigg)\Bigg].
\end{aligned}
$$

Proof.

$$
\beta_{\mathbf{p}\to\mathbf{q}}^{(t)} = e_{\mathbf{p}}^\top (-\frac{\partial\mathcal{L}}{\partial Q}) e_{\mathbf{q}} = \mathbb{E}[\mathbb{1}\{\mathbf{p}\in\mathcal{M}\}\sum_{\mathbf{r}\in\mathcal{U}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}}^{(t)} J_{\mathbf{r}}^{\mathbf{p}} K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}}]
$$

Then we let

$$
\beta_{k,\mathbf{p}\to\mathbf{q}}^{(t)} := \mathbb{E}[\mathbb{1}\{\mathbf{p}\in\mathcal{M}, k_X = k\}\sum_{\mathbf{r}\in\mathcal{U}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}}^{(t)} J_{\mathbf{r}}^{\mathbf{p}} K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}}].
$$

In the following, we denote $a_{k,\mathbf{p}} = n$ and $a_{k,\mathbf{q}} = m$ for simplicity.

Case 1: $m = n$**.** If $\mathbf{q}\in\mathcal{U}\cap\mathcal{P}_{k,n}$:

- For $\mathbf{r} = \mathbf{q}$

$$
\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_n v_{k,n}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} z_n v_{k,n} \right) \\
&= z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{q}} - (\mathbf{attn}_{\mathbf{p}\to\mathbf{q}} e_{\mathbf{q}} + \sum_{\mathbf{w}\neq\mathbf{q}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} e_{\mathbf{w}}))^\top e_{\mathbf{q}}
\end{aligned}
$$

$$= 1 - \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}.$$

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,n}$, and $\mathbf{r} \neq \mathbf{q}$

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_n v_{k,n}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} z_n v_{k,n} \right) \\
&= z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{r}} - (\mathbf{attn}_{\mathbf{p}\to\mathbf{q}} e_{\mathbf{q}} + \sum_{\mathbf{w}\neq\mathbf{q}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} e_{\mathbf{w}}))^\top e_{\mathbf{q}} \\
&= -\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}
\end{aligned}$$

  Thus

$$\begin{aligned}
&\sum_{\mathbf{r}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} J_{\mathbf{r}}^{\mathbf{p}} \cdot K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} \\
&= z_n^2 \left(1 - \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}}\right) \\
&\cdot \left(- \sum_{\mathbf{r}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} + \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} \right) \\
&= z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}^{(t)}
\end{aligned}$$

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,a}$, $a \neq n$

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_a v_{k,a}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,a}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} z_a v_{k,a} \right) \\
&= -z_a^2 \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,a}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{r}} - (\mathbf{attn}_{\mathbf{p}\to\mathbf{q}} e_{\mathbf{q}} + \sum_{\mathbf{w}\neq\mathbf{q}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} e_{\mathbf{w}}))^\top e_{\mathbf{q}} \\
&= -\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}
\end{aligned}$$

Thus

$$\sum_{\mathbf{r}\in\mathcal{U}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} J_{\mathbf{r}}^{\mathbf{p}} K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} = \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} \cdot \left(z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{a\neq n} z_a^2 \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2 \right)$$

If $\mathbf{q} \in \mathcal{M} \cap \mathcal{P}_{k,n}$:

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,n}$,

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_n v_{k,n}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} z_n v_{k,n} \right) \\
&= z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{r}} - (\mathbf{attn}_{\mathbf{p}\to\mathbf{q}} e_{\mathbf{q}} + \sum_{\mathbf{w}\neq\mathbf{q}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} e_{\mathbf{w}}))^\top e_{\mathbf{q}}
\end{aligned}$$

$$= -\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}$$

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,a}$, $a \neq n$

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_a v_{k,a}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,a}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} z_a v_{k,a} \right) \\
&= -z_a^2 \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,a}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{r}} - (\mathbf{attn}_{\mathbf{p}\to\mathbf{q}} e_{\mathbf{q}} + \sum_{\mathbf{w}\neq\mathbf{q}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} e_{\mathbf{w}}))^\top e_{\mathbf{q}} \\
&= -\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}
\end{aligned}$$

Thus

$$\begin{aligned}
&\sum_{\mathbf{r}\in\mathcal{U}} \mathbf{attn}_{\mathbf{p}\to\mathbf{r}} J_{\mathbf{r}}^{\mathbf{p}} K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} \\
&= \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} \cdot \left(z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 - z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) + \sum_{a\neq n} z_a^2 \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2 \right)
\end{aligned}$$

Putting it together,

$$\begin{aligned}
&\beta_{k,\mathbf{p}\to\mathbf{q}}^{(t)} = \mathbb{E}\left[\mathbb{1}\{\mathbf{p}\in\mathcal{M}, k_X = k\} \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} \cdot \right. \\
&\left. \left(-z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \mathbb{1}\{\mathbf{q}\in\mathcal{M}\} + z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2 \right) \right]
\end{aligned}$$

Case 2: $m \neq n$**.** Similarly, if $\mathbf{q} \in \mathcal{U} \cap \mathcal{P}_{k,m}$:

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,n}$,

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_n v_{k,n}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} z_n v_{k,n} \right) \\
&= z_n^2 (1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}) \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{r}} - \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} e_{\mathbf{q}} - \sum_{\mathbf{w}\neq\mathbf{q}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} e_{\mathbf{w}})^\top e_{\mathbf{q}} \\
&= -\mathbf{attn}_{\mathbf{p}\to\mathbf{q}}
\end{aligned}$$

- For $\mathbf{r} = \mathbf{q}$

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_m v_{k,m}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w}\in\mathcal{U}\cap\mathcal{P}_{k,m}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} z_m v_{k,m} \right) \\
&= -z_m^2 \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}} \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{q}} - \mathbf{attn}_{\mathbf{p}\to\mathbf{q}} e_{\mathbf{q}} - \sum_{\mathbf{w}\neq\mathbf{w}} \mathbf{attn}_{\mathbf{p}\to\mathbf{w}} e_{\mathbf{w}})^\top e_{\mathbf{q}} \\
&= 1 - \mathbf{attn}_{\mathbf{p}\to\mathbf{q}}
\end{aligned}$$

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,a}$, $a \neq n$, and $\mathbf{r} \neq \mathbf{q}$

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_a v_{k,a}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w} \in \mathcal{U} \cap \mathcal{P}_{k,a}} \mathbf{attn}_{\mathbf{p} \to \mathbf{w}} z_a v_{k,a} \right) \\
&= -z_a^2 \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,a}} \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{r}} - \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} e_{\mathbf{q}} - \sum_{\mathbf{w} \neq \mathbf{q}} \mathbf{attn}_{\mathbf{p} \to \mathbf{w}} e_{\mathbf{w}})^\top e_{\mathbf{q}} \\
&= -\mathbf{attn}_{\mathbf{p} \to \mathbf{q}}
\end{aligned}$$

Thus

$$\begin{aligned}
&\sum_{\mathbf{r} \in \mathcal{U}} \mathbf{attn}_{\mathbf{p} \to \mathbf{r}} J_{\mathbf{r}}^{\mathbf{p}} K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} \\
&= \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} \cdot \left(-z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}}\right) \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}} - z_m^2 \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,m}} + \sum_{a \neq n} z_a^2 \left(\mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,a}}\right)^2 \right)
\end{aligned}$$

If $\mathbf{q} \in \mathcal{M} \cap \mathcal{P}_{k,m}$:

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,n}$,

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_n v_{k,n}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w} \in \mathcal{U} \cap \mathcal{P}_{k,n}} \mathbf{attn}_{\mathbf{p} \to \mathbf{w}} z_n v_{k,n} \right) \\
&= z_n^2 (1 - \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}}) \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{r}} - \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} e_{\mathbf{q}} - \sum_{\mathbf{w} \neq \mathbf{q}} \mathbf{attn}_{\mathbf{p} \to \mathbf{w}} e_{\mathbf{w}})^\top e_{\mathbf{q}} \\
&= -\mathbf{attn}_{\mathbf{p} \to \mathbf{q}}
\end{aligned}$$

- For $\mathbf{r} \in \mathcal{U} \cap \mathcal{P}_{k,a}$, $a \neq n$

$$\begin{aligned}
J_{\mathbf{r}}^{\mathbf{p}} &= z_a v_{k,a}^\top \left(z_n v_{k,n} - \sum_{\mathbf{w} \in \mathcal{U} \cap \mathcal{P}_{k,a}} \mathbf{attn}_{\mathbf{p} \to \mathbf{w}} z_a v_{k,a} \right) \\
&= -z_a^2 \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,a}} \\
K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} &= (e_{\mathbf{r}} - \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} e_{\mathbf{q}} - \sum_{\mathbf{w} \neq \mathbf{q}} \mathbf{attn}_{\mathbf{p} \to \mathbf{w}} e_{\mathbf{w}})^\top e_{\mathbf{q}} \\
&= -\mathbf{attn}_{\mathbf{p} \to \mathbf{q}}
\end{aligned}$$

Thus

$$\begin{aligned}
&\sum_{\mathbf{r} \in \mathcal{U}} \mathbf{attn}_{\mathbf{p} \to \mathbf{r}} J_{\mathbf{r}}^{\mathbf{p}} K_{\mathbf{r}}^{\mathbf{p},\mathbf{q}} \\
&= \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} \cdot \left(-z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}}\right) \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}} + \sum_{a \neq n} z_a^2 \left(\mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,a}}\right)^2 \right).
\end{aligned}$$

Therefore

$$\begin{aligned}
\beta_{k,\mathbf{p} \to \mathbf{q}}^{(t)} = \mathbb{E}\,[&\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \mathbf{attn}_{\mathbf{p} \to \mathbf{q}} \cdot \\
&\left(-z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}}\right) \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,n}} - \mathbb{1}\{\mathbf{q} \in \mathcal{U}\} z_m^2 \mathbf{Attn}_{\mathbf{p} \to \mathcal{P}_{k,m}}\right.
\end{aligned}$$

$$+\sum_{a\neq n} z_a^2 \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\Bigg)\Bigg].$$

□

Based on the above gradient update for $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}$, we further introduce the following auxiliary quantity, which will be useful in the later proof.

$$\Upsilon^{(t+1)}_{k,\mathbf{p}\to\mathbf{q}} := \Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} + \eta\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}, \quad \text{with } \Upsilon^{(0)}_{k,\mathbf{p}\to\mathbf{q}} = 0 \tag{D.2}$$

It is easy to verify that $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}} = \sum_{k\in[K]} \Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$.

D.2 High-probability Event

We first introduce the following exponential bounds for the hypergeometric distribution Hyper (m, D, M). Hyper (m, D, M) describes the probability of certain successes (random draws for which the object drawn has a specified feature) in m draws, without replacement, from a finite population of size M that contains exactly D objects with that feature, wherein each draw is either a success or a failure.

Proposition D.5 ([15])**.** *Suppose $S \sim$ Hyper (m, D, M) with $1 \le m, D \le M$. Define $\mu_M := D/M$. Then for all $t > 0$*

$$P\left(|S - m\mu_M| > t\right) \le 2\exp\left(-\frac{t^2}{4m\mu_M + 2t}\right).$$

We then utilize this property to prove the high-probability set introduced in Appendix C.1.

Lemma D.6. *For $k \in [K]$ $n \in [N]$, define*

$$\mathcal{E}_{k,n}(\gamma, P) := \{\mathsf{M} : |\mathcal{P}_{k,n} \cap \mathcal{U}| = \Theta(C_n)\}, \tag{D.3}$$

we have

$$\mathbb{P}(\mathsf{M} \in \mathcal{E}_{k,n}) \ge 1 - 2\exp(-c_{n,1}C_n) \tag{D.4}$$

where $c_{n,0} > 0$ is some constant.

Proof. Under the random masking strategy, given $k \in [K]$ and $n \in [N]$, $Y_{k,n} = |\mathcal{U} \cap \mathcal{P}_{k,n}|$ follows the hypergeometric distribution, i.e. $Y_{k,n} \sim \text{Hyper}((1-\gamma)P, C_n, P)$. Then by tail bounds, for $t > 0$, we have:

$$\mathbb{P}[|Y_{k,n} - (1-\gamma)C_n| > t] \le 2\exp(-\frac{t^2}{4(1-\gamma)C_n + 2t})$$

Letting $t = \Theta(C_n)$, we have

$$\mathbb{P}[Y_{k,n} = \Theta(C_n)] \ge 1 - 2e^{-c_{n,1}C_n}.$$

□

We further have the following fact, which will be useful for proving the property of loss objective in the next subsection.

Lemma D.7. *For $k \in [K]$ and $n \in [N]$, we have*

$$\mathbb{P}(|\mathcal{U} \cap \mathcal{P}_{k,n}| = 0) \le \exp(-c_{n,0}C_n). \tag{D.5}$$

where $c_{n,0} > 0$ is some constant.

Proof. By the form of probability density for $\text{Hyper}((1-\gamma)P, C_n, P)$, we have

$$\mathbb{P}(|\mathcal{U} \cap \mathcal{P}_{k,n}| = 0) = \frac{\binom{C_n}{0}\binom{(P-C_n)}{(1-\gamma)P}}{\binom{P}{(1-\gamma)P}} \leq \gamma^{C_n} = \exp(-c_{n,0}C_n)).$$

□

D.3 Properties of Loss Function

Recall the training and regional reconstruction loss we consider are given by:

$$\mathcal{L}(Q) \coloneqq \frac{1}{2}\mathbb{E}\left[\sum_{\mathbf{p}\in\mathcal{P}} \mathbb{1}\{\mathbf{p} \in \mathcal{M}\} \left\|[F(\mathsf{M}(X); Q, E)]_{\mathbf{p}} - X_{\mathbf{p}}\right\|^2\right] \tag{D.6}$$

$$\mathcal{L}_{\mathbf{p}}(Q) = \frac{1}{2}\mathbb{E}\left[\mathbb{1}\{\mathbf{p} \in \mathcal{M}\} \left\|[F(\mathsf{M}(X), E)]_{\mathbf{p}} - X_{\mathbf{p}}\right\|^2\right] \tag{D.7}$$

In this part, we will present several important lemmas for such a training objective. We first single out the following lemma, which connects the loss form with the attention score.

Lemma D.8 (Loss Calculation)**.** *The population loss $L(Q)$ can be decomposed into the following form:*

$$\mathcal{L}(Q) = \sum_{\mathbf{p}\in\mathcal{P}} \mathcal{L}_{\mathbf{p}}(Q), \text{ where}$$

$$\mathcal{L}_{\mathbf{p}}(Q) = \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \cdot \left(z_{a_{k,\mathbf{p}}}^2 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\right)^2 + \sum_{a\neq a_{k,\mathbf{p}}} z_a^2 \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\right]$$

Proof.

$$\begin{aligned}
&\mathcal{L}_{\mathbf{p}}(Q) \\
&= \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \left\|[F(\mathsf{M}(X), E)]_{\mathbf{p}} - X_{\mathbf{p}}\right\|^2\right] \\
&= \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \left\|\sum_{m\in[N]} \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}} z_m v_{k,m} - z_{a_{k,\mathbf{p}}} v_{k,a_{k,\mathbf{p}}}\right\|^2\right] \\
&\overset{(i)}{=} \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \left(z_{a_{k,\mathbf{p}}}^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\right)^2 + \sum_{m\neq a_{k,\mathbf{p}}} z_m^2 \left(\mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]
\end{aligned}$$

where (i) follows since the features are orthogonal. □

We then introduce some additional crucial notations for the loss objectives.

$$\mathcal{L}^*_{\mathbf{p}} = \min_{Q\in\mathbb{R}^{d\times d}} \mathcal{L}_{\mathbf{p}}(Q), \tag{D.8a}$$

$$\mathcal{L}^{\text{low}}_{\mathbf{p}} = \frac{1}{2}(\sigma_z^2 + \frac{L^2}{N-1}) \sum_{k\in[K]} \mathbb{P}\left(|\mathcal{U} \cap \mathcal{P}_{k,z_{a_{k,\mathbf{p}}}}| = 0\right) \tag{D.8b}$$

$$\widetilde{\mathcal{L}}_{\mathbf{p}}(Q) = \sum_{k=1}^{K} \widetilde{\mathcal{L}}_{k,\mathbf{p}}(Q), \quad \text{where}$$

$$\widetilde{\mathcal{L}}_{k,\mathbf{p}}(Q) = \frac{1}{2}\mathbb{E}\Big[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k, \mathsf{M} \in \mathcal{E}_{k,z_{a_{k,\mathbf{p}}}}\} \cdot \left(z_{a_{k,\mathbf{p}}}^2 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\right)^2 + \sum_{a\neq a_{k,\mathbf{p}}} z_a^2 \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\Big] \tag{D.8c}$$

Here $\sigma_z^2 = \mathbb{E}[Z_n(X)^2]$. $\mathcal{L}_{\mathbf{p}}^\star$ denotes the minimum value of the population loss in (D.7), and $\mathcal{L}_{\mathbf{p}}^{\text{low}}$ represents the unavoidable errors for $\mathbf{p} \in \mathcal{P}$, given that all the patches in $\mathcal{P}_{k,a_{k,\mathbf{p}}}$ are masked. We will show that $\mathcal{L}_{\mathbf{p}}^{\text{low}}$ serves as a lower bound for $\mathcal{L}_{\mathbf{p}}^\star$, and demonstrate that the network trained with GD will attain nearly zero error compared to $\mathcal{L}_{\mathbf{p}}^{\text{low}}$. Our convergence will be established by the sub-optimality gap with respect to $\mathcal{L}_{\mathbf{p}}^{\text{low}}$, which necessarily implies the convergence to $\mathcal{L}_{\mathbf{p}}^\star$. (It also implies $\mathcal{L}_{\mathbf{p}}^\star - \mathcal{L}_{\mathbf{p}}^{\text{low}}$ is small.)

Lemma D.9. *For $\mathcal{L}_{\mathbf{p}}^\star$ and $\mathcal{L}_{\mathbf{p}}^{low}$ defined in (D.8a) and (D.8b), respectively, we have $\mathcal{L}_{\mathbf{p}}^{low} \leq \mathcal{L}_{\mathbf{p}}^\star$ and they are both at the order of $\Theta\Big(\exp\Big(-\big(c_1 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_2 P^{\kappa_s}\big)\Big)\Big)$ where $c_1, c_2 > 0$ are some constants.*

Proof. We first prove $\mathcal{L}_{\mathbf{p}}^{\text{low}} \leq \mathcal{L}_{\mathbf{p}}^\star$:

$$\begin{aligned}
\mathcal{L}_{\mathbf{p}}^\star &= \min_{Q\in\mathbb{R}^{d\times d}} \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\Big[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\} \cdot \left(z_{a_{k,\mathbf{p}}}^3 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\right)^2 + \sum_{a\neq a_{k,\mathbf{p}}} z_a^2 z_{a_{k,\mathbf{p}}} \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\Big] \\
&\geq \min_{Q\in\mathbb{R}^{d\times d}} \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\Big[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k\}\, \mathbb{1}\{|\mathcal{U} \cap \mathcal{P}_{k,a_{\mathbf{p},k}}| = 0\} \cdot \left(z_{a_{k,\mathbf{p}}}^3 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\right)^2 + \sum_{a\neq a_{k,\mathbf{p}}} z_a^2 z_{a_{k,\mathbf{p}}} \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\Big]
\end{aligned}$$

Notice that when all patches in $\mathcal{P}_{k,a_{k,\mathbf{p}}}$ are masked, $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}} = 0$. Moreover,

$$\sum_{m\neq a_{k,\mathbf{p}}} z_m^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}} \geq \frac{L^2}{N-1}$$

by Cauchy–Schwarz inequality. Thus

$$\mathcal{L}_{\mathbf{p}}^\star \geq \frac{1}{2}\sum_{k=1}^{K}(\sigma_z^2 + \frac{L^2}{N-1})\mathbb{P}\left(|\mathcal{U} \cap \mathcal{P}_{k,a_{k,\mathbf{p}}}| = 0\right) = \mathcal{L}_{\mathbf{p}}^{\text{low}}.$$

$\mathcal{L}_{\mathbf{p}}^{\text{low}} = \Theta\Big(\exp\Big(-\big(c_1 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_2 P^{\kappa_s}\big)\Big)\Big)$ immediately comes from Lemma D.7. Furthermore, we only need to show $\mathcal{L}_{\mathbf{p}}^\star = O\Big(\exp\Big(-\big(c_1 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_2 P^{\kappa_s}\big)\Big)\Big)$. This can be directly obtained by choosing $Q = \sigma I_d$ for some sufficiently large σ and hence omitted here. □

Lemma D.10. *Given $\mathbf{p} \in \mathcal{P}$, for any Q, we have*

$$\widetilde{\mathcal{L}}_{\mathbf{p}}(Q) \leq L_{\mathbf{p}}(Q) - \mathcal{L}_{\mathbf{p}}^{low} \leq \widetilde{\mathcal{L}}_{\mathbf{p}}(Q) + O\Big(\exp\Big(-\big(c_3 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_4 P^{\kappa_s}\big)\Big)\Big).$$

where $c_3, c_4 > 0$ are some constants.

Proof. The lower bound is directly obtained by the definition and thus we only prove the upper bound.

$$
\begin{aligned}
& L_{\mathbf{p}}(Q) - \widetilde{\mathcal{L}}_{\mathbf{p}}(Q) \\
&= \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{1}\{\mathbf{p} \in \mathcal{M}, k_X = k, \mathsf{M} \in \mathcal{E}^c_{k,z_{a_{k,\mathbf{p}}}}\} \cdot \left(z^2_{a_{k,\mathbf{p}}} \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\right)^2 + \sum_{a \neq a_{k,\mathbf{p}}} z_a^2 \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2 \right) \right] \\
&\leq \sum_{k=1}^{K} U^2 \mathbb{P}(\mathsf{M} \in \mathcal{E}^c_{k,z_{a_{k,\mathbf{p}}}}) \\
&\leq O\Big(\exp\Big(-\big(c_3 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_4 P^{\kappa_s} \big) \Big) \Big).
\end{aligned}
$$

where the last inequality follows from Lemma D.6.

□

E Overall Induction Hypotheses and Proof Plan

Our main proof utilizes the induction hypotheses. In this section, we introduce the main induction hypotheses for the positive and negative information gaps, which will later be proven to be valid throughout the entire learning process.

E.1 Positive Information Gap

We first state our induction hypothesis for the case that the information gap Δ is positive.

Induction Hypothesis E.1. For $t \leq T$, given $\mathbf{p}, \mathbf{q} \in \mathcal{P}$, for $k \in [K]$, the following holds

a. $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,a_{k,\mathbf{p}}}}$ is monotonically increasing, and $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,a_{k,\mathbf{p}}}} \in [0, \widetilde{O}(1)]$;

b. if $a_{k,\mathbf{p}} \neq 1$, then $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}$ is monotonically decreasing and $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}} \in [-\widetilde{O}(1), 0]$;

c. $|\Phi^{(t)}_{\boldsymbol{p}\to v_{k,m}}| = \widetilde{O}(\frac{1}{P^{1-\kappa_s}})$ for $m \notin \{1\} \cup \{a_{k,\mathbf{p}}\}$;

d. for $\mathbf{q} \neq \mathbf{p}$, $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}} = \widetilde{O}(\frac{1}{P^{\kappa_s}})$;

e. $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{p}} = \widetilde{O}(\frac{1}{P})$.

E.2 Negative Information Gap

Now we turn to the case that $\Delta \leq -\Omega(1)$.

Induction Hypothesis E.2. For $t \leq T$, given $\mathbf{p}, \mathbf{q} \in \mathcal{P}$, for $k \in [K]$, the following holds

a. $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,a_{k,\mathbf{p}}}}$ is monotonically increasing, and $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,a_{k,\mathbf{p}}}} \in [0, \widetilde{O}(1)]$;

b. if $a_{k,\mathbf{p}} \neq 1$, then $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}$ is monotonically decreasing and $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}} \in [-\widetilde{O}(\frac{1}{P^{-\Delta}}), 0]$;

c. $|\Phi^{(t)}_{\boldsymbol{p}\to v_{k,m}}| = \widetilde{O}(\frac{1}{P^{1-\kappa_s}})$ for $m \notin \{1\} \cup \{a_{k,\mathbf{p}}\}$;

d. for $\mathbf{q} \neq \mathbf{p}$, $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}} = \widetilde{O}(\frac{1}{P^{\kappa_s}})$;

e. $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{p}} = \widetilde{O}(\frac{1}{P})$.

E.3 Proof Outline

In both settings, we can classify the process through which transformers learn the feature attention correlation $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,a_{k,\mathbf{p}}}}$ into two distinct scenarios. These scenarios hinge on the spatial relation of the area $\boldsymbol{p}$ within the context of the k-th partition $\mathcal{D}_k$, specifically, whether $\boldsymbol{p}$ is located in the global area of the k-th cluster, i.e. whether $a_{k,\boldsymbol{p}} = 1$. The learning dynamics exhibit different behaviors of learning the local FP correlation in the local area with different Δ, while the behaviors for features located in the global area are very similar, unaffected by the value of Δ. Therefore, through Appendices F to H, we delve into the learning phases and provide technical proofs for the local area with $\Delta \geq \Omega(1)$, local area with $\Delta \leq -\Omega(1)$ and the global area respectively. Finally, we will put this analysis together to prove that the Induction Hypothesis E.1 (resp. Induction Hypothesis E.2) holds during the entire training process, thereby validating the main theorems in Appendix I.

F Analysis for the Local Area with Positive Information Gap

In this section, we focus on a specific patch $\mathbf{p} \in \mathcal{P}$ with the k-th cluster for $k \in [K]$, and present the analysis for the case that $X_{\mathbf{p}}$ is located in the local area for the k-th cluster, i.e. $a_{k,\boldsymbol{p}} > 1$. We will analyze the case that $\Delta \geq \Omega(1)$. Throughout this section, we denote $a_{k,\boldsymbol{p}} = n$ for simplicity. We will analyze the convergence of the training process via two phases of dynamics. At the beginning of each phase, we will establish an induction hypothesis, which we expect to remain valid throughout that phase. Subsequently, we will analyze the dynamics under such a hypothesis within the phase, aiming to provide proof of the hypothesis by the end of the phase.

F.1 Phase I, Stage 1

In this section, we shall discuss the initial stage of phase I. Firstly, we present the induction hypothesis in this stage.

We define the stage 1 of phase I as all iterations $t \leq T_1$, where

$$T_1 \triangleq \max\left\{t : \Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \geq -\frac{1}{U}\left(\frac{\Delta}{2} - 0.01\right)\log(P)\right\}.$$

We state the following induction hypotheses, which will hold throughout this period:

Induction Hypothesis F.1. For each $0 \leq t \leq T_1$, $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$, the following holds:

a. $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}}$ is monotonically increasing, and $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}} \in [0, O\left(\frac{\left(\frac{\Delta}{2}-0.01\right)\log(P)}{P^{0.02}}\right)]$;

b. $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}$ is monotonically decreasing and $\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}} \in [-\frac{1}{U}\left(\frac{\Delta}{2} - 0.01\right)\log(P), 0]$;

c. $|\Phi^{(t)}_{\boldsymbol{p}\to v_{k,m}}| = O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}} - \Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big)$ for $m \neq 1, n$;

d. $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}}}{C_n}\Big)$ for $a_{k,\mathbf{q}} = n$, $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{p}}| = O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}} - \Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}}{P}\Big)$;

e. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{|\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}|}{C_1}\Big) + O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}} - \Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} = 1$;

f. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}} - \Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} \neq 1, n$.

F.1.1 Property of Attention Scores

We first introduce several properties of the attention score if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold.

Lemma F.1. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold at iteration $t \le T_1$, then the following holds*

1. $1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \ge \Omega(1)$;
2. *If* $\mathsf{M} \in \mathcal{E}_{k,n}$, $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} = \Theta\Big(\frac{1}{P^{1-\kappa_s}}\Big)$;
3. *Moreover, if* $\mathsf{M} \in \mathcal{E}_{k,1}$, *we have* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = \Omega\Big(\frac{1}{P^{\frac{1-\kappa_s}{2}-0.01}}\Big)$;
4. *For* $\mathbf{q} \in \mathcal{M} \cap (\mathcal{P}_{k,n} \cup \mathcal{P}_{k,1})$, $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} = O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P}\Big)$.

Lemma F.2. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold at iteration $t \le T_1$, then for $m \ne n, 1$, the following holds:*

1. *For any* $\mathbf{q} \in \mathcal{P}_{k,m}$, $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \le O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P}\Big)$.
2. *Moreover,* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}} \le O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{N}\Big)$.

The above properties can be easily verified through direct calculations by using the definition in (2.5) and conditions in Induction Hypothesis F.1, which are omitted here for brevity.

F.1.2 Bounding the Gradient Updates for FP Correlations

Lemma F.3. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold at iteration $0 \le t \le T_1$, then $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \ge 0$ and satisfies:*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} = \Theta\Big(\frac{C_n}{P}\Big) = \Theta\Big(\frac{1}{P^{1-\kappa_s}}\Big).$$

Proof. By Lemma C.2, we have

$$\begin{aligned}
&\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \\
&= \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\ne n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&= \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}_{k,n} \cap \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\ne n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}^c_{k,n} \cap \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\ne n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&\le \mathbb{P}(\mathsf{M} \in \mathcal{E}_{k,n}) \\
&\quad \cdot \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\ne n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right) \Big| \mathcal{E}_{k,n}\right] \\
&\quad + O(1) \cdot \mathbb{P}(\mathsf{M} \in \mathcal{E}^c_{k,n}) \\
&\le O\Big(\frac{C_n}{P}\Big) + O(\exp(-c_{n,1}C_n)\Big)
\end{aligned}$$

$$\leq O\Big(\frac{C_n}{P}\Big),$$

where the second inequality invokes Lemma F.1 and Lemma D.6, and the last inequality is due to $exp(-c_{n,1}C_n) \ll \frac{C_n}{P}$. Similarly, we can show that $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \geq \Omega(\frac{C_n}{P})$.

□

Lemma F.4. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold at iteration $0 \leq t \leq T_1$, then $\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} < 0$ and satisfies*

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \geq \Omega\Big(\frac{1}{P^{2(\frac{1-\kappa_s}{2}-0.01)}}\Big) = \Omega\Big(\frac{1}{P^{0.98-\kappa_s}}\Big).$$

Proof. We first single out the following fact:

$$\begin{aligned}
&- z_1 z_n^2 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - z_1^3\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} + \sum_{a\neq 1,n} z_a^2 z_1 \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2 \\
&\leq z_1\left(\max_{a\neq 1,n} z_a^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}} - z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - z_1^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}) \\
&= -z_1(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}})\left(z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} + z_1^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} - \max_{a\neq 1,n} z_a^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right). \quad \text{(F.1)}
\end{aligned}$$

Therefore, by Lemma C.1, we have

$$\begin{aligned}
\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} &\leq \mathbb{E}\Bigg[\mathbb{1}\{k_X = k, \mathcal{E}_{k,1}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\cdot \\
&\qquad \left(-z_1(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}})\left(z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} + z_1^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} - \max_{a\neq 1,n} z_a^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)\right)\Bigg] \\
&\qquad + \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}^c_{k,1}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\cdot\sum_{a\neq 1,n} z_1^2 z_a\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right] \\
&\leq \mathbb{P}(\mathsf{M}\in\mathcal{E}_{k,1})\cdot\Big(-\big(\Omega(1)\cdot\Omega(\frac{1}{P^{2\times(\frac{1-\kappa_s}{2}-0.01)}})\big)\Big) + O(1)\cdot\mathbb{P}(\mathsf{M}\in\mathcal{E}^c_{k,1}) \\
&\leq -\Omega\Big(\frac{1}{P^{2\times(\frac{1-\kappa_s}{2}-0.01)}}\Big) = -\Omega\Big(\frac{1}{P^{0.98-\kappa_s}}\Big)
\end{aligned}$$

where the second inequality invokes Lemma F.1 and the last inequality comes from Lemma D.6. □

Lemma F.5. *At each iteration $t \leq T_1$, if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold, then for any $m > 1$ with $m \neq n$, the following holds*

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| \leq O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{N}\Big) = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big).$$

Proof. By Lemma C.1, for $m \neq n$, we have

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,m}} \leq \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\cdot\left(\sum_{a\neq m,n} z_a^2 z_m\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\right] \quad \text{(F.2)}$$

$$-\alpha^{(t)}_{\mathbf{p}\to v_{k,m}} \leq \mathbb{E}\Bigg[\mathbb{1}\{k_X = k, \mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\cdot\Big(z_m z_n^2\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}$$

$$+z_m^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\Bigg]\tag{F.3}$$

For (F.2), we have

$$\begin{aligned}
&\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}\\
&\le \mathbb{E}\left[\mathbb{1}\{k_X=k,\mathcal{E}_{k,1}\cap\mathcal{E}_{k,n}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\cdot\left(\sum_{a\neq m,n}z_a^2z_m\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\right]\\
&\quad+\mathbb{E}\left[\mathbb{1}\{k_X=k,(\mathcal{E}_{k,1}\cap\mathcal{E}_{k,n})^c\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\cdot\left(\sum_{a\neq m,n}z_a^2z_m\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\right]\\
&\le \mathbb{E}\Bigg[\mathbb{1}\{k_X=k,\mathcal{E}_{k,1}\cap\mathcal{E}_{k,n}\cap\mathbf{p}\in\mathcal{M}\}O\left(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{N}\right)\\
&\qquad\cdot\left(z_1^2z_m\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)^2+O\left(\frac{1}{N}\right)\right)\Bigg]+O(1)\cdot\mathbb{P}(\mathsf{M}\in(\mathcal{E}_{k,1}\cap\mathcal{E}_{k,n})^c)\\
&\le O\left(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{N}\right)+O(1)\cdot\mathbb{P}(\mathsf{M}\in(\mathcal{E}_{k,1}\cap\mathcal{E}_{k,n})^c)\\
&\le O\left(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{P^{1-\kappa_s}}\right)
\end{aligned}$$

where the second inequality is due to Lemma F.2, the last inequality follows from Lemma F.4 and Lemma D.6.

On the other hand, for (F.3), we can use the similar argument by invoking Lemma F.2 and Lemma F.3, and thus obtain

$$-\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}\le O\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{1-\kappa_s}}\right).$$

Putting them together, we have

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}|\le O\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}-\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\right).$$

□

F.1.3 Bounding the Gradient Updates for Positional Correlations

Lemma F.6. *For $n>1$, if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold at iteration $0\le t\le T_1$, then for $\mathbf{q}\in\mathcal{P}\setminus\{\mathbf{p}\}$ and $a_{k,\mathbf{q}}=n$, we have $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}\ge 0$ and satisfies:*

$$\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}=\Theta\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\right).$$

Furthermore, we have $|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{p}}|=O\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}-\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\right)$.

Proof. By Lemma C.2, for $\mathbf{q}\in\mathcal{P}_{k,n}$ with $\mathbf{q}\neq\mathbf{p}$, we have

$$\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}=$$

$$\underbrace{\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M},\mathbf{q}\in\mathcal{U}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}}\cdot\left(z_n^2\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2+\sum_{m\neq n}z_m^2\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]}_{H_1}$$
$$+\underbrace{\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M},\mathbf{q}\in\mathcal{M}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}}\cdot\left(-z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\right)\right]}_{H_2}$$
$$+\underbrace{\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M},\mathbf{q}\in\mathcal{M}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}}\cdot\left(\sum_{m\neq n}z_m^2\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]}_{H_3}.$$

Firstly, for H_1, notice that

$$\begin{aligned}(C_n-1)H_1&=\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\cdot\left(z_n^2\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2+\sum_{m\neq n}z_m^2\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]\\&=\Theta(\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}).\end{aligned}$$

For H_2, since $\mathbf{p},\mathbf{q}\in\mathcal{M}$, by Lemma F.1, we can upper bound $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}}$ by $O\left(\frac{1}{P}\right)$, thus

$$-H_2\leq\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M}\}O\Big(\frac{1}{P}\Big)\cdot\left(z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\right)\right]\leq O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P}\Big).$$

Further notice that H_3 can be upper bounded by $O(H_1)$, putting it together, we have

$$\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}=\Theta\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\Big).$$

Turn to $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{p}}$, when $\mathbf{q}=\mathbf{p}$,

$$\beta^{(t)}_n=\underbrace{\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{p}}\cdot\left(-z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\right)\right]}_{J_2}$$
$$+\underbrace{\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{p}}\cdot\left(\sum_{m\neq n}z_m^2\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]}_{J_3}.$$

We can bound J_2 in a similar way as H_2. Thus, we only focus on further bounding J_3:

$$\begin{aligned}J_3&\leq\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M}\}O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P}\Big)\cdot\left(\sum_{m\neq n}z_m^2\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]\\&\leq O\left(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{P}\right).\end{aligned}$$

where the first inequality holds by invoking Lemma F.1 and the last inequality follows similar arguments as analysis for (F.2). □

Lemma F.7. *For $n>1$, if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold at iteration*

$0 \le t \le T_1$, *then for* $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$ *and* $a_{k,\mathbf{q}} = 1$, *we have* $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ *satisfies:*

$$|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\left(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{P}\right) + O\left(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\right).$$

Proof. By Lemma C.2, for $\mathbf{q} \in \mathcal{P}_{k,1}$, we have

$$\begin{aligned}
&\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = \\
&- \mathbb{E}\Bigg[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}, \mathbf{q} \in \mathcal{U}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \cdot \\
&\left(z_1^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}) + z_n^2\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \sum_{a\neq 1,n} z_a^2\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\Bigg] \qquad \text{(F.4)} \\
&\underbrace{-\mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}, \mathbf{q} \in \mathcal{M}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \cdot \left(z_n^2\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\right]}_{G_2} \\
&\underbrace{+\mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}, \mathbf{q} \in \mathcal{M}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \cdot \left(\sum_{a\neq n} z_a^2\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\right]}_{G_3}.
\end{aligned}$$

For (F.4) denoted as G_1, following the direct calculations, we have

$$-(C_1 - 1)G_1 = \Theta(\alpha^{(t)}_{\mathbf{p}\to v_{k,1}})$$

We can further bound G_2 and G_3 in a similar way as H_2 and H_3 in Lemma F.6 and thus obtain

$$\begin{aligned}
-G_2 &\le O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P}\Big), \\
G_3 &\le O\left(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{P}\right).
\end{aligned}$$

which completes the proof. □

Lemma F.8. *For* $n > 1$, *if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold at iteration* $0 \le t \le T_1$, *then for* $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$ *and* $n \neq a_{k,\mathbf{q}}$, $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$ *satisfies:*

$$|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big).$$

Proof. By Lemma C.2, for $\mathbf{q} \in \mathcal{P}_{k,m}$, we have

$$\begin{aligned}
&\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = \\
&- \mathbb{E}\Bigg[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}, \mathbf{q} \in \mathcal{U}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \cdot \\
&\left(z_m^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}) + z_n^2\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \sum_{a\neq n,m} z_a^2\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\Bigg] \qquad \text{(F.5)} \\
&\underbrace{-\mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}, \mathbf{q} \in \mathcal{M}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \cdot \left(z_n^2\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\right]}_{I_2}
\end{aligned}$$

$$\underbrace{+\mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}, \mathbf{q} \in \mathcal{M}\}\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \cdot \left(\sum_{a\neq n} z_a^2 \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right)\right]}_{I_3}.$$

(F.5) can be upper bounded by $O\Big(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}|}{C_m}\Big) = O\Big(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{NC_m}\Big) = O\Big(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{P}\Big)$, where the first equality holds by invoking Lemma F.5. I_2 and I_3 can be bounded similarly as G_2 and G_3, which is omitted here. □

F.1.4 At the end of Phase I, Stage 1

Lemma F.9. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.1 hold for all $0 \le t \le T_1 = O\Big(\frac{\log(P)P^{0.98-\kappa_s}}{\eta}\Big)$, At iteration $t = T_1 + 1$, we have*

a. $\Phi^{(T_1+1)}_{\mathbf{p}\to v_{k,1}} \le -\frac{1}{U}\left(\frac{\Delta}{2} - 0.01\right)\log(P)$;

b. $\mathbf{Attn}^{(T_1+1)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = O\Big(\frac{1}{P^{(1-\kappa_c)+\frac{L}{U}(\frac{\Delta}{2}-0.01)}}\Big)$.

Proof. By comparing Lemma F.3 and Lemma F.4, we have $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \gg \alpha^{(t)}_{\mathbf{p}\to v_{k,n}}$. Then the existence of $T_{1,k} = O\Big(\frac{\log(P)P^{0.98-\kappa_s}}{\eta}\Big)$ directly follows from Lemma F.4. □

F.2 Phase I, Stage 2

During stage 1, $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ significantly decreases to decouple the FP correlations with the global feature, resulting in a decrease in $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}$, while other $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}$ with $m > 1$ remain approximately at the order of $O\left(\frac{1}{P^{1-\kappa_s}}\right)$ $\left(\Theta\left(\frac{1}{P^{1-\kappa_s}}\right)\right)$. By the end of phase I, $(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}})^2$ decreases to $O(\frac{1}{P^{1.96-2\kappa_s}})$, leading to a decrease in $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|$ as it approaches towards $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}$. At this point, stage 2 begins. Shortly after entering this phase, the prior dominant role of the decrease of $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ in learning dynamics diminishes as $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|$ reaches the same order of magnitude as $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}$.

We define stage 2 of phase I as all iterations $T_1 < t \le \widetilde{T}_1$, where

$$\widetilde{T}_1 \triangleq \max\left\{t > T_1 : \Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \le \left(\frac{\Delta}{2L} + \frac{0.01}{L} + \frac{c_1^*(1-\kappa_s)}{U}\right)\log(P)\right\}.$$

for some small constant $c_1^* > 0$.

For computational convenience, we make the following assumptions for κ_c and κ_s, which can be easily relaxed with the cost of additional calculations.

$$\frac{\Delta}{2}\Big(\frac{1}{L} - \frac{1}{U}\Big) + \frac{0.01}{L} + \frac{0.01}{U} \le \frac{c_0^*(1-\kappa_s)}{U} \tag{F.6a}$$

$$(1 - \frac{c_1^* L}{U})(1-\kappa_s) \le (1-\kappa_c) + \frac{U}{L}(\frac{\Delta}{2} + 0.01) \tag{F.6b}$$

Here c_0^* is some small. We state the following induction hypotheses, which will hold throughout this period:

Induction Hypothesis F.2. For each $T_1 < t \le \widetilde{T}_1$, $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$, the following holds:

a. $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}$ is monotonically increasing, and $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \in [0, \frac{c_0^*+c_1^*}{U}\log(P)]$;

b. $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ is monotonically decreasing and $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \in [-\frac{1}{L}\left(\frac{\Delta}{2} + 0.01\right)\log(P), -\frac{1}{U}\left(\frac{\Delta}{2} - 0.01\right)\log(P)]$;

c. $|\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big)$ for $m \neq 1, n$;

d. $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}}}{C_n}\Big)$ for $a_{k,\mathbf{q}} = n$, $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{p}}| = O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}} - \Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}}{P}\Big)$;

e. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{|\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}|}{C_1}\Big) + O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}} - \Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} = 1$.;

f. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}} - \Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} \neq 1, n$.

F.2.1 Property of Attention Scores

We first single out several properties of attention scores that will be used for the proof of Induction Hypothesis F.2.

Lemma F.10. *if Induction Hypothesis E.1 and Induction Hypothesis F.2 hold at iteration $T_1 + 1 \le t \le \widetilde{T}_1$, then the following holds*

1. $1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \ge \Omega(1)$;
2. *if* $\mathsf{M} \in \mathcal{E}_{k,n}$, $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \in \Big[\Omega\big(\frac{1}{P^{1-\kappa_s}}\big), O\big(\frac{1}{P^{(1-c_1^*-c_0^*)(1-\kappa_s)}}\big)\Big]$;
3. *Moreover,* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = O\Big(\frac{1}{P^{(1-\kappa_c)+\frac{L}{U}(\frac{\Delta}{2}-0.01)}}\Big)$; *if* $\mathsf{M} \in \mathcal{E}_{k,1}$, *we have* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = \Omega\Big(\frac{1}{P^{(1-\kappa_c)+\frac{U}{L}(\frac{\Delta}{2}+0.01)}}\Big)$;
4. *for* $\mathbf{q} \in \mathcal{M} \cap (\mathcal{P}_{k,n} \cup \mathcal{P}_{k,1})$, $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} = O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}}{P}\Big)$.

Lemma F.11. *if Induction Hypothesis E.1 and Induction Hypothesis F.2 hold at iteration $T_1 + 1 \le t \le \widetilde{T}_1$, then for $m \neq n$, the following holds:*

1. *for any* $\mathbf{q} \in \mathcal{P}_{k,m}$, $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \le O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}}{P}\Big)$;
2. *Moreover,* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}} \le O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{N}\Big)$.

F.2.2 Bounding the Gradient Updates of FP Correlations

Lemma F.12. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.2 hold at iteration $T_1 + 1 \le t \le \widetilde{T}_1$, then $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \ge 0$ and satisfies:*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} = \Omega\Big(\frac{1}{P^{1-\kappa_s}}\Big).$$

Proof. By Lemma C.2, we have

$$\begin{aligned}
&\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \\
&= \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&= \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}_{k,n} \cap \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}^c_{k,n} \cap \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&\gtrsim \mathbb{P}(\mathsf{M} \in \mathcal{E}_{k,n})
\end{aligned}$$

$$\cdot \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\Big|\mathcal{E}_{k,n}\right]$$
$$\geq \Omega\Big(\frac{C_n}{P}\Big)$$

where the last inequality invokes Lemma F.10.

□

Lemma F.13. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.2 hold at iteration $T_1 + 1 \leq t \leq \widetilde{T}_1$, then $\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} < 0$ and satisfies*

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \geq \Omega\Big(\frac{1}{P^{2(1-\kappa_c)+\frac{U}{L}(\Delta+0.02)}}\Big).$$

Proof. Following (F.1), we have

$$-z_1 z_n^2\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - z_1^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} + \sum_{a\neq 1,n} z_a^2 z_1\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2$$
$$\leq -z_1(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}})\left(z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} + z_1^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} - \max_{a\neq 1,n} z_a^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)$$

Therefore, by Lemma C.1, we obtain

$$\begin{aligned}
\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} &\leq \mathbb{E}\Big[\mathbb{1}\{k_X = k, \mathcal{E}_{k,1}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\cdot \\
&\quad \Big(-z_1(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}) \\
&\quad \cdot\left(z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} + z_1^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} - \max_{a\neq 1,n} z_a^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)\Big)\Big] \\
&\quad + \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}^c_{k,1}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\cdot\sum_{a\neq 1,n} z_1^2 z_a\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right] \\
&\leq \mathbb{P}(\mathsf{M}\in\mathcal{E}_{k,1})\cdot\left(-\Omega(1)\cdot\Omega\Big(\frac{1}{P^{2(1-\kappa_c)+\frac{2U}{L}(\frac{\Delta}{2}+0.01)}}\Big)\right) + O(1)\cdot\mathbb{P}(\mathsf{M}\in\mathcal{E}^c_{k,1}) \\
&\leq -\Omega\Big(\frac{1}{P^{2(1-\kappa_c)+\frac{U}{L}(\Delta+0.02)}}\Big)
\end{aligned}$$

where the second inequality invokes Lemma F.10 and the last inequality comes from Lemma D.6. The upper bound can be obtained by using similar arguments and invoking the upper bound for $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}$ in Lemma F.10. □

Lemma F.14. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.2 hold at iteration $T_1 + 1 \leq t \leq \widetilde{T}_1$, then for any $m > 1$ with $m \neq n$, the following holds*

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| \leq O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big).$$

The proof is similar to Lemma F.5, and thus omitted here.

F.2.3 Bounding the Gradient Updates of Positional Correlations

We then summarize the properties for gradient updates of positional correlations, which utilize the identical calculations as in Section F.1.3.

Lemma F.15. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.2 hold at iteration $T_1 + 1 \le t \le \widetilde{T}_1$, then*

a. if $a_{k,\mathbf{q}} = n$ and $\mathbf{q} \neq \mathbf{p}$, $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} \ge 0$; $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = \Theta\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\Big)$ and $|\beta^{(t)}_n| = O(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P})$.

b. if $a_{k,\mathbf{q}} = 1$, $|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big) + O\Big(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\Big)$.

c. if $a_{k,\mathbf{q}} = m$ and $m \neq 1, n$, $|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$.

F.2.4 End of Phase I, Stage 2

Lemma F.16. *Induction Hypothesis F.2 holds for all iteration $T_1 + 1 \le t \le \widetilde{T}_1 = T_1 + O\Big(\frac{\log(P)P^{1-\kappa_s}}{\eta}\Big)$, and at iteration $t = \widetilde{T}_1 + 1$, we have*

a. $\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,n}} \ge \frac{c_1^*(1-\kappa_s)\log(P)}{U}$;

b. $\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,1}} \ge -(\frac{\Delta}{2L} + \frac{0.01}{L})\log(P)$.

Proof. The existence of $\widetilde{T}_1 = T_1 + O\Big(\frac{\log(P)P^{1-\kappa_s}}{\eta}\Big)$ directly follows from Lemma F.12 and Lemma F.13. Moreover, since $\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} < 0$, then

$$\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,n}} \le \left(\frac{\Delta}{2L} + \frac{0.01}{L} + \frac{c_1^*(1-\kappa_s)}{U}\right)\log(P) - \frac{1}{U}(\frac{\Delta}{2} - 0.01) \le \frac{(c_0^* + c_1^*)(1-\kappa_s)}{U}\log(P)$$

where the last inequality invokes (F.6a). Now suppose $\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,n}} < \frac{c_1^*(1-\kappa_s)\log(P)}{U}$, then $\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,1}} < -(\frac{\Delta}{2L} + \frac{0.01}{L})\log(P)$. Denote the first time that $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ reaches $-(\frac{\Delta}{2L} + \frac{0.001}{L})\log(P)$ as $\widetilde{T}$. Note that $\widetilde{T} < \widetilde{T}_1$ since $\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}$, the change of $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$, satisfies $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \ll \log(P)$. Then for $t \ge \widetilde{T}$, the following holds:

1. $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \ge \Omega\left(\frac{1}{P^{1-\kappa_s}}\right)$;
2. $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \le O\Big(\frac{1}{P^{\frac{1-\kappa_s}{2}+0.001}}\Big)$.

Therefore,

$$\begin{aligned}
|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \le\ & \mathbb{E}\Big[\mathbb{1}\{k_X = k, \mathcal{E}_{k,1} \cap \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \cdot \\
& z_1\left(z_n^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}) + z_1^2\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}})\right)\Big] \\
& + \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}^c_{k,1} \cap \mathbf{p} \in \mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \cdot \sum_{a\neq 1,n} z_1^2 z_a \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right] \\
\le\ & O(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{\frac{1-\kappa_s}{2}+0.001}}) + \mathbb{P}(\mathsf{M} \in \mathcal{E}_{k,1}) \cdot \left(O(1) \cdot O\Big(\frac{1}{P^{\frac{1-\kappa_s}{2}+0.001}}\Big)\right) + O(1) \cdot \mathbb{P}(\mathsf{M} \in \mathcal{E}^c_{k,1}) \\
\le\ & O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{\frac{1-\kappa_s}{2}+0.001}}\Big) + O\Big(\frac{1}{P^{(1-\kappa_s)+0.002}}\Big).
\end{aligned}$$

Lemma F.12 still holds, and thus

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \le O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{0.002}}\Big).$$

Since $|\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,1}} - \Phi^{(\widetilde{T})}_{\mathbf{p}\to v_{k,1}}| \geq \Omega\left(\log(P)\right)$, we have

$$\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,n}} \geq |\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,1}} - \Phi^{(\widetilde{T})}_{\mathbf{p}\to v_{k,1}}| \cdot \Omega(P^{0.002}) + \Phi^{(\widetilde{T})}_{\mathbf{p}\to v_{k,n}} \gg \Omega(P^{0.002}\log(P)),$$

which contradicts the assumption that $\Phi^{(\widetilde{T}_1+1)}_{\mathbf{p}\to v_{k,n}} < \frac{c_1^*(1-\kappa_s)\log(P)}{U}$. □

F.3 Phase II, Stage 1

For $n > 1$, we define stage 1 of phase II as all iterations $\widetilde{T}_1 + 1 \leq t \leq T_2$, where

$$T_2 \triangleq \max\left\{t : \Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \leq \frac{(1-\kappa_s)}{L}\log(P)\right\}.$$

We state the following induction hypotheses, which will hold throughout this stage:

Induction Hypothesis F.3. For each $\widetilde{T}_1 + 1 \leq t \leq T_2$, $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$, the following holds:

a. $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}$ is monotonically increasing, and $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \in \left[\frac{c_1^*(1-\kappa_s)}{U}\log(P), \frac{(1-\kappa_s)}{L}\log(P)\right]$;

b. $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ is monotonically decreasing and

$$\Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \in \left[-\frac{1}{L}\left(\frac{\Delta}{2}+0.01\right)\log(P) - o(1), -\frac{1}{U}\left(\frac{\Delta}{2}-0.01\right)\log(P)\right];$$

c. $|\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}| = O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\right)$ for $m \neq 1, n$;

d. $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\right)$ for $a_{k,\mathbf{q}} = n$, $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{p}}| = O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\right)$;

e. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\left(\frac{|\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\right) + O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\right)$ for $a_{k,\mathbf{q}} = 1.$;

f. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\right)$ for $a_{k,\mathbf{q}} \neq 1, n$.

F.3.1 Property of Attention Scores

We first single out several properties of attention scores that will be used for the proof of Induction Hypothesis F.3.

Lemma F.17. *if Induction Hypothesis E.1 and Induction Hypothesis F.3 hold at iteration $\widetilde{T}_1 + 1 \leq t \leq T_2$, then the following holds*

1. *if $\mathsf{M} \in \mathcal{E}_{k,n}$, $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \geq \Omega\left(\frac{1}{P^{(1-\frac{c_1^* L}{U})(1-\kappa_s)}}\right)$. Moreover, if $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}$ does not reach the constant level, $1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} = \Omega(1)$; otherwise, $1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} = \Omega\left(\frac{1}{P^{(\frac{U}{L}-1)(1-\kappa_s)}}\right)$.*

2. *$\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = O\left(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P^{(1-\kappa_c)+\frac{L}{U}(\frac{\Delta}{2}-0.01)}}\right)$; if $\mathsf{M} \in \mathcal{E}_{k,1}$, we have $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = \Omega\left(\frac{1}{P^{(1-\kappa_c)+\frac{U}{L}(\frac{\Delta}{2}+0.01)}}\right)$;*

3. *for $\mathbf{q} \in \mathcal{M} \cap (\mathcal{P}_{k,n} \cup \mathcal{P}_{k,1})$, $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} = O\left(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P}\right)$*

Lemma F.18. *if Induction Hypothesis E.1 and Induction Hypothesis F.3 hold at iteration $\widetilde{T}_1 + 1 \leq t \leq T_2$, then for $m \neq n$, the following holds:*

1. *for any* $\mathbf{q} \in \mathcal{P}_{k,m}$, $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \le O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P}\Big)$.

2. *Moreover,* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \le O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{N}\Big)$.

F.3.2 Bounding the Gradient Updates of FP Correlations

Lemma F.19. *if Induction Hypothesis E.1 and Induction Hypothesis F.3 hold at iteration* $\widetilde{T}_1+1 \le t \le T_2$, *then* $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \ge 0$ *and satisfies:*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \ge \min\Big\{\Omega(\frac{1}{P^{(1-\frac{c_1^* L}{U})(1-\kappa_s)}}), \Omega\left(\frac{1}{P^{2(\frac{U}{L}-1)(1-\kappa_s)}}\right)\Big\}.$$

Proof. By Lemma C.2, we have

$$\begin{aligned}
&\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \\
&= \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&= \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}_{k,n}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}^c_{k,n}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right] \\
&\gtrsim \mathbb{P}(\mathsf{M}\in\mathcal{E}_{k,n})\cdot \\
&\quad \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \cdot \left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 z_n \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\Big|\mathcal{E}_{k,n}\right] \\
&\quad + O(1)\cdot\mathbb{P}(\mathsf{M}\in\mathcal{E}^c_{k,n}) \\
&\gtrsim \min\left\{\Omega\left(\frac{1}{P^{(1-\frac{c_1^* L}{U})(1-\kappa_s)}}\right), \Omega\left(\frac{1}{P^{2(\frac{U}{L}-1)(1-\kappa_s)}}\right)\right\}
\end{aligned}$$

where the last inequality invokes Lemma F.17 by observing that for $\mathsf{M}\in\mathcal{E}_{k,n}$,

$$\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}})^2 \ge \min\left\{\Omega\left(\frac{1}{P^{(1-\frac{c_1^* L}{U})(1-\kappa_s)}}\right)\cdot\Omega(1), \Omega(1)\cdot\Omega\left(\frac{1}{P^{2\times(\frac{U}{L}-1)(1-\kappa_s)}}\right)\right\}.$$

□

Lemma F.20. *For* $n>1$, *if Induction Hypothesis E.1 and Induction Hypothesis F.3 hold at iteration* $\widetilde{T}_1+1 \le t \le T_2$, *then* $\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} < 0$ *and satisfies*

$$\begin{aligned}
|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| &\ge \min\left\{\Omega\left(\frac{1}{P^{(1-\frac{c_1^* L}{U})(1-\kappa_s)}}\right), \Omega\left(\frac{1}{P^{(\frac{U}{L}-1)(1-\kappa_s)}}\right)\right\}\cdot\Omega\Big(\frac{1}{P^{(1-\kappa_c)+\frac{L}{U}(\frac{\Delta}{2}-0.01)}}\Big), \\
|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| &\le \max\left\{O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{(1-\kappa_c)+\frac{L}{U}(\Delta/2-0.01)}}\Big), O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{2(1-\kappa_c)+\frac{L}{U}(\Delta-0.02)-(1-\frac{c_1^* L}{U})(1-\kappa_s)}}\Big)\right\}.
\end{aligned}$$

Proof. Following (F.1), we have

$$
\begin{aligned}
&- z_1 z_n^2 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - z_1^3 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right) \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} + \sum_{a\neq 1,n} z_a^2 z_1 \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2 \\
&\leq -z_1(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}) \left(z_n^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} + z_1^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} - \max_{a\neq 1,n} z_a^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right).
\end{aligned}
$$

Therefore, by Lemma C.1, we obtain

$$
\begin{aligned}
\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} \leq\ & \mathbb{E}\Big[\mathbb{1}\{k_X = k, \mathcal{E}_{k,1} \cap \mathbf{p} \in \mathcal{M}\} \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \cdot \\
& \left(-z_1(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}) \left(z_n^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} + z_1^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} - \max_{a\neq 1,n} z_a^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)\right)\Big] \\
& + \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}^c_{k,1} \cap \mathbf{p} \in \mathcal{M}\} \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \cdot \sum_{a\neq 1,n} z_1^2 z_a \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2\right] \\
\leq\ & -\min\left\{\Omega\left(\frac{1}{P^{(1-\frac{c_1^* L}{U})(1-\kappa_s)}}\right), \Omega\left(\frac{1}{P^{(\frac{U}{L}-1)(1-\kappa_s)}}\right)\right\} \cdot \Omega\Big(\frac{1}{P^{(1-\kappa_c)+\frac{L}{U}(\frac{\Delta}{2}-0.01)}}\Big)
\end{aligned}
$$

where the second inequality invokes Lemma F.17 and (F.6b). Moreover,

$$
\begin{aligned}
|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \lesssim\ & \mathbb{E}\Big[\mathbb{1}\{k_X = k, \mathcal{E}_{k,1} \cap \mathcal{E}_{k,n} \cap \mathbf{p} \in \mathcal{M}\} \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \cdot \\
& \left(z_1 z_n^2 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} + z_1^3 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right) \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)\Big] \\
=\ & \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}_{k,1} \cap \mathcal{E}_{k,n} \cap \mathbf{p} \in \mathcal{M}\} z_1 z_n^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} \cdot \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right] \\
& + \mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}_{k,1} \cap \mathcal{E}_{k,n} \cap \mathbf{p} \in \mathcal{M}\} z_1^3 (\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}})^2 \cdot \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}}\right)\right] \\
\leq\ & \max\left\{O\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{(1-\kappa_c)+\frac{L}{U}(\frac{\Delta}{2}-0.01)}}\right), O\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{2(1-\kappa_c)+\frac{2L}{U}(\frac{\Delta}{2}-0.01)-(1-\frac{c_1^* L}{U})(1-\kappa_s)}}\right)\right\}
\end{aligned}
$$

where the second inequality invokes Lemma F.17. □

Lemma F.21. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.3 hold at iteration $\widetilde{T}_1 + 1 \leq t \leq T_2$ for any $m > 1$ with $m \neq n$, the following holds*

$$
|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| \leq O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big).
$$

The proof is similar to Lemma F.5, and thus omitted here.

F.3.3 Bounding the Gradient Updates of Positional Correlations

We then summarize the properties for gradient updates of positional correlations, which utilizes the identical calculations as in Section F.1.3.

Lemma F.22. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.3 hold at iteration $\widetilde{T}_1 + 1 \leq t \leq T_2$, then*

a. if $a_{k,\mathbf{q}} = n$ and $\mathbf{q} \neq \mathbf{p}$, $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} \geq 0$; $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = \Theta(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n})$ and $|\beta^{(t)}_n| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$.

b. if $a_{k,\mathbf{q}} = 1$, $|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big) + O\Big(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\Big)$.

c. if $a_{k,\mathbf{q}} = m$ and $m \neq 1, n$, $|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$.

F.3.4 End of Phase II, Stage 1

Lemma F.23. *Induction Hypothesis F.3 holds for all $\widetilde{T}_1 + 1 \le t \le T_2$, and at iteration $t = T_2 + 1$, we have*

a. $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} > \frac{(1-\kappa_s)}{L}\log(P)$;

b. $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} = \Omega(1)$ *if* $\mathsf{M} \in \mathcal{E}_{k,n}$.

Proof. By comparing Lemma F.19 and Lemma F.20-F.23, we have $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \gg |\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}|, |\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}|$. Then the existence of $T_2 = \widetilde{T}_1 + O\Big(\frac{\log(P)P^{\Lambda}}{\eta}\Big)$ directly follows from Lemma F.19, where

$$\Lambda = \max\Big\{(1 - \frac{c_1^* L}{U}), 2(\frac{U}{L} - 1)\Big\} \cdot (1 - \kappa_s).$$

The second statement can be directly verified by noticing that $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} > \frac{(1-\kappa_s)}{L}\log(P)$ while all other attention correlations are sufficiently small. □

F.4 Phase II, Stage 2

In this final stage, we establish that these structures indeed represent the solutions toward which the algorithm converges.

Given any $0 < \epsilon < 1$, for $n > 1$, define

$$T_2^{\epsilon} \triangleq \max\left\{t > T_2 : \Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \le \log\left(c_5\left(\left(\frac{3}{\epsilon}\right)^{\frac{1}{2}} - 1\right)N\right)\right\}. \tag{F.7}$$

where c_5 is some largely enough constant.

We state the following induction hypotheses, which will hold throughout this stage:

Induction Hypothesis F.4. For $n > 1$, suppose $\mathrm{polylog}(P) \gg \log(\frac{1}{\epsilon})$, for each $T_2 + 1 \le t \le T_2^{\epsilon}$, $\mathbf{q} \in \mathcal{P}\setminus\{\mathbf{p}\}$, the following holds:

a. $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}$ is monotonically increasing, and $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \in [\frac{(1-\kappa_s)}{L}\log(P), O(\log(P/\epsilon))]$;

b. $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ is monotonically decreasing and $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \in \Big[-\frac{1}{L}\left(\frac{\Delta}{2} + 0.01\right)\log(P) - o(1), -\frac{1}{U}\left(\frac{\Delta}{2} - 0.01\right)\log(P)\Big]$;

c. $|\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big)$ for $m \ne 1, n$;

d. $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\Big)$ for $a_{k,\mathbf{q}} = n$, $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{p}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$;

e. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{|\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\Big) + O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} = 1.$;

f. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} \ne 1, n$.

F.4.1 Property of Attention Scores

We first single out several properties of attention scores that will be used for the proof of Induction Hypothesis F.4.

Lemma F.24. *if Induction Hypothesis E.1 and Induction Hypothesis F.4 hold at iteration $T_{n,2} < t \le T_{n,2}^{\epsilon}$, then the following holds*

1. *if* $\mathsf{M} \in \mathcal{E}_{k,n}$, $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} = \Omega(1)$ *and* $(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}})^2 \ge O(\epsilon)$.

2. *Moreover,* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P^{(1-\kappa_c)+\frac{L}{U}(\frac{\Delta}{2}-0.01)}}\Big)$*; if* $\mathsf{M}\in\mathcal{E}_{k,1}$*, we have* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} = \Omega\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P^{(1-\kappa_c)+\frac{U}{L}(\frac{\Delta}{2}+0.01)}}\Big)$*;*

3. *for* $\mathbf{q}\in\mathcal{M}\cap(\mathcal{P}_{k,n}\cup\mathcal{P}_{k,1})$*,* $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} = O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P}\Big)$*.*

Lemma F.25. *if Induction Hypothesis E.1 and Induction Hypothesis F.4 hold at iteration* $T_{n,2}<t\le T^{\epsilon}_{n,2}$*,then for* $m\neq n$*, the following holds:*

1. *for any* $\mathbf{q}\in\mathcal{P}_{k,m}$*,* $\mathbf{attn}^{(t)}_{\mathbf{p}\to\mathbf{q}} \le O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{P}\Big)$*.*

2. $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} \le O\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{N}\Big)$*, and if* $\mathsf{M}\in\mathcal{E}_{k,m}$*,* $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} = \Theta\Big(\frac{1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}}{N}\Big)$*.*

F.4.2 Bounding the Gradient Updates of FP Correlations

Lemma F.26. *For* $n>1$*, if Induction Hypothesis E.1 and Induction Hypothesis F.4 hold at iteration* $T_2+1\le t\le T^{\epsilon}_2$*, then* $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}\ge 0$ *and satisfies:*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \ge \Omega(\epsilon).$$

Proof. By Lemma C.2, we have

$$\begin{aligned}
&\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}\\
&= \mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\cdot\left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2+\sum_{m\neq n}z_m^2z_n\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]\\
&= \mathbb{E}\left[\mathbb{1}\{k_X=k,\mathcal{E}_{k,n}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\cdot\left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2+\sum_{m\neq n}z_m^2z_n\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]\\
&\quad+\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathcal{E}^c_{k,n}\cap\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\cdot\left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2+\sum_{m\neq n}z_m^2z_n\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\right]\\
&\gtrsim \mathbb{P}(\mathsf{M}\in\mathcal{E}_{k,n})\cdot\\
&\mathbb{E}\left[\mathbb{1}\{k_X=k,\mathbf{p}\in\mathcal{M}\}\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\cdot\left(z_n^3\left(1-\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2+\sum_{m\neq n}z_m^2z_n\left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right)^2\right)\Big|\mathcal{E}_{k,n}\right]\\
&\quad+O(1)\cdot\mathbb{P}(\mathsf{M}\in\mathcal{E}^c_{k,n})\\
&\gtrsim \Omega(\epsilon)
\end{aligned}$$

where the last inequality invokes Lemma F.24, Lemma D.6 and the fact that

$$\epsilon \ge \exp(-\operatorname{polylog}(K)) \gg \exp\left(-c_{n,1}C_n\right).$$

□

Lemma F.27. *For* $n>1$*, if Induction Hypothesis E.1 and Induction Hypothesis F.4 hold at iteration* $T_{n,3}<t\le T^{\epsilon}_{n,4}$*, then* $\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}<0$ *and satisfies*

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| \le \max\left\{O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{(1-\kappa_c)+\frac{L}{U}(\Delta/2-0.01)}}\Big), O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{2(1-\kappa_c)+\frac{L}{U}(\Delta-0.02)-(1-\frac{c_1^*L}{U})(1-\kappa_s)}}\Big)\right\}$$

The proof follows the similar arguments Lemma F.20 by noticing that $\epsilon\gg\mathbb{P}(\mathsf{M}\in\mathcal{E}^c_{k,m})$ for any $m\neq n$.

Lemma F.28. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.4 hold at iteration $T_2 < t \le T_2^\epsilon$, then for any $m > 1$ with $m \neq n$, the following holds*

$$-O(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{1-\kappa_s}}) \le \alpha^{(t)}_{\mathbf{p}\to v_{k,m}} \le 0$$

Proof. We first note that

$$\begin{aligned}
&- z_1 z_n^2 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right) \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - z_m^3 \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right) \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,1}} + \sum_{a\neq 1,n} z_a^2 z_m \left(\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}\right)^2 \\
&\le z_m \left(\max_{a\neq m,n} z_a^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}} - z_n^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - z_m^2 \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right) \left(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}}\right) \\
&\lesssim -\Omega(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}})
\end{aligned}$$

since when $\mathsf{M} \in \mathcal{E}_{k,n}$, we have $\mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}} = \Omega(1) \gg \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,a}}$. Thus, we have

$$\begin{aligned}
0 \ge \alpha^{(t)}_{\mathbf{p}\to v_{k,m}} &\gtrsim -\mathbb{E}\left[\mathbb{1}\{k_X = k, \mathcal{E}_{k,n} \cap \mathbf{p} \in \mathcal{M}\} \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,m}} \cdot \Omega(1 - \mathbf{Attn}^{(t)}_{\mathbf{p}\to\mathcal{P}_{k,n}})\right] \\
&\ge -O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{1-\kappa_s}}\Big).
\end{aligned}$$

□

F.4.3 Bounding the Gradient Updates of Positional Correlations

We then summarize the properties for gradient updates of positional correlations, which utilizes the identical calculations as in Section F.1.3.

Lemma F.29. *For $n > 1$, if Induction Hypothesis E.1 and Induction Hypothesis F.4 hold at iteration $T_2 + 1 \le t \le T_2^\epsilon$, then*

a. if $a_{k,\mathbf{q}} = n$ and $\mathbf{q} \neq \mathbf{p}$, $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} \ge 0$; $\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = \Theta\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\Big)$ and $|\beta^{(t)}_n| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$.

b. if $a_{k,\mathbf{q}} = 1$, $|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big) + O\Big(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\Big)$.

c. if $a_{k,\mathbf{q}} = m$ and $m \neq 1, n$, $|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$.

F.4.4 End of Phase II, Stage 2

Lemma F.30. *For $n > 1$, and $0 < \epsilon < 1$, suppose $\mathrm{polylog}(P) \gg \log(\frac{1}{\epsilon})$. Then Induction Hypothesis F.4 holds for all $T_2 < t \le T_2^\epsilon = T_2 + O\Big(\frac{\log(P\epsilon^{-1})}{\eta\epsilon}\Big)$, and at iteration $t = T_2^\epsilon + 1$, we have*

1. $\widetilde{\mathcal{L}}_{k,\mathbf{p}}(Q^{T_2^\epsilon+1}) < \frac{\epsilon}{2K}$;

2. If $\mathsf{M} \in \mathcal{E}_{k,n}$, we have $(1 - \mathbf{Attn}^{(T_2^\epsilon+1)}_{\mathbf{p}\to\mathcal{P}_{k,n}})^2 \le O(\epsilon)$.

Proof. The existence of $T^\epsilon_{2,k} = T_{2,k} + O(\frac{\log(P\epsilon^{-1})}{\eta\epsilon})$ directly follows from Lemma F.26. We further derive

$$\begin{aligned}
&\widetilde{\mathcal{L}}_{k,\mathbf{p}}(Q^{T_2^\epsilon+1}) = \\
&\frac{1}{2}\mathbb{E}\left[\mathbb{1}\{k_X = k, \mathbf{p} \in \mathcal{M} \cap \mathsf{M} \in \mathcal{E}_{k,n}\} \left(z_n^2 \left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,n}}\right)^2 + \sum_{m\neq n} z_m^2 \left(\mathbf{Attn}_{n,m}\right)^2\right)\right]
\end{aligned}$$

$$\leq \frac{1}{2K} \cdot \gamma \cdot U^2 \cdot (1 + o(1)) \cdot O(\epsilon)$$
$$\leq \frac{\epsilon}{2K}$$

where the first inequality is due to direct calculations by the definition of T_2^ϵ, and the second inequality can be obtained by setting $c_{n,2}$ in (F.7) sufficiently large. □

G Analysis for Local Areas with Negative Information Gap

In this section, we focus on a specific patch $\mathbf{p} \in \mathcal{P}$ with the k-th cluster for $k \in [K]$, and present the analysis for the case that $X_\mathbf{p}$ is located in the local area for the k-th cluster, i.e. $a_{k,\boldsymbol{p}} > 1$. Throughout this section, we denote $a_{k,\boldsymbol{p}} = n$ for simplicity. When $\Delta \leq -\Omega(1)$, we can show that the gap of attention correlation changing rate for the positive case does not exist anymore, and conversely $\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \gg \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}$ from the beginning. We can reuse most of the gradient calculations in the previous section and only sketch them in this section.

Stage 1: we define stage 1 as all iterations $0 \leq t \leq T_{\text{neg},1}$, where

$$T_{\text{neg},1} \triangleq \max\left\{t : \Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \leq \frac{(1-\kappa_s)}{L}\log(P)\right\}.$$

We state the following induction hypothesis, which will hold throughout this stage:

Induction Hypothesis G.1. For each $0 \leq t \leq T_{\text{neg},1}$, $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$, the following holds:

a. $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}$ is monotonically increasing, and $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \in \left[0, \frac{(1-\kappa_s)}{L}\log(P)\right]$;

b. $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ is monotonically decreasing and $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \in \left[-O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{-\Delta}}\right), 0\right]$;

c. $|\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}| = O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\right)$ for $m \neq 1, n$;

d. $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\right)$ for $a_{k,\mathbf{q}} = n$, $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{p}}| = O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\right)$;

e. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\left(\frac{|\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\right) + O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\right)$ for $a_{k,\mathbf{q}} = 1$;

f. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\left(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\right)$ for $a_{k,\mathbf{q}} \neq 1, n$.

Through similar calculations for phase II, stage 1 in Appendix F.3, we obtain the following lemmas to control the gradient updates for attention correlations.

Lemma G.1. *If Induction Hypothesis E.2 and Induction Hypothesis G.1 hold for $0 \leq t \leq T_{\text{neg},1}$, then we have*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \geq \min\left\{\Omega\left(\frac{1}{P^{(1-\kappa_s)}}\right), \Omega\left(\frac{1}{P^{2(\frac{U}{L}-1)(1-\kappa_s)}}\right)\right\}, \tag{G.1a}$$

$$0 \geq \alpha^{(t)}_{\mathbf{p}\to v_{k,1}} \geq -O\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{-\Delta}}\right), \tag{G.1b}$$

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| \leq O\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\right) \textit{ for all } m \neq n, 1 \tag{G.1c}$$

$$\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = \Theta\left(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\right) \textit{ for } a_{k,\mathbf{q}} = n, \mathbf{q} \neq \mathbf{p} \tag{G.1d}$$

$$|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P}\Big) + O\Big(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\Big) \text{ for } a_{k,\mathbf{q}} = 1, \tag{G.1e}$$

$$|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{p}}|, |\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big) \quad \text{for all } a_{k,\mathbf{p}} \neq n, 1. \tag{G.1f}$$

Here $\Delta < 0$ implies $|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}| \ll \alpha^{(t)}_{\mathbf{p}\to v_{k,n}}$. Induction Hypothesis G.1 can be directly proved by Lemma G.1 and we have

$$T_{\text{neg},1} = O\Big(\frac{P^{\max\{1,2(\frac{U}{L}-1)\}\cdot(1-\kappa_s)}\log(P)}{\eta}\Big). \tag{G.2}$$

Stage 2: Given any $0 < \epsilon < 1$, define

$$T^{\epsilon}_{\text{neg},1} \triangleq \max\left\{t > T_1 : \Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \leq \log\left(c_6\left(\left(\frac{3}{\epsilon}\right)^{\frac{1}{2}} - 1\right)P^{1-\kappa_s}\right)\right\}. \tag{G.3}$$

where c_6 is some largely enough constant. We then state the following induction hypotheses, which will hold throughout this stage:

Induction Hypothesis G.2. For $n > 1$, suppose $\text{polylog}(P) \gg \log(\frac{1}{\epsilon})$, for $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$, and each $T_{\text{neg},1} < t \leq T^{\epsilon}_{\text{neg},1}$, the following holds:

a. $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}$ is monotonically increasing, and $\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} \in \Big[\frac{(1-\kappa_s)}{L}\log(P), O(\log(P/\epsilon))\Big]$;

b. $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ is monotonically decreasing and $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \in \Big[-O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{-\Delta}}\Big), 0\Big]$;

c. $|\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big)$ for $m \neq 1, n$;

d. $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}\Big)$ for $a_{k,\mathbf{q}} = n$, $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{p}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$;

e. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{|\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\Big) + O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} = 1$;

f. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,n}} - \Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} \neq 1, n$.

Lemma G.2. *If Induction Hypothesis E.2 and Induction Hypothesis G.2 hold for $T_{\text{neg},1} < t \leq T^{\epsilon}_{\text{neg},1}$, then we have*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} \geq \Omega(\epsilon), \tag{G.4a}$$

$$0 \geq \alpha^{(t)}_{\mathbf{p}\to v_{k,1}} \geq -O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P^{-\Delta}}\Big), \tag{G.4b}$$

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| \leq O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big) \text{ for all } m \neq n, 1 \tag{G.4c}$$

$$\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = \Theta(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{C_n}) \text{ for } a_{k,\mathbf{q}} = n, \mathbf{q} \neq \mathbf{p} \tag{G.4d}$$

$$|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}}}{P}\Big) + O\Big(\frac{|\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}|}{C_1}\Big) \text{ for } a_{k,\mathbf{q}} = 1, \tag{G.4e}$$

$$|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{p}}|, |\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,n}} - \alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big) \quad \text{for all } a_{k,\mathbf{p}} \neq n, 1. \tag{G.4f}$$

Induction Hypothesis G.2 can be directly proved by Lemma G.2. Furthermore, at the end of this stage, we will have:

Lemma G.3. *Suppose* $\text{polylog}(P) \gg \log(\frac{1}{\epsilon})$*, then Induction Hypothesis G.2 holds for all* $T_{\text{neg},1} < t \leq T^{\epsilon}_{\text{neg},1} = T_{\text{neg},1} + O\Big(\frac{\log(P\epsilon^{-1})}{\eta\epsilon}\Big)$*, and at iteration* $t = T^{\epsilon}_{\text{neg},1} + 1$*, we have*

1. $\widetilde{\mathcal{L}}_{k,\mathbf{p}}(Q^{T^{\epsilon}_{\text{neg},1}+1}) < \frac{\epsilon}{2K}$*;*
2. *If* $\mathsf{M} \in \mathcal{E}_{k,n}$ *, we have* $\Big(1 - \mathbf{Attn}^{(T^{\epsilon}_{\text{neg},1}+1)}_{\mathbf{p}\to\mathcal{P}_{k,n}}\Big)^2 \leq O(\epsilon)$.

H Analysis for the Global area

When $a_{\mathbf{p},k} = 1$, i.e. the patch lies in the global area, the analysis is much simpler and does not depend on the value of Δ. We can reuse most of the gradient calculations in Appendix F and only sketch them in this section.

For $X_{\mathbf{p}}$ in the global region $\mathcal{P}_{k,1}$, since the overall attention $\mathbf{Attn}^{(0)}_{\mathbf{p}\to\mathcal{P}_{k,1}}$ to the target feature already reaches $\Omega\Big(\frac{C_1}{P}\Big) = \Omega\Big(\frac{1}{P^{1-\kappa_c}}\Big)$ due to the large number of unmasked patches featuring $v_{k,1}$ when $\mathsf{M} \in \mathcal{E}_{k,1}$, which is significantly larger than $\mathbf{Attn}^{(0)}_{\mathbf{p}\to\mathcal{P}_{k,m}} = \Theta\Big(\frac{1}{P^{1-\kappa_s}}\Big)$ for all other $m > 1$. This results in large $\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}$ initially, and thus the training directly enters phase II.

Stage 1: we define stage 1 as all iterations $0 \leq t \leq T_{c,1}$, where

$$T_{c,1} \triangleq \max\left\{t : \Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \leq \frac{(1-\kappa_c)}{L}\log(P)\right\}.$$

We state the following induction hypotheses, which will hold throughout this stage:

Induction Hypothesis H.1. For each $0 \leq t \leq T_{c,1}$, $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$, the following holds:

a. $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ is monotonically increasing, and $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}} \in \Big[0, \frac{(1-\kappa_c)}{L}\log(P)\Big]$;

b. $\Phi_{\mathbf{p}\to v_{k,m}}$ is monotonically decreasing for $m > 1$ and $\Phi_{\mathbf{p}\to v_{k,m}} \in \Big[-O\big(\frac{\log(P)}{N}\big), 0\Big]$;

c. $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{C_1}\Big)$ for $a_{k,\mathbf{q}} = 1$, $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{p}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$;

d. $|\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big)$ for $a_{k,\mathbf{q}} \neq 1$.

Through similar calculations for phase II, stage 1 in Appendix F.3, we obtain the following lemmas to control the gradient updates for attention correlations.

Lemma H.1. *If Induction Hypothesis E.1 (or Induction Hypothesis E.2) and Induction Hypothesis H.1 hold for* $0 \leq t \leq T_{c,1}$*, then we have*

$$\alpha^{(t)}_{\mathbf{p}\to v_{k,1}} \geq \min\left\{\Omega\Big(\frac{1}{P^{(1-\kappa_c)}}\Big), \Omega\left(\frac{1}{P^{2(\frac{U}{L}-1)(1-\kappa_c)}}\right)\right\}, \tag{H.1a}$$

$$|\alpha^{(t)}_{\mathbf{p}\to v_{k,m}}| \leq O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P^{1-\kappa_s}}\Big) \quad \textit{for all } m \neq 1, \tag{H.1b}$$

$$\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}} = \Theta\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{C_1}\Big), \ \textit{for } a_{k,\mathbf{q}} = 1, \mathbf{q} \neq \mathbf{p}, \tag{H.1c}$$

$$|\beta^{(t)}_{k,\mathbf{p}\to\mathbf{p}}|, |\beta^{(t)}_{k,\mathbf{p}\to\mathbf{q}}| = O\Big(\frac{\alpha^{(t)}_{\mathbf{p}\to v_{k,1}}}{P}\Big) \quad \textit{for all } a_{k,\mathbf{q}} > 1. \tag{H.1d}$$

Induction Hypothesis H.1 can be directly proved by Lemma H.1 and we have

$$T_{c,1} = O\left(\frac{P^{\max\{1,2(\frac{U}{L}-1)\}\cdot(1-\kappa_c)}\log(P)}{\eta}\right). \tag{H.2}$$

Stage 2: Given any $0 < \epsilon < 1$, define

$$T_{c,1}^{\epsilon} \triangleq \max\left\{t > T_{c,1} : \Phi_{\mathbf{p}\to v_{k,1}}^{(t)} \leq \log\left(c_7\left(\left(\frac{3}{\epsilon}\right)^{\frac{1}{2}} - 1\right)P^{1-\kappa_c}\right)\right\}. \tag{H.3}$$

where c_7 is some largely enough constant. We then state the following induction hypotheses, which will hold throughout this stage:

Induction Hypothesis H.2. For $n > 1$, suppose $\text{polylog}(P) \gg \log(\frac{1}{\epsilon})$, $\mathbf{q} \in \mathcal{P} \setminus \{\mathbf{p}\}$, for each $T_{c,1} + 1 \leq t \leq T_{c,1}^{\epsilon}$, the following holds:

a. $\Phi_{\mathbf{p}\to v_{k,1}}^{(t)}$ is monotonically increasing, and $\Phi_{\mathbf{p}\to v_{k,1}}^{(t)} \in \left[\frac{(1-\kappa_c)}{L}\log(P), O(\log(P/\epsilon))\right]$;

b. $\Phi_{\mathbf{p}\to v_{k,m}}$ is monotonically decreasing for $n > 1$ and $\Phi_{\mathbf{p}\to v_{k,m}} \in \left[-O\left(\frac{\log(P)}{N}\right), 0\right]$;

c. $\Upsilon_{k,\mathbf{p}\to\mathbf{q}}^{(t)} = O\left(\frac{\Phi_{\mathbf{p}\to v_{k,1}}^{(t)}}{C_1}\right)$ for $a_{k,\mathbf{q}} = 1$, $|\Upsilon_{k,\mathbf{p}\to\mathbf{p}}^{(t)}| = O\left(\frac{\Phi_{\mathbf{p}\to v_{k,1}}^{(t)}}{P}\right)$;

d. $|\Upsilon_{k,\mathbf{p}\to\mathbf{q}}^{(t)}| = O\left(\frac{\Phi_{\mathbf{p}\to v_{k,1}}^{(t)}}{P}\right)$ for $a_{k,\mathbf{q}} \neq 1$.

We also have the following lemmas to control the gradient updates for attention correlations.

Lemma H.2. *If Induction Hypothesis E.1 (or Induction Hypothesis E.2) and Induction Hypothesis H.1 hold for $T_{c,1} + 1 \leq t \leq T_{c,1}^{\epsilon}$, then we have*

$$\alpha_{\mathbf{p}\to v_{k,1}}^{(t)} \geq \Omega(\epsilon), |\alpha_{\mathbf{p}\to v_{k,m}}^{(t)}| \leq O\left(\frac{\alpha_{\mathbf{p}\to v_{k,1}}^{(t)}}{P^{1-\kappa_s}}\right) \quad \textit{for all } m \neq 1 \tag{H.4a}$$

$$\beta_{k,\mathbf{p}\to\mathbf{q}}^{(t)} = \Theta\left(\frac{\alpha_{\mathbf{p}\to v_{k,1}}^{(t)}}{C_1}\right), \textit{ for } a_{k,\mathbf{q}} = 1, \mathbf{q} \neq \mathbf{p} \tag{H.4b}$$

$$|\beta_{k,\mathbf{p}\to\mathbf{p}}^{(t)}|, |\beta_{k,\mathbf{p}\to\mathbf{q}}^{(t)}| = O\left(\frac{\alpha_{\mathbf{p}\to v_{k,1}}^{(t)}}{P}\right) \quad \textit{for all } a_{k,\mathbf{q}} > 1. \tag{H.4c}$$

Induction Hypothesis H.2 can be directly proved by Lemma H.2. Furthermore, at the end of this stage, we will have:

Lemma H.3. *Suppose* $\text{polylog}(P) \gg \log(\frac{1}{\epsilon})$, *then Induction Hypothesis H.2 holds for all $T_{c,1} < t \leq T_{c,1}^{\epsilon} = T_{c,1} + O\left(\frac{\log(P\epsilon^{-1})}{\eta\epsilon}\right)$, and at iteration $t = T_{c,1}^{\epsilon} + 1$, we have*

1. $\widetilde{\mathcal{L}}_{k,\mathbf{p}}(Q^{T_{c,1}^{\epsilon}+1}) < \frac{\epsilon}{2K}$;
2. *If* $\mathsf{M} \in \mathcal{E}_{k,1}$, *we have* $\left(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,1}}^{(T_{c,1}^{\epsilon}+1)}\right)^2 \leq O(\epsilon)$.

I Proof of Main Theorems

I.1 Proof of Induction Hypotheses

We are now ready to show Induction Hypothesis E.1 (resp. Induction Hypothesis E.2) holds through the learning process.

Theorem I.1 (Positive Information Gap)**.** *For sufficiently large $P > 0$, $\eta \ll \log(P)$, $\Omega(1) \le \Delta < 1$, Induction Hypothesis E.1 holds for all iterations $t = 0, 1, \cdots, T = O\Big(\frac{e^{\text{polylog}(P)}}{\eta}\Big)$.*

Theorem I.2 (Negative Information Gap)**.** *For sufficiently large $P > 0$, $\eta \ll \log(P)$, $-0.5 < \Delta \le -\Omega(1)$, Induction Hypothesis E.2 holds for all iterations $t = 0, 1, \cdots, T = O\Big(\frac{e^{\text{polylog}(P)}}{\eta}\Big)$.*

Proof of Theorem I.1. It is easy to verify Induction Hypothesis E.1 holds at iteration $t = 0$ due to the initialization $Q^{(0)} = \mathbf{0}_{d\times d}$. At iteration $t > 0$:

- Induction Hypothesis E.1a. can be proven by Induction Hypothesis F.1-F.4 a and Induction Hypothesis H.1-H.2 a, combining with the fact that $\log(1/\epsilon) \ll \text{polylog}(P)$.
- Induction Hypothesis E.1b. can be obtained by invoking Induction Hypothesis F.1-F.4 b.
- Induction Hypothesis E.1c. can be obtained by invoking Induction Hypothesis F.1-F.4 c and Induction Hypothesis H.1-H.2 b.
- To prove Induction Hypothesis E.1d., for $\mathbf{q} \neq \mathbf{p}$, $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}} = \sum_{k=1}^{K} \Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}}$. By item d-f in Induction Hypothesis F.1-F.4 and item c-d in Induction Hypothesis H.1-H.2, we can conclude that no matter the relative areas $\mathbf{q}$ and $\mathbf{p}$ belong to for a specific cluster, for all $k \in [K]$, throughout the entire learning process, the following upper bound always holds:

$$\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{q}} \le \max_{t\in[T]}(|\Phi^{(t)}_{\boldsymbol{p}\to v_{k,n}}| + |\Phi^{(t)}_{\boldsymbol{p}\to v_{k,1}}|) \max\left\{O\Big(\frac{1}{C_1}\Big), O\Big(\frac{1}{C_n}\Big), O\Big(\frac{1}{P}\Big)\right\} \le \widetilde{O}\Big(\frac{1}{C_n}\Big).$$

  Moreover, since $K = \Theta(1)$, we then have $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}} = \widetilde{O}(\frac{1}{C_n})$, which completes the proof.
- The proof for Induction Hypothesis E.1d. is similar as before, by noticing that $\Upsilon^{(t)}_{k,\mathbf{p}\to\mathbf{p}} = \widetilde{O}(\frac{1}{P})$ for each $k \in [K]$, which is due to Induction Hypothesis F.1-F.4 d and Induction Hypothesis H.1-H.2 c.

The proof of Theorem I.2 mirrors that of Theorem I.1, with the only difference being the substitution of relevant sections with Induction Hypothesis E.2. For the sake of brevity, this part of the proof is not reiterated here.

I.2 Proof of Theorem 4.1 and Theorem 4.2 with Positive Information Gap

Theorem I.3. *Suppose $\Omega(1) \le \Delta \le 1$. For any $0 < \epsilon < 1$, suppose $\text{polylog}(P) \gg \log(\frac{1}{\epsilon})$. We apply GD to train the loss function given in (2.7) with $\eta \ll \text{poly}(P)$. Then for each $\mathbf{p} \in \mathcal{P}$, we have*

1. *The loss converges: after $T^\star = O\Big(\frac{\log(P)P^{\max\{2(\frac{U}{L}-1),1\}(1-\kappa_s)}}{\eta} + \frac{\log(P\epsilon^{-1})}{\eta\epsilon}\Big)$ iterations, $\mathcal{L}_{\mathbf{p}}(Q^{(T^\star)}) - \mathcal{L}^*_{\mathbf{p}} \le \epsilon$, where $\mathcal{L}^\star_{\mathbf{p}}$ is the global minimum of patch-level construction loss in (4.2).*
2. *Attention score concentrates: given cluster $k \in [K]$, if $X_{\mathbf{p}}$ is masked, then the one-layer transformer nearly "pays all attention" to all unmasked patches in the same area $\mathcal{P}_{k,a_{k,\mathbf{p}}}$, i.e., $\Big(1-\mathbf{Attn}^{(T^\star)}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}\Big)^2 \le O(\epsilon)$.*
3. **Local** *area learning feature attention correlation through* **two-phase***: given $k \in [K]$, if $a_{k,\mathbf{p}} > 1$, then we have*
   (a) *$\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ first quickly decrease with all other $\Phi^{(t)}_{\mathbf{p}\to v_{k,m}}$, $\Upsilon^{(t)}_{\mathbf{p}\to\mathbf{q}}$ not changing much;*
   (b) *after some point, the increase of $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ takes dominance. Such $\Phi^{(t)}_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}$ will keep growing until convergence with all other feature and positional attention correlations nearly unchanged.*
4. **Global** *area learning feature attention correlation through* **one-phase***: given $k \in [K]$, if $a_{k,\mathbf{p}} = 1$, throughout the training, the increase of $\Phi^{(t)}_{\mathbf{p}\to v_{k,1}}$ dominates, whereas all $A^{(t)}_{1,m}$ with $m \neq 1$ and position attention correlations remain close to 0.*

Proof. The first statement is obtained by letting $T^\star = \max\{T_2^\epsilon, T_{c,1}^\epsilon\} + 1$ in Lemma F.30 and Lemma H.3, combining wth Lemma D.9 and Lemma D.10, which lead to

$$\begin{aligned}
\mathcal{L}_{\mathbf{p}}(Q^{(T^\star)}) - \mathcal{L}_{\mathbf{p}}^* &\leq \mathcal{L}_{\mathbf{p}}(Q^{(T^\star)}) - \mathcal{L}_{\mathbf{p}}^{\text{low}} \\
&\leq \widetilde{\mathcal{L}}_{\mathbf{p}}(Q^{T^\star}) + O\Big(\exp\Big(-\big(c_3 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_4 P^{\kappa_s}\big)\Big)\Big) \\
&\leq K \cdot \frac{\epsilon}{2K} + O\Big(\exp\Big(-\big(c_3 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_4 P^{\kappa_s}\big)\Big)\Big) \\
&< \epsilon.
\end{aligned}$$

The second statement follows from Lemma F.30 and Lemma H.3. The third and fourth statements directly follow from the learning process described in Appendix F and Appendix H when Induction Hypothesis E.1 holds. □

I.3 Proof of Theorem 4.1 and Theorem 4.2 with Negative Information Gap

Theorem I.4. *Suppose* $-0.5 \leq \Delta \leq \Omega(1)$. *For any* $0 < \epsilon < 1$, *suppose* $\text{polylog}(P) \gg \log(\frac{1}{\epsilon})$. *We apply GD to train the loss function given in* (2.7) *with* $\eta \ll \text{poly}(P)$. *Then for each* $\mathbf{p} \in \mathcal{P}$, *we have*

1. *The loss converges: after* $T^\star = O\Big(\frac{\log(P) P^{\max\{2(\frac{U}{L}-1),1\}(1-\kappa_s)}}{\eta} + \frac{\log(P\epsilon^{-1})}{\eta\epsilon}\Big)$ *iterations,* $\mathcal{L}_{\mathbf{p}}(Q^{(T^\star)}) - \mathcal{L}_{\mathbf{p}}^* \leq \epsilon$, *where* $\mathcal{L}_{\mathbf{p}}^\star$ *is the global minimum of patch-level construction loss in* (4.2).
2. *Attention score concentrates: given cluster* $k \in [K]$, *if* $X_{\mathbf{p}}$ *is masked, then the one-layer transformer nearly "pays all attention" to all unmasked patches in the same area* $\mathcal{P}_{k,a_{k,\mathbf{p}}}$, *i.e.,* $\Big(1 - \mathbf{Attn}_{\mathbf{p}\to\mathcal{P}_{k,a_{k,\mathbf{p}}}}^{(T^\star)}\Big)^2 \leq O(\epsilon)$.
3. **All** *areas learning feature attention correlation through* **one-phase**: *given* $k \in [K]$, *throughout the training, the increase of* $\Phi_{\mathbf{p}\to v_{k,a_{k,\mathbf{p}}}}^{(t)}$ *dominates, whereas all* $\Phi_{\mathbf{p}\to v_{k,m}}^{(t)}$ *with* $m \neq 1$ *and position attention correlations* $\Upsilon_{\mathbf{p}\to\mathbf{q}}^{(t)}$ *remain close to* 0.

Proof. The first statement is obtained by letting $T^\star = \max\{T_{\text{neg},1}^\epsilon, T_{c,1}^\epsilon\} + 1$ in Lemma G.3 and Lemma H.3, combining wth Lemma D.9 and Lemma D.10, which lead to

$$\begin{aligned}
\mathcal{L}_{\mathbf{p}}(Q^{(T^\star)}) - \mathcal{L}_{\mathbf{p}}^* &\leq \mathcal{L}_{\mathbf{p}}(Q^{(T^\star)}) - \mathcal{L}_{\mathbf{p}}^{\text{low}} \\
&\leq \widetilde{\mathcal{L}}_{\mathbf{p}}(Q^{T^\star}) + O\Big(\exp\Big(-\big(c_3 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_4 P^{\kappa_s}\big)\Big)\Big) \\
&\leq K \cdot \frac{\epsilon}{2K} + O\Big(\exp\Big(-\big(c_3 P^{\kappa_c} + \mathbb{1}\left\{1 \notin \cup_{k\in[K]}\{a_{k,\mathbf{p}}\}\right\} c_4 P^{\kappa_s}\big)\Big)\Big) \\
&< \epsilon.
\end{aligned}$$

The second statement follows from Lemma G.3 and Lemma H.3. The third and fourth statements directly follow from the learning process described in Appendix G and Appendix H when Induction Hypothesis E.2 holds. □