THEORETICAL ANALYSIS OF MULTIMODAL CONTRASTIVE LEARNING: WHEN AND HOW HARD NEGATIVES HELP

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal contrastive learning has emerged as a powerful framework for learning visual and semantic representations, yet it often struggles to capture rare features. Hard negative sampling has been empirically shown to improve fine-grained recognition, but a rigorous understanding of when and how it helps remains limited. In this work, we develop a theoretical framework to analyze the training dynamics of multimodal contrastive learning with Transformer-based encoders. Our analysis reveals an implicit bias in vanilla contrastive learning toward common features, while rare features can remain weak or even be overlooked. Motivated by this finding, we show that hard negative samples can alleviate this imbalance by suppressing common features shared across negative pairs, thereby strengthening the learning of rare features. We further characterize when hard negative sampling becomes harmful, showing that excessive use of hard negatives can overly suppress common features and disrupt the balance between common and rare feature learning. Finally, we show that a simple selection strategy based on thresholding can mitigate this effect and enable more effective use of hard negative samples. Our main theoretical findings are verified by experiments on both synthetic data and standard benchmark datasets.

1 INTRODUCTION

Contrastive learning has become a powerful paradigm for representation learning by learning representations from positive and negative sample pairs (Ghosh & Lan, 2021; Zhong et al., 2022b; Xue et al., 2022; Çağatan et al., 2025; Cui et al., 2025). This approach has demonstrated remarkable success across a wide range of applications (Zhang et al., 2022; Jiang et al., 2023; Luo et al., 2023; Li et al., 2024b; Ji et al., 2024) and has become particularly important in multimodal learning (Jiang et al., 2024; Jing et al., 2024; Song et al., 2024), driving major advances in large-scale vision-language pretraining (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; 2022b). However, different features may not be learned equally during contrastive training. Samples can share broad, common features but differ in only subtle attributes that appear in a subset of the data. Such fine-grained features are therefore often less frequent across the overall data distribution, yet can provide critical information for distinguishing otherwise similar samples. In this work, we refer to these less frequent but discriminative features as rare features.

Existing efforts to better capture these rare features often rely on richer supervision or more fine-grained cross-modal alignment, for example, through longer captions or attribute descriptions (Zhang et al., 2024a; Kim & Ji, 2024; He et al., 2025), or through localized and multi-granularity cross-modal alignment (Yao et al., 2021; Zhong et al., 2022a; Li et al., 2022a; Ma et al., 2022). While effective, these approaches often rely on richer annotations or domain-specific knowledge, which may not always be readily available. A widely used strategy is hard negative sampling, which instead modifies the negative pairs used during training by identifying, selecting, or constructing more challenging negative samples. Because such negatives often share common features with the positive sample while differing in rare but discriminative features, they can encourage the model to better capture these less frequent features. Hard negative sampling has been widely adopted in practice and has shown empirical benefits for learning subtle and compositional distinctions (Yuk-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sekgonul et al., 2022; Cascante-Bonilla et al., 2023; Peng et al., 2024; Zhang et al., 2024b; Wang et al., 2025; Xie et al., 2025).

Despite the empirical success of hard negative sampling, most existing studies remain largely empirical, while its underlying theoretical mechanism remains less understood. Specifically, it remains unclear *when and how hard negative samples help the model learn rare but discriminative features*. Pioneering theoretical studies have provided important insights into feature learning in contrastive learning, including the effects of data augmentation, robustness to noisy pairs, and shared representations across modalities (Wen & Li, 2021; Nakada et al., 2023; Gui et al., 2026; Liao et al., 2026). However, these analyses typically consider randomly sampled negatives, whose effects can often be averaged out. As a result, the main learning dynamics can often be characterized primarily through positive pairs. This is no longer the case for hard negatives, since they are intentionally selected or constructed to be similar to the positive sample and can therefore exert a structured influence on which features are strengthened or suppressed during training. Consequently, the role of negative samples in feature learning remains insufficiently understood. A better theoretical understanding of this effect can not only explain when and why hard negatives are beneficial, but also provide practical guidance for designing more effective negative sampling strategies.

In this work, we provide a theoretical analysis of how multimodal contrastive learning captures different types of features and how hard negatives affect this learning process. We study a simplified but representative Transformer-MLP framework with single-head attention followed by an MLP with ReLU activations. To characterize feature learning, we consider structured data containing common features that appear frequently, such as general object-level characteristics, and rare but discriminative features that occur less often, such as fine-grained attributes. Under this framework, we characterize the training dynamics of common and rare features and reveal an implicit bias of vanilla contrastive learning toward common features. We further show how hard negatives can strengthen the learning of rare features, while overly aggressive use can excessively suppress common features. Motivated by these findings, we introduce a simple strategy based on thresholding to better balance the learning of common and rare features. Overall, our paper makes **contributions**:

First, we develop a theoretical framework for understanding how multimodal contrastive learning learns different types of features. Under a sparse-coding data model and a simplified Transformer-based architecture, we characterize how common semantic features, rare but discriminative features, and noise are learned during training. Our analysis shows that vanilla contrastive learning can effectively capture common semantic patterns and remain robust to noise, but still struggles to learn rare but discriminative features.

Second, we characterize when and how hard negative samples improve the learning of rare features. We show that hard negatives suppress the learning of common features, allowing common and rare features to be learned at more comparable strengths. Our analysis also identifies when this effect becomes excessive, showing that too many hard negatives can over-suppress common features and reverse the learning imbalance, causing rare features to dominate instead.

Third, we show that a simple thresholding-based selection strategy prevents overly hard negatives from dominating training. We further study a simple thresholding-based selection strategy that filters out hard negatives whose similarity to the anchor exceeds that of the corresponding positive sample. Our analysis shows that such a strategy can mitigate the excessive suppression of common features caused by overly hard negatives. We further provide numerical experiments that support all three theoretical insights under random negatives, hard negative sampling, and thresholding-based selection.

1.1 Related Work

Hard Negative Sampling in Contrastive Learning: Negative sample selection plays an important role in contrastive learning, as randomly sampled negatives can often provide limited learning signals. Hard negative sampling addresses this issue by selecting or constructing more challenging negative samples that are similar to positive samples but contain discriminative differences. Robinson et al. (2020) develop an unsupervised strategy for sampling hard negatives with controllable hardness, while Radenovic et al. (2023) incorporate hard negative importance sampling into vision-language pretraining. More recently, Zhang et al. (2024b) introduce intra-modal and cross-modal hard negatives to encourage models to distinguish subtle compositional differences and learn finer

image-text alignment. Other studies similarly utilize challenging negative samples to improve fine-grained vision-language understanding (Dönmez et al., 2023; Sahin et al., 2024; Oh et al., 2024). Beyond hard negative sampling, alternative approaches improve fine-grained representations through richer image-text supervision or finer cross-modal alignment (Zheng et al., 2024; Wu et al., 2024b; Lai et al., 2024; Fan et al., 2023; Urbanek et al., 2024; Li et al., 2022c; Wu et al., 2024a; Singh et al., 2023). Despite these empirical advances, how hard negative samples affect the learning dynamics of different features remains theoretically underexplored.

Convergence and Generalization Analysis of Contrastive Learning: Despite its empirical success, contrastive learning lacks a mature theoretical understanding, largely due to the complexity of its loss function. Early research investigates why augmentation is essential for the success of contrastive learning, showing that such an alignment between augmented positive pairs facilitates learning useful representations (Saunshi et al., 2019; Tian et al., 2020; Wen & Li, 2021; Saunshi et al., 2022). Some studies establish a connection between the gradients of contrastive learning and graph neural networks, highlighting interpretability through a graph-theoretic perspective (Tian et al., 2021; Wang et al., 2023). Other work also explores the connections between contrastive learning and graph theory, proposing a new loss function linked to graph spectral clustering to help explain its success (HaoChen et al., 2021). Another line of work emphasizes the necessity of data augmentation for breaking dependencies on spurious noise (Wen & Li, 2021). However, none of these works has theoretically characterized how hard negative samples affect the feature learning dynamics of multimodal contrastive learning.

Feature Learning Paradigm: The mathematical framework in this paper is closely related to the feature learning paradigm. Specifically, we assume the data follow a sparse coding model, which is a mixture of latent features, and study the training dynamics of model weights to examine how they align with these features. Most prior works focus on supervised learning (Allen-Zhu & Li, 2022; Cao et al., 2022; Zhang et al., 2023; Chowdhury et al., 2023; Li et al., 2025), where features are tied to ground-truth labels and such settings cannot be directly extended to contrastive learning. Because of the complexity of analyzing fine-grained training dynamics, existing studies are typically limited to simple one-hidden-layer neural networks, with some recent efforts exploring Transformers but still restricted to a single layer (Oymak et al., 2023; Huang & Wen, 2024; Li et al., 2024a), even under supervised settings. The most relevant works analyze the training dynamics of contrastive learning with one-hidden-layer feedforward networks (Wen & Li, 2021; Sun et al., 2025). In contrast, our paper studies Transformer architectures under a multimodal data model and provides a theoretical analysis of how hard negative samples affect the learning dynamics of common and rare features.

2 Problem Formulation and Algorithm

Contrastive Learning Framework. Let $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(L)}] \in \mathbb{R}^{d_1 \times L}$ and $\boldsymbol{Y} = [\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(L)}] \in \mathbb{R}^{d_1 \times L}$ denote the image and text input sequences with L tokens, respectively. The goal of contrastive learning is to learn two mappings, i.e.,

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^{d_1 \times L} \to \mathbb{R}^m \quad \textit{and} \quad \boldsymbol{g}_{\boldsymbol{\phi}}(\cdot) : \mathbb{R}^{d_1 \times L} \to \mathbb{R}^m, \tag{1}$$

such that semantically matched image–text pairs are mapped close to each other in the shared representation space, while mismatched pairs are well separated.

Motivated by the strong empirical performance of SigLIP (Zhai et al., 2023; Tschannen et al., 2025) in vision–language representation learning, we analyze the sigmoid-based contrastive objective, which is defined as

$$\ell(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}, \boldsymbol{X}_i, \boldsymbol{Y}_j) = -\log \frac{1}{1 + e^{-a_{i,j} \cdot \mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)}}. \tag{2}$$

where $a_{i,j} = +1$ if $\boldsymbol{X}_i$ and $\boldsymbol{Y}_j$ form a matched image-text pair, i.e., $i = j$, and $a_{i,j} = -1$ otherwise, i.e., $i \neq j$. The similarity function is defined as $\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j) := \big\langle \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{X}_i), \boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{Y}_j) \big\rangle$.

Then, let $\mathcal{B}$ denote a mini-batch of image-text pairs, the learning objective is to minimize the empirical risk with l_2-regularization as

$$\widehat{L}_{\mathrm{aug}}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}) = \frac{1}{|\mathcal{B}|} \sum_{(i,j) \in \mathcal{B}} \ell(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}, \boldsymbol{X}_i, \boldsymbol{Y}_j) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_F^2 + \frac{\lambda}{2} \|\boldsymbol{\phi}\|_F^2. \tag{3}$$

Model Architecture: Transformer-MLP. To facilitate theoretical analysis, both the image and text encoders consist of a single-head self-attention layer followed by a one-hidden-layer MLP head. Specifically, the embedding function $\boldsymbol{f_\theta}$ is defined as

$$\boldsymbol{f_\theta}(\boldsymbol{X}_i) = \big(f_1(\boldsymbol{X}_i;\boldsymbol{\theta}),\ldots,f_m(\boldsymbol{X}_i;\boldsymbol{\theta})\big)^\top \in \mathbb{R}^m, \text{with}$$
$$f_k(\boldsymbol{X}_i) = u_{f_k}\sum_{r=1}^{L}\Big[\mathrm{ReLU}\big(\langle \boldsymbol{w}_k, \mathrm{Attn}_r(\boldsymbol{X}_i)\rangle - b_{f_k}\big) - \mathrm{ReLU}\big(-\langle \boldsymbol{w}_k, \mathrm{Attn}_r(\boldsymbol{X}_i)\rangle - b_{f_k}\big)\Big], \tag{4}$$

where $\mathrm{Attn}_r(\boldsymbol{X})$ denotes the self-attention output at the r-th token position. Similarly, for the text encoder $\boldsymbol{g_\phi}$, we have

$$\boldsymbol{g_\phi}(\boldsymbol{Y}_j) = \big(g_1(\boldsymbol{Y}_j;\boldsymbol{\phi}),\ldots,g_m(\boldsymbol{Y}_j;\boldsymbol{\phi})\big)^\top \in \mathbb{R}^m, \text{with}$$
$$g_k(\boldsymbol{Y}_j) = u_{g_k}\sum_{r=1}^{L}\Big[\mathrm{ReLU}\big(\langle \boldsymbol{v}_k, \mathrm{Attn}_r(\boldsymbol{Y}_j)\rangle - b_{g_k}\big) - \mathrm{ReLU}\big(-\langle \boldsymbol{v}_k, \mathrm{Attn}_r(\boldsymbol{Y}_j)\rangle - b_{g_k}\big)\Big]. \tag{5}$$

Hard Negative Samples. Hard negative samples, which are negative examples that are highly similar to the corresponding positive samples and therefore difficult to distinguish, have been widely used to improve contrastive and multimodal representation learning. Compared with randomly sampled negatives, hard negatives typically provide more informative learning signals by requiring the model to distinguish samples that share substantial semantic content but differ in finer-grained characteristics. This makes hard negative sampling particularly relevant to fine-grained multimodal representation learning.

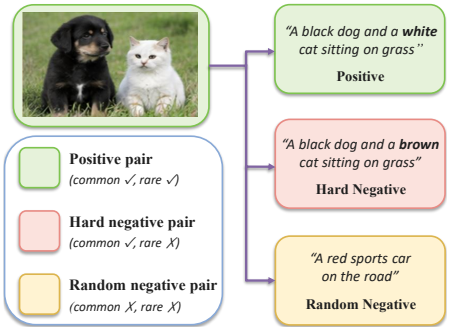

Figure 1: Examples of positive, hard negative, and random negative image-text pairs. The positive caption correctly describes both the common and rare features of the image. The hard negative caption preserves the common semantic content but differs in a rare feature, whereas the random negative caption is semantically unrelated to the image.

To better understand how hard negative samples affect contrastive learning, we distinguish them from random negatives by the features they share with their corresponding positive pair. Intuitively, a hard negative shares the common features of the positive sample but differs in its rare features, whereas a randomly sampled negative may differ from the positive sample in both common and rare features, as illustrated in Figure 1. A formal characterization of hard negative samples are provided in Assumption 3.5.

However, using hard negative samples too aggressively can overly suppress the learning of common features, causing the model to place excessive emphasis on rare features. To mitigate this effect, following existing strategies for hard negative mining (Zhang et al., 2024b), we further consider a thresholding mechanism that limits the influence of hard negative pairs during training. Specifically, at each iteration, a hard negative pair $(\boldsymbol{X}_i, \boldsymbol{Y}_j)$ is selected only when

$$|\mathrm{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_i) - \mathrm{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)| \le Th. \tag{6}$$

Intuitively, (6) shows that once the similarity gap between the positive pair and a hard-negative pair exceeds the threshold, the hard negative is no longer selected for training, thereby preventing excessive suppression of common features.

Our analysis focuses on three progressive questions. **First**, how do image and text features evolve during training under the sigmoid-based multimodal contrastive objective? **Second**, when and how do hard negative samples improve the learning of rare features and thereby enhance fine-grained recognition? **Third**, when can excessive use of hard negatives become harmful, and how can threshold-based selection mitigate this effect and balance the learning of common and rare features?

3 Main Theoretical Results

In what follows, we first summarize the main theoretical insights before presenting the technical details. These insights provide a high-level view of our formal results and facilitate their interpretation. We then introduce the assumptions underlying the analysis, followed by the formal theorem statements and theoretical guarantees.

3.1 KEY INSIGHTS OF THE FINDINGS

(K1) Contrastive learning with random negative samples favors common semantic features. With randomly sampled negatives, the model primarily learns common semantic features. Specifically, at convergence, the learned strengths of common semantic features, rare features, and noise are on the order of 1, $1/\sqrt{d_1}$, and $1/\sqrt{d_1}$, respectively (Theorem 3.1). Here, d_1 denotes the dimension of the raw input, with $d_1 \gg 1$. This separation shows that the model learns common features much more strongly than both rare features and noise components, suggesting both the success of contrastive learning in capturing dominant semantic information and its robustness to noise, but also its limitation in learning rare yet discriminative features.

(K2). Hard negative samples improve rare feature learning and suppress the learning of common features. With hard negative samples, the model can better balance the learning of common and rare features. Hard negatives share common features with the corresponding anchor while differing in rare features. As a result, the gradients induced by hard negative pairs partially offset the updates to neuron weights associated with these shared common features, reducing their effective learning rate from ϵ_1 to $(1-\rho)\epsilon_1$, where ϵ_1 and ϵ_2 denote the frequencies of common and rare features, respectively, and ρ denotes the proportion of shared common features in hard negative pairs. When $\rho = 1 - \Theta(\epsilon_2/\epsilon_1)$, common and rare features are learned at comparable strengths (Theorem 3.2). In this regime, common features no longer dominate the learned representation, allowing rare but discriminative features to be captured more effectively.

(K3). Excessive hard negative sampling can over-suppress common features, while a thresholding-based selection strategy mitigates this effect. Hard negative sampling is beneficial only when the suppression of common features is properly controlled. When too many hard negative samples are introduced such that $\rho \gg 1 - \epsilon_2/\epsilon_1$, the learning of common features becomes excessively suppressed, causing the neurons mainly to learn rare features (Theorem 3.3). To mitigate this effect, we introduce a thresholding-based selection strategy that automatically excludes overly hard negatives whose similarity to the anchor exceeds that of the corresponding positive sample. Removing these negatives weakens the negative gradient acting on shared common features, allowing common and rare features to be learned again at comparable strengths (Theorem 3.4).

3.2 ASSUMPTIONS

Data Model. Our data assumption is adopted from the sparse coding model, which has been widely used in theoretical analyses of neural networks (Wen & Li, 2021; Allen-Zhu & Li, 2022). It has also been extensively used as a modeling framework across diverse domains, including computer vision (Protter & Elad, 2008; Yang et al., 2009; Mairal et al., 2014), natural language processing (Arora et al., 2018), compressed sensing (Candès & Tao, 2010; Candes & Recht, 2012), and neuroscience (Olshausen & Field, 1997; Vinje & Gallant, 2000; Foldiak, 2003; Olshausen & Field, 2004).

Assumption 3.1 states that each token within a sample can be expressed as a weighted sum of a subset of features from the dictionary matrices $\boldsymbol{M^V}$ and $\boldsymbol{M^L}$, corrupted by additive noise $\boldsymbol{\xi}$. Here, $\boldsymbol{M^V}$ and $\boldsymbol{M^L}$ denote the dictionary matrices for the image and text modalities, respectively. $\boldsymbol{z}$ represents the latent signal, and $\boldsymbol{\xi}$ corresponds to spurious noise. Importantly, in the presence of noise, particularly when the noise level is comparable to or even exceeds the signal magnitude, no linear mapping can recover the latent signal directly from the input. This makes the model simple in form yet intrinsically challenging, thereby providing a favorable abstraction for theoretical analyses of nonlinear neural networks.

Assumption 3.1 (Sparse Coding Model)**.** *For a paired multimodal data $(\boldsymbol{X}_i, \boldsymbol{Y}_j)$, the data satisfies:*

$$\begin{aligned} \boldsymbol{X}_i &= \left[\boldsymbol{M^V}\boldsymbol{z}_i^{(1)} + \boldsymbol{\xi}_{X_i}^{(1)},\ \boldsymbol{M^V}\boldsymbol{z}_i^{(2)} + \boldsymbol{\xi}_{X_i}^{(2)},\ \ldots,\ \boldsymbol{M^V}\boldsymbol{z}_i^{(L)} + \boldsymbol{\xi}_{X_i}^{(L)}\right] \\ \boldsymbol{Y}_j &= \left[\boldsymbol{M^L}\boldsymbol{z}_j^{(1)} + \boldsymbol{\xi}_{Y_j}^{(1)},\ \boldsymbol{M^L}\boldsymbol{z}_j^{(2)} + \boldsymbol{\xi}_{Y_j}^{(2)},\ \ldots,\ \boldsymbol{M^L}\boldsymbol{z}_j^{(L)} + \boldsymbol{\xi}_{Y_j}^{(L)}\right] \end{aligned} \tag{7}$$

$\boldsymbol{z}_i^{(\ell)} \in \mathbb{R}^d$ and $\boldsymbol{z}_j^{(\ell)} \in \mathbb{R}^d$ represent the latent signals at the ℓ-th token for the image and text data respectively, while $\boldsymbol{\xi}_{X_i}^{(\ell)}$ and $\boldsymbol{\xi}_{Y_j}^{(\ell)}$ denote the corresponding additive noise terms. The matrices

$$\boldsymbol{M^V} = [\boldsymbol{M_1^V}, \ldots, \boldsymbol{M_d^V}] \in \mathbb{R}^{d_1 \times d}, \qquad \boldsymbol{M^L} = [\boldsymbol{M_1^L}, \ldots, \boldsymbol{M_d^L}] \in \mathbb{R}^{d_1 \times d}$$

are dictionary matrices for the image and text modalities respectively. Both $\boldsymbol{M}^{\boldsymbol{V}}$ *and* $\boldsymbol{M}^{\boldsymbol{L}}$ *are column-orthonormal matrices satisfying*

$$\|\boldsymbol{M}_p^{\boldsymbol{V}}\|_\infty \le \widetilde{O}\Big(\tfrac{1}{\sqrt{d_1}}\Big), \quad \|\boldsymbol{M}_q^{\boldsymbol{L}}\|_\infty \le \widetilde{O}\Big(\tfrac{1}{\sqrt{d_1}}\Big), \qquad \forall p, q \in [d].$$

We also assume $d_1 = \mathrm{poly}(d)$.

Assumption 3.2 provides a simplified mathematical abstraction of several structural properties that are commonly used to model real-world data in theoretical analysis. First, we assume that the latent signals are bounded and sparse. Sparsity is a standard assumption in prior studies, introduced primarily to facilitate the theoretical analysis, yet it also agrees with empirical observations that real-world data typically activate only a small subset of latent factors rather than spreading energy across all coordinates. Second, we assume sign consistency across tokens within the same sample, so that when a particular latent feature is active, its contribution has a consistent direction across different tokens. This captures the intuition that different parts of the same sample should generally provide coherent evidence for the same underlying feature rather than produce completely opposing signals. Finally, we distinguish common and rare features according to how frequently they occur across the data. Their occurrence probabilities are controlled by ϵ_1 and ϵ_2, respectively, with $\epsilon_1 > \epsilon_2$, so that common features appear more frequently while rare features occur less often.

Assumption 3.2 (Latent Signal)**.** *For the latent signal* $\{\boldsymbol{z}^{(i)}\}_{i=1}^L$ *with* $\boldsymbol{z}^{(i)} = (z_1^{(i)}, \ldots, z_j^{(i)}, \ldots, z_d^{(i)})^\top$*, we assume (i) all* $z_j^{(i)}$ *are bounded and symmetric around zero over all samples, and (ii)* $z_j^{(i)}$ *share the same sign across all* $i \in [L]$*. Moreover, we partition the feature dictionary into two disjoint subsets* $\mathcal{S}_1$ *and* $\mathcal{S}_2$*, where* $\mathcal{S}_1$ *denotes the set of common features and* $\mathcal{S}_2$ *denotes the set of rare features. For any* $j \in \mathcal{S}_1$ *and all* $i \in [L]$*, the sparse signal satisfies* $\Pr\left(|z_j^{(i)}| \neq 0\right) = \Theta\left(\epsilon_1 \frac{\log\log d}{d}\right)$*. For any* $j \in \mathcal{S}_2$ *and all* $i \in [L]$*, it satisfies* $\Pr\left(|z_j^{(i)}| \neq 0\right) = \Theta\left(\epsilon_2 \frac{\log\log d}{d}\right)$*, where* $\epsilon_1 > \epsilon_2$.

Remark 1 (Alternative characterization of rare features). We use feature frequency as the primary distinction for clarity and analytical convenience. However, the main theoretical insights are not specific to this choice. A similar distinction can instead be introduced through feature magnitude, e.g., by assigning smaller magnitudes to rare features, with only minor modifications to the proofs.

Assumption 3.3 states that noise follows Gaussian distributions. This is a mild condition, as no strong restriction is imposed on its variance. In particular, the noise magnitude can exceed that of the sparse signal when $d_1 \gg d$. The assumption is adopted for analytical purposes and demonstrates that contrastive learning can recover meaningful latent representations even in regimes where the signal is dominated by noise.

Assumption 3.3 (Noise)**.** *Here each noise term* $\boldsymbol{\xi}_{X_i}^{(\ell)}$ *and* $\boldsymbol{\xi}_{Y_j}^{(\ell)}$ *for* $\ell \in [L]$ *is independently drawn from the same distribution* $\boldsymbol{\xi}_{X_i}^{(\ell)}, \boldsymbol{\xi}_{Y_j}^{(\ell)} \sim \mathcal{N}(0, \sigma_\xi^2 \boldsymbol{I}_{d_1})$*, with variance* $\sigma_\xi^2 = \Theta\left(\frac{\sqrt{\log d}}{d}\right)$.

Assumption 3.4 states that a pair of positive samples shares the same latent signal across the image and text modalities. Intuitively, this means that the two samples encode the same underlying semantic structure, even though their observed token representations may differ due to modality-specific dictionary matrices and noise. In contrast, a negative pair is formed by two random samples whose latent signals are independent.

Assumption 3.4 (Positive and Random Negative Pairs)**.** *A pair of samples* $\boldsymbol{X}_i$ *and* $\boldsymbol{Y}_i$ *form a positive pair if they share the same latent signals* $\{\boldsymbol{z}_i^{(\ell)}\}_{\ell=1}^L$ *across all tokens. By contrast, a pair of samples* $\boldsymbol{X}_i$ *and* $\boldsymbol{Y}_j$ *with* $j \neq i$ *form a random negative pair, where the corresponding latent signals* $\{\boldsymbol{z}_i^{(\ell)}\}_{\ell=1}^L$ *and* $\{\boldsymbol{z}_j^{(\ell)}\}_{\ell=1}^L$ *are independent.*

Assumption 3.5 models the difference between random and hard negative pairs. For a random negative pair in Assumption 3.4, the latent features of the two samples are independent. A hard negative pair, however, can still share some semantic features even though the two samples do not form a positive pair. We capture this structure by dividing the feature coordinates into two subsets. The subset $\mathcal{S}_1$ represents common features shared by the image and text samples, while $\mathcal{S}_2$ represents rare features whose latent signals remain independent across the negative pair.

Assumption 3.5 (Hard Negative Samples)**.** *For each hard negative pair $(\boldsymbol{X}_i, \boldsymbol{Y}_j)$, we consider the common and rare feature subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ defined in Assumption 3.2. For any $p \in \mathcal{S}_1$, the image and text samples share the same latent signal, such that $z_{i,p} = z_{j,p}$. For any $p \in \mathcal{S}_2$, the corresponding latent signals are independent. We further divide the feature dictionary into common and rare features, denoted by $(\boldsymbol{M}^{\boldsymbol{V_1}}, \boldsymbol{M}^{\boldsymbol{L_1}})$ and $(\boldsymbol{M}^{\boldsymbol{V_2}}, \boldsymbol{M}^{\boldsymbol{L_2}})$, respectively.*

3.3 Formal Theoretical Results

We present our main theoretical results by characterizing feature learning in four different contrastive learning settings. (i) Theorem 3.1 shows that, under vanilla contrastive learning with $\epsilon_2 \ll \epsilon_1$, the neurons mainly learn common features, while rare features remain insufficiently learned. (ii) Theorem 3.2 shows that introducing hard negative samples reduces the effective learning rate of common features from ϵ_1 to $(1-\rho)\epsilon_1$, where ρ is correlated with the ratio of hard negative samples. When $\rho = 1-\Theta(\epsilon_2/\epsilon_1)$, this reduction balances the learning of common and rare features, allowing both types of features to be sufficiently learned. (iii) Theorem 3.3 further shows that excessive hard negative sampling, such that $(1-\rho)\epsilon_1 \ll \epsilon_2$, can overly suppress common features and cause the neurons to mainly learn rare features. (iv) Theorem 3.4 shows that thresholding mitigates the excessive suppression of common features and restores the balance condition $\rho = 1-\Theta(\epsilon_2/\epsilon_1)$, allowing both common and rare features to be sufficiently learned.

To simplify the presentation, we state the main theorem only for the image encoder, while the corresponding result for the text encoder follows analogously. The precise definitions of technical quantities appearing in the theorem statements, such as T_2 and $\mathcal{N}_k$, are deferred to the appendix.

3.3.1 Vanilla Contrastive Learning

Theorem 3.1 characterizes feature learning under vanilla contrastive learning. When $\epsilon_2 \ll \epsilon_1$, common features are learned much more strongly than rare features due to their higher occurrence frequency. As a result, the neurons mainly learn common features, while the rare feature components remain small.

Theorem 3.1 (**Vanilla Contrastive Learning**)**.** *Under vanilla contrastive learning with $\epsilon_2 \ll \epsilon_1$, for each neuron $k \in [m]$ and $t \geq T_2$, the learned neuron weight in the image encoder can be decomposed as*

$$\boldsymbol{w}_k^{(t)} = \sum_{p\in\mathcal{N}_k} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p\in\mathcal{S}_2} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p\in\mathcal{S}_1\setminus\mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p\in[d_1]\setminus[d]} \beta_{k,p} (\boldsymbol{M}^{V})_p^{\perp}, \tag{8}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega(1), \quad \alpha_{k,p}^{V_2} \leq o\left(\frac{1}{\sqrt{d_1}}\right), \quad \alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \quad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{9}$$

An analogous result holds for the neuron weights in the text encoder.

Remark 2. Theorem 3.1 characterizes how the learned neuron weights represent different latent feature directions. Depending on its initialization, each neuron $\boldsymbol{w}_k^{(t)}$ primarily learns a subset of common features, indexed by $\mathcal{N}_k$. For $p \in \mathcal{N}_k$, the coefficient along the common-feature direction $\boldsymbol{M}_p^{V_1}$ converges to constant order, while the coefficients of non-primary common features remain much smaller, on the order of $1/\sqrt{d}$. Thus, the overall latent representation of common features is distributed across different neuron weights. In contrast, the coefficients along the rare-feature directions $\boldsymbol{M}_p^{V_2}$ remain very small, at nearly the same order as the noise component $(\boldsymbol{M}^{V})^{\perp}$. Therefore, vanilla contrastive learning collectively captures common features across different neurons without substantially learning the noise directions, while rare features remain only weakly represented.

3.3.2 Contrastive Learning with Hard Negative Samples

Theorem 3.2 characterizes how hard negative samples change feature learning during contrastive learning. Hard negative samples reduce the effective learning rate of common features from ϵ_1 to $(1-\rho)\epsilon_1$, while the learning rate of rare features remains ϵ_2. Therefore, when $\rho = 1-\Theta(\epsilon_2/\epsilon_1)$, the learning rates of common and rare features become comparable, allowing the neurons to sufficiently learn both types of features.

Theorem 3.2 (**Contrastive Learning with Hard Negative Samples**). *Suppose a fraction ρ of the negative samples are hard negatives, with $\rho = 1 - \Theta(\epsilon_2/\epsilon_1)$. Then, for each neuron $k \in [m]$ and $t \geq T_2$, the learned neuron weight in the image encoder can be decomposed as*

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k \cap \mathcal{S}_1} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k \cap \mathcal{S}_2} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{10}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega(1), \quad \alpha_{k,p}^{V_2} \geq \Omega(1), \quad \alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \quad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{11}$$

An analogous result holds for the neuron weights in the text encoder.

Remark 3. Theorem 3.2 shows that each neuron $\boldsymbol{w}_k^{(t)}$ still learns an initialization-dependent subset of primary features indexed by $\mathcal{N}_k$. With an appropriate fraction of hard negative samples, however, both the common features $\boldsymbol{M}_p^{V_1}$ for $p \in \mathcal{N}_k \cap \mathcal{S}_1$ and the rare features $\boldsymbol{M}_p^{V_2}$ for $p \in \mathcal{N}_k \cap \mathcal{S}_2$ are learned at constant order. In contrast, non-primary feature directions remain small, on the order of $1/\sqrt{d}$, and the noise components remain negligible. Therefore, across different neurons, hard negative sampling enables the model to collectively capture both common and rare latent features without substantially learning noise directions.

3.3.3 Contrastive Learning with Excessive Hard Negative Samples

Theorem 3.3 characterizes feature learning when excessive hard negative samples are introduced. When $(1-\rho)\epsilon_1 \ll \epsilon_2$, the effective learning rate of common features becomes much smaller than that of rare features. As a result, the learning of common features is excessively suppressed, while the neurons mainly learn rare features.

Theorem 3.3 (**Contrastive Learning with Excessive Hard Negative Samples**). *Suppose too many hard negative samples are introduced, with $(1-\rho)\epsilon_1 \ll \epsilon_2$. Then, for each neuron $k \in [m]$ and $t \geq T_2$, the learned neuron weight in the image encoder can be decomposed as*

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{S}_1} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \in \mathcal{S}_2 \setminus \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{12}$$

where

$$\alpha_{k,p}^{V_1} \leq o\left(\frac{1}{\sqrt{d_1}}\right), \quad \alpha_{k,p}^{V_2} \geq \Omega(1), \quad \alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \quad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{13}$$

An analogous result holds for the neuron weights in the text encoder.

Remark 4. Theorem 3.3 exhibits a nearly mirrored behavior of Theorem 3.1. Under vanilla contrastive learning, the neuron weights primarily capture common features while rare features remain weak. When too many hard negatives are introduced, this behavior is reversed. Specifically, the common-feature components are suppressed to a scale comparable to noise, while each neuron learns its initialization-dependent subset of rare features with coefficients converging to constant order. Therefore, excessive hard negative sampling shifts the learned representation from being dominated by common features to being dominated by rare features, rather than achieving a balanced representation of both.

3.3.4 Contrastive Learning with Thresholded Hard Negative Samples

Theorem 3.4 characterizes how the threshold function controls the effect of hard negative samples on feature learning. By filtering out hard negative samples that would excessively suppress the shared common features, the threshold function weakens the negative gradient acting on these features. As a result, the excessive suppression of common features is mitigated, allowing the neurons to sufficiently learn both common and rare features.

Theorem 3.4 (**Contrastive Learning with Thresholded Hard Negative Samples**). *Suppose the thresholding-based selection mechanism introduced in (6) is applied. Then, for each neuron $k \in [m]$ and $t \geq T_2$, the learned neuron weight in the image encoder can be decomposed as*

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k \cap \mathcal{S}_1} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k \cap \mathcal{S}_2} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{14}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega(1), \quad \alpha_{k,p}^{V_2} \geq \Omega(1), \quad \alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \quad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{15}$$

An analogous result holds for the neuron weights in the text encoder.

Remark 5. Theorem 3.4 shows that the thresholding-based selection mechanism restores balanced learning of common and rare features. For each neuron, its initialization-dependent primary common features $\boldsymbol{M}_p^{V_1}$ and rare features $\boldsymbol{M}_p^{V_2}$ are both learned with coefficients converging to constant order, while the non-primary feature directions and noise components remain small. Therefore, across different neurons, the model collectively captures both common and rare latent features without substantially learning noise directions. In particular, compared with the excessive hard negative setting in Theorem 3.3, thresholding prevents common features from being overly suppressed and recovers a more balanced representation of the two types of features.

4 Numerical Experiments

Experiments on ARO, VALSE, VSR, and FOIL-COCO. Table 1 reports the results on ARO, VALSE, VSR, and FOIL-COCO, comparing three fine-tuning strategies based on the pretrained SigLIP2 model: random negative samples (SigLIP2 + Random), hard negative samples (SigLIP2 + HN), and hard negative samples with thresholding (SigLIP2 + HN + Thresholding). Random negative samples serve as the baseline, while hard negative samples encourage the model to distinguish fine-grained differences between image-text pairs, and thresholding controls their effect during fine-tuning. Additional experiments on synthetic data that further validate our theoretical insights under controlled settings are deferred to Appendix A.2.

We report accuracy (%) on all benchmarks, where a prediction is correct if the model assigns a higher similarity score to the correct image-text pair than to the corresponding negative pair. ARO evaluates compositional understanding of relations and attributes; VALSE evaluates linguistic phenomena including actions, coreference, Foil-it!, plurality, and spatial relations; VSR evaluates visual spatial reasoning; and FOIL-COCO evaluates the ability to distinguish correct image descriptions from incorrect ones. Hard negative samples consistently improve performance over random negative samples across all benchmarks, while thresholding provides further improvements. These results support the effectiveness of hard negative samples for fine-grained understanding and show that thresholding further improves performance by controlling their effect during fine-tuning.

Table 1: Accuracy (%) on ARO, VALSE, VSR, and FOIL-COCO. We compare fine-tuning with random negative samples (Random), hard negative samples (HN), and hard negative samples with thresholding (HN + Thresholding).

Model	ARO		VALSE					VSR	FOIL-COCO
	Relation	Attribute	Actions	Coreference	Foil-it!	Plurality	Sp.rel.		
SigLIP2 + HN + Thresholding	64.72	50.42	58.90	51.55	70.63	56.05	54.21	79.09	75.47
SigLIP2 + HN	63.53	50.16	57.32	49.72	70.31	55.70	52.90	72.68	72.86
SigLIP2 + Random	62.65	50.04	56.38	48.87	69.78	52.88	51.59	70.99	71.46

5 Conclusion

Our work provides a theoretical analysis of the feature learning dynamics of multimodal contrastive learning with Transformer-based encoders. Specifically, we show that vanilla contrastive learning preferentially learns frequently occurring common features, while rare but discriminative features can remain insufficiently learned. We further theoretically characterize when and how hard negative samples improve rare feature learning, showing that they suppress the learning of common features shared across negative pairs and thereby balance the learning of common and rare features. However, introducing too many hard negative samples can excessively suppress common features and reverse this learning imbalance. Motivated by this theoretical characterization, we further show that thresholding can control the influence of hard negative samples and restore balanced feature learning. Our theoretical findings are supported by numerical experiments on both synthetic data and standard fine-grained vision-language benchmarks. An important direction for future work is to develop a broader theoretical understanding of alternative approaches beyond hard negative sampling that can promote the learning of rare features.

AI USE STATEMENT

We used ChatGPT for language polishing of this paper. For numerical experiments, we used ChatGPT and Codex to assist with checking our code for potential bugs.

ETHICS STATEMENT

This work is primarily theoretical and does not involve human subjects, personal data, or other sensitive information. We do not identify any immediate ethical concerns beyond the broader considerations associated with the development and use of machine learning methods.

REPRODUCIBILITY STATEMENT

For the theoretical results, we present the assumptions and main results in the main text, while proof sketches and complete proofs are included in the appendix to facilitate reproduction of our main theoretical findings. For the numerical experiments, we provide the experimental settings, implementation details, and hyperparameter configurations in the appendix. A demo is also included in the supplementary material to further support reproducibility.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd annual symposium on foundations of computer science (FOCS)*, pp. 977–988. IEEE, 2022.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

Ömer Veysel Çağatan, Ömer Faruk Tal, and M Emre Gursoy. Adversarial robustness of discriminative self-supervised learning in vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2313–2324, 2025.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE transactions on information theory*, 56(5):2053–2080, 2010.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20155–20165, 2023.

Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6074–6114. PMLR, 2023.

Jingyi Cui, Yi-Ge Zhang, Hengyu Liu, and Yisen Wang. An inclusive theoretical framework of robust supervised contrastive loss against label noise. *arXiv preprint arXiv:2501.01130*, 2025.

Esra Dönmez, Pascal Tilli, Hsiu-Yu Yang, Ngoc Thang Vu, and Carina Silberer. Hnc: Leveraging hard negative captions towards models with fine-grained visual-linguistic comprehension capabilities. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 364–388, 2023.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

Peter Foldiak. Sparse coding in the primate cortex. *The handbook of brain theory and neural networks*, 2003.

Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2697–2702. IEEE, 2021.

Yu Gui, Cong Ma, and Zongming Ma. Multi-modal contrastive learning adapts to intrinsic dimensions of shared latent variables. *Advances in Neural Information Processing Systems*, 38: 122364–122409, 2026.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 89927–89944, 2025.

Yu Huang and Zixin Wen. Yuejie chi, and yingbin liang. transformers provably learn feature-position correlations in masked image modeling. *arXiv preprint arXiv:2403.02233*, 2024.

Chen Ji, Su Yang, Hongyu Sun, and Yuqing Zhang. Applying contrastive learning to code vulnerability type classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11942–11952, 2024.

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

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27036. IEEE, 2024.

Yangqin Jiang, Chao Huang, and Lianghao Huang. Adaptive graph contrastive learning for recommendation. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4252–4261, 2023.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024.

Jeonghwan Kim and Heng Ji. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6187–6207, 2024.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pp. 111–127. Springer, 2024.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? *arXiv preprint arXiv:2402.15607*, 2024a.

Hongkang Li, Yihua Zhang, Pin-Yu Chen, Sijia Liu, Meng Wang, et al. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *International Conference on Learning Representations*, volume 2025, pp. 101273–101305, 2025.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022a.

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

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

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022c.

Mingxin Li, Richong Zhang, and Zhijie Nie. Towards better understanding of contrastive sentence representation learning: A unified paradigm for gradient. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14506–14521, 2024b.

Haixu Liao, Yating Zhou, Songyang Zhang, Meng Wang, and Shuai Zhang. Theoretical analysis of contrastive learning under imbalanced data: From training dynamics to a pruning solution. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=DUXG9E8dEO`.

Dongsheng Luo, Wei Cheng, Yingheng Wang, Dongkuan Xu, Jingchao Ni, Wenchao Yu, Xuchao Zhang, Yanchi Liu, Yuncong Chen, Haifeng Chen, et al. Time series contrastive learning with information-aware augmentations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 4534–4542, 2023.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 638–647, 2022.

Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4348–4380. PMLR, 2023.

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multi-modal capabilities of pre-trained vlms for improving vision-linguistic compositionality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19060–19076, 2024.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, pp. 26724–26768. PMLR, 2023.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13279–13288. IEEE, 2024.

Matan Protter and Michael Elad. Image sequence denoising via sparse and redundant representations. *IEEE transactions on Image Processing*, 18(1):27–35, 2008.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6967–6977. IEEE, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5551–5561. IEEE, 2024.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pp. 5628–5637. PMLR, 2019.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 869–893, 2023.

Qi Song, Tianxiang Gong, Shiqi Gao, Haoyi Zhou, and Jianxin Li. Quest: quadruple multimodal contrastive learning with constraints and self-penalization. *Advances in Neural Information Processing Systems*, 37:28889–28919, 2024.

Jiawei Sun, Shuai Zhang, Hongkang Li, and Meng Wang. Contrastive learning with data misalignment: Feature purity, training dynamics and theoretical generalization guarantees. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26690–26699. IEEE, 2024.

William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

Yeyuan Wang, Dehong Gao, Lei Yi, Linbo Jin, Jinxia Zhang, Libin Yang, and Xiaoyan Cai. Enhancing fine-grained vision-language pretraining with negative augmented samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8060–8068, 2025.

Yifei Wang, Qi Zhang, Tianqi Du, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. A message passing perspective on learning dynamics of contrastive learning. *arXiv preprint arXiv:2303.04435*, 2023.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *International conference on learning representations*, volume 2024, pp. 52484–52503, 2024a.

Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. Lotlip: Improving language-image pre-training for long text understanding. *Advances in Neural Information Processing Systems*, 37:64996–65019, 2024b.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.

Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pp. 24851–24871. PMLR, 2022.

Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *2009 IEEE Conference on computer vision and pattern recognition*, pp. 1794–1801. Ieee, 2009.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pp. 310–325. Springer, 2024a.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13774–13784. IEEE, 2024b.

Shuai Zhang, Meng Wang, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Miao Liu. Joint edge-model sparse learning is provably efficient for graph neural networks. *arXiv preprint arXiv:2302.02922*, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pp. 73–90. Springer, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16772–16782. IEEE, 2022a.

Yuanyi Zhong, Haoran Tang, Junkun Chen, Jian Peng, and Yu-Xiong Wang. Is self-supervised learning more robust than supervised learning? *arXiv preprint arXiv:2206.05259*, 2022b.

A Overview of the Appendix and Proof Sketch

The appendices are organized systematically to provide supporting materials for the main text.

Appendix A summarizes the proof sketches and presents additional experiments on synthetic data to further support our theoretical claims.

Appendix B introduces the key notation, definitions, and preliminary results used throughout the theoretical analysis, which we recommend readers become familiar with before proceeding to the main proofs.

Appendix C presents a simplified theoretical framework for analyzing the training dynamics and convergence of sigmoid-based contrastive learning with Transformer–MLP encoders under random negative sampling, without distinguishing between common and rare features. This analysis serves as the foundation for the subsequent proofs. The remaining theoretical results follow the same overall proof framework, with modifications to several key steps to account for the different negative-sampling mechanisms and feature structures considered in the main theorems. We therefore recommend that readers review this appendix before proceeding to the main theoretical results.

Building on this general framework, Appendix D analyzes vanilla contrastive learning (Theorem 3.1) and contrastive learning with hard negative samples (Theorem 3.2), characterizing how hard negative samples balance the learning of common and rare features. Appendix E further analyzes contrastive learning with excessive hard negative samples (Theorem 3.3) and shows how thresholding mitigates the excessive suppression of common features and restores balanced feature learning (Theorem 3.4).

Finally, Appendix F provides the detailed proofs for the general training dynamics and convergence results presented in Appendix C, which constitute the majority of the appendix. We recommend that readers first consult the corresponding proof sketches before turning to the full technical proofs.

A.1 Proof Sketch

We first analyze the training dynamics of vanilla contrastive learning. Specifically, we divide the training process into two stages and characterize how the feature and non-feature components of the neuron weights evolve throughout training, followed by the convergence guarantee.

In Stage 1, we analyze how the neuron weights in the image and text modalities reinforce each other along the corresponding feature directions while remaining suppressed along the non-feature directions. For a positive image-text pair, the gradient of the image-side neuron weight $\boldsymbol{w}_k$ along feature direction $\boldsymbol{M}_p^V$ is determined by the alignment of the corresponding text-side neuron weight $\boldsymbol{v}_k$ with $\boldsymbol{M}_p^L$, and vice versa. This cross-modal dependence produces an incremental term of order $\epsilon_p \frac{\eta \log\log d}{d}$, so features with larger frequency ϵ_p grow faster during training. In contrast, standard negative pairs have zero expected gradient along the feature directions because their latent signals are independent, while the expected gradients along the non-feature directions are also zero due to the orthogonality between the feature and non-feature subspaces. As a result, the feature components gradually dominate the neuron weights, while the non-feature components remain essentially unchanged up to negligible variation (see Appendix F.1, Lemmas F.1 and F.2). At iteration T_1, the neuron weights become concentrated in the feature subspace, and lucky neurons develop significantly stronger alignment with their associated feature directions (see Appendix F.1, Theorem F.1).

In Stage 2, the feature directions established in Stage 1 are further strengthened. In particular, if neuron k is a lucky neuron for feature p, its projection onto $\boldsymbol{M}_p^V$ grows to at least $\Omega(1)\|\boldsymbol{w}_k^{(T_2)}\|_2$, while the projection of an ordinary neuron in $\mathcal{M}_p$ remains bounded by a constant multiple of the corresponding projection of a lucky neuron (see Appendix F.2, Theorem F.2). After T_2, each neuron can be decomposed into a set of learned feature components indexed by $\mathcal{N}_k$, weak components along the remaining feature directions, and negligible components along the non-feature directions (see Appendix F.2, Theorem F.3). Building on this feature decomposition, Appendix F.3 constructs an optimal learner with pseudo loss $O(e^{-\alpha})$ and establishes the final convergence guarantee.

We next analyze how hard negative samples change the learning dynamics of common and rare features. Under vanilla contrastive learning, when $\epsilon_2 \ll \epsilon_1$, common features are learned much more strongly than rare features. In contrast to standard negative pairs, whose latent signals are

independent, hard negative pairs can share common features with the corresponding positive pairs. For the same feature coordinate, we divide the features into two subsets S_1 and S_2, where features in S_1 are shared across the hard negative pair, while the corresponding latent signals for features in S_2 remain independent. Let ρ denote the proportion of shared common features in S_1. The negative gradient associated with these shared features suppresses the positive learning signal along common feature directions, reducing their effective learning rate from ϵ_1 to $(1-\rho)\epsilon_1$. Consequently, when $\rho = 1-\Theta(\epsilon_2/\epsilon_1)$, the learning strengths of common and rare features become comparable, allowing both types of features to be sufficiently learned (see Appendix D, Theorem D.2).

Finally, we analyze the effect of thresholding when too many hard negative samples are introduced. When $(1-\rho)\epsilon_1 \ll \epsilon_2$, hard negative sampling excessively suppresses common features, causing rare features to dominate the learned representations while common feature components remain weak. To control this effect, we retain hard negative samples only when

$$|\mathrm{sim}(\boldsymbol{X}_i, \boldsymbol{Y}_i) - \mathrm{sim}(\boldsymbol{X}_i, \boldsymbol{Y}_j)| \leq Th.$$

Thresholding filters out hard negative samples that would otherwise excessively suppress shared common features, thereby reducing the effective proportion ρ of shared common features involved in hard-negative training. As a result, the effective learning rate $(1-\rho)\epsilon_1$ of common features increases. When $\rho = 1 - \Theta(\epsilon_2/\epsilon_1)$, the learning strengths of common and rare features become comparable again. Therefore, thresholding prevents common features from being excessively suppressed and restores the balance between common and rare feature learning (see Appendix E, Theorem E.2).

A.2 Synthetic Experimental Settings and Results

Experiment 1 (Feature alignment dynamics). We evaluate how hard negative samples affect the learning dynamics of common and rare features. Specifically, we measure the squared cosine alignment between the learned neuron weights and the corresponding common and rare feature directions throughout training. As shown in Figure 2, with random negative samples, the alignment with common features continuously increases and reaches 0.3301, whereas the alignment with rare features remains close to zero, with a final value of only 0.0003. This demonstrates that vanilla contrastive learning preferentially learns common features while rare features remain weak. In contrast, after introducing hard negative samples, the final common feature alignment decreases substantially to 0.0715, while the rare feature alignment increases to 0.1946. These results demonstrate that hard negative samples suppress the learning of common features shared across negative pairs and promote the learning of rare features, providing empirical support for our theoretical characterization of how hard negatives balance common and rare feature learning.

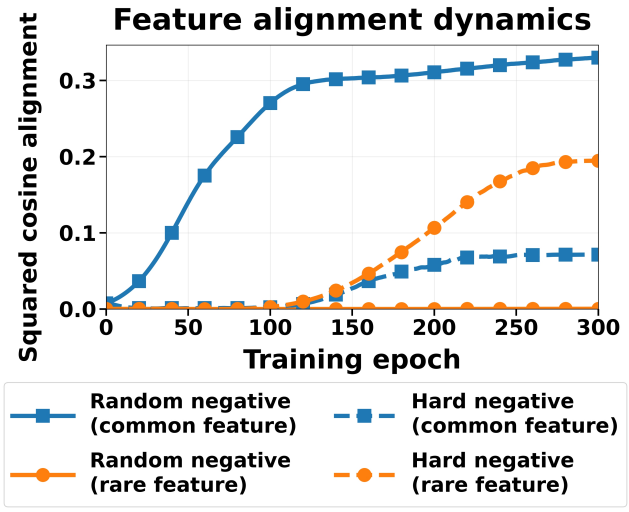

Figure 2: Feature alignment dynamics under random and hard negative sampling.

Experiment 2 (Feature subspace learning dynamics). We further evaluate how hard negative samples affect the learning dynamics of the entire feature subspace. Specifically, we measure the

proportion of each neuron's energy projected onto the full feature subspace and report the average over neurons and random trials. As shown in Figure 3, both methods start with approximately 20% of their energy in the feature subspace. With hard negative samples, the feature-subspace ratio increases much faster, reaching 94.72% at epoch 50 and eventually converging to 99.77%. In comparison, random negative sampling reaches only 49.14% at epoch 50 and gradually converges to 96.47%. These results show that although both methods eventually learn representations dominated by the true feature subspace, hard negative samples substantially accelerate this learning process and produce representations that are more strongly concentrated in the feature subspace.

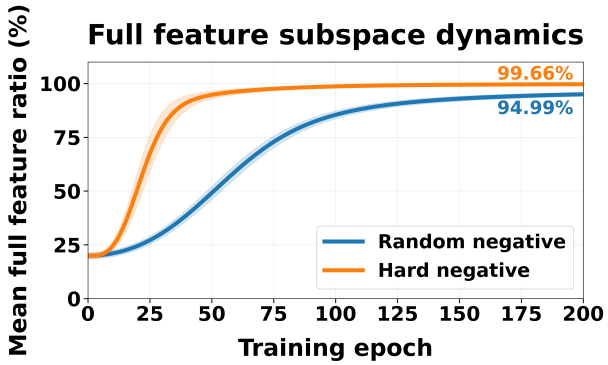

Figure 3: Feature subspace learning dynamics under random and hard negative sampling.

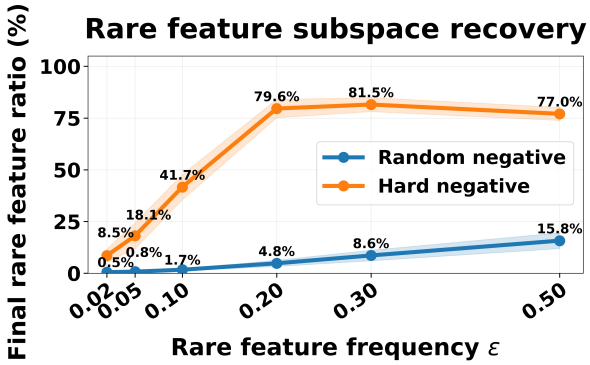

Figure 4: Rare feature subspace recovery under different rare feature frequencies.

Experiment 3 (Rare feature subspace recovery). We evaluate how the frequency of rare features affects their recovery under random and hard negative sampling. Specifically, we vary the rare-feature frequency ϵ_2 and measure the final proportion of neuron energy projected onto the rare-feature subspace. As shown in Figure 4, with random negative samples, the rare feature subspace ratio increases only gradually from 0.5% to 15.8% as ϵ_2 increases from 0.02 to 0.5, indicating that rare features remain weakly learned even as they become more frequent. In contrast, hard negative sampling substantially improves rare feature recovery across different frequencies: the ratio increases from 8.5% at $\epsilon_2 = 0.02$ to 41.7% at $\epsilon_2 = 0.1$, and reaches around 80% when ϵ_2 is further increased to 0.2–0.3. These results demonstrate that rare features become easier to learn as their frequency increases, while hard negative samples substantially strengthen their recovery, providing empirical support for our theoretical characterization of rare feature learning.

Experiment 4 (Downstream classification). We evaluate whether the improved learning of rare features leads to better downstream fine grained classification performance. Specifically, we freeze the pretrained image encoder and train a linear classifier using different numbers of labeled downstream samples. As shown in Figure 5, hard negative sampling consistently achieves higher test accuracy than random negative sampling across all training sample sizes. With 40 training samples,

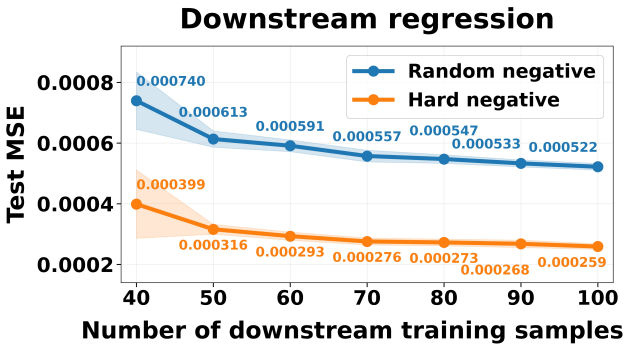

Figure 5: Downstream classification performance under random and hard negative sampling.

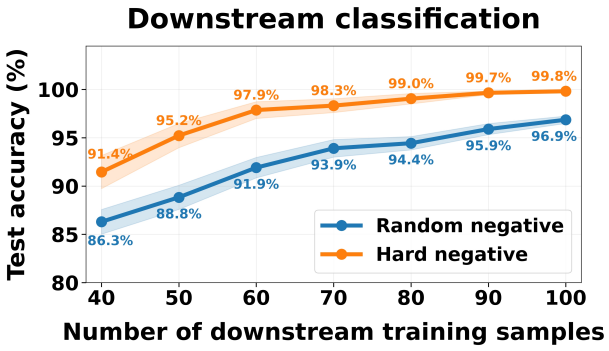

Figure 6: Downstream regression performance under random and hard negative sampling.

the accuracy improves from 86.3% to 91.4%, while with 60 samples, it increases from 91.9% to 97.9%. As the number of training samples further increases, both methods achieve higher accuracy, while hard negative sampling consistently maintains better performance, reaching 99.8% with 100 samples compared with 96.9% using random negative sampling. These results demonstrate that the representations learned with hard negative samples contain richer rare feature information, leading to consistently better fine grained downstream classification performance.

Experiment 5 (Downstream regression). We further evaluate whether the improved learning of rare features leads to better downstream regression performance. Specifically, we freeze the pre-trained image encoder and train a linear regressor using different numbers of labeled downstream samples, where the regression targets are determined by the rare features. As shown in Figure 6, hard negative sampling consistently achieves lower test MSE than random negative sampling across all training sample sizes. With 40 training samples, the test MSE decreases from 7.40×10^{-4} to 3.99×10^{-4}, while with 60 samples, it decreases from 5.91×10^{-4} to 2.93×10^{-4}. With 100 samples, hard negative sampling achieves a test MSE of 2.59×10^{-4}, compared with 5.22×10^{-4} using random negative sampling. These results demonstrate that the representations learned with hard negative samples better preserve the continuous information carried by rare features, leading to consistently better downstream regression performance.

Experiment 6 (Neuron feature alignment). We visualize how individual neurons align with common and rare feature directions before and after training with random negative sampling. Specifically, Figure 7 reports the squared cosine alignment between all 30 image side neurons and the 10 ground truth features, where the first five dimensions correspond to common features and the last five correspond to rare features. At initialization, the alignment values are generally small and exhibit no clear structure across feature directions. After training, strong alignment emerges almost exclusively along the common feature directions, with several neurons developing large projections onto particular common features, while the alignment with rare features remains close to zero for most neurons. These results directly illustrate the implicit bias of vanilla contrastive learning toward common features and show that rare features remain weakly learned when only random negative samples are used.

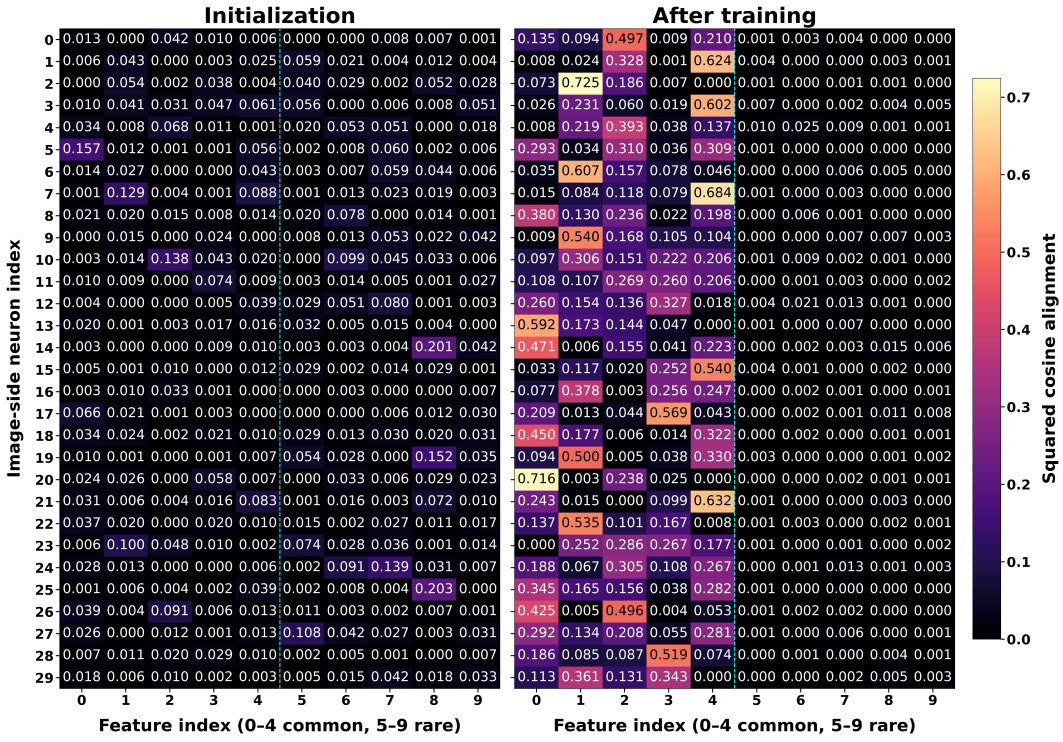

Figure 7: Neuron feature alignment before and after training with random negative sampling.

Experiment 7 (Neuron alignment dynamics). We visualize how the alignment between individual neurons and common and rare feature directions evolves during training with random negative sampling. Figure 8 shows the squared cosine alignment between all 30 image side neurons and the 10 ground truth features at different training epochs, where the first five dimensions correspond to common features and the last five correspond to rare features. At initialization, the alignment values are generally small and exhibit no clear structure. As training proceeds, strong alignment gradually emerges along the common feature directions, with several neurons developing increasingly large projections onto particular common features. This structure becomes pronounced by epoch 100 and remains stable throughout the later stages of training. In contrast, the alignment with rare features remains weak and gradually approaches zero for most neurons. These dynamics provide a direct visualization of the implicit bias of vanilla contrastive learning toward common features, showing that common features are progressively strengthened during training while rare features remain insufficiently learned.

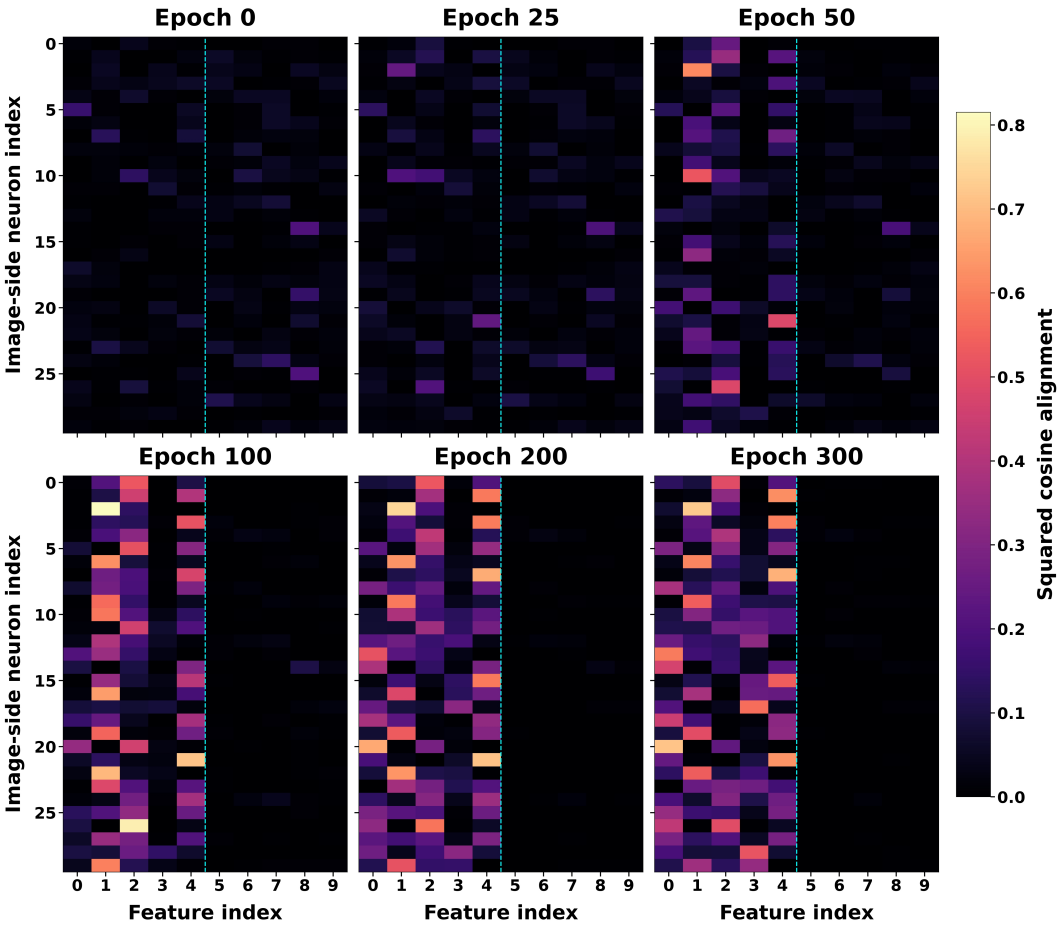

Figure 8: Neuron alignment dynamics during training with random negative sampling.

Experiment 8 (Neuron feature alignment with hard negatives). We visualize how individual neurons align with common and rare feature directions before and after training with hard negative sampling. Figure 9 reports the squared cosine alignment between all 30 image side neurons and the 10 ground truth features, where the first five dimensions correspond to common features and the last five correspond to rare features. At initialization, the alignment values are generally small and exhibit no clear structure. After training, strong alignment emerges along both common and rare feature directions, with different neurons developing substantial projections onto features from both groups. In contrast to random negative sampling, where the learned neurons are predominantly aligned with common features, hard negative sampling preserves strong common feature components while substantially strengthening the alignment with rare features. These results directly illustrate that hard negative samples alleviate the learning imbalance between common and rare features and enable the model to learn a more balanced representation of both types of features.

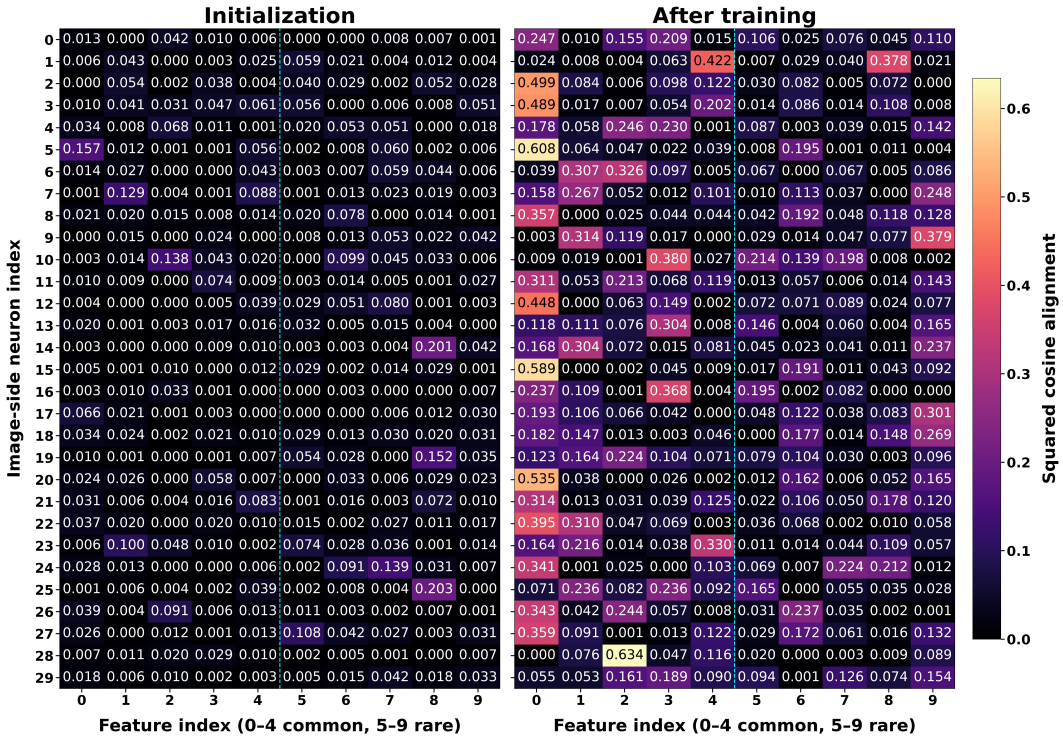

Figure 9: Neuron feature alignment before and after training with hard negative sampling.

Experiment 9 (Neuron alignment dynamics with hard negatives). We visualize how the alignment between individual neurons and common and rare feature directions evolves during training with hard negative sampling. Figure 10 shows the squared cosine alignment between all 30 image side neurons and the 10 ground truth features at different training epochs, where the first five dimensions correspond to common features and the last five correspond to rare features. At initialization, the alignment values are generally small and exhibit no clear structure. During the early stages of training, strong alignment first emerges along the common feature directions. As training proceeds, alignment with the rare feature directions gradually becomes stronger, and multiple neurons develop substantial responses to different rare features in the later stages of training. In contrast to random negative sampling, where rare feature alignment remains weak, hard negative sampling enables both common and rare feature directions to be progressively learned and retained. These dynamics illustrate how hard negative samples alleviate the learning imbalance between common and rare features and lead to more balanced feature representations.

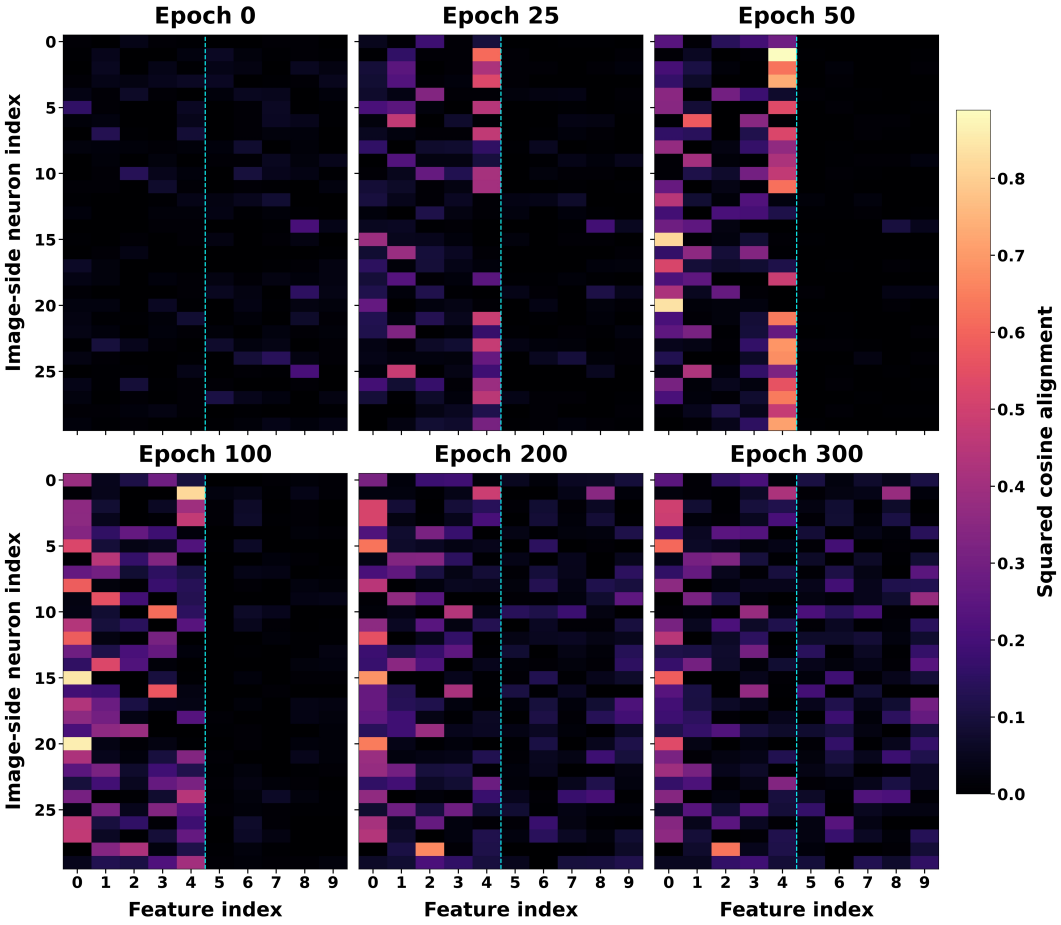

Figure 10: Neuron alignment dynamics during training with hard negative sampling.

B Notations and Lemmas

To streamline the presentation, we begin by introducing the key notations, definitions, and preliminary results that will serve as the foundation for the subsequent analysis.

Notations. First, we introduce the notations that will appear in the appendix.

Let

$$\boldsymbol{h}_{\boldsymbol{X}}^{(r)} := \mathrm{Attn}_r(\boldsymbol{X}), \qquad \text{and} \qquad \boldsymbol{h}_{\boldsymbol{Y}}^{(s)} := \mathrm{Attn}_r(\boldsymbol{Y}). \tag{16}$$

denote the transformer representations of the r-th image token and the s-th text tokenrespectively. $\mathcal{P}^+$ and $\mathcal{P}^-$ denote the distributions of pair indices (i, j) corresponding to positive pairs ($j = i$) and negative pairs ($j \neq i$) respectively.

Population Gradient. The gradient of the population risk with respect to the neuron weight $\boldsymbol{w}_k^{(t)}$ at iteration t is given by

$$\begin{aligned}&\nabla_{\boldsymbol{w}_k} L(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}) \\ =&\mathbb{E}_{(i,j)\sim\mathcal{P}^{\mathrm{all}}}\left[\left(-a_{i,j}\frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\right].\end{aligned} \tag{17}$$

The gradient of the population risk with respect to the neuron weight $\boldsymbol{v}_k^{(t)}$ is similarly given by

$$\begin{aligned}&\nabla_{\boldsymbol{v}_k} L(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}) \\ =&\mathbb{E}_{(i,j)\sim\mathcal{P}^{\mathrm{all}}}\left[\left(-a_{i,j}\frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ f_k(\boldsymbol{X}_i)\sum_{s=1}^{L}\mathbf{1}_{|\langle \boldsymbol{v}_k, \boldsymbol{h}_{\boldsymbol{Y}}^{(s)}\rangle|\geq b_{g_k}}\boldsymbol{h}_{\boldsymbol{Y}}^{(s)}\right].\end{aligned} \tag{18}$$

The population risk is defined as

$$L(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}) = \mathbb{E}_{(i,j)\sim\mathcal{P}^{\mathrm{all}}}\left[\ell(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}, \boldsymbol{X}_i, \boldsymbol{Y}_j))\right]. \tag{19}$$

Empirical Gradient. The gradient of the empirical risk function $\widehat{L}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}})$ with respect to the neuron weight $\boldsymbol{w}_k^{(t)}$ at iteration t is given by

$$\begin{aligned}&\nabla_{\boldsymbol{w}_k} \widehat{L}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}) \\ =&\frac{1}{|\mathcal{B}|}\sum_{i,j}^{|\mathcal{B}|}\left[\left(-a_{i,j}\frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\right].\end{aligned} \tag{20}$$

Similarly, the gradient of the empirical risk with respect to the neuron weight $\boldsymbol{v}_k^{(t)}$ is given by

$$\begin{aligned}&\nabla_{\boldsymbol{v}_k} \widehat{L}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}) \\ =&\frac{1}{|\mathcal{B}|}\sum_{i,j}^{|\mathcal{B}|}\left[\left(-a_{i,j}\frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ f_k(\boldsymbol{X}_i)\sum_{s=1}^{L}\mathbf{1}_{|\langle \boldsymbol{v}_k, \boldsymbol{h}_{\boldsymbol{Y}}^{(s)}\rangle|\geq b_{g_k}}\boldsymbol{h}_{\boldsymbol{Y}}^{(s)}\right].\end{aligned} \tag{21}$$

Concentration Bound. The following lemma shows that, given a sufficiently large batch size, the approximation error between the empirical gradient and the population gradient remains bounded with high probability. Building on this principle, we will first analyze the training dynamics under the population gradient, and subsequently account for the deviation arising from the empirical gradient.

Lemma B.1 (Approximation of empirical gradients by population gradients)**.** *with high probability at least $1 - n^{-q}$, the empirical gradients uniformly approximate the population gradients for every iteration t:*

$$\left\|\nabla_{\boldsymbol{w}_k}\widehat{L}_{\mathrm{aug}}(\boldsymbol{f}_\theta, \boldsymbol{g}_\phi) - \nabla_{\boldsymbol{w}_k} L_{\mathrm{aug}}(\boldsymbol{f}_\theta, \boldsymbol{g}_\phi)\right\|_2 \leq \widetilde{O}\left(\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right), \qquad \forall k \in [m], \tag{22}$$

and similarly,

$$\left\|\nabla_{\boldsymbol{v}_k}\widehat{L}_{\mathrm{aug}}(\boldsymbol{f}_\theta, \boldsymbol{g}_\phi) - \nabla_{\boldsymbol{v}_k} L_{\mathrm{aug}}(\boldsymbol{f}_\theta, \boldsymbol{g}_\phi)\right\|_2 \leq \widetilde{O}\left(\|\boldsymbol{w}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right), \qquad \forall k \in [m]. \tag{23}$$

Proof: Lemma B.1. The proof of Lemma B.1 follows from standard concentration inequalities for sub-Gaussian random variables. In particular, we will use the tail bound

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i,j}^{n} X_{i,j} - \mathbb{E}[X]\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2}\right), \tag{24}$$

where

$$X_{i,j} := \left(-a_{i,j}\frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right) g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\boldsymbol{h}_{\boldsymbol{X}}^{(r)} \tag{25}$$

are independent sub-Gaussian random variables with variance proxy σ^2.

We now choose t such that the right-hand side is bounded by n^{-q} for some constant $q > 0$. Specifically, we require

$$2\exp\left(-\frac{nt^2}{2\sigma^2}\right) \leq n^{-q}. \tag{26}$$

Taking logarithms on both sides and rearranging gives

$$\begin{aligned} \frac{nt^2}{2\sigma^2} &\geq q\log n \\ t^2 &\geq \frac{2\sigma^2(q\log n)}{n} \\ t &= \sigma\sqrt{\frac{2(q\log n)}{n}}. \end{aligned} \tag{27}$$

Substituting this choice back into the concentration inequality gives

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i,j}^{n} X_{i,j} - \mathbb{E}[X]\right| \geq \sigma\sqrt{\frac{2(q\log n)}{n}}\right) \leq n^{-q}. \tag{28}$$

Therefore, with probability at least $1 - n^{-q}$,

$$\left|\frac{1}{n}\sum_{i,j}^{n} X_{i,j} - \mathbb{E}[X]\right| \leq \sigma\sqrt{\frac{2(q\log n)}{n}}. \tag{29}$$

Next, we derive an upper bound for σ.

$$\begin{aligned} \sigma &\leq C\max_{i,j}\|X_{i,j}\|_2 \\ &\leq C\max_{i,j}\|\left(-a_{i,j}\frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right) g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\|_2 \\ &\leq C\|\frac{1}{2}\cdot C_1\cdot \boldsymbol{v}_k\cdot L\cdot C_2\|_2 \\ &\leq O(L\cdot\|\boldsymbol{v}_k\|_2) \end{aligned} \tag{30}$$

Finally, substituting the expressions of $X_{i,j}$ and the bound for σ into the above inequality yields

$$\left\|\nabla_{\boldsymbol{w}_k}\widehat{L}_{\mathrm{aug}}(\boldsymbol{f}_\theta,\boldsymbol{g}_\phi) - \nabla_{\boldsymbol{w}_k}L_{\mathrm{aug}}(\boldsymbol{f}_\theta,\boldsymbol{g}_\phi)\right\|_2 \leq \widetilde{O}\left(\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right) \qquad \forall k\in[m], \tag{31}$$

with probability at least $1 - n^{-q}$. □

Characterization of Neurons. Definition B.1 divides neurons into two categories, ordinary neurons and lucky neurons, based on the initial joint alignment of the image-side and text-side neuron weights with the corresponding feature directions $\boldsymbol{M}_p^V$ and $\boldsymbol{M}_p^L$. These sets will serve as the foundation for our subsequent analysis of the coupled training dynamics.

Definition B.1 (Characterization of Neurons)**.** *We define the following sets of neurons according to their initial joint alignment with the corresponding image and text feature directions.*

(a) For each $p \in [d]$, we define the set of ordinary neurons $\mathcal{M}_p \subseteq [m]$ *as*

$$\mathcal{M}_p := \left\{ k \in [m] : \frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^V \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^L \rangle\Big) \geq \frac{\sqrt{c_2 \log d}}{\sqrt{d}} \sqrt{\sum_{q \in [d]} \left[\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_q^V \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_q^L \rangle\Big)\right]^2} \right\}. \tag{32}$$

(b) For each $p \in [d]$, we define the set of lucky neurons $\mathcal{M}_p^\star \subseteq [m]$ *as*

$$\mathcal{M}_p^\star := \left\{ k \in [m] : \begin{array}{c} \frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^V \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^L \rangle\Big) \geq \frac{\sqrt{c_1 \log d}}{\sqrt{d}} \sqrt{\sum_{q \in [d]} \left[\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_q^V \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_q^L \rangle\Big)\right]^2}, \\ \frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_{p'}^V \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_{p'}^L \rangle\Big) \leq \frac{\sqrt{c_2 \log d}}{\sqrt{d}} \sqrt{\sum_{q \in [d]} \left[\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_q^V \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_q^L \rangle\Big)\right]^2}, \\ \forall p' \in [d],\ p' \neq p \end{array} \right\}. \tag{33}$$

where

$$c_1 = 2(1+\gamma), \qquad c_2 = 2(1-\gamma), \qquad \gamma \text{ is a small constant}. \tag{34}$$

Properties at initialization. At initialization ($t = 0$), we establish several key properties of the neurons that will be used in the subsequent analysis of the SGD trajectory. Before presenting Lemma B.2, we briefly summarize its main implications: (a) the image-side and text-side neuron weights $\boldsymbol{w}_k^{(0)}$ and $\boldsymbol{v}_k^{(0)}$ have squared norms of order $\sigma_0^2 d_1$; (b) their projections onto the corresponding feature subspaces have squared norms of order $\sigma_0^2 d$; (c) for each feature $p \in [d]$, there exist sufficiently many lucky neurons, while the number of ordinary neurons remains controlled; (d) for each neuron $k \in [m]$, the number of features with which it has sufficiently strong joint alignment is limited.

Lemma B.2. *At initialization ($t = 0$), the following properties hold:*

(a) With high probability, for every $k \in [m]$,

$$\|\boldsymbol{w}_k^{(0)}\|_2^2 \in \left[\sigma_0^2 d_1 \left(1 - \widetilde{O}\Big(\tfrac{1}{\sqrt{d_1}}\Big)\right),\ \sigma_0^2 d_1 \left(1 + \widetilde{O}\Big(\tfrac{1}{\sqrt{d_1}}\Big)\right)\right], \tag{35}$$

and similarly,

$$\|\boldsymbol{v}_k^{(0)}\|_2^2 \in \left[\sigma_0^2 d_1 \left(1 - \widetilde{O}\Big(\tfrac{1}{\sqrt{d_1}}\Big)\right),\ \sigma_0^2 d_1 \left(1 + \widetilde{O}\Big(\tfrac{1}{\sqrt{d_1}}\Big)\right)\right]. \tag{36}$$

(b) With high probability, for every $k \in [m]$,

$$\left\|\boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \in \left[\sigma_0^2 d \left(1 - \widetilde{O}\Big(\tfrac{1}{\sqrt{d}}\Big)\right),\ \sigma_0^2 d \left(1 + \widetilde{O}\Big(\tfrac{1}{\sqrt{d}}\Big)\right)\right], \tag{37}$$

and similarly,

$$\left\|\boldsymbol{M}^L (\boldsymbol{M}^L)^\top \boldsymbol{v}_k^{(0)}\right\|_2^2 \in \left[\sigma_0^2 d \left(1 - \widetilde{O}\Big(\tfrac{1}{\sqrt{d}}\Big)\right),\ \sigma_0^2 d \left(1 + \widetilde{O}\Big(\tfrac{1}{\sqrt{d}}\Big)\right)\right]. \tag{38}$$

(c) Let $m = d^{C_m}$ be the number of neurons. With probability at least $1 - o\big(\frac{1}{d^4}\big)$, for each $p \in [d]$,

$$|\mathcal{M}_p^\star| \geq \Omega(d^{\omega_1}) =: \Xi_1, \qquad |\mathcal{M}_p| \leq O(d^{\omega_2}) =: \Xi_2, \tag{39}$$

where

$$\omega_1 = C_m - (1+\gamma), \qquad \omega_2 = C_m - (1-\gamma). \tag{40}$$

(d) For each $k \in [m]$, there are at most $O\left(d^\gamma\right)$ indices $p \in [d]$ such that $k \in \mathcal{M}_p$.

C Preliminary Training Dynamics of sigmoid-based contrastive learning

C.1 Feature Decomposition under Vanilla Contrastive Learning

We first present the baseline feature decomposition under vanilla contrastive learning, which serves as the foundation for the more specific feature decomposition results in the subsequent appendices. In particular, when the common and rare features have comparable frequencies, i.e., $\epsilon_1 = \Theta(\epsilon_2)$, the learned neuron weights can be characterized without distinguishing the learning dynamics of common and rare features. Theorem C.1 gives this baseline decomposition for the image-side neuron weights. The results in Appendices D and E extend this theorem to settings with imbalanced common and rare feature frequencies, hard negative samples, excessive hard negative samples, and thresholding. The feature decomposition is established in the proof of Theorem F.3 in Appendix F.2

Theorem C.1 (Vanilla Feature Decomposition). *Suppose that $\epsilon_1 = \Theta(\epsilon_2)$. For each neuron $k \in [m]$ and $t \geq T_2$, the image-side neuron weight learns the following set of features:*

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k} \alpha_{k,p} \boldsymbol{M}_p^V + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^V + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^\perp, \tag{41}$$

where

$$\alpha_{k,p} \geq \Omega(1), \qquad \alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{42}$$

C.2 Stage 1

Theorem C.2 shows two main effects of contrastive learning in the first training stage: (a) the neuron weights in the image and text modalities reinforce each other along the corresponding feature directions, and the growth rate along feature p depends on its frequency ϵ_p, and (b) the neuron weights are suppressed along the corresponding non-feature directions. In particular, the growth of $\boldsymbol{w}_k^{(t)}$ along an image feature direction $\boldsymbol{M}_p^{\boldsymbol{V}}$ depends on the alignment of $\boldsymbol{v}_k^{(t)}$ with the corresponding text feature $\boldsymbol{M}_p^{\boldsymbol{L}}$, and vice versa. The strength of this cross-modal reinforcement is proportional to the feature frequency ϵ_p, so features with larger ϵ_p grow faster, while features with smaller ϵ_p are harder to capture in the first training stage. In contrast, the neuron weights do not receive such a growth signal along the non-feature directions and therefore remain suppressed during training. We can find the proof of Theorem C.2 in Appendix F.1, where the corresponding update equations are given in Eq. 148, Eq. 149, Eq. 150, and Eq. 151.

Theorem C.2 (Stage 1). *During the first training stage, the update of neuron weights $\boldsymbol{w}_k^{(t)}$ and $\boldsymbol{v}_k^{(t)}$ can be bounded for all $t \in [0, T_1]$ as follows, where $T_1 = \Theta\left(\frac{d \log d_1}{\eta \log \log d}\right)$.*

For the corresponding feature directions, we have

$$|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle| \geq (1 - \eta\lambda) |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle| + \epsilon_p \frac{\eta \log \log d}{d} |\langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle| - \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right), \tag{43}$$

and

$$|\langle \boldsymbol{v}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle| \geq (1 - \eta\lambda) |\langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle| + \epsilon_p \frac{\eta \log \log d}{d} |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle| - \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right). \tag{44}$$

For the corresponding non-feature directions, we have

$$|\langle \boldsymbol{w}_k^{(t+1)}, (\boldsymbol{M}^{\boldsymbol{V}})^\perp \rangle| \leq (1 - \eta\lambda) |\langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^{\boldsymbol{V}})^\perp \rangle| + \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right), \tag{45}$$

and

$$|\langle \boldsymbol{v}_k^{(t+1)}, (\boldsymbol{M}^{\boldsymbol{L}})^\perp \rangle| \leq (1 - \eta\lambda) |\langle \boldsymbol{v}_k^{(t)}, (\boldsymbol{M}^{\boldsymbol{L}})^\perp \rangle| + \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right). \tag{46}$$

C.3 STAGE 2

Before presenting the theoretical results in Stage 2, we first categorize neurons into two groups. The *ordinary neurons* $\mathcal{M}_p$ strongly align with a certain feature direction, while the *lucky neurons* $\mathcal{M}_p^\star$ form a special subset that aligns with only one feature direction (see Appendix B for the formal definition). In Stage 2: (a) lucky neurons in $\mathcal{M}_p^\star$ grow significantly in alignment with the corresponding feature directions $\boldsymbol{M}_p^V$ and $\boldsymbol{M}_p^L$, such that their projections onto the corresponding feature directions become comparable to their weight norms; (b) ordinary neurons in $\mathcal{M}_p$ are bounded by the feature components of lucky neurons up to a constant factor. The proof of Theorem C.3 is provided in Appendix F.2, as part of the proof of Theorem F.2.

Theorem C.3 (Stage 2). *During the second training stage, the neuron weights $\boldsymbol{w}_k^{(t)}$ and $\boldsymbol{v}_k^{(t)}$ can be bounded for all $t \in [T_1, T_2]$ as follows, where $T_2 = T_1 + \Theta\left(\frac{d \log d}{\eta \log \log d}\right)$.*

(a) For each $p \in [d]$, if $k \in \mathcal{M}_p^\star$, then:

$$|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^V \rangle| \geq \Omega(1) \|\boldsymbol{w}_k^{(T_2)}\|_2, \qquad |\langle \boldsymbol{v}_k^{(T_2)}, \boldsymbol{M}_p^L \rangle| \geq \Omega(1) \|\boldsymbol{v}_k^{(T_2)}\|_2. \tag{47}$$

(b) For each $p \in [d]$, if $k' \in \mathcal{M}_p$ and $k \in \mathcal{M}_p^\star$, then:

$$|\langle \boldsymbol{w}_{k'}^{(T_2)}, \boldsymbol{M}_p^V \rangle| \leq O\left(|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^V \rangle|\right), \qquad |\langle \boldsymbol{v}_{k'}^{(T_2)}, \boldsymbol{M}_p^L \rangle| \leq O\left(|\langle \boldsymbol{v}_k^{(T_2)}, \boldsymbol{M}_p^L \rangle|\right). \tag{48}$$

C.4 CONVERGENCE

Theorem C.4 establishes the convergence. In particular, (49) shows that the algorithm converges with bounded training error after T_2. Moreover, (50) and (52) characterize the structure of the learned neuron weights in the image and text modalities: upon convergence, they become strongly aligned with a subset of features within $\mathcal{N}_k$, weakly aligned with the remaining features, and remain small in the non-feature directions. The size of $\mathcal{N}_k$ is bounded as in (54), and for each feature, there exist a sufficient number of neurons that specialize in learning that single feature. The feature decomposition is established in the proof of Theorem F.3 in Appendix F.2, and the convergence analysis is provided in Appendix F.3.

Theorem C.4 (Convergence). *Suppose we train the neural networks $\boldsymbol{f_\theta}$ and $\boldsymbol{g_\phi}$ via contrastive learning. Starting from iteration T_2, for any T satisfying $T\eta \geq m$, the following guarantees hold:*

$$\frac{1}{T} \sum_{t=T_2}^{T_2+T-1} L_{\text{aug}}(f_t) \leq O\left(e^{-\alpha}\right). \tag{49}$$

Moreover, for each neuron $k \in [m]$ and $t \geq T_2$, the image-side neuron weight learns the following set of features:

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k} \alpha_{k,p} \boldsymbol{M}_p^V + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^V + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^\perp, \tag{50}$$

where

$$\alpha_{k,p} \geq \Omega(1), \qquad \alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{51}$$

Similarly, the text-side neuron weight learns the following set of features:

$$\boldsymbol{v}_k^{(t)} = \sum_{p \in \mathcal{N}_k} \widetilde{\alpha}_{k,p} \boldsymbol{M}_p^L + \sum_{p \notin \mathcal{N}_k} \widetilde{\alpha}'_{k,p} \boldsymbol{M}_p^L + \sum_{p \in [d_1] \setminus [d]} \widetilde{\beta}_{k,p} (\boldsymbol{M}^L)_p^\perp, \tag{52}$$

where

$$\widetilde{\alpha}_{k,p} \geq \Omega(1), \qquad \widetilde{\alpha}'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\widetilde{\beta}_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{53}$$

Furthermore, the size of $\mathcal{N}_k$ is bounded as

$$|\mathcal{N}_k| = O(d^\gamma). \tag{54}$$

Finally, for each $\boldsymbol{M}_p^V$ and $\boldsymbol{M}_p^L$, there are at least $\Omega\left(m \cdot d^{-(1+\gamma)}\right)$ neurons $k \in [m]$ such that $|\mathcal{N}_k| = 1$.

D Training Dynamics with Random Negatives and Hard Negatives

D.1 Theorem D.1: Vanilla Contrastive Learning

Theorem D.1 characterizes the feature decomposition learned by each neuron under vanilla contrastive learning. When $\epsilon_2 \ll \epsilon_1$, the neurons mainly learn common features, while the rare feature components remain small. This result corresponds to Theorem 3.1 in the main text.

Theorem D.1 (**Vanilla Contrastive Learning**). *Under vanilla contrastive learning, suppose that*

$$\epsilon_2 \ll \epsilon_1. \tag{55}$$

Then, for each neuron $k \in [m]$ and $t \geq T_2$, the learned weight can be decomposed as

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{S}_2} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \in \mathcal{S}_1 \setminus \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{56}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega(1), \qquad \alpha_{k,p}^{V_2} \leq o\left(\frac{1}{\sqrt{d_1}}\right), \tag{57}$$

and

$$\alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{58}$$

D.2 Theorem D.2: Contrastive Learning with Hard Negative Samples

Theorem D.2 shows how hard negative samples change the feature decomposition learned by each neuron during contrastive learning. By introducing hard negative samples, the effective learning rate of common features is reduced from ϵ_1 to $(1-\rho)\epsilon_1$, while the learning rate of rare features remains ϵ_2. Therefore, hard negative samples reduce the learning imbalance between common and rare features and allow the neurons to sufficiently learn both types of features. This result corresponds to Theorem 3.2 in the main text.

Theorem D.2 (**Contrastive Learning with Hard Negative Samples**). *Under Assumption D.1, let ρ denote the proportion of shared common features in hard negative pairs.*

After introducing hard negative samples, suppose that

$$\rho = 1 - \Theta(\epsilon_2/\epsilon_1). \tag{59}$$

Then, for each neuron $k \in [m]$ and $t \geq T_2$, the learned weight can be decomposed as

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k \cap \mathcal{S}_1} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k \cap \mathcal{S}_2} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{60}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega(1), \qquad \alpha_{k,p}^{V_2} \geq \Omega(1), \tag{61}$$

and

$$\alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{62}$$

D.3 Proof of Theorem D.1 and Theorem D.2

The proofs of Theorems D.1 and D.2 follow the general proof framework established in Appendices F.1, F.2, and F.3. Therefore, we focus on the key differences introduced by hard negative samples and present the corresponding modifications to the analysis below.

We first recall the dependence structure of the latent variables z. For positive pairs, $z_{i,p}$ and $z_{i,q}$ are independent for any $p \neq q$. For standard negative pairs, $z_{i,p}$ is independent of both $z_{j,p}$ and $z_{j,q}$.

We now consider the hard negative pair case with $j \neq i$. It is straightforward that $z_{i,p}$ and $z_{j,q}$ remain independent for $p \neq q$. Therefore, it remains to analyze the case $p = q$, i.e., the dependence between $z_{i,p}$ and $z_{j,p}$.

Assumption D.1 (Hard Negative Sample)**.** *For each hard negative pair $(\boldsymbol{X}_i, \boldsymbol{Y}_j)$, we consider the common and rare feature subsets S_1 and S_2 defined in Assumption 3.2. For any $p \in S_1$, the image and text samples share the same latent signal, such that $z_{i,p} = z_{j,p}$. For any $p \in S_2$, the corresponding latent signals are independent. We further divide the feature dictionary into common and rare features, denoted by $(\boldsymbol{M}^{\boldsymbol{V_1}}, \boldsymbol{M}^{\boldsymbol{L_1}})$ and $(\boldsymbol{M}^{\boldsymbol{V_2}}, \boldsymbol{M}^{\boldsymbol{L_2}})$, respectively. Let ρ denote the proportion of shared common features in hard negative pairs.*

Therefore, we have:

$$
\begin{aligned}
&\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[g_k(\boldsymbol{Y}_j)\langle\nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i), \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&\frac{1}{L^2}\sum_{q\in[d]}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\mathbb{E}_N\left[\hat{\boldsymbol{z}}_{j,q}\hat{\boldsymbol{z}}_{i,p}\right] + \frac{1}{L^2}\sum_{q\in[d]}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\mathbb{E}_{HN}\left[\hat{\boldsymbol{z}}_{j,q}\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{1}{L^2}\sum_{q\neq p}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\mathbb{E}_{HN}\left[\hat{\boldsymbol{z}}_{j,q}\hat{\boldsymbol{z}}_{i,p}\right] + \frac{1}{L^2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L}}\rangle\mathbb{E}_{HN}\left[\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{1}{L^2}\sum_{S_1}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle\mathbb{E}_{HN}\left[\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right] + \frac{1}{L^2}\sum_{S_2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_2}}\rangle\mathbb{E}_{HN}\left[\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{\rho}{L^2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle\mathbb{E}\left[\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right].
\end{aligned}
\tag{63}
$$

Combining the contributions from positive and negative pairs, we obtain:

$$
\begin{aligned}
&\mathbb{E}_{(i,j)\sim\mathcal{P}^{\text{all}}}\left[\left(a_{i,j}\frac{e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&p_+\cdot\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[\left(a_{i,i}\frac{e^{a_{i,i}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_i)\right)}}{1+e^{a_{i,i}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_i)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
+&p_-\cdot\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[\left(a_{i,j}\frac{e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&\frac{1}{2L^2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L}}\rangle\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[\hat{\boldsymbol{z}}_{i,p}\hat{\boldsymbol{z}}_{i,p}\right] - \frac{\rho}{2L^2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{1}{2L^2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[\hat{\boldsymbol{z}}_{i,p}\hat{\boldsymbol{z}}_{i,p}\right] - \frac{\rho}{2L^2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right] \\
+&\frac{1}{2L^2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_2}}\rangle\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[\hat{\boldsymbol{z}}_{i,p}\hat{\boldsymbol{z}}_{i,p}\right].
\end{aligned}
\tag{64}
$$

Substituting the above result into the gradient update, we obtain:

$$
\begin{aligned}
&\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\
=&\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle - \eta \langle \nabla_{\boldsymbol{w}_k} L_{\mathrm{aug}}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}), \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \quad \pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}} \right) \\
\overset{①}{=}&(1-\eta\lambda) \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\
&+\eta \mathbb{E}_{(i,j)\sim \mathcal{P}^{all}} \left[\left(a_{i,j} \frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}}{1 + e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right) \cdot \ g_k(\boldsymbol{Y}_j) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_{\boldsymbol{X}}^{(r)} \rangle| \geq b_{f_k}} \langle \boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \right] \\
&\pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}} \right) \\
=&\left(1-\eta\lambda\right) \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + (1-\rho)\epsilon_1 \frac{\eta \log\log d}{d} \langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L_1}} \rangle + \epsilon_2 \frac{\eta \log\log d}{d} \langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L_2}} \rangle \\
&\pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}} \right).
\end{aligned} \tag{65}
$$

Case 1: We first consider the case where, after introducing hard negative samples,

$$
\rho = 1 - \Theta(\epsilon_2/\epsilon_1).
$$

First, we have:

$$
\begin{aligned}
&\|\boldsymbol{w}_k^{(T)}\|_2^2 \\
=&\|\boldsymbol{M}^{\boldsymbol{V_1}}(\boldsymbol{M}^{\boldsymbol{V_1}})^\top \boldsymbol{w}_k^{(T)}\|_2^2 + \|\boldsymbol{M}^{\boldsymbol{V_2}}(\boldsymbol{M}^{\boldsymbol{V_2}})^\top \boldsymbol{w}_k^{(T)}\|_2^2 + \|(\boldsymbol{M}^{\boldsymbol{V}})^\perp((\boldsymbol{M}^{\boldsymbol{V}})^\perp)^\top \boldsymbol{w}_k^{(T)}\|_2^2 \\
=&\left\|\boldsymbol{M}^{\boldsymbol{V_1}}(\boldsymbol{M}^{\boldsymbol{V_1}})^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \left(1 + (1-\rho)\epsilon_1 \frac{\eta \log\log d}{d}\right)^{2T} \\
&+\left\|\boldsymbol{M}^{\boldsymbol{V_2}}(\boldsymbol{M}^{\boldsymbol{V_2}})^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \left(1 + \epsilon_2 \frac{\eta \log\log d}{d}\right)^{2T} + \|\boldsymbol{w}_k^{(0)}\|_2^2 \\
=&\left\|\boldsymbol{M}^{\boldsymbol{V_1}}(\boldsymbol{M}^{\boldsymbol{V_1}})^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \Theta\left(e^{(1-\rho)\epsilon_1 \frac{\eta \log\log d}{d} \cdot T}\right) \\
&+\left\|\boldsymbol{M}^{\boldsymbol{V_2}}(\boldsymbol{M}^{\boldsymbol{V_2}})^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \Theta\left(e^{\epsilon_2 \frac{\eta \log\log d}{d} \cdot T}\right) + \|\boldsymbol{w}_k^{(0)}\|_2^2 \\
=&\frac{d}{d_1}\Theta\left(e^{(1-\rho)\epsilon_1 \frac{\eta \log\log d}{d} \cdot T}\right) \|\boldsymbol{w}_k^{(0)}\|_2^2 + \frac{d}{d_1}\Theta\left(e^{\epsilon_2 \frac{\eta \log\log d}{d} \cdot T}\right) \|\boldsymbol{w}_k^{(0)}\|_2^2.
\end{aligned} \tag{66}
$$

Then, we have:

$$
\begin{aligned}
\|\boldsymbol{M}^{\boldsymbol{V}}(\boldsymbol{M}^{\boldsymbol{V}})^\top \boldsymbol{w}_k^{(T)}\|_2^2 =&\|\boldsymbol{w}_k^{(T)}\|_2^2 - \|(\boldsymbol{M}^{\boldsymbol{V}})^\perp((\boldsymbol{M}^{\boldsymbol{V}})^\perp)^\top \boldsymbol{w}_k^{(T)}\|_2^2 \\
\geq&\|\boldsymbol{w}_k^{(T)}\|_2^2 - (1+\epsilon)\|(\boldsymbol{M}^{\boldsymbol{V}})^\perp((\boldsymbol{M}^{\boldsymbol{V}})^\perp)^\top \boldsymbol{w}_k^{(0)}\|_2^2 \\
\geq&\|\boldsymbol{w}_k^{(T)}\|_2^2 - (1+\epsilon)\|\boldsymbol{w}_k^{(0)}\|_2^2 \\
\geq&\|\boldsymbol{w}_k^{(T)}\|_2^2 - \frac{(1+\epsilon)\|\boldsymbol{w}_k^{(T)}\|_2^2}{\frac{d}{d_1}\Theta\left(e^{(1-\rho)\epsilon_1 \frac{\eta \log\log d}{d} \cdot T}\right)} - \frac{(1+\epsilon)\|\boldsymbol{w}_k^{(T)}\|_2^2}{\frac{d}{d_1}\Theta\left(e^{\epsilon_2 \frac{\eta \log\log d}{d} \cdot T}\right)} \\
\geq&\frac{1}{2}\|\boldsymbol{w}_k^{(T)}\|_2^2.
\end{aligned} \tag{67}
$$

Therefore, it suffices to choose

$$
T_1 \geq \frac{d}{(1-\rho)\epsilon_1 \eta \log\log d} \log\left(4(1+\epsilon)\frac{d_1}{d}\right), \tag{68}
$$

which guarantees that

$$\frac{(1+\epsilon)}{\frac{d}{d_1}\Theta\left(e^{(1-\rho)\epsilon_1\frac{\eta\log\log d}{d}}T_1\right)} \leq \frac{1}{4}, \qquad \frac{(1+\epsilon)}{\frac{d}{d_1}\Theta\left(e^{\epsilon_2\frac{\eta\log\log d}{d}}T_1\right)} \leq \frac{1}{4}. \tag{69}$$

For each $p \in [d]$, and each $k \in \mathcal{M}_p^\star$,

$$|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^{\boldsymbol{V_1}}\rangle| \geq \sqrt{1+\gamma}\frac{\sqrt{\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T_1)}\|_2, \tag{70}$$

and

$$|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle| \geq \sqrt{1+\gamma}\frac{\sqrt{\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T_1)}\|_2. \tag{71}$$

After entering the second stage, we separately consider the common features and rare features under gradient descent.

For $\boldsymbol{M}_p^{\boldsymbol{V_1}}$ such that $k \in \mathcal{M}_p^\star$, at iteration $t+1$,

$$\begin{aligned}\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V_1}}\rangle \geq &\left(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V_1}}\rangle - \operatorname{sign}\left(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V_1}}\rangle\right) b_{f_k}^{(t)}\right) \\ &\cdot \left(1 + (1-\rho)\epsilon_1\frac{\eta\log\log d}{d}\left(1 - \frac{1}{\operatorname{polylog}(d)}\right)\right).\end{aligned} \tag{72}$$

For $\boldsymbol{M}_p^{\boldsymbol{V_2}}$ such that $k \in \mathcal{M}_p^\star$, at iteration $t+1$,

$$\begin{aligned}\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle \geq &\left(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle - \operatorname{sign}\left(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle\right) b_{f_k}^{(t)}\right) \\ &\cdot \left(1 + \epsilon_2\frac{\eta\log\log d}{d}\left(1 - \frac{1}{\operatorname{polylog}(d)}\right)\right).\end{aligned} \tag{73}$$

Following the same proof as before, for all iterations $t \in (T_1, T_2]$, the neurons $k \in [m]$ satisfy the following properties:

For $p \in [d]$, if $k \in \mathcal{M}_p^\star$, then

$$\left|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V_1}}\rangle\right| \geq \sqrt{1+\gamma} b_{f_k}^{(t)}, \tag{74}$$

and

$$\left|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle\right| \geq \sqrt{1+\gamma} b_{f_k}^{(t)}. \tag{75}$$

Finally, we obtain our key result. For each neuron $k \in [m]$, the following conditions hold at iteration $t = T_2$:

For each $p \in [d]$, if $k \in \mathcal{M}_p^\star$, then

$$\left|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^{\boldsymbol{V_1}}\rangle\right| \geq \Omega(1)\|\boldsymbol{w}_k^{(T_2)}\|_2, \tag{76}$$

and

$$\left|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle\right| \geq \Omega(1)\|\boldsymbol{w}_k^{(T_2)}\|_2. \tag{77}$$

Finally, we obtain the feature decomposition result. For each neuron $k \in [m]$ and $t \geq T_2$, the weight learns the following set of features:

$$\boldsymbol{w}_k^{(t)} = \sum_{p\in\mathcal{N}_k}\alpha_{k,p}^{V_1}\boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p\in\mathcal{N}_k}\alpha_{k,p}^{V_2}\boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p\notin\mathcal{N}_k}\alpha'_{k,p}\boldsymbol{M}_p^{V} + \sum_{p\in[d_1]\setminus[d]}\beta_{k,p}(\boldsymbol{M}^V)_p^\perp, \tag{78}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega(1), \qquad \alpha_{k,p}^{V_2} \geq \Omega(1), \tag{79}$$

$$\alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{80}$$

Case 2: We next consider the vanilla contrastive learning setting without hard negative samples, where

$$\epsilon_2 \ll \epsilon_1.$$

First, we have:

$$\begin{aligned}
\|\boldsymbol{M}^{\boldsymbol{V}}(\boldsymbol{M}^{\boldsymbol{V}})^\top \boldsymbol{w}_k^{(T)}\|_2^2 =& \|\boldsymbol{w}_k^{(T)}\|_2^2 - \|(\boldsymbol{M}^{\boldsymbol{V}})^\perp((\boldsymbol{M}^{\boldsymbol{V}})^\perp)^\top \boldsymbol{w}_k^{(T)}\|_2^2 \\
\geq& \|\boldsymbol{w}_k^{(T)}\|_2^2 - (1+\epsilon)\,\|(\boldsymbol{M}^{\boldsymbol{V}})^\perp((\boldsymbol{M}^{\boldsymbol{V}})^\perp)^\top \boldsymbol{w}_k^{(0)}\|_2^2 \\
\geq& \|\boldsymbol{w}_k^{(T)}\|_2^2 - (1+\epsilon)\,\|\boldsymbol{w}_k^{(0)}\|_2^2 \\
\geq& \|\boldsymbol{w}_k^{(T)}\|_2^2 - \frac{(1+\epsilon)\,\|\boldsymbol{w}_k^{(T)}\|_2^2}{\frac{d}{d_1}\Theta\left(e^{\epsilon_1 \frac{\eta \log\log d}{d}\cdot T}\right)} \\
\geq& \frac{1}{2}\|\boldsymbol{w}_k^{(T)}\|_2^2.
\end{aligned} \tag{81}$$

Therefore, it suffices to choose

$$T_1 \geq \frac{d}{\epsilon_1 \eta \log\log d}\log\left(2(1+\epsilon)\frac{d_1}{d}\right), \tag{82}$$

which guarantees that

$$\frac{(1+\epsilon)}{\frac{d}{d_1}\Theta\left(e^{\epsilon_1 \frac{\eta \log\log d}{d} T_1}\right)} \leq \frac{1}{2}. \tag{83}$$

Therefore, the analysis of $\boldsymbol{w}_k$ along $\boldsymbol{M}^{\boldsymbol{V_1}}$ remains the same as in Case 1. In the following, we focus on the behavior of $\boldsymbol{w}_k$ along $\boldsymbol{M}^{\boldsymbol{V_2}}$.

$$\begin{aligned}
\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle =& \left(1-\eta\lambda+\epsilon_2\frac{\eta\log\log d}{d}\right)^T \langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle \pm \widetilde{O}\left(T\eta \max_{0\leq t<T}\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
\leq& \left(1-\eta\lambda+\epsilon_2\frac{\eta\log\log d}{d}\right)^T \langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle + \widetilde{O}(\epsilon)\|\boldsymbol{w}_k^{(0)}\|_2 \\
\leq& O\left(\sqrt{\frac{\log d}{d_1}}\left(1-\eta\lambda+\epsilon_2\frac{\eta\log\log d}{d}\right)^T\right)\|\boldsymbol{w}_k^{(0)}\|_2 + \widetilde{O}(\epsilon)\|\boldsymbol{w}_k^{(0)}\|_2 \\
\leq& O\left(\sqrt{\left[1+(\frac{d_1}{d})^{\frac{\epsilon_2}{\epsilon_1}}\right]\frac{\log d}{d_1}}\right)\|\boldsymbol{w}_k^{(T)}\|_2 \\
\leq& O\left(\sqrt{\frac{\log d}{d_1}}\right)\|\boldsymbol{v}_k^{(T)}\|_2.
\end{aligned} \tag{84}$$

Therefore, we obtain the conclusion of the first stage. For each $p \in [d]$, and each $k \in \mathcal{M}_p^\star$, the components along $\boldsymbol{M}_p^{\boldsymbol{V_1}}$ and $\boldsymbol{M}_p^{\boldsymbol{V_2}}$ satisfy

$$|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^{\boldsymbol{V_1}}\rangle| \geq \sqrt{1+\gamma}\frac{\sqrt{\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T_1)}\|_2, \tag{85}$$

and

$$|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^{\boldsymbol{V_2}}\rangle| \leq O\left(\sqrt{\frac{\log d}{d_1}}\right)\|\boldsymbol{w}_k^{(T_1)}\|_2. \tag{86}$$

After entering the second stage, we continue our analysis from gradient descent. The analysis along $\boldsymbol{M}^{\boldsymbol{V_1}}$ remains the same as before. For $\boldsymbol{M}_p^{\boldsymbol{V_2}}$ such that $k \in \mathcal{M}_p^\star$, at iteration $t+1$,

$$\begin{aligned}\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V_2}} \rangle \leq & \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V_2}} \rangle \left(1 + \epsilon_2 \frac{\eta \log\log d}{d} \left(1 - \frac{1}{\operatorname{polylog}(d)}\right)\right) \\ \leq & \langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^{\boldsymbol{V_2}} \rangle \left(1 + \epsilon_2 \frac{\eta \log\log d}{d} \left(1 - \frac{1}{\operatorname{polylog}(d)}\right)\right)^{t-T_1} \\ \leq & O\left(\sqrt{\frac{\log d}{d_1}}\right) \|\boldsymbol{w}_k^{(T_1)}\|_2 \\ \leq & O\left(\sqrt{\frac{\log d}{d_1}}\right) \|\boldsymbol{w}_k^{(t+1)}\|_2.\end{aligned} \tag{87}$$

Finally, we obtain our key result. For each neuron $k \in [m]$, the following conditions hold at iteration $t = T_2$:

For each $p \in [d]$, if $k \in \mathcal{M}_p^\star$, then

$$\left|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^{\boldsymbol{V_1}} \rangle\right| \geq \Omega(1) \|\boldsymbol{w}_k^{(T_2)}\|_2, \tag{88}$$

and

$$\left|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^{\boldsymbol{V_2}} \rangle\right| \leq \widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{w}_k^{(T_2)}\|_2}{\sqrt{d_1}}\right). \tag{89}$$

Finally, we obtain the feature decomposition result. For each neuron $k \in [m]$ and $t \geq T_2$, the weight learns the following set of features:

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^\perp, \tag{90}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega(1), \qquad \alpha_{k,p}^{V_2} \leq o\left(\frac{1}{\sqrt{d_1}}\right), \tag{91}$$

$$\alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{92}$$

E Training Dynamics with and without Threshold

E.1 Theorem E.1: Contrastive Learning with Excessive Hard Negative Samples

Theorem E.1 shows the feature decomposition learned by each neuron when too many hard negative samples are introduced. When $(1-\rho)\epsilon_1 \ll \epsilon_2$, the learning of common features is excessively suppressed, while the neurons mainly learn rare features. This result corresponds to Theorem 3.3 in the main text.

Theorem E.1 (Contrastive Learning with Excessive Hard Negative Samples). *Without the threshold function, suppose that too many hard negative samples are introduced such that*

$$(1-\rho)\epsilon_1 \ll \epsilon_2. \tag{93}$$

Then, for each neuron $k \in [m]$ and $t \geq T_2$, the learned weight can be decomposed as

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{S}_1} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \in \mathcal{S}_2 \setminus \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^\perp, \tag{94}$$

where

$$\alpha_{k,p}^{V_1} \leq o\left(\frac{1}{\sqrt{d_1}}\right), \qquad \alpha_{k,p}^{V_2} \geq \Omega(1), \tag{95}$$

and

$$\alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{96}$$

E.2 Theorem E.2: Contrastive Learning with Thresholded Hard Negative Samples

Theorem E.2 shows how the threshold function changes the feature decomposition learned by each neuron when too many hard negative samples are introduced. By filtering out hard negative samples that would otherwise excessively suppress shared common features, the threshold function reduces the effective proportion ρ of shared common features involved in hard-negative training. This increases the effective learning rate of common features, while the learning rate of rare features remains ϵ_2. As a result, the balance between common and rare feature learning is restored, allowing the neurons to sufficiently learn both types of features. This result corresponds to Theorem 3.4 in the main text.

Theorem E.2 (Contrastive Learning with Thresholded Hard Negative Samples). *After introducing the threshold function, suppose that*

$$\rho = 1 - \Theta(\epsilon_2/\epsilon_1). \tag{97}$$

Then, for each neuron $k \in [m]$ and $t \geq T_2$, the learned weight can be decomposed as

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k \cap \mathcal{S}_1} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k \cap \mathcal{S}_2} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{98}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega(1), \qquad \alpha_{k,p}^{V_2} \geq \Omega(1), \tag{99}$$

and

$$\alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{100}$$

E.3 Proof of Theorem E.1 and Theorem E.2

The proofs of Theorems E.1 and E.2 follow the general proof framework established in Appendices F.1, F.2, and F.3. Therefore, we focus on the key differences introduced by hard negative samples and present the corresponding modifications to the analysis below.

Although hard negative samples improve the learning of rare features, using too many hard negative samples may excessively reduce the learning of common features. As a result, the neurons may focus too much on rare features while learning less from common features. To control this effect, we introduce a threshold function into hard negative sampling.

Following the same analysis as in Theorem F.1, we consider the negative gradient along the feature direction $\boldsymbol{M}_p^{\boldsymbol{V}}$. After introducing the threshold function, we have

$$\begin{aligned}
&\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[\mathbf{1}_{(|\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{Y}_i)-\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{Y}_j)|\leq Th)} g_k(\boldsymbol{Y}_j)\langle\nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i), \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&\frac{1}{L^2}\sum_{S_1}\langle \boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle \mathbb{E}_{HN}\left[\mathbf{1}_{(|\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{Y}_i)-\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{Y}_j)|\leq Th)}\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{\rho}{L^2}\langle \boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle \mathbb{E}\left[\mathbf{1}_{(|\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{Y}_i)-\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{Y}_j)|\leq Th)}\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right].
\end{aligned} \tag{101}$$

Here, ρ denotes the effective proportion of shared common features involved in hard-negative training under the current hard negative sampling setting. Since thresholding retains only a subset of the hard negative samples, the effective proportion ρ is reduced after thresholding. Consequently, the negative gradient acting on shared common feature directions is weakened.

Therefore, following the same positive and negative gradient decomposition as in the proof of Theorem F.1, we have

$$\begin{aligned}
&\mathbb{E}_{(i,j)\sim\mathcal{P}^{\mathrm{all}}}\left[\left(a_{i,j}\frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\langle \boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&\frac{1}{2L^2}\langle \boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle \mathbb{E}_{(i)\sim\mathcal{P}^+}\left[\hat{\boldsymbol{z}}_{i,p}\hat{\boldsymbol{z}}_{i,p}\right] - \frac{\rho}{2L^2}\langle \boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_1}}\rangle \mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[\mathbf{1}_{(|\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{Y}_i)-\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{Y}_j)|\leq Th)}\hat{\boldsymbol{z}}_{j,p}\hat{\boldsymbol{z}}_{i,p}\right] \\
&+\frac{1}{2L^2}\langle \boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L_2}}\rangle \mathbb{E}_{(i)\sim\mathcal{P}^+}\left[\hat{\boldsymbol{z}}_{i,p}\hat{\boldsymbol{z}}_{i,p}\right].
\end{aligned} \tag{102}$$

Then, the update of $\boldsymbol{w}_k^{(t)}$ along the feature direction $\boldsymbol{M}_p^{\boldsymbol{V}}$ is given by

$$\begin{aligned}
&\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\
=&\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle - \eta \langle \nabla_{\boldsymbol{w}_k} L_{\mathrm{aug}}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}), \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
=&(1-\eta\lambda)\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + (1-\rho)\epsilon_1 \frac{\eta \log\log d}{d} \langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L_1}} \rangle + \epsilon_2 \frac{\eta \log\log d}{d} \langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L_2}} \rangle \\
&\pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right).
\end{aligned} \tag{103}$$

Since thresholding reduces the effective proportion ρ of shared common features involved in hard-negative training, the factor $1-\rho$ increases. Therefore, the effective learning rate $(1-\rho)\epsilon_1$ of common features increases, preventing common feature learning from being excessively suppressed.

We finally consider the effect of the threshold function by comparing the following two cases.

Case 1: We first consider the case where too many hard negative samples are introduced, such that

$$(1-\rho)\epsilon_1 \ll \epsilon_2. \tag{104}$$

In this case, the learning of common features is excessively suppressed by hard negative samples. Following the same analysis as above, we obtain the feature decomposition result. For each neuron $k \in [m]$ and $t \geq T_2$, the weight learns the following set of features:

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{105}$$

where

$$\alpha_{k,p}^{V_1} \leq o\left(\frac{1}{\sqrt{d_1}}\right), \qquad \alpha_{k,p}^{V_2} \geq \Omega\left(1\right), \tag{106}$$

$$\alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{107}$$

Case 2: We next consider the case where the threshold function is introduced and

$$\rho = 1 - \Theta(\epsilon_2/\epsilon_1). \tag{108}$$

Since thresholding reduces the effective proportion ρ of shared common features involved in hard-negative training, the excessive suppression of common features is mitigated. Following the same analysis, we obtain the feature decomposition result. For each neuron $k \in [m]$ and $t \geq T_2$, the weight learns the following set of features:

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_1} \boldsymbol{M}_p^{\boldsymbol{V_1}} + \sum_{p \in \mathcal{N}_k} \alpha_{k,p}^{V_2} \boldsymbol{M}_p^{\boldsymbol{V_2}} + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^{V} + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{109}$$

where

$$\alpha_{k,p}^{V_1} \geq \Omega\left(1\right), \qquad \alpha_{k,p}^{V_2} \geq \Omega\left(1\right), \tag{110}$$

$$\alpha'_{k,p} \leq o\left(\frac{1}{\sqrt{d}}\right), \qquad |\beta_{k,p}| \leq o\left(\frac{1}{\sqrt{d_1}}\right). \tag{111}$$

Therefore, when too many hard negative samples are introduced, the learning of common features can be excessively suppressed. By introducing the threshold function, the effective proportion ρ of shared common features involved in hard-negative training is reduced, which increases the effective learning rate $(1-\rho)\epsilon_1$ of common features. When $\epsilon_2 = \Theta(1)(1-\rho)\epsilon_1$, the effective learning rates of common and rare features become comparable, allowing the neurons to sufficiently learn both types of features.

F Detailed proofs of results in Appendix C

F.1 Stage 1

In this section we analyze the training process at the initial stage. Here we define the stage transition time

$$T_1 = \Theta\left(\frac{d}{\eta \log\log d}\log\left(2(1+\epsilon)\frac{d_1}{d}\right)\right) \tag{112}$$

to be the iteration when

$$\|\boldsymbol{M}^{\boldsymbol{V}}(\boldsymbol{M}^{\boldsymbol{V}})^\top \boldsymbol{w}_k^{(t)}\|_2^2 \geq \tfrac{1}{2}\|\boldsymbol{w}_k^{(t)}\|_2^2, \tag{113}$$

where the neuron weights are more concentrated in the image feature space.

F.1.1 Theorem F.1

Before stating Theorem F.1, we give a short description of its parts: (a) For all neurons, most of the weights lie in the image feature subspace; (b) Lucky neurons are strongly aligned with their associated feature directions; (c) Neurons not in the set $\mathcal{M}_p$ have only weak alignment with feature p; and (d) All neuron weights have only small components in non-feature directions.

Theorem F.1 (Initial feature decoupling)**.** *At iteration $t = T_1$, we have the following results:*

(a) For all $k \in [m]$,

$$\|\boldsymbol{M}^{\boldsymbol{V}}(\boldsymbol{M}^{\boldsymbol{V}})^\top \boldsymbol{w}_k^{(T_1)}\|_2^2 \geq \tfrac{1}{2}\|\boldsymbol{w}_k^{(T_1)}\|_2^2. \tag{114}$$

(b) For each $p \in [d]$, and each $k \in \mathcal{M}_p^\star$,

$$|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle| \geq \sqrt{1+\gamma}\,\frac{\sqrt{\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T_1)}\|_2. \tag{115}$$

(c) For each $p \in [d]$, and each $k \notin \mathcal{M}_p$,

$$|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle| \leq \sqrt{1-\gamma}\,\frac{\sqrt{\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T_1)}\|_2. \tag{116}$$

(d) For each $k \in [m]$ and $p \in [d_1] \setminus [d]$,

$$|\langle \boldsymbol{w}_k^{(T_1)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^\perp\rangle| \leq \mathcal{O}\left(\sqrt{\tfrac{\log d}{d_1}}\right)\|\boldsymbol{w}_k^{(T_1)}\|_2. \tag{117}$$

F.1.2 Useful Lemmas

In Lemma F.1, we characterize the population-gradient signals along the feature and non-feature directions during the initial stage. For positive pairs, the gradient along each image feature direction is determined by the alignment of the corresponding text-side neuron weight with the associated text feature. In contrast, negative pairs produce zero expected gradient along the feature directions, while both positive and negative pairs produce zero expected gradient along the non-feature directions.

Lemma F.1 (Gradient characterization)**.** *Let $f_k(\cdot)$ and $g_k(\cdot)$ denote the k-th neurons at iteration $t \leq T_1$. Then the following hold:*

(a) For each $p \in [d]$, under the positive-pair distribution $\mathcal{P}^+$,

$$\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[g_k(\boldsymbol{Y}_i)\left\langle \nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i), \boldsymbol{M}_p^V\right\rangle\right] = \frac{1}{L^2}\left\langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^L\right\rangle \mathbb{E}\left[\hat{\boldsymbol{z}}_{i,p}\hat{\boldsymbol{z}}_{i,p}\right]. \tag{118}$$

(b) For each $p \in [d]$, under the negative-pair distribution $\mathcal{P}^-$,

$$\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[g_k(\boldsymbol{Y}_j)\left\langle \nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i), \boldsymbol{M}_p^V\right\rangle\right] = 0. \tag{119}$$

(c) For the orthogonal component $(\boldsymbol{M}^V)^\perp$, under both the positive-pair and negative-pair distributions, we have

$$\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[g_k(\boldsymbol{Y}_i)\left\langle\nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i),(\boldsymbol{M}^V)^\perp\right\rangle\right]=0, \tag{120}$$

and

$$\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[g_k(\boldsymbol{Y}_j)\left\langle\nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i),(\boldsymbol{M}^V)^\perp\right\rangle\right]=0. \tag{121}$$

In Lemma F.2, we show that for each neuron $k\in[m]$, the components of the neuron weights $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ in the corresponding non-feature subspaces remain essentially unchanged, up to negligible variation.

Lemma F.2. *For all $t\le T_1$, the following properties hold:*

(a) For the image-side neuron weight $\boldsymbol{w}_k^{(t)}$,

$$\left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top\boldsymbol{w}_k^{(t)}\right\|_2^2\le(1+\epsilon)\left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top\boldsymbol{w}_k^{(0)}\right\|_2^2. \tag{122}$$

(b) For the text-side neuron weight $\boldsymbol{v}_k^{(t)}$,

$$\left\|(\boldsymbol{M}^L)^\perp((\boldsymbol{M}^L)^\perp)^\top\boldsymbol{v}_k^{(t)}\right\|_2^2\le(1+\epsilon)\left\|(\boldsymbol{M}^L)^\perp((\boldsymbol{M}^L)^\perp)^\top\boldsymbol{v}_k^{(0)}\right\|_2^2. \tag{123}$$

F.1.3 Proof of Theorem F.1

Proof of Theorem F.1(a): The result in (a) follows directly from the orthogonal-component bound established in Lemma F.2. We have

$$\begin{aligned}\|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top\boldsymbol{w}_k^{(T)}\|_2^2=&\|\boldsymbol{w}_k^{(T)}\|_2^2-\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top\boldsymbol{w}_k^{(T)}\|_2^2\\ \ge&\|\boldsymbol{w}_k^{(T)}\|_2^2-(1+\epsilon)\,\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top\boldsymbol{w}_k^{(0)}\|_2^2\\ \ge&\|\boldsymbol{w}_k^{(T)}\|_2^2-(1+\epsilon)\,\|\boldsymbol{w}_k^{(0)}\|_2^2\\ \ge&\|\boldsymbol{w}_k^{(T)}\|_2^2-\frac{(1+\epsilon)\,\|\boldsymbol{w}_k^{(T)}\|_2^2}{\frac{d}{d_1}\Theta(e^{\frac{\eta\log\log d}{d}\cdot T})}\\ \ge&\frac{1}{2}\|\boldsymbol{w}_k^{(T)}\|_2^2.\end{aligned} \tag{124}$$

Therefore, it suffices to choose

$$T_1\ge\frac{d}{\eta\log\log d}\log\left(2(1+\epsilon)\frac{d_1}{d}\right), \tag{125}$$

which guarantees that

$$\frac{(1+\epsilon)}{\frac{d}{d_1}\Theta\left(e^{\frac{\eta\log\log d}{d}T_1}\right)}\le\frac{1}{2}. \tag{126}$$

Similarly, for the text-side neuron weight $\boldsymbol{v}_k$, we have

$$\|\boldsymbol{M}^L(\boldsymbol{M}^L)^\top\boldsymbol{v}_k^{(T)}\|_2^2\ge\frac{1}{2}\|\boldsymbol{v}_k^{(T)}\|_2^2. \tag{127}$$

This completes the proof. □

Proof of Theorem F.1(b):

$$\begin{aligned}
\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle &= \frac{1}{2}\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\Big) \\
&= \left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\Big) \\
&\geq \left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T\frac{\sqrt{c_1\log d}}{\sqrt{d}}\sqrt{\sum_{p\in[d]}\left[\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\Big)\right]^2} \\
&= \frac{\sqrt{c_1\log d}}{\sqrt{d}}\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T\sqrt{\sum_{p\in[d]}\left[\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\Big)\right]^2} \\
&= \frac{\sqrt{c_1\log d}}{\sqrt{d}}\sqrt{\sum_{p\in[d]}\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle^2} \\
&= \frac{\sqrt{c_1\log d}}{\sqrt{d}}\|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T)}\|_2 \\
&\geq \frac{1}{\sqrt{2}}\frac{\sqrt{c_1\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T)}\|_2 \\
&\geq \sqrt{1+\gamma}\frac{\sqrt{\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T)}\|_2
\end{aligned} \tag{128}$$

□

Proof of Theorem F.1(c):

$$\begin{aligned}
\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle &= \frac{1}{2}\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\Big) \\
&= \left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\Big) \\
&\leq \left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T\frac{\sqrt{c_2\log d}}{\sqrt{d}}\sqrt{\sum_{p\in[d]}\left[\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\Big)\right]^2} \\
&= \frac{\sqrt{c_2\log d}}{\sqrt{d}}\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T\sqrt{\sum_{p\in[d]}\left[\frac{1}{2}\Big(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\Big)\right]^2} \\
&= \frac{\sqrt{c_2\log d}}{\sqrt{d}}\sqrt{\sum_{p\in[d]}\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle^2} \\
&= \frac{\sqrt{c_2\log d}}{\sqrt{d}}\|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T)}\|_2 \\
&\leq \frac{\sqrt{c_2\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T)}\|_2 \\
&\leq \sqrt{1-\gamma}\frac{\sqrt{\log d}}{\sqrt{d}}\|\boldsymbol{w}_k^{(T)}\|_2
\end{aligned} \tag{129}$$

□

Proof of Theorem F.1(d): At initialization we have

$$\boldsymbol{w}_k^{(0)} \sim \mathcal{N}(0, \sigma_0^2 I_{d_1}). \tag{130}$$

Hence for any unit vector $(\boldsymbol{M}^V)_p^\perp$, the projection satisfies

$$\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp \rangle \sim \mathcal{N}(0, \sigma_0^2). \tag{131}$$

By the standard Gaussian tail bound (sub-Gaussian with parameter σ_0),

$$\mathbb{P}\Big(\Big|\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp \rangle\Big| > \sigma_0 \sqrt{2 \log d}\Big) \le 2 \exp\Big(-\frac{t^2}{2\sigma_0^2}\Big) = \frac{2}{d}. \tag{132}$$

Therefore, with high probability,

$$\Big|\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp \rangle\Big| \le \sigma_0 \, O(\sqrt{\log d}). \tag{133}$$

Moreover, since

$$\|\boldsymbol{w}_k^{(0)}\|_2 = \Theta(\sigma_0 \sqrt{d_1}) \tag{134}$$

with high probability, the above bound is equivalently

$$\Big|\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp \rangle\Big| \le O\left(\sqrt{\frac{\log d}{d_1}}\right) \|\boldsymbol{w}_k^{(0)}\|_2. \tag{135}$$

We have

$$\begin{aligned}
\langle \boldsymbol{w}_k^{(T)}, (\boldsymbol{M}^V)_p^\perp \rangle &= (1-\eta\lambda)^T \langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp \rangle \pm \widetilde{O}\left(T\eta \max_{0 \le t < T} \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&\le (1-\eta\lambda)^T \langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp \rangle + \widetilde{O}(\epsilon) \|\boldsymbol{w}_k^{(0)}\|_2 \\
&\le O\left(\sqrt{\frac{\log d}{d_1}}\right) \|\boldsymbol{w}_k^{(0)}\|_2 \\
&\le O\left(\sqrt{\frac{\log d}{d_1}}\right) \|\boldsymbol{w}_k^{(T)}\|_2.
\end{aligned} \tag{136}$$

Similarly,

$$\begin{aligned}
\langle \boldsymbol{v}_k^{(T)}, (\boldsymbol{M}^L)_p^\perp \rangle &= (1-\eta\lambda)^T \langle \boldsymbol{v}_k^{(0)}, (\boldsymbol{M}^L)_p^\perp \rangle \pm \widetilde{O}\left(T\eta \max_{0 \le t < T} \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&\le (1-\eta\lambda)^T \langle \boldsymbol{v}_k^{(0)}, (\boldsymbol{M}^L)_p^\perp \rangle + \widetilde{O}(\epsilon) \|\boldsymbol{v}_k^{(0)}\|_2 \\
&\le O\left(\sqrt{\frac{\log d}{d_1}}\right) \|\boldsymbol{v}_k^{(0)}\|_2 \\
&\le O\left(\sqrt{\frac{\log d}{d_1}}\right) \|\boldsymbol{v}_k^{(T)}\|_2.
\end{aligned} \tag{137}$$

□

F.1.4 PROOF OF LEMMA F.1

Proof of Lemma F.1(a): For simplicity of notation, we omit $\boldsymbol{\theta}$ and ϕ. For the positive pair case, we consider $j = i$. Then,

$$
\begin{aligned}
&\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[g_k(\boldsymbol{Y}_i)\langle\nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i), \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&\mathbb{E}\left[g_k(\boldsymbol{Y}_i)\left\langle\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq 0}\cdot\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\right\rangle\right] \\
=&\mathbb{E}\left[\sum_{s=1}^{L}\langle\boldsymbol{v}_k, \boldsymbol{h}_{\boldsymbol{Y}}^{(s)}\rangle\left(\sum_{r=1}^{L}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\left(\sum_{s=1}^{L}\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\tilde{\boldsymbol{z}}_i^{(s)} + \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}\rangle\right)\left(\sum_{r=1}^{L}\langle\boldsymbol{M}^{\boldsymbol{V}}\tilde{\boldsymbol{z}}_i^{(r)} + \tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\left(\sum_{s=1}^{L}\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\tilde{\boldsymbol{z}}_i^{(s)}\rangle\right)\left(\sum_{r=1}^{L}\langle\boldsymbol{M}^{\boldsymbol{V}}\tilde{\boldsymbol{z}}_i^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_i\rangle\cdot\langle\boldsymbol{M}^{\boldsymbol{V}}\hat{\boldsymbol{z}}_i, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_i\rangle\cdot\left\langle\sum_{p'\in[d]}\boldsymbol{M}_{p'}^{\boldsymbol{V}}\hat{\boldsymbol{z}}_{i,p'}, \boldsymbol{M}_p^{\boldsymbol{V}}\right\rangle\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_i\rangle\cdot\sum_{p'\in[d]}\langle\boldsymbol{M}_{p'}^{\boldsymbol{V}}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\hat{\boldsymbol{z}}_{i,p'}\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_i\rangle\cdot\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\sum_{q\in[d]}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\hat{\boldsymbol{z}}_{i,q}\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{1}{L^2}\sum_{q\in[d]}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\mathbb{E}\left[\hat{\boldsymbol{z}}_{i,q}\hat{\boldsymbol{z}}_{i,p}\right].
\end{aligned}
\tag{138}
$$

In particular, since the entries of the latent vectors $\boldsymbol{z_i}$ are independent, we have

$$
\frac{1}{L^2}\sum_{q\in[d]}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\mathbb{E}\left[\hat{\boldsymbol{z}}_{i,q}\hat{\boldsymbol{z}}_{i,p}\right] = \frac{1}{L^2}\langle\boldsymbol{v}_k, \boldsymbol{M}_p^{\boldsymbol{L}}\rangle\mathbb{E}\left[\hat{\boldsymbol{z}}_{i,p}\hat{\boldsymbol{z}}_{i,p}\right]. \tag{139}
$$

This completes the proof. □

Proof of Lemma F.1(b): For the negative pair case, we consider $j \neq i$. Then,

$$
\begin{aligned}
&\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[g_k(\boldsymbol{Y}_j)\langle\nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i), \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&\mathbb{E}\left[g_k(\boldsymbol{Y}_j)\left\langle\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq 0}\cdot\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\right\rangle\right] \\
=&\mathbb{E}\left[\sum_{s=1}^{L}\langle\boldsymbol{v}_k, \boldsymbol{h}_{\boldsymbol{Y}}^{(s)}\rangle\left(\sum_{r=1}^{L}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\left(\sum_{s=1}^{L}\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\tilde{\boldsymbol{z}}_j^{(s)}+\tilde{\boldsymbol{\xi}}_{Y_j}^{(s)}\rangle\right)\left(\sum_{r=1}^{L}\langle\boldsymbol{M}^{\boldsymbol{V}}\tilde{\boldsymbol{z}}_i^{(r)}+\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\left(\sum_{s=1}^{L}\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\tilde{\boldsymbol{z}}_j^{(s)}\rangle\right)\left(\sum_{r=1}^{L}\langle\boldsymbol{M}^{\boldsymbol{V}}\tilde{\boldsymbol{z}}_i^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_j\rangle\cdot\langle\boldsymbol{M}^{\boldsymbol{V}}\hat{\boldsymbol{z}}_i, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_j\rangle\cdot\left\langle\sum_{p'\in[d]}\boldsymbol{M}_{p'}^{\boldsymbol{V}}\hat{\boldsymbol{z}}_{i,p'}, \boldsymbol{M}_p^{\boldsymbol{V}}\right\rangle\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_j\rangle\cdot\sum_{p'\in[d]}\langle\boldsymbol{M}_{p'}^{\boldsymbol{V}}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\hat{\boldsymbol{z}}_{i,p'}\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k, \boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_j\rangle\cdot\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\sum_{q\in[d]}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\hat{\boldsymbol{z}}_{j,q}\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{1}{L^2}\sum_{q\in[d]}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\mathbb{E}\left[\hat{\boldsymbol{z}}_{j,q}\hat{\boldsymbol{z}}_{i,p}\right].
\end{aligned}
\tag{140}
$$

Since $j \neq i$, the image sample $\boldsymbol{X}_i$ and the text sample $\boldsymbol{Y}_j$ are independent negative pairs. Therefore, for every $q \in [d]$ and $p \in [d]$,

$$
\mathbb{E}\left[\hat{\boldsymbol{z}}_{j,q}\hat{\boldsymbol{z}}_{i,p}\right] = 0. \tag{141}
$$

Hence,

$$
\frac{1}{L^2}\sum_{q\in[d]}\langle\boldsymbol{v}_k, \boldsymbol{M}_q^{\boldsymbol{L}}\rangle\mathbb{E}\left[\hat{\boldsymbol{z}}_{j,q}\hat{\boldsymbol{z}}_{i,p}\right] = 0. \tag{142}
$$

This completes the proof. □

Proof of Lemma F.1(c): For the orthogonal component $(\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}$, we first consider the positive pair case when $j = i$.

$$
\begin{aligned}
&\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[g_k(\boldsymbol{Y}_i)\langle\nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i), (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\rangle\right] \\
=&\mathbb{E}\left[g_k(\boldsymbol{Y}_i)\left\langle\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq 0}\cdot\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\right\rangle\right] \\
=&\mathbb{E}\left[g_k(\boldsymbol{Y}_i)\cdot\left(\sum_{r=1}^{L}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\rangle\right)\right] \\
=&\mathbb{E}\left[\sum_{s=1}^{L}\langle\boldsymbol{v}_k,\boldsymbol{h}_{\boldsymbol{Y}}^{(s)}\rangle\cdot\left(\sum_{r=1}^{L}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\left(\sum_{s=1}^{L}\langle\boldsymbol{v}_k,\boldsymbol{M}^{\boldsymbol{L}}\tilde{\boldsymbol{z}}_i^{(s)}+\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}\rangle\right)\left(\sum_{r=1}^{L}\langle\boldsymbol{M}^{\boldsymbol{V}}\tilde{\boldsymbol{z}}_i^{(r)}+\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\left(\sum_{s=1}^{L}\langle\boldsymbol{v}_k,\boldsymbol{M}^{\boldsymbol{L}}\tilde{\boldsymbol{z}}_i^{(s)}\rangle\right)\left(\sum_{r=1}^{L}\langle\boldsymbol{M}^{\boldsymbol{V}}\tilde{\boldsymbol{z}}_i^{(r)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\rangle\right)\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k,\boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_i\rangle\cdot\langle\boldsymbol{M}^{\boldsymbol{V}}\hat{\boldsymbol{z}}_i, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\rangle\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k,\boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_i\rangle\cdot\left\langle\sum_{p\in[d]}\boldsymbol{M}_p^{\boldsymbol{V}}\hat{\boldsymbol{z}}_{i,p}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\right\rangle\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k,\boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_i\rangle\cdot\sum_{p\in[d]}\langle\boldsymbol{M}_p^{\boldsymbol{V}}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\rangle\hat{\boldsymbol{z}}_{i,p}\right] \\
=&\frac{1}{L^2}\mathbb{E}\left[\langle\boldsymbol{v}_k,\boldsymbol{M}^{\boldsymbol{L}}\hat{\boldsymbol{z}}_i\rangle\cdot 0\right] \\
=&0.
\end{aligned}
\tag{143}
$$

The last equality follows from the orthogonality condition

$$
\langle\boldsymbol{M}_p^{\boldsymbol{V}}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\rangle = 0, \qquad \forall p\in[d]. \tag{144}
$$

This completes the proof. □

When we consider the negative-pair case with $j \neq i$, we similarly have

$$
\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[g_k(\boldsymbol{Y}_j)\left\langle\nabla_{\boldsymbol{w}_k} f_k(\boldsymbol{X}_i), (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}\right\rangle\right] = 0. \tag{145}
$$

The proof follows the same argument as in the positive pair case.

F.1.5 Proof of Lemma F.2

Proof: First, we must determine the precise gradient expression for each feature $\boldsymbol{M}_p^{\boldsymbol{V}}$ and $(\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp}$. The gradient descent update for the projection of $\boldsymbol{w}_k^{(t)}$ onto $\boldsymbol{M}_p^{\boldsymbol{V}}$ can be written as

$$
\begin{aligned}
&\langle\boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle \\
=&\langle\boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle - \eta\langle\nabla_{\boldsymbol{w}_k} L_{\text{aug}}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}), \boldsymbol{M}_p^{\boldsymbol{V}}\rangle \quad \pm\widetilde{O}\left(\eta\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
\overset{①}{=}&(1-\eta\lambda)\langle\boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle + \eta\mathbb{E}_{(i,j)\sim\mathcal{P}^{all}}\left[\left(a_{i,j}\frac{e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)}, \boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right] \\
&\pm\widetilde{O}\left(\eta\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right).
\end{aligned}
\tag{146}
$$

① is because λ is the coefficient of the regularization term, as well as the gradient formula obtained earlier. $\widetilde{O}\Big(\eta\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\Big)$ is due to the approximation between population gradients and empirical gradients.

We can use Lemma F.1 (a) and Lemma F.1 (b), and note that $\|\boldsymbol{w}_k\|_2, \|\boldsymbol{v}_k\|_2 \approx 0$, to obtain that:

$$
\begin{aligned}
&\mathbb{E}_{(i,j)\sim\mathcal{P}^{\text{all}}}\left[\left(a_{i,j}\frac{e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)},\boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right]\\
=&p_+\cdot\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[\left(a_{i,i}\frac{e^{a_{i,i}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_i)\right)}}{1+e^{a_{i,i}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_i)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)},\boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right]\\
+&p_-\cdot\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\left[\left(a_{i,j}\frac{e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}{1+e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right)\cdot\ g_k(\boldsymbol{Y}_j)\sum_{r=1}^{L}\mathbf{1}_{|\langle\boldsymbol{w}_k,\boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle|\geq b_{f_k}}\langle\boldsymbol{h}_{\boldsymbol{X}}^{(r)},\boldsymbol{M}_p^{\boldsymbol{V}}\rangle\right]\\
=&\frac{1}{2L^2}\langle\boldsymbol{v}_k,\boldsymbol{M}_p^{\boldsymbol{L}}\rangle\mathbb{E}_{(i)\sim\mathcal{P}^+}\left[\hat{\boldsymbol{z}}_{i,p}\hat{\boldsymbol{z}}_{i,p}\right].
\end{aligned}
\tag{147}
$$

Putting all the above calculations together, we have

$$
\langle\boldsymbol{w}_k^{(t+1)},\boldsymbol{M}_p^{\boldsymbol{V}}\rangle=(1-\eta\lambda)\langle\boldsymbol{w}_k^{(t)},\boldsymbol{M}_p^{\boldsymbol{V}}\rangle+\epsilon_p\frac{\eta\log\log d}{d}\langle\boldsymbol{v}_k,\boldsymbol{M}_p^{\boldsymbol{L}}\rangle\pm\ \widetilde{O}\left(\eta\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right).
\tag{148}
$$

Similarly, we have

$$
\langle\boldsymbol{v}_k^{(t+1)},\boldsymbol{M}_p^{\boldsymbol{L}}\rangle=(1-\eta\lambda)\langle\boldsymbol{v}_k^{(t)},\boldsymbol{M}_p^{\boldsymbol{L}}\rangle+\epsilon_p\frac{\eta\log\log d}{d}\langle\boldsymbol{w}_k,\boldsymbol{M}_p^{\boldsymbol{V}}\rangle\pm\widetilde{O}\left(\eta\|\boldsymbol{w}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right).
\tag{149}
$$

Prior to the induction step, we establish, by a similar method, the stochastic gradient descent update of $\boldsymbol{w}_k$ along the dense feature direction $(\boldsymbol{M}^{\boldsymbol{V}})_p^\perp$. Specifically, we obtain the following update equation:

$$
\begin{aligned}
&\langle\boldsymbol{w}_k^{(t+1)},(\boldsymbol{M}^{\boldsymbol{V}})_p^\perp\rangle\\
=&\langle\boldsymbol{w}_k^{(t)},(\boldsymbol{M}^{\boldsymbol{V}})_p^\perp\rangle-\eta\langle\nabla_{\boldsymbol{w}_k}L_{\text{aug}}(\boldsymbol{f}_\theta,\boldsymbol{g}_\phi),(\boldsymbol{M}^{\boldsymbol{V}})_p^\perp\rangle\pm\widetilde{O}\left(\eta\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right)\\
=&(1-\eta\lambda)\langle\boldsymbol{w}_k^{(t)},(\boldsymbol{M}^{\boldsymbol{V}})_p^\perp\rangle-\eta\Big\langle\nabla_{\boldsymbol{w}_k}L(\boldsymbol{f}_\theta,\boldsymbol{g}_\phi),(\boldsymbol{M}^{\boldsymbol{V}})_p^\perp\Big\rangle\pm\widetilde{O}\left(\eta\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right)\\
=&(1-\eta\lambda)\langle\boldsymbol{w}_k^{(t)},(\boldsymbol{M}^{\boldsymbol{V}})_p^\perp\rangle\pm\widetilde{O}\left(\eta\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right).
\end{aligned}
\tag{150}
$$

Similarly, we have

$$
\langle\boldsymbol{v}_k^{(t+1)},(\boldsymbol{M}^{\boldsymbol{L}})_p^\perp\rangle=(1-\eta\lambda)\langle\boldsymbol{v}_k^{(t)},(\boldsymbol{M}^{\boldsymbol{L}})_p^\perp\rangle\pm\widetilde{O}\left(\eta\|\boldsymbol{w}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right).
\tag{151}
$$

□

The one step gradient descent updates have been derived above. We now analyze the accumulated dynamics over T iterations by solving the coupled update equations from iteration 0 to T.

Proof: Recall that we have the following coupled update equations:

$$\begin{aligned}\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle &= (1-\eta\lambda)\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \epsilon_p \frac{\eta \log\log d}{d} \langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right), \\ \langle \boldsymbol{v}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle &= (1-\eta\lambda)\langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle + \epsilon_p \frac{\eta \log\log d}{d} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right).\end{aligned} \tag{152}$$

Ignoring the lower-order perturbation terms temporarily, the recursion can be written as

$$\begin{bmatrix} \langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\ \langle \boldsymbol{v}_k^{(t+1)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \end{bmatrix} = \begin{bmatrix} 1-\eta\lambda & \epsilon_p \frac{\eta \log\log d}{d} \\ \epsilon_p \frac{\eta \log\log d}{d} & 1-\eta\lambda \end{bmatrix} \begin{bmatrix} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\ \langle \boldsymbol{v}_k^{(t)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \end{bmatrix}. \tag{153}$$

The transition matrix has eigenvectors $(1,1)^\top$ and $(1,-1)^\top$, with corresponding eigenvalues

$$1-\eta\lambda+\epsilon_p \frac{\eta \log\log d}{d}, \qquad 1-\eta\lambda-\epsilon_p \frac{\eta \log\log d}{d}. \tag{154}$$

Therefore, the initialization can be decomposed as

$$\begin{bmatrix} \langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\ \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \end{bmatrix} = \frac{\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle - \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{155}$$

Applying the recursion from time 0 to time T gives

$$\begin{aligned}\begin{bmatrix} \langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\ \langle \boldsymbol{v}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \end{bmatrix} &= \frac{\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle}{2} \left(1-\eta\lambda+\epsilon_p \frac{\eta \log\log d}{d}\right)^T \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &\quad + \frac{\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle - \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle}{2} \left(1-\eta\lambda-\epsilon_p \frac{\eta \log\log d}{d}\right)^T \begin{bmatrix} 1 \\ -1 \end{bmatrix}.\end{aligned} \tag{156}$$

Taking the first coordinate, we obtain

$$\begin{aligned}\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle &= \frac{1}{2}\left[\left(1-\eta\lambda+\epsilon_p \frac{\eta \log\log d}{d}\right)^T + \left(1-\eta\lambda-\epsilon_p \frac{\eta \log\log d}{d}\right)^T\right] \langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\ &\quad + \frac{1}{2}\left[\left(1-\eta\lambda+\epsilon_p \frac{\eta \log\log d}{d}\right)^T - \left(1-\eta\lambda-\epsilon_p \frac{\eta \log\log d}{d}\right)^T\right] \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle.\end{aligned} \tag{157}$$

Taking the second coordinate, we obtain

$$\begin{aligned}\langle \boldsymbol{v}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle &= \frac{1}{2}\left[\left(1-\eta\lambda+\epsilon_p \frac{\eta \log\log d}{d}\right)^T - \left(1-\eta\lambda-\epsilon_p \frac{\eta \log\log d}{d}\right)^T\right] \langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\ &\quad + \frac{1}{2}\left[\left(1-\eta\lambda+\epsilon_p \frac{\eta \log\log d}{d}\right)^T + \left(1-\eta\lambda-\epsilon_p \frac{\eta \log\log d}{d}\right)^T\right] \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle.\end{aligned} \tag{158}$$

Finally, adding back the perturbation terms, and using that the additional multiplication by $\epsilon_p \frac{\eta \log\log d}{d}$ is lower order, we have

$$\begin{aligned}\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle &= \frac{1}{2}\left[\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T + \left(1-\eta\lambda-\epsilon_p\frac{\eta\log\log d}{d}\right)^T\right]\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\ &\quad + \frac{1}{2}\left[\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T - \left(1-\eta\lambda-\epsilon_p\frac{\eta\log\log d}{d}\right)^T\right]\langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \\ &\quad \pm \widetilde{O}\left(T\eta \max_{0\le t<T} \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right),\end{aligned} \tag{159}$$

and

$$\begin{aligned}\langle \boldsymbol{v}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle &= \frac{1}{2}\left[\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T - \left(1-\eta\lambda-\epsilon_p\frac{\eta\log\log d}{d}\right)^T\right]\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle \\ &\quad + \frac{1}{2}\left[\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T + \left(1-\eta\lambda-\epsilon_p\frac{\eta\log\log d}{d}\right)^T\right]\langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \\ &\quad \pm \widetilde{O}\left(T\eta \max_{0\le t<T} \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right).\end{aligned} \tag{160}$$

Therefore, when $T \gg \frac{d}{\eta \log\log d}$, we have $\left(1-\eta\lambda-\epsilon_p\frac{\eta\log\log d}{d}\right)^T = 0$. Substituting this equation into the previous expressions yields

$$\begin{aligned}\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle &= \frac{1}{2}\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T \left(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\right) \\ &\quad \pm \widetilde{O}\left(T\eta \max_{0\le t<T} \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right),\end{aligned} \tag{161}$$

and

$$\begin{aligned}\langle \boldsymbol{v}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle &= \frac{1}{2}\left(1-\eta\lambda+\epsilon_p\frac{\eta\log\log d}{d}\right)^T \left(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle\right) \\ &\quad \pm \widetilde{O}\left(T\eta \max_{0\le t<T} \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right).\end{aligned} \tag{162}$$

This completes the proof. □

Proof: Recall that

$$\begin{aligned}\langle \boldsymbol{w}_k^{(t+1)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle &= (1-\eta\lambda)\langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle \pm \widetilde{O}\left(\eta\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\ \langle \boldsymbol{v}_k^{(t+1)}, (\boldsymbol{M}^{\boldsymbol{L}})_p^{\perp} \rangle &= (1-\eta\lambda)\langle \boldsymbol{v}_k^{(t)}, (\boldsymbol{M}^{\boldsymbol{L}})_p^{\perp} \rangle \pm \widetilde{O}\left(\eta\|\boldsymbol{w}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right).\end{aligned} \tag{163}$$

Ignoring the lower-order perturbation term temporarily and recursively applying the above update from time 0 to time T, we obtain

$$\langle \boldsymbol{w}_k^{(T)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle = (1-\eta\lambda)^T \langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle. \tag{164}$$

Adding back the perturbation terms, and noting that the perturbation accumulates linearly over T iterations, yields

$$\langle \boldsymbol{w}_k^{(T)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle = (1-\eta\lambda)^T \langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle \pm \widetilde{O}\left(T\eta \max_{0\le t<T} \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}} \right). \tag{165}$$

Similarly,

$$\langle \boldsymbol{v}_k^{(T)}, (\boldsymbol{M}^{\boldsymbol{L}})_p^{\perp} \rangle = (1-\eta\lambda)^T \langle \boldsymbol{v}_k^{(0)}, (\boldsymbol{M}^{\boldsymbol{L}})_p^{\perp} \rangle \pm \widetilde{O}\left(T\eta \max_{0\le t<T} \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}} \right). \tag{166}$$

This completes the proof. □

Next, we consider the sum and difference of the two modalities along the corresponding feature and orthogonal directions. The first equation shows that the sum of the projections of $\boldsymbol{w}_k^{(t)}$ and $\boldsymbol{v}_k^{(t)}$ onto their corresponding feature directions grows during training. The second equation shows that the difference between the projections of $\boldsymbol{w}_k^{(t)}$ and $\boldsymbol{v}_k^{(t)}$ onto their corresponding feature directions decreases, and thus the two projections become approximately equal when T is sufficiently large. The last two equations show that both the sum and difference of the projections onto the corresponding orthogonal directions remain controlled during training.

Proof:

$$\begin{aligned}\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle = &\left(1 - \eta\lambda + \epsilon_p \frac{\eta \log\log d}{d}\right)^T \left(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle + \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \right) \\ &\pm \widetilde{O}\left(T\eta \max_{0\le t<T} \left(\|\boldsymbol{w}_k^{(t)}\|_2 + \|\boldsymbol{v}_k^{(t)}\|_2 \right) \sqrt{\frac{qL}{|\mathcal{B}|}} \right).\end{aligned} \tag{167}$$

$$\begin{aligned}\langle \boldsymbol{w}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle - \langle \boldsymbol{v}_k^{(T)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle = &\left(1 - \eta\lambda - \epsilon_p \frac{\eta \log\log d}{d}\right)^T \left(\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{V}} \rangle - \langle \boldsymbol{v}_k^{(0)}, \boldsymbol{M}_p^{\boldsymbol{L}} \rangle \right) \\ &\pm \widetilde{O}\left(T\eta \max_{0\le t<T} \left(\|\boldsymbol{w}_k^{(t)}\|_2 + \|\boldsymbol{v}_k^{(t)}\|_2 \right) \sqrt{\frac{qL}{|\mathcal{B}|}} \right).\end{aligned} \tag{168}$$

$$\begin{aligned}\langle \boldsymbol{w}_k^{(T)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle + \langle \boldsymbol{v}_k^{(T)}, (\boldsymbol{M}^{\boldsymbol{L}})_p^{\perp} \rangle =&(1-\eta\lambda)^T \left(\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle + \langle \boldsymbol{v}_k^{(0)}, (\boldsymbol{M}^{\boldsymbol{L}})_p^{\perp} \rangle \right) \\ &\pm \widetilde{O}\left(T\eta \max_{0\le t<T} \left(\|\boldsymbol{w}_k^{(t)}\|_2 + \|\boldsymbol{v}_k^{(t)}\|_2 \right) \sqrt{\frac{qL}{|\mathcal{B}|}} \right).\end{aligned} \tag{169}$$

$$\begin{aligned}\langle \boldsymbol{w}_k^{(T)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle - \langle \boldsymbol{v}_k^{(T)}, (\boldsymbol{M}^{\boldsymbol{L}})_p^{\perp} \rangle =&(1-\eta\lambda)^T \left(\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^{\boldsymbol{V}})_p^{\perp} \rangle - \langle \boldsymbol{v}_k^{(0)}, (\boldsymbol{M}^{\boldsymbol{L}})_p^{\perp} \rangle \right) \\ &\pm \widetilde{O}\left(T\eta \max_{0\le t<T} \left(\|\boldsymbol{w}_k^{(t)}\|_2 + \|\boldsymbol{v}_k^{(t)}\|_2 \right) \sqrt{\frac{qL}{|\mathcal{B}|}} \right).\end{aligned} \tag{170}$$

□

Proof of Lemma F.2: Finally we give an upper bound of

$$\left\| (\boldsymbol{M}^V)^{\perp} ((\boldsymbol{M}^V)^{\perp})^{\top} \boldsymbol{w}_k^{(T)} \right\|_2^2. \tag{171}$$

We can calculate similarly, by

$$
\begin{aligned}
&\left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(T)}\right\|_2^2 \\
&= \sum_{p\in[d_1]\setminus[d]} \left[\langle \boldsymbol{w}_k^{(T)}, (\boldsymbol{M}^V)_p^\perp\rangle\right]^2 \\
&= \sum_{p\in[d_1]\setminus[d]} \left[(1-\eta\lambda)^T\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp\rangle \pm \widetilde{O}\left(T\eta \max_{0\le t<T}\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right)\right]^2 \\
&\le \sum_{p\in[d_1]\setminus[d]} \left[\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp\rangle + \widetilde{O}\left(T\eta \max_{0\le t<T}\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right)\right]^2 \\
&= \sum_{p\in[d_1]\setminus[d]} \langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp\rangle^2 + \sum_{p\in[d_1]\setminus[d]} 2\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp\rangle \widetilde{O}\left(T\eta \max_{0\le t<T}\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&\quad + \sum_{p\in[d_1]\setminus[d]} \widetilde{O}\left(T^2\eta^2 \max_{0\le t<T}\|\boldsymbol{v}_k^{(t)}\|_2^2\frac{qL}{|\mathcal{B}|}\right) \\
&\le \left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 + \max_{p\in[d_1]\setminus[d]}\left|\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp\rangle\right| \widetilde{O}\left(d_1 T\eta \max_{0\le t<T}\|\boldsymbol{v}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&\quad + \widetilde{O}\left(d_1 T^2\eta^2 \max_{0\le t<T}\|\boldsymbol{v}_k^{(t)}\|_2^2\frac{qL}{|\mathcal{B}|}\right) \\
&\overset{①}{\le} \left(1+\widetilde{O}\left(\frac{d_1 T\eta}{\sqrt{d_1}}\sqrt{\frac{qL}{|\mathcal{B}|}}\right)\right)\left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 + \widetilde{O}(\epsilon)\,\|\boldsymbol{w}_k^{(0)}\|_2^2 \\
&\overset{②}{\le} (1+\epsilon)\left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2.
\end{aligned}
\tag{172}
$$

① is because at initialization, we have (1) $\left|\langle \boldsymbol{w}_k^{(0)}, (\boldsymbol{M}^V)_p^\perp\rangle\right| \le \widetilde{O}\left(\frac{\|\boldsymbol{w}_k^{(0)}\|_2}{\sqrt{d_1}}\right)$, (2) $\|\boldsymbol{w}_k^{(0)}\|_2^2 \le O\left(\left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2\right)$, (3) using the boundedness of the training dynamics, $\max_{0\le t<T}\|\boldsymbol{v}_k^{(t)}\|_2 \le O(1)\|\boldsymbol{v}_k^{(0)}\|_2 = O(1)\|\boldsymbol{w}_k^{(0)}\|_2$, and (4) the empirical-gradient perturbation satisfies $d_1T^2\eta^2\frac{qL}{|\mathcal{B}|} \le \epsilon$. ② is because at initialization, (1) $\|\boldsymbol{w}_k^{(0)}\|_2^2 \le O\left(\left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2\right)$, and (2) $\widetilde{O}\left(\frac{d_1T\eta}{\sqrt{d_1}}\sqrt{\frac{qL}{|\mathcal{B}|}}\right) \le \epsilon$. Similarly, for the text-side weight $\boldsymbol{v}_k$, we have

$$
\left\|(\boldsymbol{M}^L)^\perp((\boldsymbol{M}^L)^\perp)^\top \boldsymbol{v}_k^{(T)}\right\|_2^2 \le (1+\epsilon)\left\|(\boldsymbol{M}^L)^\perp((\boldsymbol{M}^L)^\perp)^\top \boldsymbol{v}_k^{(0)}\right\|_2^2. \tag{173}
$$

Thus, the orthogonal components remain controlled. This completes the proof. □

Finally, we bound the squared ℓ_2-norm of $\boldsymbol{w}_k^{(t)}$ as follows:

Proof.

$$
\|\boldsymbol{w}_k^{(t)}\|_2^2 = \|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(t)}\|_2^2 + \|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(t)}\|_2^2 \overset{(1)}{\le} O(1)\|\boldsymbol{w}_k^{(0)}\|_2^2 \overset{(2)}{\le} \frac{1}{\mathrm{poly}(d_1)}. \tag{174}
$$

(1) is because:

$$\begin{aligned}\|\boldsymbol{w}_k^{(T)}\|_2^2 &= \|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T)}\|_2^2 + \|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(T)}\|_2^2 \\ &= \left\|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \left(1+\frac{\eta\log\log d}{d}\right)^{2T} + (1+\epsilon)\left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \\ &= \left\|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \Theta\Big(e^{\frac{\eta\log\log d}{d}\cdot T}\Big) + \left\|(\boldsymbol{M}^V)^\perp((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(0)}\right\|_2^2 \\ &= \frac{d}{d_1}\|\boldsymbol{w}_k^{(0)}\|_2^2\Theta\Big(e^{\frac{\eta\log\log d}{d}\cdot T}\Big) + \|\boldsymbol{w}_k^{(0)}\|_2^2 \\ &= \frac{d}{d_1}\Theta(e^{\frac{\eta\log\log d}{d}\cdot T})\|\boldsymbol{w}_k^{(0)}\|_2^2.\end{aligned} \tag{175}$$

(2) is because:

$$\begin{aligned}\|\boldsymbol{w}_k^{(0)}\|_2^2 &\le \sigma_0^2 d_1\left(1+\widetilde{O}\left(\frac{1}{\sqrt{d_1}}\right)\right) \\ &\le \Theta\left(\frac{1}{\mathrm{poly}(d_1)}\right) d_1\left(1+\widetilde{O}\left(\frac{1}{\sqrt{d_1}}\right)\right) \\ &= \frac{1}{\mathrm{poly}(d_1)}.\end{aligned} \tag{176}$$

□

F.2 Stage 2

The second stage is defined as the iterations $t \ge T_1$ but $t \le T_2$, where

$$T_2 = T_1 + \Theta\left(\frac{d\log d}{\eta\log\log d}\right) \tag{177}$$

is defined as the iteration when one of the neurons $k \in [m]$ satisfies

$$\left\|\boldsymbol{w}_k^{(T_2)}\right\|_2^2 \ge d\left\|\boldsymbol{w}_k^{(T_1)}\right\|_2^2. \tag{178}$$

F.2.1 Theorem F.2

Before stating Theorem F.2, we briefly describe its main results: (a) If a neuron k is a lucky neuron for feature p, then the projection of $\boldsymbol{w}_k^{(T_2)}$ onto $\boldsymbol{M}_p^V$ is on the same order as the full neuron weight $\|\boldsymbol{w}_k^{(T_2)}\|_2$. (b) The bias term $b_{f_k}^{(T_2)}$ grows with the neuron weight and reaches at least $\frac{\mathrm{polylog}(d)}{\sqrt{d}}\|\boldsymbol{w}_k^{(T_2)}\|_2$ at iteration T_2. (c) For each feature p, the projection of every ordinary neuron in $\mathcal{M}_p$ onto $\boldsymbol{M}_p^V$ is bounded by a constant multiple of the largest projection among the lucky neurons in $\mathcal{M}_p^\star$.

Theorem F.2 (Emergence of singletons)**.** *For each neuron $k \in [m]$, the following conditions hold at iteration $t = T_2$:*

(a) For each $p \in [d]$, if $k \in \mathcal{M}_p^\star$, then

$$\left|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^V\rangle\right| \ge \Omega(1)\|\boldsymbol{w}_k^{(T_2)}\|_2. \tag{179}$$

(b)

$$b_{f_k}^{(T_2)} \ge \frac{\mathrm{polylog}(d)}{\sqrt{d}}\|\boldsymbol{w}_k^{(T_2)}\|_2. \tag{180}$$

(c) Let

$$\alpha_p^\star = \max_{k\in\mathcal{M}_p^\star}\left|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^V\rangle\right|. \tag{181}$$

then there exists a constant $C_p = \Theta(1)$ such that

$$\left|\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p^V\rangle\right| \le C_p\alpha_p^\star, \quad \forall k \in \mathcal{M}_p. \tag{182}$$

F.2.2 THEOREM F.3

Theorem F.3 (Feature decomposition). *For each neuron $k \in [m]$ and $t \geq T_2$, the weight learns the following set of features:*

$$\boldsymbol{w}_k^{(t)} = \sum_{p \in \mathcal{N}_k} \alpha_{k,p} \boldsymbol{M}_p^V + \sum_{p \notin \mathcal{N}_k} \alpha'_{k,p} \boldsymbol{M}_p^V + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^{\perp}, \tag{183}$$

where

$$\alpha_{k,p} \geq \Omega(1), \qquad \alpha'_{k,p} \leq o\Big(\frac{1}{\sqrt{d}}\Big), \qquad |\beta_{k,p}| \leq o\Big(\frac{1}{\sqrt{d_1}}\Big). \tag{184}$$

Furthermore, the size of $\mathcal{N}_k$ is bounded as

$$|\mathcal{N}_k| = O(d^{\gamma}). \tag{185}$$

Finally, for each $\boldsymbol{M}_p^V$, there are at least $\Omega\big(m \cdot d^{-(1+\gamma)}\big)$ neurons $k \in [m]$ such that $\mathcal{N}_k = \{p\}$.

F.2.3 NOTATIONS

Definition F.1 (Notations). *For simpler presentation, we define the following notations: given $\boldsymbol{h}_X^{(r)} = \frac{1}{L}(\boldsymbol{M}^V \tilde{z}_n^{(r)} + \tilde{\boldsymbol{\xi}}_X^{(r)}) \sim \mathcal{D}_{\boldsymbol{h}_X}$, $\boldsymbol{h}_Y^{(s)} = \frac{1}{L}(\boldsymbol{M}^L \tilde{z}_n^{(s)} + \tilde{\boldsymbol{\xi}}_Y^{(s)}) \sim \mathcal{D}_{\boldsymbol{h}_Y}$, we let (for each $p \in [d]$):*

$$h_X^{(r)\setminus p} := \frac{1}{L}\left(\sum_{\substack{p' \neq p \\ p' \in [d]}} \boldsymbol{M}_{p'}^V \tilde{z}_{n,p'}^{(r)} + \tilde{\boldsymbol{\xi}}_X^{(r)}\right), \quad h_Y^{(s)\setminus p} := \frac{1}{L}\left(\sum_{\substack{p' \neq p \\ p' \in [d]}} \boldsymbol{M}_{p'}^L \tilde{z}_{n,p'}^{(s)} + \tilde{\boldsymbol{\xi}}_Y^{(s)}\right) \tag{186}$$

$$S_{k,t}^{(r)\setminus p} := \langle \boldsymbol{w}_k^{(t)}, h_X^{(r)\setminus p} \rangle, \quad S_{k,t}^{(s)\setminus p} := \langle \boldsymbol{v}_k^{(t)}, h_Y^{(s)\setminus p} \rangle \tag{187}$$

$$S_{k,t}^{(r,s)\setminus p} := \tfrac{1}{2}\left(S_{k,t}^{(r)\setminus p} + S_{k,t}^{(s)\setminus p}\right), \quad \bar{S}_{k,t}^{(r,s)\setminus p} := \tfrac{1}{2}\left(S_{k,t}^{(s)\setminus p} - S_{k,t}^{(r)\setminus p}\right) \tag{188}$$

$$\alpha_{k,p}^{(t)} := \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \tag{189}$$

Whenever the neuron index $k \in [m]$ is clear from the context, we drop the subscript k and the time index t for notational simplicity.

F.2.4 USEFUL LEMMAS

Lemma F.3 (Gradient for sparse features). *Suppose Lemma F.5 holds at iteration $t \geq 0$. For $p \in [d]$, define the events*

$$\begin{aligned}
A_1 &:= \left\{ \mathcal{S}_{k,t}^{\setminus p} \geq b_{f_k} - \alpha_p C_{\tilde{z}} \right\}, \\
A_2 &:= \left\{ \bar{\mathcal{S}}_{k,t}^{\setminus p} \geq b_{f_k} - (\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)})\alpha_p \right\}, \\
A_3 &:= \left\{ \left| \bar{\mathcal{S}}_{k,t}^{\setminus p} + (\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)})\alpha_p \right| \geq \frac{1}{2}\left(\alpha_p C_{\tilde{z}} - b_{f_k}\right) \right\}, \\
A_4 &:= \left\{ \mathcal{S}_{k,t}^{\setminus p} \geq \frac{1}{2}\left(\alpha_p C_{\tilde{z}} - b_{f_k}\right) \right\},
\end{aligned} \tag{190}$$

and the quantities L_1, L_2, L_3, L_4 as

$$L_1 := \sqrt{\frac{\mathbb{E}\Big[|\bar{\mathcal{S}}_{k,t}^{\setminus p}|^2(\mathbf{1}_{A_1}+\mathbf{1}_{A_2})\Big]}{\mathbb{E}[\langle \boldsymbol{w}_k^{(t)}, \tilde{\boldsymbol{\xi}}_{X_i}\rangle^2]}}, \qquad L_2 := \Pr(A_1),$$
$$L_3 := \sqrt{\frac{\mathbb{E}\Big[|\bar{\mathcal{S}}_{k,t}^{\setminus p}|^2(\mathbf{1}_{A_3}+\mathbf{1}_{A_4})\Big]}{\mathbb{E}[\langle \boldsymbol{w}_k^{(t)}, \tilde{\boldsymbol{\xi}}_{X_i}\rangle^2]}}, \qquad L_4 := \Pr(A_3). \tag{191}$$

Then we have the following results:

***(a) (all features)** For all $k \in [m]$, if $\alpha_p \geq 0$, we have (when $\alpha_p \leq 0$ the opposite inequality holds)*

$$\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{P}^+}\left[g_k(\boldsymbol{Y}_i)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right] \\
&= \frac{1}{L}\alpha_p \cdot \mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\tilde{z}_{i,p}^{(s)}\sum_{r=1}^{L}\tilde{z}_{i,p}^{(r)}\mathbf{1}_{\left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X+\boldsymbol{h}_Y}{2}\right\rangle\right|\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X-\boldsymbol{h}_Y}{2}\right\rangle\right|}\right] \\
&\pm O(\alpha_p)\cdot\mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\frac{|\tilde{z}_{i,p}^{(r)}|+|\tilde{z}_{i,p}^{(s)}|}{2}|\tilde{z}_{i,p}^{(r)}|\right]\cdot O(L_1+L_2).
\end{aligned} \tag{192}$$

***(b) (lucky features)** If $\alpha_p > b_{f_k}$, we have*

$$\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{P}^+}\left[g_k(\boldsymbol{Y}_i)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right] \\
&= \frac{1}{L}(\alpha_p - b_{f_k})\cdot \mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\tilde{z}_{i,p}^{(s)}\sum_{r=1}^{L}\tilde{z}_{i,p}^{(r)}\mathbf{1}_{\left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X+\boldsymbol{h}_Y}{2}\right\rangle\right|\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X-\boldsymbol{h}_Y}{2}\right\rangle\right|}\right] \\
&\pm O(\alpha_p)\cdot\mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\frac{|\tilde{z}_{i,p}^{(r)}|+|\tilde{z}_{i,p}^{(s)}|}{2}|\tilde{z}_{i,p}^{(r)}|\right]\cdot O(L_3+L_4).
\end{aligned} \tag{193}$$

If $\alpha_p < -b_{f_k}$, then the opposite inequality holds with $(\alpha_p - b_{f_k})$ replaced by $(\alpha_p + b_{f_k})$.

Lemma F.4 (Gradient from dense signals)**.** *Let $k \in [m]$ and $p \in [d]$. Suppose Lemma F.5 holds for the current iteration t. Then*

$$\left|\mathbb{E}\left[g_k(\boldsymbol{Y}_i)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{z}_X^{(r)}\rangle|\geq b_{f_k}}\left\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, \boldsymbol{M}_p^V\right\rangle\right]\right| \leq \widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{w}_k\|_2}{d^2}\right)\cdot\Pr\big(g_k(\boldsymbol{Y}_i)\neq 0\big). \tag{194}$$

For dense features $(\boldsymbol{M}_p^V)^\perp$, $p \in [d_1]\setminus[d]$, we have a similar result:

$$\left|\mathbb{E}\left[g_k(\boldsymbol{Y}_i)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{z}_X^{(r)}\rangle|\geq b_{f_k}}\left\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, (\boldsymbol{M}_p^V)^\perp\right\rangle\right]\right| \leq \widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{w}_k\|_2}{d\sqrt{d_1}}\right)\cdot\Pr\big(g_k(\boldsymbol{Y}_i)\neq 0\big). \tag{195}$$

Next, we discuss Lemma F.5. This lemma describes the alignment of neurons with the feature and non-feature directions during the second stage. (a) Lucky neurons have large projection on their feature direction, which means they can be activated and are clearly aligned with that feature. (b) If a neuron does not belong to $\mathcal{M}_p$, then its projection on feature p stays small, which means it cannot be activated and has only weak alignment. (c) The projection of a neuron weight on non-feature directions is very small, which means the neuron does not learn the non-feature directions.

Lemma F.5. *For all iterations $t \in (T_1, T_2]$, the neurons $k \in [m]$ satisfy the following properties:*

(a) For $p \in [d]$, if $k \in \mathcal{M}_p^\star$, then

$$\left|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle\right| \geq \sqrt{1+\gamma}\, b_{f_k}^{(t)}. \tag{196}$$

(b) For $p \in [d]$, if $k \notin \mathcal{M}_p$, then

$$\left|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle\right| \le \sqrt{1-\gamma}\, b_{f_k}^{(t)}, \tag{197}$$

and furthermore,

$$\left|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle\right| \le \widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{w}_k^{(t)}\|_2}{\sqrt{d}}\right). \tag{198}$$

(c) For each $k \in [m]$, and for all $p \in [d_1] \setminus [d]$,

$$\left|\langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)_p^{\perp} \rangle\right| \le \widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{w}_k^{(t)}\|_2}{\sqrt{d_1}}\right). \tag{199}$$

Lemma F.6. *For each $k \in [m]$, define*

$$\Lambda_k := \left\{p \in [d] : |\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^V \rangle| \le \tfrac{\sigma_0}{d}\right\} \subseteq [d]. \tag{200}$$

Then

$$|\Lambda_k| = O\Big(\tfrac{d}{\mathrm{polylog}(d)}\Big). \tag{201}$$

F.2.5 Proof of Theorem F.2

Proof of Theorem F.2 (a): We follow similar analysis as in the proof of Lemma F.5. In order to prove (a)-(c), we have to discuss the two substages of the learning process below.

When all $\|\boldsymbol{w}_k^{(t)}\|_2 \le O(1)\|\boldsymbol{w}_k^{(T_1)}\|_2$: From similar analysis in the proof of Lemma F.5, the iteration complexity for a neuron $k \in [m]$ to reach $\|\boldsymbol{w}_k^{(t)}\|_2 \ge \Omega(1)\|\boldsymbol{w}_k^{(T_1)}\|_2$ is no smaller than

$$T'_{k,1} := \max\left\{T_1 + \Omega\left(\frac{d\log d}{\eta \log\log d}\right), T_2\right\}. \tag{202}$$

When some $\|\boldsymbol{w}_k^{(t)}\|_2 \ge \Omega(1)\|\boldsymbol{w}_k^{(T_1)}\|_2$.

We first prove Theorem F.2 (a). In the first stage, for $p \notin \mathcal{N}_k$, we have

$$\begin{aligned}\sum_{p\in[d],\, p\notin\mathcal{N}_k} \langle \boldsymbol{w}_k^{(T'_{k,1})}, \boldsymbol{M}_p^V\rangle^2 &\le \sum_{p\in[d],\, p\notin\mathcal{N}_k} \langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V\rangle^2 \left(1 + \epsilon_p \frac{O(\eta)}{d\ \mathrm{polylog}(d)}\right)^{T_2} + \widetilde{O}\left(\frac{\|\boldsymbol{w}_k^{(T_1)}\|_2^2}{d^{3/2}}\right) \\ &\le O(1)\left\|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T_1)}\right\|_2^2,\end{aligned} \tag{203}$$

where we used the fact that $\|\boldsymbol{w}_k^{(T_1)}\|_2 \lesssim \|\boldsymbol{M}^V(\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T_1)}\|_2$.

For $p \in [d_1] \setminus [d]$ we have

$$\begin{aligned}&\sum_{p\in[d_1]\setminus[d]} \langle \boldsymbol{w}_k^{(T'_{k,1})}, (\boldsymbol{M}^V)_p^{\perp}\rangle^2 \\ \le &\sum_{p\in[d_1]\setminus[d]} \langle \boldsymbol{w}_k^{(T_1)}, (\boldsymbol{M}^V)_p^{\perp}\rangle^2 + O\left(\frac{\eta(T'_{k,1}-T_1)}{d}\right) e^{-\Omega(\log^{1/4} d)} \max_{t'\in[T_1, T'_{k,1}]} \|\boldsymbol{w}_k^{(t')}\|_2^2 \\ \le &O(1)\left\|(\boldsymbol{M}^V)^{\perp}((\boldsymbol{M}^V)^{\perp})^\top \boldsymbol{w}_k^{(T_1)}\right\|_2^2.\end{aligned} \tag{204}$$

Typically, if $k \in \mathcal{M}_p^\star$, there exists $t \le T_2$ such that $\|\boldsymbol{w}_k^{(t)}\|_2 \ge (1 + \frac{\varepsilon_p}{\varepsilon_{\max}})\|\boldsymbol{w}_k^{(T_1)}\|_2$, as we have argued in the proof of Lemma F.5. Thus, we have

$$\begin{aligned}|\langle \boldsymbol{w}_k^{(T'_{k,1})}, \boldsymbol{M}_p^V\rangle|^2 &\ge \|\boldsymbol{w}_k^{(T'_{k,1})}\|_2^2 - \sum_{p\in[d],\, p\notin\mathcal{N}_k} \langle \boldsymbol{w}_k^{(T'_{k,1})}, \boldsymbol{M}_p^V\rangle^2 - \sum_{p\in[d_1]\setminus[d]} \langle \boldsymbol{w}_k^{(T'_{k,1})}, (\boldsymbol{M}^V)_p^\perp\rangle^2 \\ &\ge \Omega(1)\|\boldsymbol{w}_k^{(T_1)}\|_2^2 - (1+o(1))\|\boldsymbol{w}_k^{(T_1)}\|_2^2 \\ &\ge \Omega(1)\|\boldsymbol{w}_k^{(T_1)}\|_2^2,\end{aligned} \tag{205}$$

which proves the claim.

In the second stage, if $k \in \mathcal{M}_p^\star$, then from similar calculations as above, we can prove by induction that starting from $t = T'_{k,1}$, it holds:

$$\begin{aligned}|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V\rangle| &\ge |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle| \left(1 + \Omega\left(\epsilon_p \frac{\eta \log\log d}{d}\right)\right) \\ &\ge \|\boldsymbol{w}_k^{(t)}\|_2 \left(1 + \Omega\left(\epsilon_p \frac{\eta \log\log d}{d}\right)\right) \\ \sum_{p'\in[d],\, p'\neq p} \langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_{p'}^V\rangle^2 &\le \sum_{p'\in[d],\, p'\neq p} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_{p'}^V\rangle^2 \left(1 + \epsilon_p \frac{O(\eta)}{d\ \mathrm{polylog}(d)}\right)^2 \\ \sum_{p\in[d_1]\setminus[d]} \langle \boldsymbol{w}_k^{(t+1)}, (\boldsymbol{M}^V)_p^\perp\rangle^2 &\le \sum_{p\in[d_1]\setminus[d]} \langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)_p^\perp\rangle^2 \left(1 + \frac{O(\eta)}{d\ \mathrm{polylog}(d)}\right)^2,\end{aligned} \tag{206}$$

which implies

$$|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V\rangle| \ge |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle| \cdot \frac{\|\boldsymbol{w}_k^{(t+1)}\|_2}{\|\boldsymbol{w}_k^{(t)}\|_2} \ge (1-o(1))\|\boldsymbol{w}_k^{(t+1)}\|_2. \tag{207}$$

□

Proof of Theorem F.2 (b): Next, we prove Theorem F.2 (b). In the first stage, the bias growth is large, i.e.,

$$\begin{aligned}b_{f_k}^{(T'_{k,1})} &\ge b_{f_k}^{(T_1)}(1+\frac{\eta}{d})^{T'_{k,1}-T_1} \ge b_{f_k}^{(T_1)} \cdot \mathrm{polylog}(d) \\ &\ge \frac{\mathrm{polylog}(d)}{\sqrt{d}}\|\boldsymbol{w}_k^{(T_1)}\|_2 \ge \frac{\mathrm{polylog}(d)}{\sqrt{d}}\|\boldsymbol{w}_k^{(T'_{k,1})}\|_2.\end{aligned} \tag{208}$$

In the second stage, the bias is large consistently, i.e.,

$$b_{f_k}^{(t+1)} \ge b_{f_k}^{(t)} \cdot \frac{\|\boldsymbol{w}_k^{(t+1)}\|_2}{\|\boldsymbol{w}_k^{(t)}\|_2} \ge \frac{\mathrm{polylog}(d)}{\sqrt{d}}\|\boldsymbol{w}_k^{(t+1)}\|_2 \ge \frac{1}{4}\|\boldsymbol{w}_k^{(T'_{k,1})}\|_2. \tag{209}$$

□

Proof of Theorem F.2 (c): Finally, we prove Theorem F.2 (c): Assuming $\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle > 0$ (the opposite case is similar), from $t = T_1$, for $k \in \mathcal{M}_p^\star$, we have

$$\begin{aligned}\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V\rangle &= \left(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle - b_{f_k}^{(t)}\right)\left(1 + \epsilon_p \frac{\eta \log\log d}{d}\right) \pm O\left(\frac{\eta|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle|}{d\ \mathrm{polylog}(d)}\right) \\ &\ge \Omega(1)\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle \left(1 + \epsilon_p \frac{\eta \log\log d}{d}\left(1 - \frac{1}{\mathrm{polylog}(d)}\right)\right) \\ &\ge \Omega(1)\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V\rangle \left(1 + \epsilon_p \frac{\eta \log\log d}{d}\left(1 - \frac{1}{\mathrm{polylog}(d)}\right)\right)^{t+1-T_1}.\end{aligned} \tag{210}$$

which implies that after certain iteration $t = T_1 + T'$, where $T' = \Theta\left(\frac{d}{\eta}\right)$, we shall have

$$|\langle \boldsymbol{w}_k^{(T_1+T')}, \boldsymbol{M}_p^V \rangle| \geq \log\log d \cdot |\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V \rangle| \geq b_{f_k}^{(T_1)} \cdot \log\log d. \tag{211}$$

However, at iteration $t = T_1 + \Theta\left(\frac{d}{\eta}\right)$, we can see from previous analysis that $\|\boldsymbol{w}_k^{(t)}\|_2 \leq (1 + o(1))\|\boldsymbol{w}_k^{(T_1)}\|_2$, so the bias growth can be bounded as

$$\begin{aligned} b_{f_k}^{(t)} &\leq b_{f_k}^{(T_1)} \left(1 + \frac{\eta}{d}\right)^{\Theta(\frac{d}{\eta})} \cdot \max\left\{\frac{\|\boldsymbol{w}_k^{(t)}\|_2}{\|\boldsymbol{w}_k^{(T_1)}\|_2}, 1\right\} \\ &\leq b_{f_k}^{(T_1)} \left(1 + \frac{\eta}{d} \cdot \Theta\left(\frac{d}{\eta}\right)\right) \cdot \max\{(1 + o(1)), 1\} \\ &\leq O\Big(b_{f_k}^{(T_1)}\Big). \end{aligned} \tag{212}$$

Now from our initialization properties in Lemma F.6, we have that $\langle \boldsymbol{w}_{k'}^{(0)}, \boldsymbol{M}_p^V \rangle^2 \leq O(\sigma_0^2 \log d)$ for all $k' \in [m]$. Thus via similar arguments, we also have

$$|\langle \boldsymbol{w}_{k'}^{(t)}, \boldsymbol{M}_p^V \rangle| \leq |\langle \boldsymbol{w}_{k'}^{(0)}, \boldsymbol{M}_p^V \rangle| \left(1 + \epsilon_p \frac{\eta \log\log d}{d} \left(1 \pm \frac{1}{\mathrm{polylog}(d)}\right)\right)^t. \tag{213}$$

holds for all $k' \in [m]$. Now it is easy to see that for $t \leq T_2 = T_1 + \Theta\left(\frac{d \log d}{\eta \log\log d}\right)$, we have

$$\begin{aligned} \frac{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|}{|\langle \boldsymbol{w}_{k'}^{(t)}, \boldsymbol{M}_p^V \rangle|} &\geq \Omega(1) \cdot \frac{|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V \rangle| \left(1 + \epsilon_p \frac{\eta \log\log d}{d} \left(1 - \frac{\eta}{\mathrm{polylog}(d)}\right)\right)^{t-T_1}}{|\langle \boldsymbol{w}_{k'}^{(T_1)}, \boldsymbol{M}_p^V \rangle| \left(1 + \epsilon_p \frac{\eta \log\log d}{d} + \frac{\eta}{d\,\mathrm{polylog}(d)}\right)^{t-T_1}} \\ &\geq \left(1 - O\left(\epsilon_p \frac{\eta \log\log d}{d\,\mathrm{polylog}(d)}\right)\right)^{t-T_1} \geq \Omega(1). \end{aligned} \tag{214}$$

Thus, the last claim is proved. □

F.2.6 Proof of Theorem F.3

Proof of Theorem F.3. We have

$$\begin{aligned} \boldsymbol{w}_k^{(t)} &= \sum_{p \in \mathcal{N}_k,\, p \in [d]} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \boldsymbol{M}_p^V + \sum_{p \notin \mathcal{N}_k,\, p \in [d]} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \boldsymbol{M}_p^V + \sum_{p \in [d_1] \setminus [d]} \langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)_p^\perp \rangle (\boldsymbol{M}^V)_p^\perp \\ &\leq \sum_{p \in \mathcal{N}_k,\, p \in [d]} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \boldsymbol{M}_p^V + \sum_{p \notin \mathcal{N}_k,\, p \in [d]} \widetilde{O}\left(\frac{\|\boldsymbol{w}_k^{(t)}\|_2}{\sqrt{d}}\right) \boldsymbol{M}_p^V + \sum_{p \in [d_1] \setminus [d]} \widetilde{O}\left(\frac{\|\boldsymbol{w}_k^{(t)}\|_2}{\sqrt{d_1}}\right) (\boldsymbol{M}^V)_p^\perp \\ &= \sum_{p \in \mathcal{N}_k,\, p \in [d]} \alpha_{k,p} \boldsymbol{M}_p^V + \sum_{p \notin \mathcal{N}_k,\, p \in [d]} \alpha'_{k,p} \boldsymbol{M}_p^V + \sum_{p \in [d_1] \setminus [d]} \beta_{k,p} (\boldsymbol{M}^V)_p^\perp. \end{aligned} \tag{215}$$

According to Theorem F.2 (a) (c), we can obtain the range of $\alpha_{k,p}$ is $\alpha_{k,p} \geq \Omega(1)$. Meanwhile, according to Lemma F.5 (b) (c), we can obtain $\alpha'_{k,p} \leq o(\frac{1}{\sqrt{d}})$ and $\beta_{k,p} \leq o(\frac{1}{\sqrt{d_1}})$. □

F.2.7 Proof of Lemma F.3

Proof of Lemma F.3 (a). In the proof we will make the following simplification of notations: we drop the time superscript $^{(t)}$, and also the subscript for neuron index k. We recall that $\alpha_{k,p}^{(t)} :=$

$\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle$ in (189). We first consider the case $0 < \alpha_p < b_k^{(t)}$ and rewrite the expectation as follows:

$$
\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{P}+}\left[g_k(\boldsymbol{Y}_i)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
=&\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\Big(\mathrm{ReLU}\Big(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle - b_{g_k}\Big) - \mathrm{ReLU}\Big(-\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle - b_{g_k}\Big)\Big)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
=&\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\Big(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle - b_{g_k}\Big)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right.\\
&\left.-\sum_{s=1}^{L}\Big(-\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle - b_{g_k}\Big)\mathbf{1}_{-\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
=&\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right.\\
&-b_{g_k}\sum_{s=1}^{L}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\\
&+\sum_{s=1}^{L}\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\mathbf{1}_{-\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\\
&\left.+b_{g_k}\sum_{s=1}^{L}\mathbf{1}_{-\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
=&\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\mathbf{1}_{|\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle|\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
&-\mathbb{E}_{i\sim\mathcal{P}+}\left[b_{g_k}\sum_{s=1}^{L}\Big(\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}-\mathbf{1}_{-\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\Big)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right].
\end{aligned}
\tag{216}
$$

Thus the expectation can be expanded as:

$$
\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{P}+}\left[g_k(\boldsymbol{Y}_i)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
=&\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\mathbf{1}_{|\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle|\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
&-\mathbb{E}_{i\sim\mathcal{P}+}\left[b_{g_k}\sum_{s=1}^{L}\Big(\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}-\mathbf{1}_{-\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\Big)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
=&\frac{1}{L}\mathbb{E}_{i\sim\mathcal{P}+}\left[\alpha_p\sum_{s=1}^{L}\tilde{z}_{i,p}^{(s)}\sum_{r=1}^{L}\tilde{z}_{i,p}^{(r)}\mathbf{1}_{\left|\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}\right]\\
&+\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\Big(S^{(s)\setminus p}-b_{g_k}\Big)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
&+\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\Big(S^{(s)\setminus p}+b_{g_k}\Big)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
&+\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\Big(\frac{1}{L}\alpha_p\tilde{z}_{i,p}^{(s)}+S^{(s)\setminus p}-b_{g_k}\Big)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
&+\mathbb{E}_{i\sim\mathcal{P}+}\left[\sum_{s=1}^{L}\Big(\frac{1}{L}\alpha_p\tilde{z}_{i,p}^{(s)}+S^{(s)\setminus p}+b_{g_k}\Big)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
=&J_1+J_2+J_3.
\end{aligned}
\tag{217}
$$

Now we need to obtain absolute bounds for both J_2 and J_3. We start with J_2, where

$$\begin{aligned}
J_2 &= \mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(S^{(s)\setminus p}-b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
&+\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(S^{(s)\setminus p}+b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\tilde{z}_{i,p}^{(r)}\right]\\
&=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(S^{(s)\setminus p}-b_{g_k}\right)\sum_{r=1}^{L}\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}\tilde{z}_{i,p}^{(r)}\right]\\
&+\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(S^{(s)\setminus p}+b_{g_k}\right)\sum_{r=1}^{L}\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\leq -b_{f_k}-\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}\tilde{z}_{i,p}^{(r)}\right].
\end{aligned}\tag{218}$$

We proceed with the first term

$$J_{2.1}=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}(S^{(s)\setminus p}-b_{g_k})\sum_{r=1}^{L}\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}\tilde{z}_{i,p}^{(r)}\right].\tag{219}$$

First, from a trivial calculation conditioned on the randomness of $\tilde{z}_{i,p}^{(r)}$, we have

$$\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}(S^{(s)\setminus p}-b_{g_k})\sum_{r=1}^{L}\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}\tilde{z}_{i,p}^{(r)}\right]\\
&=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}(S^{(s)\setminus p}-b_{g_k})\sum_{r=1}^{L}\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\geq b_{f_k}+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|}\tilde{z}_{i,p}^{(r)}\right]\\
&=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}(S^{(s)\setminus p}-b_{g_k})\sum_{r=1}^{L}\mathbf{1}_{\mathcal{S}^{(r,s)\setminus p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|}\tilde{z}_{i,p}^{(r)}\right]\\
&=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}(S^{(s)\setminus p}-b_{g_k})\sum_{r=1}^{L}|\tilde{z}_{i,p}^{(r)}|\Big(\mathbf{1}_{\mathcal{S}^{(r,s)\setminus p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|}\right.\\
&\qquad\left.-\mathbf{1}_{\mathcal{S}^{(r,s)\setminus p}\geq b_{f_k}+\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}-\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|}\Big)\right].
\end{aligned}\tag{220}$$

Now define:

$$\begin{aligned}
Z&=\frac{1}{2}\left(\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}-\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|\right),\\
Z'&=\frac{1}{2}\left(\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|-\left|\bar{\mathcal{S}}^{(r,s)\setminus p}-\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|\right).
\end{aligned}\tag{221}$$

In this case, we always have $|Z'| \leq |\alpha_p^{(r,s)}| \left| \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right|$, and

$$\begin{aligned} &\Big| \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} \geq b_{f_k} - \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left| \bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right|} \\ &- \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} \geq b_{f_k} + \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left| \bar{\mathcal{S}}^{(r,s)\setminus p} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right|} \Big| \\ &= \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} - b_{f_k} - Z \in [-|\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} - Z'|,\ |\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} - Z'|]}. \end{aligned} \tag{222}$$

Equality Proof: The original two threshold values are

$$\begin{aligned} T_1 &= b_{f_k} - \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left| \bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right|, \\ T_2 &= b_{f_k} + \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left| \bar{\mathcal{S}}^{(r,s)\setminus p} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right|. \end{aligned} \tag{223}$$

We can rewrite these two terms as

$$\begin{aligned} T_1 &= b_{f_k} + \left(-\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left| \bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right| \right), \\ T_2 &= b_{f_k} + \left(\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left| \bar{\mathcal{S}}^{(r,s)\setminus p} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right| \right). \end{aligned} \tag{224}$$

Then the midpoint between these two terms is

$$\text{Midpoint} = b_{f_k} + Z. \tag{225}$$

The distance between them is

$$\begin{aligned} &T_2 - T_1 \\ &= 2\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left(\left| \bar{\mathcal{S}}^{(r,s)\setminus p} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right| - \left| \bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right| \right) \\ &= 2 \left(\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} - Z' \right). \end{aligned} \tag{226}$$

Therefore, shifting from the midpoint $b_{f_k} + Z$ by $\left| \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} - Z' \right|$ in both directions will exactly span the distance between T_1 and T_2. The indicator function differs by 1 if and only if

$$T_1 \leq \mathcal{S}^{(r,s)\setminus p} < T_2. \tag{227}$$

Therefore we have:

$$\begin{aligned} &\Big| \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} \geq b_{f_k} - \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left| \bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right|} \\ &\qquad - \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} \geq b_{f_k} + \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \left| \bar{\mathcal{S}}^{(r,s)\setminus p} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \right|} \Big| \\ &= \left| \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} \geq T_1} - \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} \geq T_2} \right| \\ &= \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} - \frac{T_1+T_2}{2} \in \left[-\frac{T_2-T_1}{2}, \frac{T_2-T_1}{2} \right]} \\ &= \mathbf{1}_{\mathcal{S}^{(r,s)\setminus p} - b_{f_k} - Z \in \left[-\left| \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} - Z' \right|, \left| \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} - Z' \right| \right]}. \end{aligned} \tag{228}$$

Next, we have

$$\begin{aligned}|Z'| =&\frac{1}{2}\big||\bar{\mathcal{S}}^{(r,s)\backslash p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}|-|\bar{\mathcal{S}}^{(r,s)\backslash p}-\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}|\big|\\ \leq&\frac{1}{2}\cdot 2|\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}|=|\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}|=\left|\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|\end{aligned} \tag{229}$$

$$\begin{aligned}|Z| =&\frac{1}{2}\big||\bar{\mathcal{S}}^{(r,s)\backslash p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}|+|\bar{\mathcal{S}}^{(r,s)\backslash p}-\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}|\big|\\ \leq&\frac{1}{2}\cdot 2(|\bar{\mathcal{S}}^{(r,s)\backslash p}|+|\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}|)=|\bar{\mathcal{S}}^{(r,s)\backslash p}|+\left|\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|.\end{aligned} \tag{230}$$

Which allows us to proceed as follows:

$$\begin{aligned}&\left|\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}(S^{(s)\backslash p}-b_{g_k})\sum_{r=1}^{L}\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}\tilde{z}_{i,p}^{(r)}\right]\right|\\
=&\left|\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}(S^{(s)\backslash p}-b_{g_k}-Z+Z)\sum_{r=1}^{L}|\tilde{z}_{i,p}^{(r)}|\Big(\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}+\left|\bar{\mathcal{S}}^{(r,s)\backslash p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|}\right.\right.\right.\\
&\left.\left.\left.-\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}\geq b_{f_k}+\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}+\left|\bar{\mathcal{S}}^{(r,s)\backslash p}-\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|}\Big)\mathbf{1}_{\mathcal{S}^{\backslash p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}}\right]\right|\\
\leq&\frac{1}{L}\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}(\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2}\right|+2|\alpha_p^{(r,s)}|\left|\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2}\right||\tilde{z}_{i,p}^{(r)}|\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}}\right]\\
&+\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}|\bar{\mathcal{S}}^{(r,s)\backslash p}||\tilde{z}_{i,p}^{(r)}|\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}-b_{f_k}-Z\in[-|\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}-Z'|,\,|\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}-Z'|]}\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}}\right]\\
=&\frac{1}{L}\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}(\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2}\right|+2|\alpha_p^{(r,s)}|\left|\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2}\right||\tilde{z}_{i,p}^{(r)}|\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}}\right]\\
&+\sum_{s=1}^{L}\sum_{r=1}^{L}\sqrt{\mathbb{E}_{i\sim\mathcal{P}^+}\left[|\bar{\mathcal{S}}^{(r,s)\backslash p}|^2(\tilde{z}_{i,p}^{(r)})^2\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}}\right]}\\
&\cdot\sqrt{\mathbb{E}_{i\sim\mathcal{P}^+}\left[\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}-b_{f_k}-Z\in[-|\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}-Z'|,\,|\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}-Z'|]}\right]}\\
\overset{①}{\leq}&\frac{1}{L}\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}(\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2}\right|+2|\alpha_p^{(r,s)}|\left|\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2}\right||\tilde{z}_{i,p}^{(r)}|\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}}\right]\\
&+\sum_{s=1}^{L}\sum_{r=1}^{L}\sqrt{\mathbb{E}_{i\sim\mathcal{P}^+}\left[|\bar{\mathcal{S}}^{(r,s)\backslash p}|^2(\tilde{z}_{i,p}^{(r)})^2\mathbf{1}_{\mathcal{S}^{(r,s)\backslash p}\geq b_{f_k}-\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}}\right]}\cdot\sqrt{\frac{\mathbb{E}_{i\sim\mathcal{P}^+}\left[(\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}+|\alpha_p^{(r,s)}|\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L})^2\right]}{\mathbb{E}\left[\left\langle \boldsymbol{w}_k,\frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}+\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L}\right\rangle^2\right]}}\\
\leq&O(\alpha_p)\mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\frac{|\tilde{z}_{i,p}^{(r)}|+|\tilde{z}_{i,p}^{(s)}|}{2}|\tilde{z}_{i,p}^{(r)}|\right]\\
&\cdot\left(\Pr(\mathcal{S}^{\backslash p}\geq b_{f_k}-\alpha_pC_{\tilde{z}})+\sqrt{\frac{\mathbb{E}\left[|\bar{\mathcal{S}}^{\backslash p}|^2\mathbf{1}_{\mathcal{S}^{\backslash p}\geq b_{f_k}-\alpha_pC_{\tilde{z}}}\right]}{\mathbb{E}\left[\left\langle \boldsymbol{w}_k,\frac{\tilde{\boldsymbol{\xi}}_{X_i}+\tilde{\boldsymbol{\xi}}_{Y_i}}{2L}\right\rangle^2\right]}}\right).\end{aligned} \tag{231}$$

where in ① we have used the randomness of $\frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}+\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L}$ in the following manner. Fixing the randomness of $\tilde{z}_{i,p}$, we have $\mathcal{S}^{(r,s)\setminus p} - Z$ is a random variable depending solely on the randomness of $\frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}+\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L}$, and thus we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\mathcal{S}^{(r,s)\setminus p}-b_{f_k}-Z\in[-|\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}-Z'|,\ |\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}-Z'|]}\right] \\
\leq &\mathbb{E}\left[\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}+\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L}\right\rangle-\left|\langle \boldsymbol{w}_k,\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}\rangle\right|\in[|(\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)})\alpha_p^{(r,s)}|-(\alpha_p+|(\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)})\alpha_p^{(r,s)}|),\ \alpha_p+2|(\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)})\alpha_p^{(r,s)}|]}\right] \\
= &\mathbb{E}\left[\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}}{2L}\right\rangle+\left\langle \boldsymbol{w}_k,\frac{\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L}\right\rangle-\left|\left\langle \boldsymbol{w}_k,\frac{\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{L}\right\rangle\right|\in[-O(\alpha_p+|(\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)})\alpha_p^{(r,s)}|),\ O(\alpha_p+|(\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)})\alpha_p^{(r,s)}|)]}\right] \\
\leq &\mathbb{E}\left[\frac{O\left((\alpha_p\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}+|\alpha_p^{(r,s)}|\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L})^2\right)}{\left\langle \boldsymbol{w}_k,\frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}+\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L}\right\rangle^2}\right].
\end{aligned}
\tag{232}
$$

Simultaneously, from similar analysis as above, we have for the second term in J_2:

$$
\begin{aligned}
J_{2.2} &= \left|\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(S^{(s)\setminus p}+b_{g_k}\right)\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\leq -b_{f_k}-\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}\tilde{z}_{i,p}^{(r)}\right]\right| \\
&= \left|\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(S^{(s)\setminus p}+b_{g_k}\right)\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\leq -b_{f_k}-\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|}\tilde{z}_{i,p}^{(r)}\right]\right| \\
&\leq O(\alpha_p)\,\mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\frac{|\tilde{z}_{i,p}^{(r)}|+|\tilde{z}_{i,p}^{(s)}|}{2}|\tilde{z}_{i,p}^{(r)}|\right]\sqrt{\frac{\mathbb{E}\left[|\bar{\mathcal{S}}\setminus p|^2\mathbf{1}_{\mathcal{S}^{\setminus p}\geq b_{f_k}-\alpha_p}\right]}{\mathbb{E}\left[\left\langle \boldsymbol{w}_k,\frac{\tilde{\boldsymbol{\xi}}_{X_i}+\tilde{\boldsymbol{\xi}}_{Y_i}}{2L}\right\rangle^2\right]}}.
\end{aligned}
\tag{233}
$$

Now we turn to J_3:

$$
\begin{aligned}
J_3 = &\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(\frac{1}{L}\alpha_p\tilde{z}_{i,p}^{(s)}+S^{(s)\setminus p}-b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\tilde{z}_{i,p}^{(r)}\right] \\
&+\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(\frac{1}{L}\alpha_p\tilde{z}_{i,p}^{(s)}+S^{(s)\setminus p}+b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right] \\
&\overset{①}{=}\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(\frac{1}{L}\alpha_p\tilde{z}_{i,p}^{(s)}+S^{(s)\setminus p}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\tilde{z}_{i,p}^{(r)}\right] \\
&+\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(\frac{1}{L}\alpha_p\tilde{z}_{i,p}^{(s)}+S^{(s)\setminus p}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\tilde{z}_{i,p}^{(r)}\right].
\end{aligned}
\tag{234}
$$

① is from the symmetry of $\sum_{r=1}^{L} \boldsymbol{h}_X^{(r)}$ and $\sum_{s=1}^{L} \boldsymbol{h}_Y^{(s)}$ over the randomness of $\boldsymbol{h}_X, \boldsymbol{h}_Y$. Since $\sum_{r=1}^{L} \boldsymbol{h}_X^{(r)} \overset{d}{=} \sum_{s=1}^{L} \boldsymbol{h}_Y^{(s)}$, we observe

$$\sum_{s=1}^{L}\sum_{r=1}^{L} f(\boldsymbol{h}_X^{(r)}, \boldsymbol{h}_Y^{(s)}) \overset{d}{=} \sum_{s=1}^{L}\sum_{r=1}^{L} f(\boldsymbol{h}_Y^{(s)}, \boldsymbol{h}_X^{(r)}),$$

$$\mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L} f(\boldsymbol{h}_X^{(r)}, \boldsymbol{h}_Y^{(s)})\right] = \mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L} f(\boldsymbol{h}_Y^{(s)}, \boldsymbol{h}_X^{(r)})\right],$$

$$\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L} b_{g_k} \mathbf{1}_{\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)}\rangle \geq b_{f_k}} \mathbf{1}_{\langle \boldsymbol{v}_k, \boldsymbol{h}_Y^{(s)}\rangle \leq -b_{g_k}} \tilde{z}_{i,p}^{(r)}\right] = \mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L} b_{g_k} \mathbf{1}_{\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)}\rangle \leq -b_{f_k}} \mathbf{1}_{\langle \boldsymbol{v}_k, \boldsymbol{h}_Y^{(s)}\rangle \geq b_{g_k}} \tilde{z}_{i,p}^{(r)}\right], \tag{235}$$

which allows us to drop the b_{g_k} terms in J_3. The analysis of the rest of J_3 is somewhat similar.

First we observe that whenever $\mathbf{1}_{\langle \boldsymbol{v}_k, \boldsymbol{h}_Y^{(s)}\rangle \geq b_{g_k}} \mathbf{1}_{\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)}\rangle \leq -b_{f_k}} \neq 0$, we have $\langle \boldsymbol{v}_k, \boldsymbol{h}_Y^{(s)}\rangle \geq b_{g_k}$ and $\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)}\rangle \leq -b_{f_k}$. Thus, we can get $\bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} \geq b_{g_k} + \left|\mathcal{S}^{(r,s)\setminus p} + \alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L}\right|$. When this inequality holds, we always have $\bar{\mathcal{S}}^{(r,s)\setminus p} \geq b_{g_k} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}$.

Together with all the above observations, we proceed to compute as:

$$\begin{aligned}
&\left|\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(\frac{1}{L}\alpha_p \tilde{z}_{i,p}^{(s)} + S^{(s)\setminus p}\right)\mathbf{1}_{\langle \boldsymbol{v}_k, \boldsymbol{h}_Y^{(s)}\rangle \geq b_{g_k}} \mathbf{1}_{\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)}\rangle \leq -b_{f_k}} \tilde{z}_{i,p}^{(r)}\right]\right| \\
&= \left|\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(\frac{1}{L}\alpha_p \tilde{z}_{i,p}^{(s)} + S^{(s)\setminus p}\right)\mathbf{1}_{\alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L} + \bar{\mathcal{S}}^{(r,s)\setminus p} \geq b_{g_k} + \left|\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + \mathcal{S}^{(r,s)\setminus p}\right|} \tilde{z}_{i,p}^{(r)}\right]\right| \\
&\leq \mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left|\bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}\right| |\tilde{z}_{i,p}^{(r)}| \mathbf{1}_{\bar{\mathcal{S}}^{(r,s)\setminus p} \geq b_{g_k} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}}\right. \\
&\qquad \left. \cdot \mathbf{1}_{\bar{\mathcal{S}}^{(r,s)\setminus p} \in [\, b_{g_k} - \alpha_p \left|\frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L}\right| + \left|\bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}\right|,\ b_{g_k} + \alpha_p \left|\frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L}\right| + \left|\bar{\mathcal{S}}^{(r,s)\setminus p} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}\right| \,]}\right] \\
&\leq \sum_{s=1}^{L}\sum_{r=1}^{L} \sqrt{\mathbb{E}\left[|\bar{\mathcal{S}}^{(r,s)\setminus p}|^2 |\tilde{z}_{i,p}^{(r)}|^2\right] \mathbf{1}_{\bar{\mathcal{S}}^{(r,s)\setminus p} \geq b_{g_k} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}}} \\
&\qquad \cdot \sqrt{\mathbb{E}\left[\mathbf{1}_{\bar{\mathcal{S}}^{(r,s)\setminus p} \in [\, b_{g_k} - \alpha_p \left|\frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L}\right| + \left|\bar{\mathcal{S}}^{(r,s)\setminus p} + \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}\right|,\ b_{g_k} + \alpha_p \left|\frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L}\right| + \left|\bar{\mathcal{S}}^{(r,s)\setminus p} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}\right| \,]}\right]} \\
&+ \frac{1}{L}\mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L} \alpha_p^{(r,s)} \left|\frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2}\right| |\tilde{z}_{i,p}^{(r)}| \mathbf{1}_{\bar{\mathcal{S}}^{(r,s)\setminus p} \geq b_{g_k} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}}\right] \\
&\leq \sum_{s=1}^{L}\sum_{r=1}^{L} \sqrt{\mathbb{E}\left[|\bar{\mathcal{S}}^{(r,s)\setminus p}|^2 |\tilde{z}_{i,p}^{(r)}|^2\right] \mathbf{1}_{\bar{\mathcal{S}}^{(r,s)\setminus p} \geq b_{g_k} - \alpha_p^{(r,s)} \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}}} \\
&\qquad \cdot \sqrt{\mathbb{E}\left[\mathbf{1}_{\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)} - \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L}\rangle \in [\, -\left|\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + |\alpha_p^{(r,s)}| \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}\right|,\ \left|\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + |\alpha_p^{(r,s)}| \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}\right| \,]}\right]} \\
&\leq \frac{1}{L} O(\alpha_p)\, \mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L} \frac{|\tilde{z}_{i,p}^{(r)}| + |\tilde{z}_{i,p}^{(s)}|}{2} |\tilde{z}_{i,p}^{(r)}|\right] \sqrt{\frac{\mathbb{E}\left[|\bar{\mathcal{S}}^{\setminus p}|^2 \mathbf{1}_{\bar{\mathcal{S}}^{\setminus p} \geq b_{g_k} - (\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)})\alpha_p^{(r,s)}}\right]}{\mathbb{E}\left[\left\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)} + \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L}\right\rangle^2\right]}} \\
&+ \frac{1}{L}\left(O(\alpha_p)\, \mathbb{E}\left[\sum_{s=1}^{L}\sum_{r=1}^{L} \frac{|\tilde{z}_{i,p}^{(r)}| + |\tilde{z}_{i,p}^{(s)}|}{2} |\tilde{z}_{i,p}^{(r)}|\right]\right) \Pr(\bar{\mathcal{S}}^{\setminus p} \geq b_{g_k} - (\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)})\alpha_p^{(r,s)}).
\end{aligned} \tag{236}$$

The last inequality holds because $\left\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)} - \tilde{\boldsymbol{\xi}}_{X_i}^{(r)}}{2L} \right\rangle$ has the same distribution as $\left\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)} + \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L} \right\rangle$, i.e., $\left\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)} - \tilde{\boldsymbol{\xi}}_{X_i}^{(r)}}{2L} \right\rangle \overset{d}{=} \left\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)} + \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L} \right\rangle$. We use the randomness to obtain that

$$\begin{aligned}
&\mathbb{E}_{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}} \left[\mathbf{1}_{\left\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)} - \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L} \right\rangle \in [-|\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + |\alpha_p^{(r,s)}| \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}|,\ |\alpha_p \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L} + |\alpha_p^{(r,s)}| \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}|] } \right] \\
&\leq \frac{(\alpha_p | \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L}| + |\alpha_p^{(r,s)}| | \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}|)^2}{\mathbb{E}_{\xi} \left[\left\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{Y_i}^{(s)} - \tilde{\boldsymbol{\xi}}_{X_i}^{(r)}}{2L} \right\rangle^2 \right]} = \frac{(\alpha_p | \frac{\tilde{z}_{i,p}^{(r)} + \tilde{z}_{i,p}^{(s)}}{2L}| + |\alpha_p^{(r,s)}| | \frac{\tilde{z}_{i,p}^{(r)} - \tilde{z}_{i,p}^{(s)}}{2L}|)^2}{\mathbb{E}_{\xi} \left[\left\langle \boldsymbol{w}_k, \frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)} + \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2L} \right\rangle^2 \right]}.
\end{aligned} \tag{237}$$

The second term of J_3 can be similarly bounded by the same quantity. Now by combining all the results of J_1, J_2, J_3 above, we have the desired result for (a).

$$\begin{aligned}
&\mathbb{E}_{i \sim \mathcal{P}^+} \left[g_k(\boldsymbol{Y}_i) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)} \rangle| \geq b_{f_k}} \tilde{z}_{i,p}^{(r)} \right] \\
&\leq \frac{1}{L} \alpha_p \cdot \mathbb{E}_{i \sim \mathcal{P}^+} \left[\sum_{s=1}^{L} \tilde{z}_{i,p}^{(s)} \sum_{r=1}^{L} \tilde{z}_{i,p}^{(r)} \mathbf{1}_{\left| \left\langle \boldsymbol{w}_k, \frac{\boldsymbol{h}_X + \boldsymbol{h}_Y}{2} \right\rangle \right| \geq b_{f_k} + \left| \left\langle \boldsymbol{w}_k, \boldsymbol{h}_X - \frac{\boldsymbol{h}_X + \boldsymbol{h}_Y}{2} \right\rangle \right|} \right] \\
&\pm O(\alpha_p) \cdot \mathbb{E} \left[\sum_{s=1}^{L} \sum_{r=1}^{L} \frac{|\tilde{z}_{i,p}^{(r)}| + |\tilde{z}_{i,p}^{(s)}|}{2} |\tilde{z}_{i,p}^{(r)}| \right] \cdot O(L_1 + L_2).
\end{aligned} \tag{238}$$

□

Proof of Lemma F.3 (b). This proof is extremely similar to the above proof of Lemma F.3 (a), we describe the differences here and sketch the remaining. First we need to decompose the expectation as follows:

$$\begin{aligned}
&\mathbb{E}_{i \sim \mathcal{P}^+} \left[g_k(\boldsymbol{Y}_i) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)} \rangle| \geq b_{f_k}} \left(\sum_{\ell=1}^{L} \delta_{i,\ell}^{(r)} z_{i,p}^{(\ell)} \right) \right] \\
&= \mathbb{E}_{i \sim \mathcal{P}^+} \left[\frac{1}{L} (\alpha_p - b_{g_k}) \sum_{s=1}^{L} \tilde{z}_{i,p}^{(s)} \sum_{r=1}^{L} \tilde{z}_{i,p}^{(r)} \mathbf{1}_{\left| \left\langle \boldsymbol{w}_k, \frac{\boldsymbol{h}_X^{(r)} + \boldsymbol{h}_Y^{(s)}}{2} \right\rangle \right| \geq b_{f_k} + \left| \left\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)} + \boldsymbol{h}_Y^{(s)}}{2} \right\rangle \right|} \right] \\
&+ \mathbb{E}_{i \sim \mathcal{P}^+} \left[\sum_{s=1}^{L} S^{(s) \setminus p} \sum_{r=1}^{L} \tilde{z}_{i,p}^{(r)} \mathbf{1}_{\left| \left\langle \boldsymbol{w}_k, \frac{\boldsymbol{h}_X^{(r)} + \boldsymbol{h}_Y^{(s)}}{2} \right\rangle \right| \geq b_{f_k} + \left| \left\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)} + \boldsymbol{h}_Y^{(s)}}{2} \right\rangle \right|} \right] \\
&+ \mathbb{E}_{i \sim \mathcal{P}^+} \left[\sum_{s=1}^{L} \left(\frac{1}{L} \alpha_p \tilde{z}_{i,p}^{(s)} + S^{(s) \setminus p} \right) \sum_{r=1}^{L} \tilde{z}_{i,p}^{(r)} \mathbf{1}_{\left| \left\langle \boldsymbol{w}_k, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)} + \boldsymbol{h}_Y^{(s)}}{2} \right\rangle \right| \geq b_{f_k} + \left| \left\langle \boldsymbol{w}_k, \frac{\boldsymbol{h}_X^{(r)} + \boldsymbol{h}_Y^{(s)}}{2} \right\rangle \right|} \right] \\
&= J_1 + J_2 + J_3.
\end{aligned} \tag{239}$$

where we have used the following facts:

$$\sum_{s=1}^{L}\sum_{r=1}^{L}\mathbb{E}_{i\sim\mathcal{P}^+}\left[b_{g_k}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\tilde{z}_{i,p}^{(r)}\right]=-\sum_{s=1}^{L}\sum_{r=1}^{L}\mathbb{E}_{i\sim\mathcal{P}^+}\left[b_{g_k}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}\tilde{z}_{i,p}^{(r)}\right];$$
$$\sum_{s=1}^{L}\sum_{r=1}^{L}\mathbb{E}_{i\sim\mathcal{P}^+}\left[b_{g_k}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}\tilde{z}_{i,p}^{(r)}\right]=\sum_{s=1}^{L}\sum_{r=1}^{L}\mathbb{E}_{i\sim\mathcal{P}^+}\left[b_{g_k}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\tilde{z}_{i,p}^{(r)}\right];$$
$$\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}=\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|};$$
$$\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}=\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\leq -b_{f_k}-\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|};$$
$$\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\geq b_{f_k}}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}=\mathbf{1}_{\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle\leq -b_{f_k}}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}=\mathbf{1}_{\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}. \tag{240}$$

Now observe that J_2 can be dealt with as follows: define events $A_3 := \left\{\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+(\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)})\alpha_p^{(r,s)}\right|\geq\frac{\alpha_p C_{\tilde{z}}-b_{g_k}}{2}\right\}$ and $A_4 := \left\{\bar{\mathcal{S}}^{(r,s)\setminus p}\geq\frac{\alpha_p C_{\tilde{z}}-b_{g_k}}{2}\right\}$, and notice that $\mathbf{1}\leq\mathbf{1}_{A_3}+\mathbf{1}_{A_4}$, we can compute

$$\begin{aligned}
J_2&=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}S^{(s)\setminus p}\sum_{r=1}^{L}\mathbf{1}_{\left|\left\langle \boldsymbol{w}_k,\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}-\frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|}\tilde{z}_{i,p}^{(r)}\right]\\
&=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}S^{(s)\setminus p}\sum_{r=1}^{L}|\tilde{z}_{i,p}^{(r)}|\left(\mathbf{1}_{A_3}+\mathbf{1}_{A_4}\mathbf{1}_{A_3^c}\right)\right.\\
&\qquad\left.\cdot\,\mathbf{1}_{\mathcal{S}^{(r,s)\setminus p}\in[\,b_{f_k}-\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}\right|+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|,\,b_{f_k}+\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}\right|+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}-\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|\,]}\right]\\
&\overset{(1)}{\leq}\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\sum_{r=1}^{L}\left(O(\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}\right|+|\alpha_p^{(r,s)}|\left|\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|)|\tilde{z}_{i,p}^{(r)}|\mathbf{1}_{A_3}+|S^{(s)\setminus p}||\tilde{z}_{i,p}^{(r)}|(\mathbf{1}_{A_3}+\mathbf{1}_{A_4})\right)\right.\\
&\qquad\left.\cdot\,\mathbf{1}_{\mathcal{S}^{(r,s)\setminus p}\in[\,\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}\right|-b_{f_k}+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}+\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|,\,b_{f_k}+\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}\right|+\left|\bar{\mathcal{S}}^{(r,s)\setminus p}-\alpha_p^{(r,s)}\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|\,]}\right]\\
&\leq\frac{1}{L}\mathbb{E}\left[\sum_{s=1}^{L}\frac{|\tilde{z}_{i,p}^{(r)}|+|\tilde{z}_{i,p}^{(s)}|}{2}\sum_{r=1}^{L}|\tilde{z}_{i,p}^{(r)}|\right]\cdot O(\alpha_p)\cdot\left(\Pr(A_3)+\sqrt{\frac{\mathbb{E}[|\bar{\mathcal{S}}^{\setminus p}|^2(\mathbf{1}_{A_3}+\mathbf{1}_{A_4})]}{\mathbb{E}[\langle \boldsymbol{w}_k,\xi_p\rangle^2]}}\right).
\end{aligned}\tag{241}$$

where ① relies on the fact that whenever $\mathbf{1}_{\left\langle \boldsymbol{w}_k,\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}\right\rangle\geq b_{f_k}+\left|\left\langle \boldsymbol{w}_k,\frac{\tilde{z}_{i,p}^{(r)}}{L}-\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}\right\rangle\right|}\neq 0$, we have $|S^{(s)\setminus p}|\leq b_{f_k}+\alpha_p\left|\frac{\tilde{z}_{i,p}^{(r)}+\tilde{z}_{i,p}^{(s)}}{2L}\right|+|\alpha_p^{(r,s)}|\left|\frac{\tilde{z}_{i,p}^{(r)}-\tilde{z}_{i,p}^{(s)}}{2L}\right|+|\bar{\mathcal{S}}^{(r,s)\setminus p}|$, and also that we have assumed $b_{f_k}<\alpha_p C_{\tilde{z}}$. The term J_3 can be bounded from similar analysis as in the proof of Lemma F.3 (a), but changing the factor from $O(L_1+L_2)$ to $O(L_3+L_4)$. Combining these calculations gives the desired results. □

F.2.8 Proof of Lemma F.4

Proof of Lemma F.4. Again in this proof we omit the time superscript $^{(t)}$. First we deal with the case where the feature under consideration is $\boldsymbol{M}_p^V$, $p\in[d]$. $\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}$ and $\tilde{\boldsymbol{\xi}}_{Y_i}^{(r)}$ are independent. Now we

can write as follows:

$$\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{P}^+}\left[g_k(\boldsymbol{Y}_i)\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle\right]\\
&=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle-b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle\right]\\
&+\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle+b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\leq -b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle\right]\\
&=I_1+I_2.
\end{aligned}\tag{242}$$

For the first term on the RHS, we have

$$I_1=\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle-b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle\right].\tag{243}$$

where $\{\boldsymbol{M}_p^V\}_{p\in[d]}$ is a basis for $\mathbb{R}^{d_1}$ satisfying $\|\boldsymbol{M}_p^V\|_\infty\leq O(1/\sqrt{d_1})$.

$$\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle-b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle\right]\\
&\leq\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle-b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\Big(\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle+2(\alpha-\alpha')|\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle||\geq b_{f_k}}\right.\\
&\left.-\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle-2(\alpha-\alpha')|\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle||\geq b_{f_k}}\Big)|\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle|\right]\\
&\leq\mathbb{E}_{(\boldsymbol{I}-\boldsymbol{M}_p^V(\boldsymbol{M}_p^V)^\top)\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}}\left[\sum_{s=1}^{L}\left(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle-b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\left|\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle\right|\right]\\
&\times\mathbb{E}_{(\boldsymbol{I}-\boldsymbol{M}_p^V(\boldsymbol{M}_p^V)^\top)\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}}\left[\left|\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle+2(\alpha-\alpha')|\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle||\geq b_{f_k}}-\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle-2(\alpha-\alpha')|\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle||\geq b_{f_k}}\right|\right]\\
&\leq O(1)\cdot\mathbb{E}\left[(\alpha_p+|\bar{\alpha}_p|)^2\sum_{s=1}^{L}\sum_{r=1}^{L}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle^3\cdot\frac{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle-b_{g_k}}{\mathbb{E}[\langle \boldsymbol{w}_k,\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}\rangle^2]}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\right].
\end{aligned}\tag{244}$$

where in the last inequality we have used the randomness of $(\boldsymbol{I}-\boldsymbol{M}_p^V(\boldsymbol{M}_p^V)^\top)\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}$, which allows us to obtain the denominator $\Omega(1)\mathbb{E}\left[|\langle \boldsymbol{w}_k,\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}\rangle|^2\right]$. Noticing that $\sum_{p\in[d]}(|\alpha_p|^2+|\bar{\alpha}_p|^2)=\mathcal{O}(\|\boldsymbol{w}_k\|_2^2)$, coupled with the Cauchy–Schwarz inequality, we can similarly obtain:

$$\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{P}^+}\left[\sum_{s=1}^{L}\left(\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle-b_{g_k}\right)\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\sum_{r=1}^{L}\mathbf{1}_{|\langle \boldsymbol{w}_k,\boldsymbol{h}_X^{(r)}\rangle|\geq b_{f_k}}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle\right]\\
&\leq O(1)\cdot\mathbb{E}\left[(\alpha_p+|\bar{\alpha}_p|)^2\sum_{s=1}^{L}\sum_{r=1}^{L}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle^3\cdot\frac{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle-b_{g_k}}{\mathbb{E}[\langle \boldsymbol{w}_k,\tilde{\boldsymbol{\xi}}_{X_i}^{(r)}\rangle^2]}\mathbf{1}_{\langle \boldsymbol{v}_k,\boldsymbol{h}_Y^{(s)}\rangle\geq b_{g_k}}\right]\\
&\leq O(1)\cdot\mathbb{E}\left[(\alpha_p+|\bar{\alpha}_p|)^2\sum_{r=1}^{L}\langle\tilde{\boldsymbol{\xi}}_{X_i}^{(r)},\boldsymbol{M}_p^V\rangle^3\right]\cdot\frac{1}{\|\boldsymbol{w}_k\|_2^2/d}\times\|\boldsymbol{v}_k\|_2^2\cdot\Pr(g_{k,t}(\boldsymbol{Y}_i)\neq 0)\\
&\leq\widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{w}_k\|_2}{d^2}\right)\cdot\Pr(g_{k,t}(\boldsymbol{Y}_i)\neq 0).
\end{aligned}\tag{245}$$

where in the second inequality we have used the following arguments: first we can compute

$$\begin{aligned}\Big(\langle \boldsymbol{v}_k, \boldsymbol{h}_Y^{(s)}\rangle - b_{g_k}\Big) &\le \sum_{p\in\mathcal{N}_k} \big|\langle \boldsymbol{v}_k, \boldsymbol{M}_p^L\rangle \tilde{z}_{i,p}^{(s)}\big| + \left|\sum_{p\notin\mathcal{N}_k} \langle \boldsymbol{v}_k, \boldsymbol{M}_p^L\rangle \tilde{z}_{i,p}^{(s)}\right| + \big|\langle \boldsymbol{v}_k, \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}\rangle\big| \\ &\le I_{1.1} + I_{1.2} + I_{1.3}.\end{aligned} \tag{246}$$

And from Lemma F.5, we have $|I_{1.1}| \ge \Omega(\|\boldsymbol{v}_k^{(t)}\|_2)$ with probability at least $1 - \widetilde{\mathcal{O}}(1/d)$, and $|I_{1.2}|, |I_{1.3}| \le \widetilde{\mathcal{O}}\Big(\frac{\|\boldsymbol{v}_k^{(t)}\|_2}{\sqrt{d}}\Big)$ w.h.p.

For the second term I_2 on the RHS, we can prove it in a similar way. Summing up over I_1 and I_2, we obtain the desired bound. For the dense feature $(\boldsymbol{M}_p^V)^\perp$, the analysis is similar and we omit it for brevity. □

F.2.9 Proof of Lemma F.5

Proof of Lemma F.5 (a). At iteration $t = T_1$, we have verified all the above properties in Theorem F.1. Suppose all the properties hold for $t < T_2$. We will show that they continue to hold at iteration $t+1$. To calculate the gradient $\nabla_{\boldsymbol{w}_k} L_{\text{aug}}$ along each feature $\boldsymbol{M}_p^V$ and $(\boldsymbol{M}_p^V)^\perp$, we apply Lemma F.3, Lemma F.4, and Lemma F.1. We first verify the conditions required in Lemma F.3 (a) and (b). In particular, we recall the definition of A_1, A_2, A_3 and A_4 in (190) and the proof of lemma C.4 in Allen-Zhu & Li (2022). We have

$$|\bar{S}_{k,t}^{(r,s)\setminus p}|^2 = \left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_Y^{(s)\setminus p} - \boldsymbol{h}_X^{(r)\setminus p}}{2}\right\rangle\right|^2 \le \widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{w}_k^{(t)}\|_2^2}{d}\right). \tag{247}$$

$$\Pr(A_1), \Pr(A_2) \le e^{-\Omega(\log^{1/4} d)} \qquad \text{when } |\alpha_{k,p}^{(t)}| \le \sqrt{1-\gamma}\, b_{f_k}^{(t)}. \tag{248}$$

$$\Pr(A_3), \Pr(A_4) \le e^{-\Omega(\log^{1/4} d)} \qquad \text{when } |\alpha_{k,p}^{(t)}| \ge \sqrt{1+\gamma}\, b_{f_k}^{(t)}. \tag{249}$$

This further implies that when $|\alpha_{k,p}^{(t)}| \le \sqrt{1-\gamma}\, b_{f_k}^{(t)}$,

$$\sqrt{\mathbb{E}\left[|\bar{S}_{k,t}^{(r,s)\setminus p}|^2(\mathbf{1}_{A_1} + \mathbf{1}_{A_2})\right]} \le \sqrt{\widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{w}_k^{(t)}\|_2^2}{d}\right)(\Pr(A_1) + \Pr(A_2))} \overset{(1)}{\le} \frac{\|\boldsymbol{w}_k^{(t)}\|_2}{\sqrt{d}\cdot \operatorname{polylog}(d)}. \tag{250}$$

Here (1) follows from $e^{-\Omega(\log^{1/4} d)} \le 1/\operatorname{polylog}(d)$.

For L_1, we have

$$L_1 := \sqrt{\frac{\mathbb{E}\left[|\bar{S}_{k,t}^{(r,s)\setminus p}|^2(\mathbf{1}_{A_1} + \mathbf{1}_{A_2})\right]}{\mathbb{E}\left[\left\langle \boldsymbol{w}_k^{(t)}, \frac{\tilde{\boldsymbol{\xi}}_{X_i}^{(r)} + \tilde{\boldsymbol{\xi}}_{Y_i}^{(s)}}{2}\right\rangle^2\right]}} \le \sqrt{\frac{\frac{\|\boldsymbol{w}_k^{(t)}\|_2^2}{d\cdot\operatorname{polylog}(d)}}{\Theta\left(\frac{\|\boldsymbol{w}_k^{(t)}\|_2^2\sqrt{\log d}}{d}\right)}} \le \frac{1}{\operatorname{polylog}(d)}. \tag{251}$$

Moreover, $L_2 = \Pr(A_2)$ satisfies the same bound, and hence

$$L_1, L_2 \le \frac{1}{\operatorname{polylog}(d)}. \tag{252}$$

Similarly, when $|\alpha_{k,p}^{(t)}| \ge \sqrt{1+\gamma}\, b_{f_k}^{(t)}$, we also have $L_3, L_4 \le 1/\operatorname{polylog}(d)$.

Now we separately discuss three cases:

(a) When $k \in \mathcal{M}_p^\star$, if $\tilde{z}_{i,p}^{(s)}$ and $\tilde{z}_{i,p}^{(r)}$ are nonzero, say $\frac{\tilde{z}_{i,p}^{(s)}+\tilde{z}_{i,p}^{(r)}}{2} = C_{\tilde{z}}^{(r,s)}$, we simply have

$$\begin{aligned}
&\Pr\left(\left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right| \geq b_{f_k}^{(t)} + \left|\left\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|\right) \\
\geq 1 - \Pr&\left(\left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)\setminus p}+\boldsymbol{h}_Y^{(s)\setminus p}}{2}\right\rangle\right| \geq b_{f_k}^{(t)} - \alpha_{k,p}^{(t)}\frac{\tilde{z}_{i,p}^{(s)}+\tilde{z}_{i,p}^{(r)}}{2} + \left|\left\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right|\right).
\end{aligned} \tag{253}$$

From the observations that: (1) $\left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)\setminus p}+\boldsymbol{h}_Y^{(s)\setminus p}}{2}\right\rangle\right| \leq \frac{\gamma}{2} b_{f_k}^{(t)}$ with probability at least $1 - e^{-\Omega(\log^{1/4} d)}$ (The proof is similar to the proof of lemma C.4 in Allen-Zhu & Li (2022)); (2) $\left|\left\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right| \leq O(\|\boldsymbol{w}_k^{(t)}\|_2 \sigma_\xi)$ with probability at least $1 - e^{-\Omega(\log^{1/2} d)}$.

Therefore, we can calculate that

$$\begin{aligned}
\left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\right\rangle\right| &= \left|\alpha_{k,p}^{(t)}\frac{\tilde{z}_{n,p}^{(r)}+\tilde{z}_{n,p}^{(s)}}{2} + \left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)\setminus p}+\boldsymbol{h}_Y^{(s)\setminus p}}{2}\right\rangle\right| \\
&\geq \left|\alpha_{k,p}^{(t)}\frac{\tilde{z}_{n,p}^{(r)}+\tilde{z}_{n,p}^{(s)}}{2}\right| - \left|\left\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)\setminus p}+\boldsymbol{h}_Y^{(s)\setminus p}}{2}\right\rangle\right| \\
&\geq \sqrt{1+\gamma}\, b_{f_k}^{(t)} - \frac{\gamma}{2} b_{f_k}^{(t)} \\
&\geq b_{f_k}^{(t)}
\end{aligned} \tag{254}$$

So it can be easily verified that:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{L}\tilde{z}_{n,p}^{(s)}\sum_{r=1}^{L}\tilde{z}_{n,p}^{(r)}\mathbf{1}_{|\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\rangle| \geq b_{f_k}^{(t)} + |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\rangle|}\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{L}\tilde{z}_{n,p}^{(s)}\sum_{r=1}^{L}\tilde{z}_{n,p}^{(r)}\right]\Pr\left(|\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\rangle| \geq b_{f_k}^{(t)} + |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)}+\boldsymbol{h}_Y^{(s)}}{2}\rangle|\right) \\
&= \epsilon_p \frac{L^2 \log\log d}{d}\left(1 - \frac{1}{\mathrm{polylog}(d)}\right).
\end{aligned} \tag{255}$$

Now we can compute as follows: for $\boldsymbol{M}_p^V$ such that $k \in \mathcal{M}_p^\star$, at iteration $t+1$:

$$
\begin{aligned}
&\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle \\
=&\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle - \eta \langle \nabla_{\boldsymbol{w}_k} L_{\mathrm{aug}}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}), \boldsymbol{M}_p^V \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
=&(1-\eta\lambda)\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&+\eta \mathbb{E}_{(i,j)\sim \mathcal{P}^{all}} \left[\left(a_{i,j} \frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}}{1 + e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right) g_k(\boldsymbol{Y}_j) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} \rangle| \geq b_{f_k}^{(t)}} \langle \boldsymbol{h}_X^{(r)}, \boldsymbol{M}_p^V \rangle \right] \\
\overset{①}{=}&(1-\eta\lambda)\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&+\eta \mathbb{E}_{(i,j)\sim \mathcal{P}^{all}} \left[\left(a_{i,j} \frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}}{1 + e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right) g_k(\boldsymbol{Y}_j) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} \rangle| \geq b_{f_k}^{(t)}} \left(\tilde{z}_{n,p}^{(r)} + \langle \tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, \boldsymbol{M}_p^V \rangle \right) \right] \\
=&(1-\eta\lambda)\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&+\eta \mathbb{E}_{(i,j)\sim \mathcal{P}^{all}} \left[\left(a_{i,j} \frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}}{1 + e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right) g_k(\boldsymbol{Y}_j) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} \rangle| \geq b_{f_k}^{(t)}} \tilde{z}_{n,p}^{(r)} \right] \\
&+\eta \mathbb{E}_{(i,j)\sim \mathcal{P}^{all}} \left[\left(a_{i,j} \frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}}{1 + e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right) g_k(\boldsymbol{Y}_j) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} \rangle| \geq b_{f_k}^{(t)}} \langle \tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, \boldsymbol{M}_p^V \rangle \right] \\
\overset{②}{\geq}&\left(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle - \mathrm{sign}\left(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \right) b_{f_k}^{(t)} \right) \left(1 + \epsilon_p \frac{\eta \log\log d}{d} \left(1 - \frac{1}{\mathrm{polylog}(d)} \right) \right).
\end{aligned}
\tag{256}
$$

① is because

$$
\left\langle \boldsymbol{M}^V \tilde{\boldsymbol{z}}_n^{(r)}, \boldsymbol{M}_p^V \right\rangle = \left\langle \sum_{p'=1}^{d} \tilde{z}_{n,p'}^{(r)} \boldsymbol{M}_{p'}^V, \boldsymbol{M}_p^V \right\rangle = \sum_{p'=1}^{d} \tilde{z}_{n,p'}^{(r)} \left\langle \boldsymbol{M}_{p'}^V, \boldsymbol{M}_p^V \right\rangle = \tilde{z}_{n,p}^{(r)}. \tag{257}
$$

② is because Lemma F.3, Lemma F.4 and.

$$
\sum_{k\in[m]} \|\boldsymbol{w}_k^{(t)}\|_2^2 \leq \sum_{k\in[m]} \|\boldsymbol{w}_k^{(T_2)}\|_2^2 \leq \sum_{k\in[m]} d\, \|\boldsymbol{w}_k^{(T_1)}\|_2^2 \leq \frac{1}{\mathrm{poly}(d)}. \tag{258}
$$

which follows from our definition of iteration T_2 and the properties at iteration T_1 in Theorem F.1.

Next we compare this growth to the growth of bias $b_{f_k}^{(t+1)}$. Since we raise our bias by

$$
b_{f_k}^{(t+1)} = \max\left\{ b_{f_k}^{(t)} \left(1 + \frac{\eta}{d}\right),\, b_{f_k}^{(t)} \frac{\|\boldsymbol{w}_k^{(t+1)}\|_2}{\|\boldsymbol{w}_k^{(t)}\|_2} \right\}, \tag{259}
$$

The last inequality in (256)

$$
\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle \geq \left(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle - \mathrm{sign}(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle) \cdot b_{f_k}^{(t)} \right) \cdot \left(1 + \frac{\eta}{d}\right) \tag{260}
$$

has already verified the case of $\left(1 + \frac{\eta}{d}\right)$. As long as

$$
\frac{\|\boldsymbol{w}_k^{(t+1)}\|_2}{\|\boldsymbol{w}_k^{(t)}\|_2} \leq \frac{|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle|}{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|}, \tag{261}
$$

we can obtain the desired result.

After establishing the bounds of growth speed for each feature, we now calculate the proportions they contribute to each neuron weight $k \in [m]$. Namely, we need to prove that when Lemma F.5 holds at iteration $t \in [T_1, T_2]$, we have

$$\frac{|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle|}{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|} \geq \frac{\|\boldsymbol{w}_k^{(t+1)}\|_2}{\|\boldsymbol{w}_k^{(t)}\|_2} \quad \text{for } k \in \mathcal{M}_p^\star. \tag{262}$$

we argue as follows: from previous calculations we have:

$$\begin{aligned}
&\sum_{p' \in [d],\, p' \neq p} \langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_{p'}^V \rangle^2 + \sum_{p' \in [d_1] \setminus [d]} \langle \boldsymbol{w}_k^{(t+1)}, (\boldsymbol{M}^V)_{p'}^\perp \rangle^2 \\
\leq &\sum_{p' \in [d],\, p' \neq p} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_{p'}^V \rangle^2 \left(1 + \epsilon_{p'} \frac{O(\eta)}{d\,\mathrm{polylog}(d)}\right)^2 \\
+ &\sum_{p' \in [d_1] \setminus [d]} \langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)_{p'}^\perp \rangle^2 + \widetilde{O}\Big(\frac{\eta}{d}\Big)\, e^{-\Omega(\log^{1/4} d)} \|\boldsymbol{w}_k^{(t)}\|_2^2.
\end{aligned} \tag{263}$$

Therefore by adding $\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle^2$ to the LHS we have:

$$\begin{aligned}
\|\boldsymbol{w}_k^{(t+1)}\|_2^2 \leq \|\boldsymbol{w}_k^{(t)}\|_2^2 &\left(1 + \frac{O(\eta)}{d \cdot \mathrm{polylog}(d)}\right)^2 \\
&+ \left(\frac{|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle|}{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|} - \frac{O(\eta)}{d \cdot \mathrm{polylog}(d)}\right) |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|^2.
\end{aligned} \tag{264}$$

which implies

$$\frac{\|\boldsymbol{w}_k^{(t+1)}\|_2^2}{\|\boldsymbol{w}_k^{(t)}\|_2^2} \leq \left(1 + \frac{O(\eta)}{d \cdot \mathrm{polylog}(d)}\right)^2 + \left(\frac{|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle|}{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|}\right) \frac{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|^2}{\|\boldsymbol{w}_k^{(t)}\|_2^2}. \tag{265}$$

Therefore, $\frac{|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle|}{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|} \geq \frac{\|\boldsymbol{w}_k^{(t+1)}\|_2}{\|\boldsymbol{w}_k^{(t)}\|_2}$ is as desired.

□

Proof of Lemma F.5 (b). When $k \notin \mathcal{M}_p^\star$, we can similarly obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{L} \tilde{z}_{n,p}^{(s)} \sum_{r=1}^{L} \tilde{z}_{n,p}^{(r)} \mathbf{1}_{|\langle \boldsymbol{w}_k^{(t)}, \frac{\boldsymbol{h}_X^{(r)} + \boldsymbol{h}_Y^{(s)}}{2} \rangle| \geq b_{f_k}^{(t)} + |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} - \frac{\boldsymbol{h}_X^{(r)} + \boldsymbol{h}_Y^{(s)}}{2} \rangle|}\right] \\
\leq &\epsilon_p \frac{L^2 \log\log d}{d} \left(\frac{1}{\mathrm{polylog}(d)}\right) \\
= &O\left(\epsilon_p \frac{L^2}{d \cdot \mathrm{polylog}(d)}\right).
\end{aligned} \tag{266}$$

And similarly we can compute the gradient descent dynamics as follows:

For $p \in [d]$ such that $|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle| \geq \frac{\|\boldsymbol{w}_k^{(t)}\|_2 d}{\sqrt{d_1}}$, we have (assume here $\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle > 0$, the opposite is similar)

$$\begin{aligned}
&\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle \\
=&\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle - \eta \langle \nabla_{\boldsymbol{w}_k} L_{\text{aug}}(\boldsymbol{f}_\theta, \boldsymbol{g}_\phi), \boldsymbol{M}_p^V \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
\leq&\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \left(1 - \eta\lambda + \epsilon_p \frac{O(\eta)}{d \cdot \text{polylog}(d)}\right) \pm \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
\leq&\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle \left(1 + \epsilon_p \frac{O(\eta)}{d \cdot \text{polylog}(d)}\right) \pm \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right).
\end{aligned} \tag{267}$$

Since from our update rule

$$b_{f_k}^{(t+1)} \geq b_{f_k}^{(t)} \left(1 + \frac{\eta}{d}\right) \tag{268}$$

we know that

$$b_{f_k}^{(t+1)} \geq \left(1 + \frac{\eta}{d}\right) b_{f_k}^{(t)} \geq \left(1 + \frac{O(\eta)}{d \cdot \text{polylog}(d)}\right) b_{f_k}^{(t)} \geq \frac{|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle|}{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|} b_{f_k}^{(t)}. \tag{269}$$

Thus, if $|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle| \leq \sqrt{1-\gamma}\, b_{f_k}^{(t)}$ at iteration t, we have

$$|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle| \leq \frac{b_{f_k}^{(t+1)} |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|}{b_{f_k}^{(t)}} \leq \sqrt{1-\gamma}\, b_{f_k}^{(t+1)} \quad \text{if } |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle| \geq \frac{\|\boldsymbol{w}_k^{(t)}\|_2 d}{\sqrt{d_1}} \text{ at iteration } t;$$

$$|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle| \leq \sqrt{1-\gamma}\, b_{f_k}^{(t+1)} \quad \text{if } |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle| \leq \frac{\|\boldsymbol{w}_k^{(t)}\|_2 d}{\sqrt{d_1}} \text{ at iteration } t.$$

It is also worth noting that similar calculations also lead to a lower bound:

$$|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle| \geq |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle|(1 - \eta\lambda) - \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right). \tag{270}$$

We leave the part of proving $|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle| \leq \widetilde{O}\left(\frac{\|\boldsymbol{w}_k^{(t+1)}\|_2}{\sqrt{d}}\right)$ to later.

To prove $|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle| \leq \widetilde{O}\left(\frac{\|\boldsymbol{w}_k^{(t+1)}\|_2}{\sqrt{d}}\right)$ if $k \notin \mathcal{M}_p^\star$, we first use inequality to compute

$$\begin{aligned}
&\|\boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(t+1)}\|_2 \\
\geq&\|\boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T_1)}\|_2 (1 - \eta\lambda)^{t - T_1 + 1} - \widetilde{O}\left(\eta(t - T_1 + 1) \max_{t' \in [T_1, t+1]} \|\boldsymbol{w}_k^{(t')}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
\geq&\|\boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T_1)}\|_2 (1 - \eta\lambda)^{t - T_1 + 1} - \widetilde{O}\left(\eta(t - T_1) \|\boldsymbol{w}_k^{(T_1)}\|_2 \sqrt{\frac{dqL}{|\mathcal{B}|}}\right) \\
\geq&\|\boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T_1)}\|_2 (1 - o(1)) \quad \text{for} \quad t \leq \Theta\left(\frac{d \log d}{\eta \log\log d}\right).
\end{aligned} \tag{271}$$

Notice that

$$|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V \rangle| \leq \sqrt{1-\gamma} \frac{\sqrt{\log d}}{\sqrt{d}} \|\boldsymbol{w}_k^{(T_1)}\|_2 \leq O\left(\sqrt{\frac{\log d}{d}}\right) \left\|\boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T_1)}\right\|_2. \tag{272}$$

From Theorem F.1. Suppose it also holds for iteration t, we have

$$
\begin{aligned}
&|\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{M}_p^V \rangle| \\
\leq &|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V \rangle| \left(1 + \epsilon_p \frac{O(\eta)}{d \ \mathrm{polylog}(d)}\right) + \widetilde{O}\left(\eta \|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
\leq &|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V \rangle| \left(1 + \epsilon_p \frac{O(\eta)}{d \ \mathrm{polylog}(d)}\right)^{t-T_1} + \widetilde{O}\left(\eta (t-T_1) \|\boldsymbol{w}_k^{(T_1)}\|_2 \sqrt{\frac{dqL}{|\mathcal{B}|}}\right) \\
&\text{(because } \|\boldsymbol{w}_k^{(t)}\|_2 \leq \sqrt{d} \|\boldsymbol{w}_k^{(T_1)}\|_2 \text{ by the definition of } T_2) \\
\leq &|\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V \rangle| (1 + o(1)) + \widetilde{O}\left(\eta (t-T_1) \|\boldsymbol{w}_k^{(T_1)}\|_2 \sqrt{\frac{dqL}{|\mathcal{B}|}}\right) \\
\overset{①}{\leq} &O\left(\sqrt{\frac{\log d}{d}}\right) \left\| \boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(T_1)} \right\|_2 (1 + o(1)) \\
\overset{②}{\leq} &O\left(\sqrt{\frac{\log d}{d}}\right) \left\| \boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(t+1)} \right\|_2 (1 + o(1)) \\
\leq &O\left(\sqrt{\frac{\log d}{d}}\right) \|\boldsymbol{w}_k^{(t+1)}\|_2 (1 + o(1)).
\end{aligned}
\tag{273}
$$

① and ② use the above inequality respectively.

□

Proof of Lemma F.5 (c). Next we consider the learning dynamics for the dense features. We can use Lemma F.4 to calculate its dynamics by

$$
\begin{aligned}
&\langle \boldsymbol{w}_k^{(t+1)}, (\boldsymbol{M}^V)^\perp \rangle \\
= &\langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)^\perp \rangle - \eta \langle \nabla_{\boldsymbol{w}_k} L_{\mathrm{aug}}(\boldsymbol{f}_\theta, \boldsymbol{g}_\phi), (\boldsymbol{M}^V)^\perp \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
= &(1 - \eta\lambda) \langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)^\perp \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&+ \eta \mathbb{E}_{(i,j) \sim \mathcal{P}^{all}} \left[\left(a_{i,j} \frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_\theta, \boldsymbol{g}_\phi}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}}{1 + e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_\theta, \boldsymbol{g}_\phi}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right) g_k(\boldsymbol{Y}_j) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} \rangle| \geq b_{f_k}^{(t)}} \langle \boldsymbol{h}_X^{(r)}, (\boldsymbol{M}^V)^\perp \rangle \right] \\
= &(1 - \eta\lambda) \langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)^\perp \rangle \pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
&+ \eta \mathbb{E}_{(i,j) \sim \mathcal{P}^{all}} \left[\left(a_{i,j} \frac{e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_\theta, \boldsymbol{g}_\phi}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}}{1 + e^{a_{i,j}\left(-\mathrm{sim}_{\boldsymbol{f}_\theta, \boldsymbol{g}_\phi}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right) g_k(\boldsymbol{Y}_j) \sum_{r=1}^{L} \mathbf{1}_{|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_X^{(r)} \rangle| \geq b_{f_k}^{(t)}} \langle \tilde{\boldsymbol{\xi}}_{X_i}^{(r)}, (\boldsymbol{M}^V)^\perp \rangle \right] \\
= &(1 - \eta\lambda) \langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)^\perp \rangle + \widetilde{O}\left(\frac{\eta \|\boldsymbol{w}_k^{(t)}\|_2}{d\sqrt{d_1}}\right) \Pr\big(g_k(\boldsymbol{Y}_j) \neq 0\big) \pm \widetilde{O}\left(\eta \|\boldsymbol{v}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
\leq &\langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)^\perp \rangle + O\left(\frac{\eta \|\boldsymbol{w}_k^{(t)}\|_2}{d\sqrt{d_1}} e^{-\Omega(\log^{1/4} d)}\right).
\end{aligned}
\tag{274}
$$

For the dense features, we can compute as follows:

$$
\begin{aligned}
&|\langle \boldsymbol{w}_k^{(t+1)}, (\boldsymbol{M}^V)_p^{\perp} \rangle| \\
\leq &|\langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)_p^{\perp} \rangle| + O\left(\frac{\eta \|\boldsymbol{w}_k^{(t)}\|_2}{d\sqrt{d_1}} e^{-\Omega(\log^{1/4} d)} \right) \\
\leq &|\langle \boldsymbol{w}_k^{(T_1)}, (\boldsymbol{M}^V)_p^{\perp} \rangle| + \sum_{t'=T_1}^{t} O\left(\frac{\eta \|\boldsymbol{w}_k^{(t')}\|_2}{d\sqrt{d_1}} e^{-\Omega(\log^{1/4} d)} \right) \\
\leq &O\left(\sqrt{\frac{\log d}{d_1}} \right) \|\boldsymbol{w}_k^{(T_1+1)}\|_2 \left(1 + \frac{O(\eta)}{d \ \mathrm{polylog}(d)} \right) + \sum_{t'=T_1+1}^{t} O\left(\frac{\eta \|\boldsymbol{w}_k^{(t')}\|_2}{d\sqrt{d_1}} e^{-\Omega(\log^{1/4} d)} \right) \\
\leq &O\left(\sqrt{\frac{\log d}{d_1}} \right) \|\boldsymbol{w}_k^{(t+1)}\|_2 \left(1 + \frac{O(\eta)}{d \ \mathrm{polylog}(d)} \right)^{(t-T_1)} \\
\leq &O\left(\sqrt{\frac{\log d}{d_1}} \right) \|\boldsymbol{w}_k^{(t+1)}\|_2.
\end{aligned} \tag{275}
$$

where we have used the assumption that $\|\boldsymbol{w}_k^{(t)}\|_2 \leq \|\boldsymbol{w}_k^{(t+1)}\|_2 \left(1 + \frac{O(\eta)}{d \ \mathrm{polylog}(d)}\right)$ for all $k \in [m]$ and all $t \leq T_2 = T_1 + \Theta\left(\frac{d \log d}{\eta \log \log d}\right)$.

Which we prove here: First of all, from previous calculations we have

$$
\left\| \boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(t+1)} \right\|_2 \geq \left\| \boldsymbol{M}^V (\boldsymbol{M}^V)^\top \boldsymbol{w}_k^{(t)} \right\|_2 (1 - \eta\lambda) - O\left(\frac{\eta \|\boldsymbol{w}_k^{(t)}\|_2}{d^{3/2}} e^{-\Omega(\log^{1/4} d)} \right). \tag{276}
$$

Also the trajectory of $\left\| (\boldsymbol{M}^V)^\perp ((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(t+1)} \right\|_2$ can be lower bounded as

$$
\left\| (\boldsymbol{M}^V)^\perp ((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(t+1)} \right\|_2 \geq \left\| (\boldsymbol{M}^V)^\perp ((\boldsymbol{M}^V)^\perp)^\top \boldsymbol{w}_k^{(t)} \right\|_2 (1 - \eta\lambda) - O\left(\frac{\eta}{d}\right) e^{-\Omega(\log^{1/4} d)} \|\boldsymbol{w}_k^{(t)}\|_2. \tag{277}
$$

Thus by combining the change of $\boldsymbol{w}_k^{(t)}$ over two subspaces, we have

$$
\begin{aligned}
\|\boldsymbol{w}_k^{(t+1)}\|_2^2 &\geq \|\boldsymbol{w}_k^{(t)}\|_2^2 (1 - \eta\lambda)^2 - O\left(\frac{\eta}{d}\right) e^{-\Omega(\log^{1/4} d)} \|\boldsymbol{w}_k^{(t)}\|_2^2 \\
&\geq \|\boldsymbol{w}_k^{(t)}\|_2^2 \left(1 - \frac{O(\eta)}{d \ \mathrm{polylog}(d)} \right).
\end{aligned} \tag{278}
$$

Therefore, we can write:

$$
\|\boldsymbol{w}_k^{(t)}\|_2 \leq \|\boldsymbol{w}_k^{(t+1)}\|_2 \left(1 + \frac{O(\eta)}{d \cdot \mathrm{polylog}(d)} \right). \tag{279}
$$

Which gives the desired bound. □

F.2.10 Proof of Lemma F.6

Proof of Lemma F.6: We have

$$
\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^V \rangle \sim \mathcal{N}(0, \sigma_0^2),
$$

and we want to control the event

$$
\left| \left\langle \boldsymbol{w}_k^{(0)}, \boldsymbol{M}_p^V \right\rangle \right| \leq \frac{\sigma_0}{d}. \tag{280}
$$

To obtain Lemma F.6, we use the standard Gaussian anti-concentration property near the mean. □

F.2.11 TIME T2

In the proof above, we have depended on the crucial assumption that $T_2 := \min\left\{t \in \mathbb{N} : \exists k \in [m] \text{ s.t. } \|\boldsymbol{w}_k^{(t)}\|_2^2 \geq d\|\boldsymbol{w}_k^{(T_1)}\|_2^2\right\}$ is of order $T_1 + \Theta\left(\frac{d\log d}{\eta \log\log d}\right)$. Now we verify it as follows. If $k \in \mathcal{M}_p^\star$ for some $p \in [d]$ (which also means $p' \notin \mathcal{N}_k$ for $p' \neq p$), for some $t = T_1 + O\left(\frac{d\log d}{\eta\log\log d}\right)$, we have

$$\begin{aligned} |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle| &\geq |\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V\rangle| \left(1 + \Omega\left(\epsilon_p \frac{\eta\log\log d}{d}\right)\right)^{t-T_1} \\ &\geq d\sqrt{\frac{2\log d}{d}}\|\boldsymbol{w}_k^{(T_1)}\|_2. \end{aligned} \tag{281}$$

Thus for some $t = T_1 + O\left(\frac{d\log d}{\eta\log\log d}\right)$, we have $|\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle|^2 \geq d\|\boldsymbol{w}_k^{(T_1)}\|_2^2$, which proves that $T_2 \leq T_1 + O\left(\frac{d\log d}{\eta\log\log d}\right)$.

These results are to verify that $\|\boldsymbol{w}_k^{(t)}\|_2^2 \geq d\|\boldsymbol{w}_k^{(T_1)}\|_2^2$ holds under the order of $T_2 = T_1 + \Theta\left(\frac{d\log d}{\eta\log\log d}\right)$.

$$\begin{aligned} \left(1 + \Omega\left(\epsilon_j \frac{\eta\log\log d}{d}\right)\right)^{\Theta\left(\frac{d\log d}{\eta\log\log d}\right)} &= \exp\left(\Omega\left(\epsilon_j\frac{\eta\log\log d}{d}\right)\cdot\Theta\left(\frac{d\log d}{\eta\log\log d}\right)\right) \\ &= \exp(\Omega(\epsilon_j \log d)) \\ &= d^{\Omega(1)}. \end{aligned} \tag{282}$$

Conversely, we also have for all $t \leq T_1 + O\left(\frac{d\log d}{\eta\log\log d}\right)$

$$\begin{aligned} &\sum_{p'\in[d]:p'\neq p} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_{p'}^V\rangle^2 + \sum_{p'\in[d_1]\setminus[d]} \langle \boldsymbol{w}_k^{(t)}, (\boldsymbol{M}^V)_{p'}^\perp\rangle^2 \\ &\leq \|\boldsymbol{w}_k^{(T_1)}\|_2^2 \left(1 + \epsilon_{p'}\frac{O(\eta)}{d\ \mathrm{polylog}(d)}\right)^{t-T_1} + \max_{t'\leq t} O\left(\frac{\eta(t-T_1)}{d}\right) e^{-\Omega(\log^{1/4} d)}\|\boldsymbol{w}_k^{(t')}\|_2^2 \\ &\leq o\left(d\|\boldsymbol{w}_k^{(T_1)}\|_2^2\right). \end{aligned} \tag{283}$$

Except for the principal direction $\boldsymbol{M}_p^V$ (i.e., the alignment direction of neuron k), the total growth of squared weights along all other directions remains far below the target scale $d \cdot \|\boldsymbol{w}_k^{(T_1)}\|_2^2$.

And also

$$\begin{aligned} |\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{M}_p^V\rangle| &\leq |\langle \boldsymbol{w}_k^{(T_1)}, \boldsymbol{M}_p^V\rangle| \left(1 + \epsilon_p\frac{\eta\log\log d}{d}\left(1 - \frac{1}{\mathrm{polylog}(d)}\right)\right)^{t-T_1} \\ &\leq O\left(\sqrt{\frac{\log d}{d}}\|\boldsymbol{w}_k^{(T_1)}\|_2\right)\left(1 + \epsilon_p\frac{\eta\log\log d}{d}\left(1 - \frac{1}{\mathrm{polylog}(d)}\right)\right)^{t-T_1}. \end{aligned} \tag{284}$$

Therefore we at least need $T_1 + \frac{d\log\left(\Omega\left(\sqrt{d}\sqrt{\frac{d}{\log d}}\right)\right)}{\eta\log\log d}(1 - o(1))$ iteration to let any neuron $k \in [m]$ reach $\|\boldsymbol{w}_k^{(t)}\|_2^2 \geq d\|\boldsymbol{w}_k^{(T_1)}\|_2^2$, which proves that $T_2 = T_1 + \Theta\left(\frac{d\log d}{\eta\log\log d}\right)$.

F.3 CONVERGENCE

F.3.1 OPTIMAL LEARNER

Definition F.2 (Optimal Learner)**.** *We define a learner network that we deem as the optimal feature map for this task. we define $u_f^\star := \{u_{f,k}^\star\}_{k\in[m]}$ as follows:*

$$u_{f,k}^\star = \begin{cases} \dfrac{\alpha}{|\mathcal{M}_p^\star| d} \cdot \operatorname{sign}\big(\langle \boldsymbol{w}_k^{(T_2)}, \boldsymbol{M}_p\rangle\big), & \text{if } k \in \mathcal{M}_p^\star, \\ 0, & \text{if } k \notin \bigcup_{p\in[d]} \mathcal{M}_p^\star. \end{cases} \tag{285}$$

Furthermore, we define the optimal feature map $f_t^\star$ as follows. For $k \in [m]$, the k-th neuron of f_{t,u_f} given weight $u_{f,k} \in \mathbb{R}$ is

$$f_{t,u_f,k}(\boldsymbol{X}_i) \;=\; u_{f,k} \sum_{r=1}^{L} \Big[\big(\langle \boldsymbol{w}_k, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle - \boldsymbol{b}_k\big)\mathbf{1}_{\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle \geq \boldsymbol{b}_k} - \big(-\langle \boldsymbol{w}_k, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle - \boldsymbol{b}_k\big)\mathbf{1}_{-\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{h}_{\boldsymbol{X}}^{(r)}\rangle \geq \boldsymbol{b}_k} \Big]. \tag{286}$$

Finally, we write f_{t,u_f} as the concatenation

$$f_{t,u_f}(\cdot) \;=\; \big(f_{t,u_f,1}(\cdot), \ldots, f_{t,u_f,m}(\cdot)\big)^\top. \tag{287}$$

F.3.2 OPTIMALITY

Lemma F.7 (Optimality)**.** *Let $\{u_{f,k}^\star\}_{k\in[m]}$ and f_{t,u_f} be defined as in Definition F.2. When Theorem F.2, define the pseudo loss function*

$$\widetilde{L}(f_{t,u_f^\star}, g_t) := \mathbb{E}\left[-\log \frac{1}{1 + e^{a_{i,j}\left(-\operatorname{sim}_{\boldsymbol{f_\theta}, \boldsymbol{g_\phi}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right]. \tag{288}$$

we obtain the following loss guarantee:

$$\widetilde{L}(f_{t,u_f^\star}, g_t) \leq \; O\big(e^{-\alpha}\big). \tag{289}$$

Proof.

$$\begin{aligned} \widetilde{L}(f_{t,u_f^\star}, g_t) =& \mathbb{E}\left[-\log \frac{1}{1 + e^{a_{i,j}\left(-\operatorname{sim}_{\boldsymbol{f_\theta}, \boldsymbol{g_\phi}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}} \right] \\ =& \mathbb{E}\left[\log(1 + e^{a_{i,j}\left(-\operatorname{sim}_{\boldsymbol{f_\theta}, \boldsymbol{g_\phi}}(\boldsymbol{X}_i, \boldsymbol{Y}_j)\right)}) \right] \\ \leq& \mathbb{E}\left[\log(1 + e^{-\langle \boldsymbol{f_\theta}(\boldsymbol{X}_i), \boldsymbol{g_\phi}(\boldsymbol{Y}_j)\rangle}) \right] \\ \overset{①}{\leq}& \mathbb{E}\left[\log(1 + e^{-\sum_{k\in\mathcal{M}_p^\star} \sum_{p\in[d]} f_{t,u_f^\star,k}(\boldsymbol{X}_i) g_{t,u_g,k}(\boldsymbol{Y}_j)}) \right] \\ \overset{②}{\leq}& O\big(e^{-\alpha}\big) \end{aligned} \tag{290}$$

① follows from the fact that among the m neurons, only $|\mathcal{M}_p^\star|$ neurons are active, and the summation is taken over all $p \in [d]$. ② is obtained by substituting (285) and (286) into the above expression.

□

F.3.3 CONVERGENCE

Proof. We start with the proof of convergence. Denote $u_f^{(t)} = (u_{f,1}^{(t)}, \ldots, u_{f,m}^{(t)})$, since our update is

$$\begin{aligned} u_f^{(t+1)} =& u_f^{(t)} - \eta \nabla_{u_f} \widehat{L}_{\text{aug}}(f_t) \\ =& u_f^{(t)} - \eta \nabla_{u_f} L_{\text{aug}}(f_t) \pm \widetilde{O}\left(\|\boldsymbol{w}_k^{(t)}\|_2 \sqrt{\frac{qL}{|\mathcal{B}|}} \right) \end{aligned} \tag{291}$$

we have

$$\begin{aligned}
&\eta\langle\nabla_{u_f} L_{\text{aug}}(f_t), u_f^{(t)} - u_f^\star\rangle \\
\leq&\tfrac{\eta^2}{2}\|\nabla_{u_f} L_{\text{aug}}(f_t)\|_F^2 + \tfrac{1}{2}\|u_f^{(t)} - u_f^\star\|_F^2 - \tfrac{1}{2}\|u_f^{(t+1)} - u_f^\star\|_F^2 + \widetilde{O}\left(\eta\|\boldsymbol{w}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right) \\
\leq&\eta^2 m + \tfrac{1}{2}\|u_f^{(t)} - u_f^\star\|_F^2 - \tfrac{1}{2}\|u_f^{(t+1)} - u_f^\star\|_F^2 + \widetilde{O}\left(\eta\|\boldsymbol{w}_k^{(t)}\|_2\sqrt{\frac{qL}{|\mathcal{B}|}}\right),
\end{aligned} \tag{292}$$

where the inequality comes from

$$\left\|\nabla_{u_f} L_{\text{aug}}(f_t)\right\|_F^2 = \sum_{k=1}^{m}\left\|\nabla_{u_{f,k}} L_{\text{aug}}(f_t)\right\|^2. \tag{293}$$

Now we will use the tools from online learning to obtain a loss guarantee: define a pseudo objective for parameter θ

$$\begin{aligned}
\widetilde{L}_{\text{aug}_t}(\theta) &:= \widetilde{L}(f_{t,u_f}, g_t) + \tfrac{\lambda}{2}\sum_{k\in[m]}\|u_{f,k}\|_2^2 \\
&= \mathbb{E}\left[-\log\frac{1}{1+e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right] + \tfrac{\lambda}{2}\sum_{k\in[m]}\|u_{f,k}\|_2^2 + \tfrac{\lambda}{2}\sum_{k\in[m]}\|w_k^{(T_2)}\|_2^2.
\end{aligned} \tag{294}$$

The above objective is convex with respect to u_f. This follows from three observations: (i) for fixed f_t and g_t, the $a_{i,j}\left(-t\,\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)+b\right)$ of the softplus function is affine in u_f; (ii) the softplus function, i.e., $\text{softplus}(x) = \log(1+e^x)$, is convex; and (iii) the regularization term is convex. Therefore, by the composition rule of convex functions, $\widetilde{L}(f_{t,u_f}, g_t)$ is convex with respect to u_f.

Moreover, we have

$$\widetilde{L}_{\text{aug}_t}(u_f^{(t)}) = L_{\text{aug}}(f_t), \tag{295}$$

and

$$\nabla_{u_{f,k}}\widetilde{L}_{\text{aug}_t}(u_{f,k}^{(t)}) = \nabla_{u_{f,k}} L_{\text{aug}}(f_t). \tag{296}$$

Thus we have

$$\begin{aligned}
&\eta\langle\nabla_{u_f} L_{\text{aug}}(f_t), u_f^{(t)} - u_f^\star\rangle \\
=&\eta\langle\nabla_\theta\widetilde{L}_{\text{aug}_t}(u_f^{(t)}), u_f^{(t)} - u_f^\star\rangle \\
\overset{①}{\geq}&\widetilde{L}_{\text{aug}_t}(u_f^{(t)}) - \widetilde{L}_{\text{aug}_t}(u_f^\star) \\
\geq&\widetilde{L}_{\text{aug}_t}(u_f^{(t)}) - \mathbb{E}\left[-\log\frac{1}{1+e^{a_{i,j}\left(-\text{sim}_{\boldsymbol{f_\theta},\boldsymbol{g_\phi}}(\boldsymbol{X}_i,\boldsymbol{Y}_j)\right)}}\right] - \tfrac{\lambda}{2}\sum_{k\in[m]}\|u_{f,k}^\star\|_2^2 - \tfrac{\lambda}{2}\sum_{k\in[m]}\|w_k^{(T_2)}\|_2^2. \\
\overset{②}{\geq}&\widetilde{L}_{\text{aug}_t}(u_f^{(t)}) - O\left(e^{-\alpha}\right) - \sum_{k\in[m]} O(\lambda\|u_{f,k}^\star\|_2^2) - \sum_{k\in[m]} O(\lambda\|w_k^{(T_2)}\|_2^2) \\
\geq&L_{\text{aug}}(f_t) - O\left(e^{-\alpha}\right).
\end{aligned} \tag{297}$$

① is because the surrogate objective function $\widetilde{L}_{\text{aug}_t}$ is a convex function with respect to u_f, so we can use a first-order convex lower bound: $f(u_f) - f(u_f') \leq \langle\nabla f(u_f), u_f - u_f'\rangle$. ② is because $\sum_{k\in[m]}\lambda\|u_{f,k}^\star\|_2^2 = \sum_{p\in[d]}\sum_{k\in\mathcal{M}_p^\star}\lambda\|u_{f,k}^\star\|_2^2 = \sum_{j\in[d]}\sum_{i\in\mathcal{M}_j^\star}\lambda(\frac{1}{|\mathcal{M}_p^\star|d\log d})^2 < e^{-\alpha}$ and $\sum_{k\in[m]}\lambda\|w_k^{(T_2)}\|_2^2 < e^{-\alpha}$. By a telescoping summation, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=T_2}^{T_2+T-1}\big(L_{\mathrm{aug}}(f_t)-O(e^{-\alpha})\big) &\le \frac{1}{T}\sum_{t=T_2}^{T_2+T-1}\eta\langle\nabla_{u_f}L_{\mathrm{aug}}(f_t),u_f^{(t)}-u_f^\star\rangle\\
&\le \frac{O(\|u_f^{(T_2)}-u_f^\star\|_F^2)}{T\eta}\\
&= \frac{O\Big(\|u_f^{(T_2)}\|_F^2+\|u_f^\star\|_F^2-2\,\mathrm{Tr}((u_f^{(T_2)})^\top u_f^\star)\Big)}{T\eta}\\
&\le \frac{O\Big(\|u_f^{(T_2)}\|_F^2+\|u_f^\star\|_F^2\Big)}{T\eta}\\
&\le \frac{O\Big(m\|u_f^\star\|_2^2\Big)}{T\eta}\\
&\le O\Big(\tfrac{\alpha m}{T\eta}\Big).
\end{aligned}
\tag{298}
$$

Since $T\eta \ge m$, this proves the claim.

□