

Adaptive Progressive Fine-Tuning of VLMs for Long-Tailed Multimodal Retrieval

Farid Alijani[✉], Elina Late[✉], Sanna Kumpulainen[✉]

Faculty of Information Technology and Communication Sciences, Tampere University
Tampere, Finland

{farid.aliyani, elina.late, sanna.kumpulainen}@tuni.fi

Abstract—Adapting large VLMs to specialized, long-tailed domains requires a careful balance between performance and the preservation of pretrained knowledge. Although full parameter fine-tuning is powerful, it is resource-intensive and can easily overfit on imbalanced data. We propose Adaptive Progressive Fine-Tuning (APFT), a strategy that automates this complex process. APFT employs a staged layer unfreezing process guided by an event-triggered mechanism; instead of relying on a fixed schedule, phase transitions are automatically initiated based on real-time training stability metrics like loss volatility and performance plateaus. Upon transition, a cosine annealing scheduler is re-initialized, and weight decay is adaptively increased to regularize the newly trainable parameters. Experiments on the long-tailed HISTORY-X4 archival dataset indicate that APFT significantly outperforms all baselines, including full fine-tuning and LoRA. The advantage is most pronounced on tailed labels, where our APFT method achieves a 19.9% relative improvement in text-to-image $mAP@10$ over the strongest baseline, demonstrating its ability to effectively adapt to new domains while preserving foundational knowledge.

Index Terms—Vision-Language Models, Long-Tailed Recognition, Adaptive Fine-Tuning, Progressive Unfreezing, Multimodal Retrieval, Computer Vision.

I. INTRODUCTION

Large Vision-Language Models (VLMs) such as CLIP [1] and ALIGN [2] have revolutionized vision-language pretraining since they benefit from contrastive learning on massive paired web data which can further simplify learning rich and generalizable representations from web-scale data represented in digital archives. These developments are especially needed for historical image collections where traditional text-based retrieval systems struggle because of rich yet challenging contents with varying quality, evolving terminology, and incomplete metadata. [3]–[5]. VLMs remarkable zero-shot capabilities often fall short on downstream task datasets and when applied to specialized domains that exhibit a significant departure from the pretraining distribution. Achieving state-of-the-art performance in such domains, particularly those involving historical photographs with unique visual challenges and long-tailed conceptual distributions, necessitates fine-tuning [6].

However, the choice of a fine-tuning strategy presents a critical trade-off [7]. The standard approach, full fine-tuning, updates all model parameters but is computationally expensive and risks “catastrophic forgetting,” where the model’s invaluable general-purpose knowledge is overwritten by domain-

specific biases. This is especially problematic on imbalanced datasets, where the model can easily overfit to high-frequency *head* labels. On the other end of the spectrum, parameter-efficient fine-tuning (PEFT) methods such as LoRA [8] or Linear Probing [9] offer a lightweight alternative by updating only a small subset of parameters, but their constrained capacity can limit the final performance.

Progressive unfreezing [10] offers a compelling middle ground, beginning by training only the final layers and gradually unfreezing deeper layers over time. This allows the model to first adapt its task-specific head before cautiously tuning its foundational representations. However, existing implementations typically rely on a rigid, pre-defined schedule of layer unfreezing and hyperparameter changes. This approach is suboptimal, as the ideal moment to introduce new trainable parameters depends heavily on the dynamic learning state. A fixed schedule is a blind guess that can either unfreeze layers too early, causing instability, or too late, wasting computation on a plateaued model.

To address this gap, we introduce Adaptive Progressive Fine-Tuning (APFT), a novel framework that intelligently automates the fine-tuning process. At its core, APFT replaces fixed schedules with an event-triggered mechanism that monitors training stability by analyzing the validation loss trend, volatility, and rate of improvement. A phase transition, which unfreezes the next group of layers, is initiated only when the model shows clear signs of having converged or stagnated with its current set of trainable parameters. Crucially, each phase transition is coupled with a dynamic adaptation of hyperparameters. To better explore the updated optimization environment, the learning rate scheduler is reinitialized, and weight decay is raised to mitigate overfitting as the capacity of the model for learning expands.

We demonstrate the efficacy of APFT on HISTORY-X4, a large-scale and highly imbalanced archival photograph dataset where the long-tailed distribution makes standard fine-tuning particularly challenging. Our contributions are threefold: 1) we propose the APFT framework, which automates the layer unfreezing process based on training dynamics rather than a fixed schedule. 2) we introduce a novel multi-criteria decision mechanism for triggering these phase transitions and a corresponding method for dynamic hyperparameter adaptation. 3) we provide a comprehensive analysis showing that APFT significantly outperforms pretrained baselines and conven-

tional fine-tuning strategies, particularly in improving retrieval performance for rare, tail-end labels.

II. RELATED WORK

The challenge of adapting large pretrained models to downstream tasks has spurred extensive research, particularly for Vision-Language Models (VLMs) such as [1]. The most direct approach, full fine-tuning, updates all model parameters but is computationally demanding and can lead to catastrophic forgetting, where the model loses its powerful general-purpose representations. To mitigate this, a suite of the PEFT methods has emerged. These methods freeze the original model weights and introduce a small number of new, trainable parameters. Prominent PEFT strategies include Adapters [11]–[14], which insert lightweight modules between transformer layers, and Low-Rank Adaptation (LoRA) [8], [15]–[19], which adapts frozen weight matrices by training a low-rank decomposition of their update. While these methods significantly reduce the training cost, their constrained capacity can sometimes create a performance ceiling below that of a fully adapted model, motivating the need for strategies that offer a more flexible trade-off.

An alternative paradigm that bridges the gap between full and parameter-efficient tuning is progressive unfreezing. This technique was popularized in Natural Language Processing by ULMFiT [10], which proposed gradually unfreezing model layers from the output inwards while using discriminative learning rates for different layers. This approach allows the model to first adapt its task-specific head before cautiously tuning its more general, foundational feature extractors, thereby preserving pretrained knowledge. The core idea of layer-wise adaptation has historical precedent [20] and has proven effective. However, existing implementations of progressive unfreezing typically rely on a rigid, pre-defined schedule of layer transitions and hyperparameter changes. Such fixed schedules are suboptimal, as they fail to account for the unique learning dynamics of a given model-dataset combination, potentially unfreezing layers too early or too late. Our work directly addresses this limitation by introducing a mechanism to automate these transitions based on real-time training stability.

Our research is also situated within the context of long-tailed recognition, a critical challenge for real-world datasets like historical archives. A significant body of work has focused on addressing label imbalance through data-level or loss-level interventions. Data-based methods often involve re-sampling strategies to either over-sample tail labels or under-sample *head* labels. Loss-based methods, which are more common, aim to re-weight the loss function to give more importance to tail labels, such as through label-balanced losses [21] or label-distribution-aware margin adjustments [22]. Another influential paradigm involves decoupling the training process into separate stages for representation learning and classifier training [23]. These effective methods primarily address the classification objective. In contrast, our APFT method tackles the long-tail problem from an architectural and optimization

perspective, controlling the model’s capacity and learning dynamics to ensure that representations for both *head* and *tail* labels are learned effectively within a unified, end-to-end framework.

III. ADAPTIVE PROGRESSIVE FINE-TUNING

To address the challenges of fine-tuning the VLMs on long-tailed domain-specific datasets, we propose Adaptive Progressive Fine-Tuning (APFT). This strategy avoids the pitfalls of full fine-tuning by gradually introducing domain-specific knowledge while preserving the robust, general-purpose features learned during pre-training. APFT contains three core components: (1) a structured, hierarchical layer unfreezing schedule, (2) an automated mechanism for detecting training plateaus to trigger phase transitions, and (3) a dynamic hyperparameter adaptation scheme that adjusts the learning rate and weight decay in response to these transitions.

A. Framework Overview

The APFT training process is structured into a series of phases. Compared with conventional progressive unfreezing methods which rely on a fixed, pre-defined number of epochs per phase, APFT dynamically determines the phase transition. As outlined in Alg. 1, the model trains within a given phase until its performance, monitored on a validation set, shows signs of stagnation or instability. Upon detection of such a condition, a phase transition is triggered, which involves unfreezing a subsequent group of layers and adapting the hyperparameters of the optimizer (learning rate and weight decay) to the new, more complex training state. This cycle continues until all specified layers are unfrozen or a global early stopping criterion is met.

B. Structured Layer Grouping and Unfreezing

The foundation of our APFT approach is a structured partitioning of the CLIP model into functional groups, enabling a hierarchical unfreezing process. We categorize the model parameters into five distinct groups: *visual frontend*, *visual transformer*, *text frontend*, *text transformer*, and *projections*. Our unfreezing strategy proceeds from the output layers inwards. The *projections* group, which maps features to the shared embedding space, is made trainable from the first phase. In subsequent phases, we progressively unfreeze layers of the *text transformer* and *visual transformer* blocks, starting from the final block and moving towards the input. The *frontend* layers are typically unfrozen last.

C. Automated Phase Transition Detection

A key novelty of APFT is its ability to automatically decide when to transition between phases. Instead of relying on a fixed schedule, we monitor the model stability and progress on the validation set over a sliding window of the last w epochs which triggers a transition if any of the following conditions are met:

Algorithm 1 Adaptive Progressive Fine-Tuning (APFT)

Require: $M, D_{train}, D_{val}, \eta_0, \lambda_0, S, \varepsilon_{max}$

- 1: $phase \leftarrow 0, epochs_in_phase \leftarrow 0$
- 2: Initialize optimizer O & scheduler S with η_0, λ_0
- 3: Initialize EarlyStopper ε
- 4: **for** $epoch = 1, \dots, \varepsilon_{max}$ **do**
- 5: **if** $phase$ changed or $epoch = 1$ **then**
- 6: Unfreeze layers for $phase$ according to S
- 7: Update O with trainable parameters of M
- 8: Re-initialize S with current LR and WD
- 9: **end if**
- 10: Train M on D_{train} for one epoch
- 11: Evaluate \mathcal{L}_{val} on D_{val}
- 12: $epochs_in_phase \leftarrow epochs_in_phase + 1$
- 13: **if** should stop **then**
- 14: **break**
- 15: **end if**
- 16: **if** phase transition condition **then**
- 17: $phase \leftarrow$ new phase
- 18: $\eta, \lambda \leftarrow$ new values
- 19: $epochs_in_phase \leftarrow 0$
- 20: $\varepsilon.reset()$
- 21: **end if**
- 22: **end for**
- 23: Restore best weights from E

1) *High Loss Volatility*: makes the training process unstable. We measure this using the coefficient of variation (CV) of the \mathcal{L}_{val} within the window. A transition is triggered if: $\frac{\sigma(\mathcal{L}_{val_window})}{\mu(\mathcal{L}_{val_window})} > \tau_{vol}$, where σ and μ are the standard deviation and mean, respectively, and τ_{vol} is a volatility threshold.

2) *Worsening Loss Trend*: model is no longer improving. We compute the slope m of the best-fit line for the validation losses in the window. A transition is triggered if the slope indicates a worsening trend: $m > \tau_{slope}$.

3) *Stagnated Improvement*: learning progress has diminished significantly. We calculate the average pairwise improvement Δ_{pair} between consecutive epochs in the window. A transition is triggered if this value falls below a threshold τ_{imp} , provided the model is not already near its best-observed performance: $\Delta_{pair} < \tau_{imp}$ and $|\mathcal{L}_{current} - \mathcal{L}_{best}| > \delta_{min}$.

D. Dynamic Hyperparameter Adaptation

When a phase transition is triggered, APFT dynamically adapts the learning rate (η) and weight decay (λ) to suit the new architectural state.

The new learning rate, η_{new} , for phase $p + 1$ is calculated as a product of the initial learning rate η_0 and several adaptive factors: $\eta_{new} = \eta_0 \cdot f_{phase} \cdot f_{stability} \cdot f_{window}$, where f_{phase} is an exponential decay factor based on the phase progress, ensuring a general downward trend in LR. $f_{stability}$, calculated as the ratio of the current validation loss to the best-recorded loss, modulates the LR aggressively if performance is poor

and conservatively if it is near optimal. f_{window} is a minor adjustment factor based on the analysis window size.

Simultaneously, the weight decay is increased to apply stronger regularization to the larger set of trainable parameters. The new weight decay, λ_{new} , increases proportionally to the phase progress, preventing overfitting as the model gains more freedom. Following this adaptation, the optimizer state is reset, and the learning rate scheduler is re-initialized to begin a fresh cycle for the new phase. This holistic adaptation prevents training divergence and allows for stable learning throughout the progressive unfreezing process.

IV. EXPERIMENTS

A. Dataset and Preprocessing

All experiments are performed on HISTORY-X4, a large-scale, single-label dataset containing wartime photographs. Following [6], the dataset is curated from four public digital archives including the U.S. National Archives Catalog [24], Europeana [25], World War Photos [26], and SMU Libraries Digital Collections [27]. The dataset is partitioned into a training set of 133,172 images and a validation set of 71,709 images. As illustrated in Fig.1, the 67 distinct labels exhibit a severe long-tailed distribution, a characteristic challenge of real-world archival data. The most frequent label, 'aircraft', appears over 31,000 times, whereas tail-end labels such as 'treaty of versailles' and 'battle bulge' have fewer than 50 samples each. We categorize the labels into a *head* of 13 highly frequent labels, a *torso* of 23 mid-frequency labels, and a long *tail* of 31 rare labels. This extreme imbalance makes the dataset an ideal testbed for evaluating a model's ability to learn from both frequent and rare examples. For preprocessing, all images are resized to 336x336 pixels, center-cropped, and normalized using the mean and standard deviation computed from the training set.

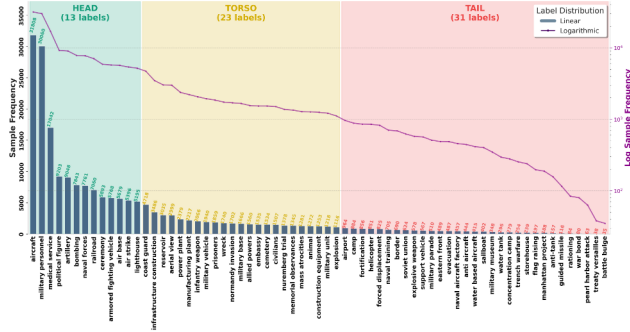
B. Baselines and Compared Methods

We evaluate the performance of APFT against three strong and widely-used baselines to provide a comprehensive comparison.

1) *Zero-Shot Pretrained CLIP*: We utilize the pretrained $ViT - L/14@336px$ model as a zero-shot baseline. This method performs retrieval without any fine-tuning on the target dataset, serving as a reference point to measure the domain gap and the performance gains achieved through adaptation.

2) *Full Fine-Tuning (Full-FT)*: This is a standard transfer learning where all parameters of the pretrained CLIP model are unfrozen and trained end-to-end on the HISTORY-X4 dataset. It represents the upper bound of model adaptation in terms of trainable parameters but is prone to overfitting and catastrophic forgetting.

3) *LoRA Fine-Tuning (LoRA-FT)*: As a state-of-the-art PEFT method, we include LoRA. This baseline injects low-rank adaptation matrices into the transformer layers of the CLIP model, training only these new parameters while keeping the original model weights frozen. It provides a computationally efficient alternative to full fine-tuning.



(a) Label frequency distribution of the HISTORY-X4 dataset, sorted by prevalence, highlighting the severe long-tail imbalance.



(b) Qualitative samples from HISTORY-X4, for given distributions of high-frequency *head*, mid-frequency *torso*, and low-frequency *tail* labels, illustrating the visual and conceptual diversity of the dataset.

Fig. 1: HISTORY-X4 characteristics.

C. Implementation Details

In our comparison, all fine-tuning strategies are built upon the same $ViT-L-14@336px$ pretrained CLIP model. All training runs shared an initial learning rate of $1e-5$, a weight decay of $1e-2$, a batch size of 32, and used the *AdamW* optimizer. A key distinction lies in the learning rate scheduling: for Full-FT and LoRA-FT, we employed a *OneCycleLR* scheduler, a standard and powerful choice for training with a fixed number of epochs. This choice is fundamental to our adaptive approach, as *OneCycleLR* is incompatible with a dynamic training duration; its schedule is pre-calculated for a fixed number of total steps. The ability of *CosineAnnealingWarmRestarts* to be re-initialized

at each phase transition is essential for APFT, allowing the learning rate to adapt to the changing model architecture.

The LoRA-FT baseline was configured with a rank of 64 and an alpha of 128. Our APFT method’s adaptive behavior was governed by the following early stopping and phase transition parameters: a stability analysis window of 10 epochs, a patience of 5 epochs, and a minimum of 10 epochs before any early stopping could occur. The thresholds for triggering transitions were set to a volatility of 15%, a slope of $1e-4$, and a pairwise improvement of $1e-4$. All experiments were conducted on a single NVIDIA Tesla V100 GPU with 32GB of VRAM, utilizing automatic mixed precision training via *GradScaler* for efficiency.

D. Evaluation Protocol

We evaluate all methods on two core zero-shot retrieval tasks: image-to-text (I2T), where a query image is used to retrieve the most relevant text label from the set of 67 unique label names, and text-to-image (T2I), where a text query (a label name) is used to retrieve relevant images from the entire validation set. Performance is quantified using three standard retrieval metrics: mean Precision ($mP@K$), mean Average Precision ($mAP@K$), and *Recall@K*. Each metric is calculated at different cutoff values $K \in 1, \dots, 20$ to provide a thorough assessment of ranking quality at various retrieval depths along with a comprehensive analysis of strengths and weaknesses for each model.

V. RESULTS

A. Quantitative Evaluation

The primary retrieval performance of our proposed APFT method against all baselines is summarized in Table I. The results clearly demonstrate that APFT consistently and significantly outperforms the Zero-Shot, Full Full-FT, and LoRA-FT approaches across both I2T and T2I retrieval tasks.

Notably, in the I2T task, APFT achieves a $mAP@10$ of 75.9%, a substantial improvement of 15.7 percentage points over the strongest baseline, LoRA-FT (60.2%). The advantage is even more pronounced in the more challenging T2I retrieval task, where APFT achieves a $mAP@10$ of 83.8%, surpassing LoRA-FT by 13.9 percentage points. The performance of the Pretrained model (and the identical Linear-Probe) highlights the significant domain gap, particularly in T2I recall (1.6%). While Full-FT and LoRA-FT provide considerable gains, they are ultimately surpassed by the architectural and optimization flexibility of APFT’s dynamic adaptation. The low absolute T2I recall values for all methods underscore the difficulty of retrieving specific images from a large, diverse validation set, yet APFT still nearly doubles the performance of the next-best method.

To provide a more granular view of performance, Fig. 2 shows the mAP and *Recall* retrieval metrics as a function of the retrieval depth, K . It illustrates that the advantage is not confined to $K = 10$ but is maintained across all retrieval depths, $K \in 1, \dots, 20$. This consistent superiority suggests that our method not only improves the ranking of the single

TABLE I: Retrieval metrics for performance comparison on HISTORY-X4 validation set.

Method	I2T		T2I	
	mAP@10	Recall@10	mAP@10	Recall@10
Pretrained	0.278	0.629	0.462	0.016
Full-FT	0.512	0.800	0.573	0.021
LoRA-FT	0.602	0.878	0.699	0.028
Linear-Probe	0.278	0.629	0.462	0.016
APFT (Ours)	0.759	0.899	0.838	0.052

best match but also enhances the quality of the entire ranked list of retrieved items.

B. Analysis of APFT Dynamics

To get a deeper understanding of the APFT mechanisms, we visualize its training dynamics in Fig. 3 which provides a holistic view of how the automated phase transitions and dynamic hyperparameter adjustments guide the learning process.

The validation loss curve (Fig. 3a) serves as the primary driver for the adaptive process. The initial learning in Phase 0 shows a rapid decrease in loss, which then begins to plateau around epoch 9. Our automated detection mechanism correctly identifies this stagnation (improvement of only -2.23% at the transition) and triggers the first phase transition. This intervention introduces new trainable parameters, causing a temporary and expected spike in validation loss as the model adapts, followed by a new phase of learning. This cycle of stabilization and adaptive intervention repeats at epoch 20 and epoch 28. The entire process is terminated by the early stopping criterion at epoch 34, demonstrating that the system autonomously determines both the phase structure and the overall training duration.

Each transition is coupled with a precise set of interventions. The unfreezing heatmap (Fig. 3d) reveals the architectural changes: while the Projections group is trainable from the start, tranches of the Visual and Text Transformer blocks are progressively unfrozen in phases 1, 2, and 3. The plot also highlights the method’s efficiency, showing that a near-optimal state was reached after using only four of the eight planned phases. Concurrently, the hyperparameter plot (Fig. 3b) illustrates the dynamic adaptation: at each transition, the learning rate is reset to a new, lower maximum for a fresh annealing cycle, while the weight decay is incrementally increased to regularize the growing set of trainable parameters.

Finally, the phase efficiency analysis (Fig. 3c) quantifies the impact of this strategy. The percentage of trainable parameters increases from a mere 0.43% in Phase 0 to 34.28% by Phase 3. Crucially, we observe a trend of diminishing returns in learning efficiency (% improvement per epoch). Phase 0 is the most efficient (0.510 %/epoch), rapidly adapting the model’s output space. Subsequent phases, which fine-tune deeper features, yield smaller but still vital improvements. This confirms that APFT correctly identifies points of diminishing returns and ex-

plores new model capacity precisely when needed, effectively balancing rapid adaptation with stable fine-tuning.

C. Ablation Study and Sensitivity Analysis

To better understand the contributions of APFT’s components, we discuss key ablations and parameter sensitivities. The thresholds for phase transitions (τ_{vol} , τ_{slope} , τ_{imp}) were determined empirically based on preliminary experiments on a small hold-out validation set to find a balance between responsiveness and stability.

Sensitivity to Thresholds: We observed that the system is reasonably robust to minor variations in these thresholds. For example, decreasing the volatility threshold (τ_{vol}) from 15% to 10% caused phase transitions to trigger more frequently, leading to minor instability and slightly worse final performance. Conversely, increasing it to 20% delayed transitions, slowing convergence in a manner similar to a fixed-schedule approach. This suggests our chosen values represent a stable operating point for this dataset.

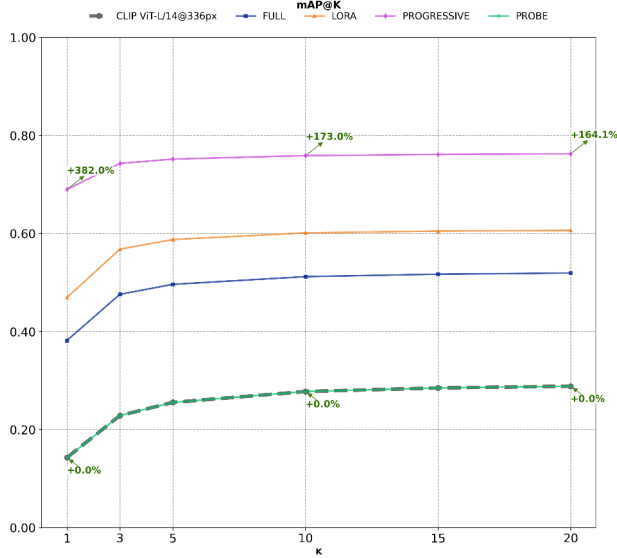
Component Importance: The core components of APFT are synergistic. Disabling the dynamic hyperparameter adaptation (i.e., keeping a constant LR) while retaining automated unfreezing led to training divergence after the second phase transition, as the optimizer could not adapt to the increased model capacity. Similarly, using the dynamic scheduler resets but with a fixed, pre-scheduled unfreezing interval resulted in suboptimal performance, as the transitions were not aligned with the model’s actual learning plateaus. This indicates that both the event-triggered unfreezing and the co-adaptation of hyperparameters are critical to the success of APFT.

D. Long-Tailed Distribution Performance

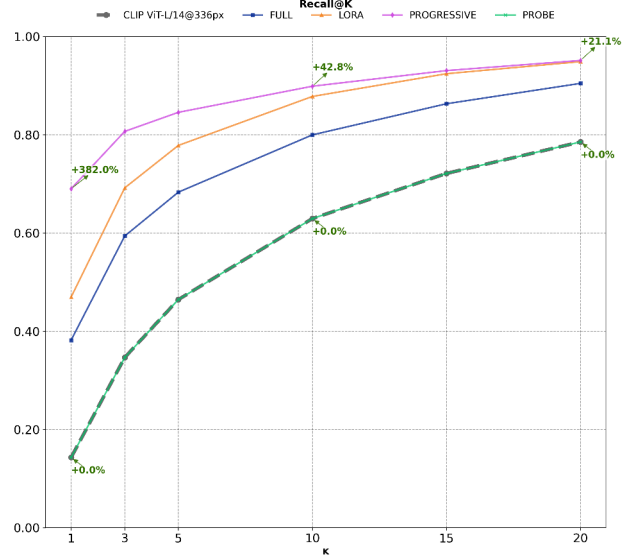
We evaluate the strength of APFT to handle the severe label imbalance of the HISTORY-X4 dataset by analyzing its performance on the tail-end labels. Figure 4 provides a striking qualitative example for T2I retrieval task with the rare query *pearl harbor attack*.

The performance of the Pretrained model (Fig. 4a) is poor; its top-ranked result is incorrectly labeled *Naval forces*, indicating a failure to grasp the specific event. Full-FT model (Fig. 4b) indicates marginal improvement but still ranks a generic *Bombing* image highly. LoRA-FT (Fig. 4c) performs strongly, correctly retrieving relevant images of the attack. However, our APFT method (Fig. 4d) demonstrates the most complex understanding. It not only retrieves correct images of the attack but also assigns a very high score to an incorrect but contextually relevant image of a destroyed *Water based aircraft* from the event, showcasing a deeper semantic grasp of the query.

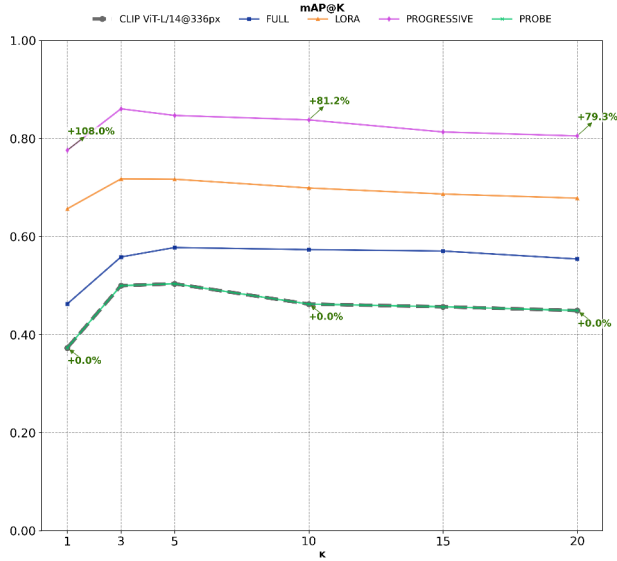
The quantitative results in Figure 5 highlight the limitations of baseline methods on the long tail. While most fine-tuning approaches achieve strong performance on frequent *head* and *torso* labels, they struggle significantly with rare *tail* labels. Specifically, the Pretrained, Full-FT, and Probe models cluster at 64.5% *Recall@10*, while the competitive LoRA-FT method reaches only 67.7%.



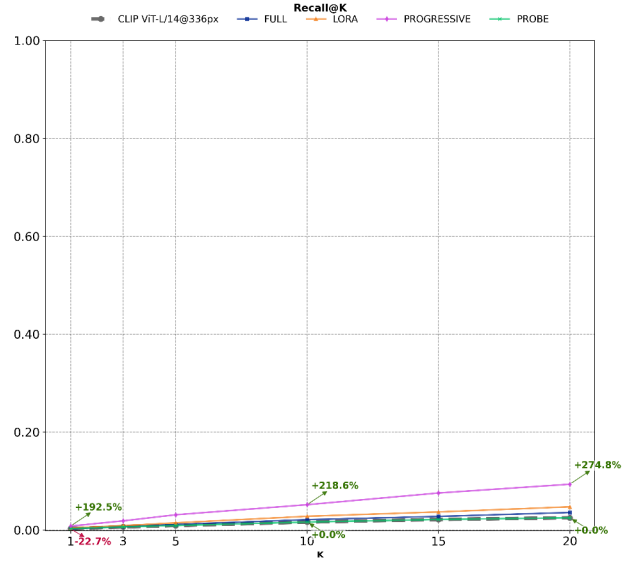
(a) mAP@K for I2T Retrieval



(b) Recall@K for I2T Retrieval



(c) mAP@K for T2I Retrieval



(d) Recall@K for T2I Retrieval

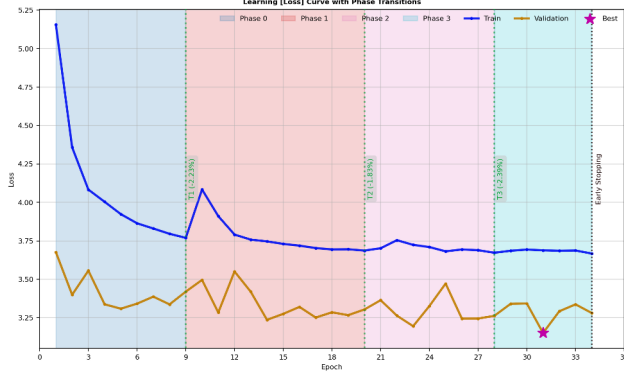
Fig. 2: Retrieval performance curves across various methods and metrics. The plots compare (a-b) I2T and (c-d) T2I performance for both mAP and Recall as a function of K. Our APFT method consistently outperforms all baselines across all tasks and retrieval depths.

In contrast, our APFT method achieves a $Recall@10$ of 96.8% on the tail labels. This represents a dramatic improvement of 29.1 percentage points over the next best method, LoRA-FT, and a performance lift of over 32 percentage points compared to Full-FT. This confirms that our adaptive strategy, by carefully managing model capacity and optimization, effectively mitigates the model’s inherent bias towards high-frequency data. It is uniquely capable of learning robust

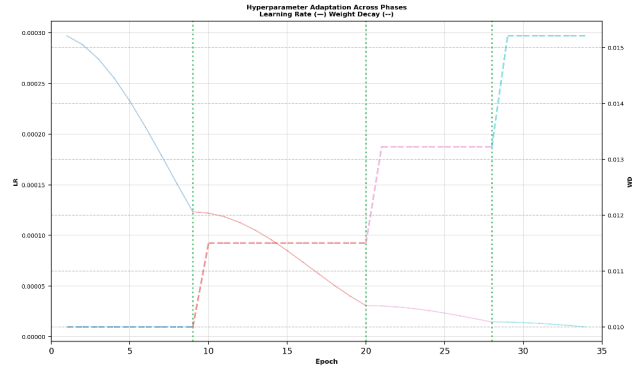
representations for even the rarest and most specific labels in the dataset, directly addressing the core challenge of long-tailed recognition.

VI. CONCLUSION

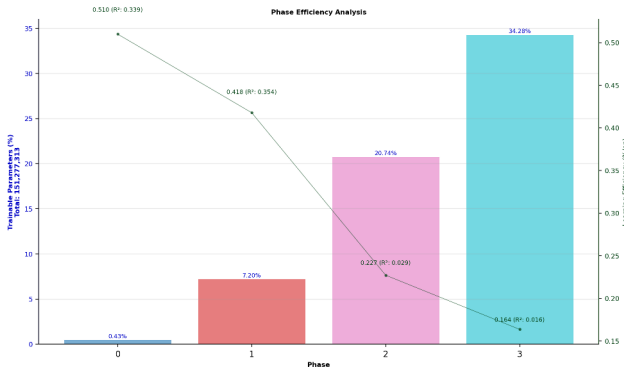
We introduced APFT, a novel framework that enhances traditional progressive unfreezing with automated phase transitions and dynamic hyperparameter scheduling. By allowing the model’s own learning state to dictate the fine-tuning process,



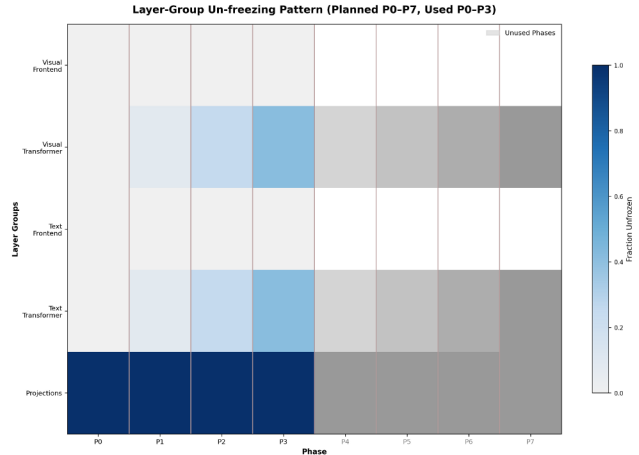
(a) Learning curve with automated phase transitions.



(b) Dynamic LR and Weight Decay adaptation.



(c) Phase efficiency and trainable parameter growth.



(d) Layer-group unfreezing pattern per phase.

Fig. 3: Training dynamics of our APFT method. The plots illustrate (a) how validation loss plateaus trigger phase transitions (dashed green lines), (b) the corresponding dynamic adaptation of the learning rate (LR) and weight decay (WD), (c) the resulting growth in trainable parameters and decreasing learning efficiency per phase, and (d) the specific layer groups being unfrozen at each phase.

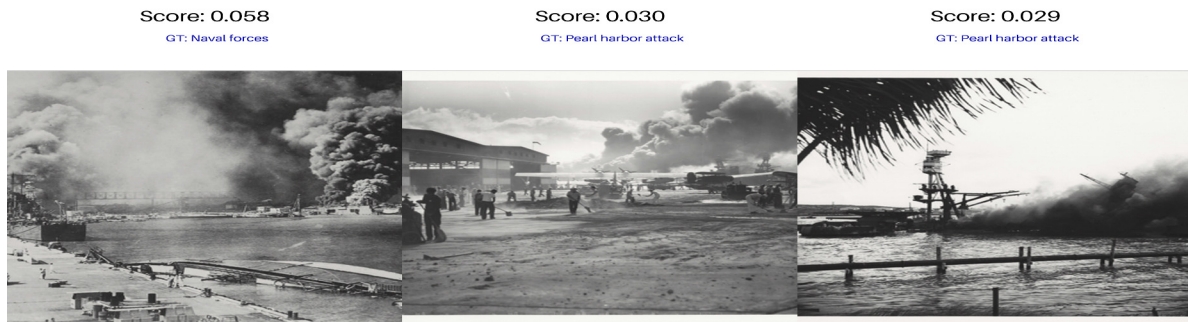
APFT avoids the rigidity of fixed schedules and provides a more stable and efficient path to high performance.

Our detailed experiments on the long-tailed HISTORY-X4 dataset indicate that this adaptive approach is highly effective. APFT significantly outperforms standard baselines, including full fine-tuning and LoRA, in both I2T and T2I retrieval. Crucially, its greatest advantage is in recognizing rare concepts, where it achieves a *Recall@10* of 96.8% on tail classes—a performance lift of over 29 percentage points compared to the next-best method. This confirms that APFT effectively mitigates the model bias towards high-frequency data, providing a robust and resource-conscious solution for adapting large pretrained vision-language models to specialized domains.

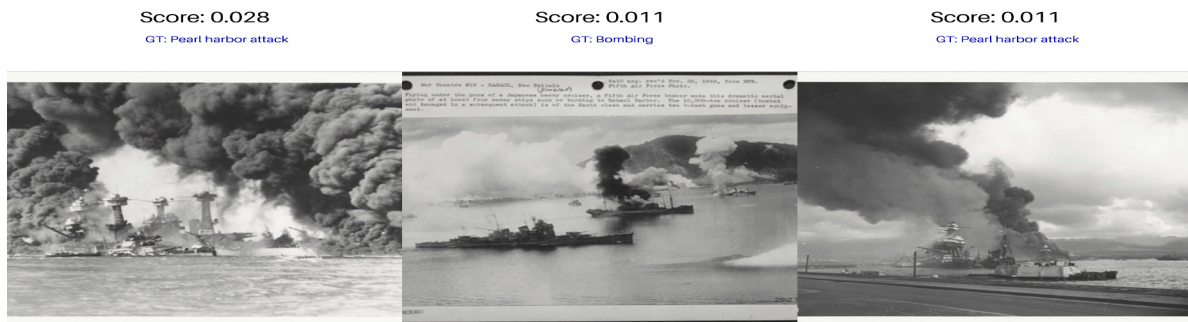
We acknowledge certain limitations which point towards potential areas for subsequent work. This study was conducted on a single, albeit large and challenging, archival dataset. While APFT demonstrates clear advantages in this domain, future research should validate its generalizability on stan-

dard long-tailed benchmarks (e.g., ImageNet-LT) and across different VLM architectures. Furthermore, our evaluation was limited to text queries corresponding to class names; exploring performance with more descriptive, free-form text queries would better reflect real-world retrieval scenarios.

One limitation of this study is the single-label nature of our dataset, which can under-represent the full semantic content of complex historical images. Future work should therefore explore multi-label settings to better capture this complexity. Additionally, addressing the inherent label noise common in archival collections would be crucial. Finally, user studies are needed to evaluate the real-world utility of such systems in digital humanities, where hybrid systems combining APFT’s automated retrieval with original scholarly annotations [3] hold significant potential for enhancing discovery in cultural heritage collections.



(a) Zero-Shot CLIP [1] with pretrained ViT - L - 14@336px



(b) Full-FT



(c) LoRA-FT [8]



(d) APFT (Ours)

Fig. 4: Qualitative T2I retrieval results for the rare (*tail*) query *pearl harbor attack*. APFT (d) successfully retrieves specific and contextually relevant images of the event, demonstrating a more complex understanding than Zero-Shot CLIP (a), Full-FT (b), and LoRA-FT (c) baseline methods.

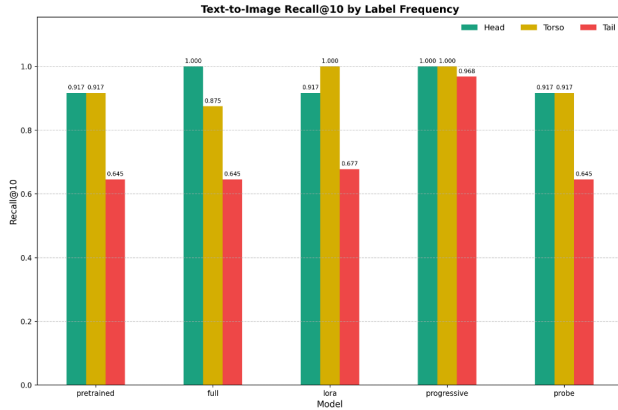


Fig. 5: T2I *Recall@10* performance disaggregated by label frequency (*head*, *torso* and *tail*). APFT outperforms all the other fine-tuning methods on the challenging *tail* labels, demonstrating its effectiveness at mitigating data imbalance.

ACKNOWLEDGMENT

This study was sponsored by the Research Council of Finland with grant number 351247. The authors wish to also acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

REFERENCES

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.
- [2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021.
- [3] E. Late, H. Ruotsalainen, and S. Kumpulainen, "Image searching in an open photograph archive: search tactics and faced barriers in historical research," *Int. J. Digit. Libr.*, vol. 25, pp. 715–728, 2024.
- [4] Elina Late, Hille Ruotsalainen, and Sanna Kumpulainen, "In a perfect world: Exploring the desires and realities for digitized historical image archives," *Proceedings of the Association for Information Science and Technology*, vol. 60, no. 1, pp. 244–254, Oct. 2023.
- [5] E. Late, H. Ruotsalainen, M. Seker, J. Raitoharju, A. Männistö, and S. Kumpulainen, "From textual to visual image searching: User experience of advanced image search tool," in *Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPD L 2023, Zadar, Croatia, September 26–29, 2023, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 277–283.
- [6] F. Aljani, E. Late, and S. Kumpulainen, "Historyclip: Adaptive multimodal retrieval of imbalanced long-tailed archival data," in *Linking Theory and Practice of Digital Libraries*. Cham: Springer Nature Switzerland, 2026, pp. 245–262.
- [7] J. Xing, J. Liu, J. Wang, L. Sun, X. Chen, X. Gu, and Y. Wang, "A survey of efficient fine-tuning methods for vision-language models — prompt and adapter," *Computers & Graphics*, vol. 119, p. 103885, 2024.
- [8] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.
- [9] A. Tomihari and I. Sato, "Understanding linear probing then fine-tuning language models from ntk perspective," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2025.

- [10] J. Howard and S. Ruder, "Fine-tuned language models for text classification," *CoRR*, vol. abs/1801.06146, 2018.
- [11] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2790–2799.
- [12] Y. Bai, H. Zhao, Z. Lin, A. Kale, J. Gu, T. Yu, S. Kim, and Y. Fu, "Advancing vision-language models with adapter ensemble strategies," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 15 702–15 720.
- [13] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.
- [14] H. Lu, Y. Huo, G. Yang, Z. Lu, W. Zhan, M. Tomizuka, and M. Ding, "Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling," in *The Twelfth International Conference on Learning Representations*, 2024.
- [15] Z. Liu, S. Kundu, A. Li, J. Wan, L. Jiang, and P. Beerel, "AFLoRA: Adaptive freezing of low rank adaptation in parameter efficient fine-tuning of large models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 161–167.
- [16] Y. Ji, Y. Liu, Z. Zhang, Z. Zhang, Y. Zhao, X. Hao, G. Zhou, X. Zhang, and X. Zheng, "Enhancing adversarial robustness of vision-language models through low-rank adaptation," in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ser. ICMR '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 550–559.
- [17] R. Pan, X. Liu, S. Diao, R. Pi, J. Zhang, C. Han, and T. Zhang, "LISA: Layerwise importance sampling for memory-efficient large language model fine-tuning," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [18] R. Qiang, R. Zhang, and P. Xie, "BiloRA: A bi-level optimization framework for low-rank adapters," 2024.
- [19] S. Wang, L. Chen, J. Jiang, B. Xue, L. Kong, and C. Wu, "LoRA meets dropout under a unified framework," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1995–2008.
- [20] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1615–1625.
- [21] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2019, pp. 9260–9269.
- [22] H. E. C. Cao, R. Sarlin, and A. Jung, "Learning explainable decision rules via maximum satisfiability," *IEEE Access*, vol. 8, pp. 218 180–218 185, 2020.
- [23] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *International Conference on Learning Representations*, 2020.
- [24] U.S. National Archives and Records Administration, "National archives catalog," 2024. [Online]. Available: <https://catalog.archives.gov/>
- [25] Europeana Foundation, "Europeana," 2024. [Online]. Available: <https://www.europeana.eu/en>
- [26] World War Photos, "World war photos - over 18,000 original photos from the second world war," 2024. [Online]. Available: <https://www.worldwarphotos.info/>
- [27] Southern Methodist University Libraries, "Degolyer library, smu," 2024. [Online]. Available: <https://www.smu.edu/libraries/digitalcollections>