From Prototypes to General Distributions: An Efficient Curriculum for Masked Image Modeling

Jinhong Lin* Cheng-En Wu* Huanran Li Jifan Zhang Yu Hen Hu Pedro Morgado
University of Wisconsin–Madison
{jlin398, cwu356, hli488, jzhang769, yhhu, pmorgado}@wisc.edu

Abstract

Masked Image Modeling (MIM) has emerged as a powerful self-supervised learning paradigm for visual representation learning, enabling models to acquire rich visual representations by predicting masked portions of images from their visible regions. While this approach has shown promising results, we hypothesize that its effectiveness may be limited by optimization challenges during early training stages, where models are expected to learn complex image distributions from partial observations before developing basic visual processing capabilities. To address this limitation, we propose a prototype-driven curriculum learning framework that structures the learning process to progress from prototypical examples to more complex variations in the dataset. Our approach introduces a temperature-based annealing scheme that gradually expands the training distribution, enabling more stable and efficient learning trajectories. Through extensive experiments on ImageNet-1K, we demonstrate that our curriculum learning strategy significantly improves both training efficiency and representation quality while requiring substantially fewer training epochs compared to standard Masked Auto-Encoding. Our findings suggest that carefully controlling the order of training examples plays a crucial role in self-supervised visual learning, providing a practical solution to the early-stage optimization challenges in MIM.

1. Introduction

Masked Image Modeling (MIM), initially explored in [34, 44] and popularized in [16], has emerged as a powerful self-supervised learning paradigm for visual representation learning. By learning to predict masked portions of images from their visible regions, these models can acquire rich visual representations without requiring explicit labels. However, despite their success, current MIM approaches face a fundamental challenge that has received limited attention: the optimization difficulties during early training stages.

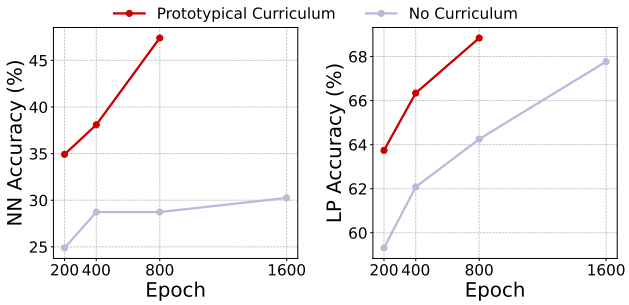

Figure 1. Our prototypical curriculum (red line) significantly improves MAE downstream performance on ImageNet-1k compared to the standard MAE training strategy (purple line). We show both the nearest neighbor (left) and linear probe (right) accuracy of models trained for different numbers of epochs.

The core challenge lies in the immediate expectation for models to learn a complex mapping from partial observations to the full distribution of natural images, even before developing basic visual processing capabilities. This is analogous to expecting an art student to reproduce complex masterpieces before mastering fundamental sketching techniques. This difficulty is compounded by the inherent complexity and high dimensionality of natural image distributions, making the initial learning phase particularly challenging and potentially inefficient.

In this paper, we propose a novel curriculum learning framework for MIM that addresses these optimization challenges while maintaining the core reconstruction objective. Our approach is motivated by the observation that human visual learning often progresses from simple to complex patterns [31, 38], and that certain prototypical examples can serve as effective "templates" for learning basic visual concepts. Rather than immediately exposing the model to the full complexity of natural image distributions, we structure the learning

*Equal Contribution

process to begin with a simplified distribution of prototypical examples that capture essential visual patterns for each semantic category. Our key insight is that by carefully controlling the expansion of the training distribution from prototypical examples to more complex variations, we can create a more efficient and stable learning trajectory, ultimately leading to more robust representation learning.

Through extensive experiments on the ImageNet-1K dataset, we demonstrate that incorporating a prototype-driven curriculum strategy significantly improves masked image modeling, both in terms of training efficiency and representation quality (Fig. 1). By gradually transitioning from prototypical to diverse examples during training, our approach achieves superior performance while requiring substantially fewer training epochs compared to standard Masked Auto-Encoding (MAE) [17]. Extensive ablation studies also reveal that the success of our method stems from careful prototype identification and a progressive learning schedule, suggesting that the order in which examples are presented plays a crucial role in self-supervised visual learning.

2. Related Work

Masked Image Modeling Masked Image Modeling has emerged from early self-supervised learning approaches in computer vision, gaining prominence after the success of BERT in natural language processing. While initial attempts at masking in vision were made in [34, 44] using convolutional architectures for denoising and inpainting tasks, modern MIM approaches took shape with the advent of Vision Transformers. Two influential concurrent works, MAE [17] and SimMIM [49], established the core framework by introducing high-ratio random masking and transformer-based architectures. This foundation sparked numerous innovations: some methods focused on improving the masking strategy through attention-guided [21] or semantically-driven approaches [24], while others explored alternative learning objectives by combining reconstruction with contrastive learning [7] or incorporating knowledge distillation [19]. MIM has since evolved beyond static images to handle various modalities including video [42, 46], point clouds [32, 50], and multi-modal data [2, 18, 29], demonstrating its versatility as a self-supervised learning paradigm. Our work builds on this foundation by proposing a novel data selection strategy that accelerates MIM training by emphasizing common data early in the learning process.

Dataset Selection Dataset selection refers to the process of selecting a subset of the data that best represents the entire dataset, enabling efficient training without sacrificing performance. Data selection techniques in supervised learning focus on identifying key subsets of training data that have the greatest impact on model performance. One research avenue, most related to our method, explores proxy models to locate informative subsets. For instance, studies such as [3, 9, 14, 40] employ k-means clustering or uncertainty estimates from proxy models to select data based on feature space distance. Other methods leverage sample difficulty and uncertainty. [41] found that frequently forgotten examples during training are more valuable for model refinement. Gradient-based selection techniques have also been used in several studies, including [22, 23, 28, 35, 36], to identify data points that closely approximate the overall gradient, facilitating efficient subset selection based on gradient insights. In general, these methods show that to retain performance in supervised learning, the selected subset must be representative of the visual diversity in the dataset, often undersampling easy to learn "prototypical" samples, which we show are beneficial in the early phases of MIM training to accelerate learning.

In self-supervised learning (SSL), dataset selection has been less explored, though recent advances are addressing this gap. Geometry-based methods, such as those proposed by [20], leverage augmentation similarities among data points within a dataset to enhance data efficiency in contrastive learning. However, this method suffers from scalability issues and relies on a strongly pre-trained model to extract fine-grained features. In contrast, our proposed approach eliminates this dependency, allowing for data selection based solely on low-level features. Dynamic selection methods, such as [43], adapt data selection dynamically during training by refining the selection process from coarse to fine-grained as the model evolves. While effective for contrastive learning, this approach necessitates frequent and substantial computational overhead. Our method, on the other hand, is tailored specifically for Masked Image Modeling (MIM) and avoids additional computational demands during training.

Importance Sampling Techniques Importance sampling has emerged as a promising technique for data selection in stochastic optimization. Early methods by [1, 22, 26, 37] prioritized samples with higher importance, often determined by large gradient norms or high loss values. These techniques aim to reduce variance in stochastic gradients, occasionally improving convergence rates. However, they generally lack rigorous convergence guarantees or substantial speedups when applied to overparameterized models, such as deep neural networks. More recent work by [27] refines these concepts by selecting non-noisy, non-redundant, and

task-relevant examples that minimize loss on a holdout validation set. While this approach accelerates training, it depends on a validation set and lacks theoretical convergence guarantees for non-convex models, limiting its utility in deep learning.

In gradient approximation, the CRAIG method [28] selects sample subsets that approximate the total gradient sum by maximizing a submodular function. While CRAIG guarantees convergence for strongly convex models, its effectiveness declines in non-convex settings like deep learning. GradMatch [23] and Glister [23] further explore gradient-based subset selection. GradMatch uses orthogonal matching pursuit (OMP) to select minimal training subsets, while Glister employs a validation set for gradient alignment. Both methods are effective in smaller scenarios but face scalability challenges with large datasets.

3. Method

We propose a curriculum learning strategy to enhance training of Masked Image Modeling methods, in particular, Masked Auto Encoders (MAE) [17]. An overview of our strategy is illustrated in Figure 2.

3.1. [Background] Masked Image Modeling

Masked Image Modeling (MIM) is a self-supervised learning framework that seeks to reconstruct images from partial observations. Given an input image $x \in \Re^{H\times W\times 3}$, MIM procedures first divide it into N non-overlapping patches, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$, each of size $\boldsymbol{x}_i \in \Re^{P\times P\times 3}$. Then, given a pre-specified mask ratio m (typically 75%), $\boldsymbol{X}$ is randomly partitioned into two disjoint sets of visible $\mathcal{V}$ and masked patches $\mathcal{M}$, where $|\mathcal{M}| = Nm$ and $|\mathcal{V}| = N(1-m)$. The visible patches are processed by an encoder model $\mathcal{E}(\cdot)$ to encode them into visible latent representations $\boldsymbol{Z}_\mathcal{V} = \mathcal{E}(\boldsymbol{X}_\mathcal{V}; \theta_\mathcal{E}) \in \Re^{N(1-m)\times d}$, where d denotes the feature dimensionality and $\theta_\mathcal{E}$ the encoder parameters. The visible representations $\boldsymbol{Z}_\mathcal{V}$ together with the positions of masked patches $p_\mathcal{M}$ are then used to reconstruct the full sequence of patches by a decoder model, $\boldsymbol{Y} = \mathcal{D}(\boldsymbol{Z}_\mathcal{V}, p_\mathcal{M}; \theta_\mathcal{D}) \in \Re^{N\times(P\times P\times 3)}$ parameterized by $\theta_\mathcal{D}$. Both encoder and decoder are trained to minimize the mean squared error between the predicted and actual pixel values of masked patches

$$\mathcal{L} = \mathbb{E}_\mathcal{X}\left[\|\boldsymbol{X}_\mathcal{M} - \boldsymbol{Y}_\mathcal{M}\|_2^2\right]. \tag{1}$$

After pre-training, only the encoder $\mathcal{E}$ is retained for downstream tasks, while the decoder $\mathcal{D}$ is discarded.

3.2. Improving MIM through Prototypical Curriculum Learning

Despite the success of MIM approaches in self-supervised visual representation learning, these models face significant optimization challenges during the early stages of training. This challenge stems from attempting to learn the complex mapping from partial observations to the full distribution of natural images, even before the model has developed basic visual processing capabilities. This is analogous to asking a student to reproduce complex artwork before mastering basic sketching techniques. To address this limitation, we propose a curriculum learning approach that, without changing the MIM reconstruction objective, aids the learning process by starting from a simplified distribution of training images, and gradually broadening the training distribution to include more complex examples.

Our key insight is that the training process can be made more efficient by initially focusing on prototypical examples that capture essential visual patterns for each semantic category (Fig. 2). These prototypical images serve as "templates" that are easier to reconstruct, allowing the model to first learn basic visual concepts before tackling more complex variations. As training progresses, we gradually expand the distribution to include examples that deviate further from these prototypes, eventually covering the full range of natural image variations. As we will show empirically, this structured learning trajectory leads to more stable optimization, and improved training efficiency through structured exploration of the data distribution.

3.2.1. Identifying Prototypical Examples

Prototypical images can be identified through clustering in feature space. Specifically, we cluster image representations $\boldsymbol{z}_i \in \Re^d$ using the mini-batch K-means clustering, and leverage the resulting K cluster centroids $\boldsymbol{\mu}_k \in \Re^d$ to locate likely prototypical images, *i.e.*, images whose representations $\boldsymbol{z}_i$ are close to each centroid $\boldsymbol{\mu}_k$.

Choosing the Number of Clusters To automate the choice of K, we vary it in the range $2 \leq K \leq K_{max}$, and use the Davies-Bouldin index $\mathcal{DB}$ to assess the clustering quality. $\mathcal{DB}$ measures both how compact each cluster is and how well-separated they are from other clusters [10].

Representation space The choice of representation space can significantly impact the selection of prototypical images. To study its impact on the curriculum, we investigate representations across different levels of semantic abstraction, from low-level visual descriptors such as traditional computer vision features (e.g., SIFT,

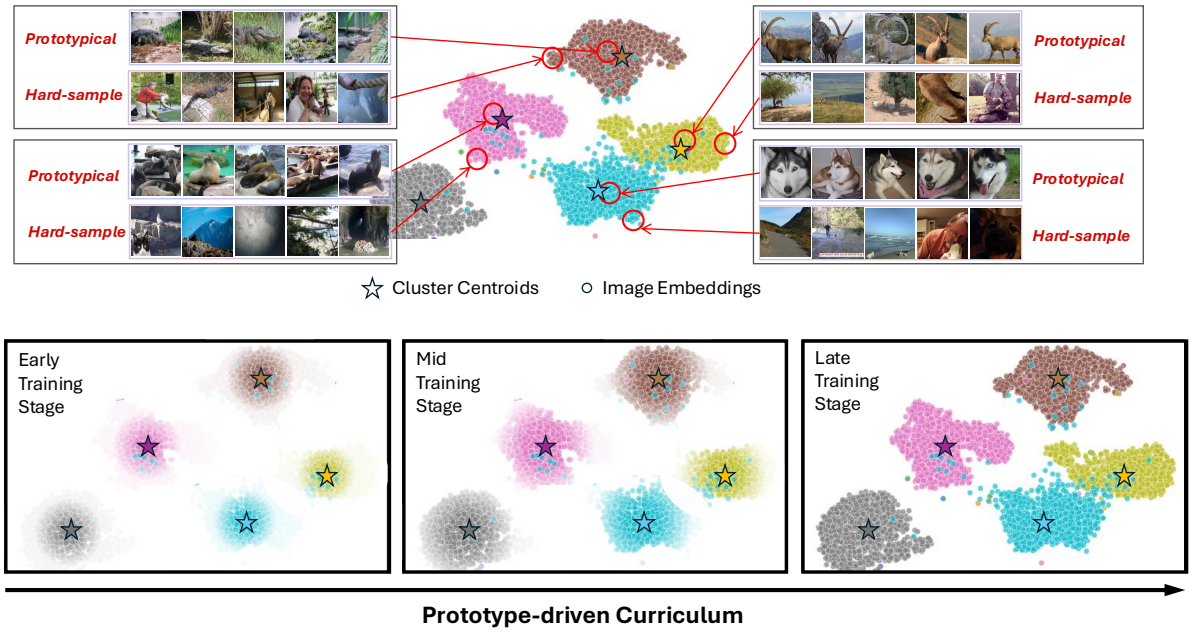

Figure 2. Framework Overview. We introduce a curriculum learning framework for Masked Autoencoders (MAE) that prioritizes sampling specific images early in training and progressively expands to cover the entire dataset. K-means clustering is applied within a visual feature space to identify clusters that exhibit prototypical visible semantics. The framework selects images near the cluster centers as prototypes, representing basic and common visual features, while more challenging examples are sampled from the cluster boundaries. Additional visualizations are provided in Fig.4.

HOG) to representations obtained from pretrained models like DINO [5] or MAE [17]. We find that while prototype-driven curricula provide benefits regardless of feature space, higher-level semantic features from self-supervised models like DINO are the most effective.

3.2.2. Prototypical Curriculum

The proposed curriculum controls the emphasis on prototypical examples through a temperature-based sampling mechanism. For each data point $\boldsymbol{z}_i$, we first quantify its "prototypicality" based on the distance to the nearest centroid ($\boldsymbol{\mu}_k$).

$$d(\boldsymbol{z}_i, \boldsymbol{\mu}_k) = \|\boldsymbol{z}_i - \boldsymbol{\mu}_k\|_2 \tag{2}$$

However, the divergence within each clusters may vary, with some clusters being tightly packed around their center (low divergence) and others more spread out (high divergence). To prevent sampling from being biased by variations in cluster divergences, we normalize the distances within each cluster using min-max normalization

$$\hat{d}_i = \frac{d(\boldsymbol{z}_i, \boldsymbol{\mu}_k) - d_{\min}}{d_{\max} - d_{\min}}, \tag{3}$$

where $d_{\max}$ and $d_{\min}$ are the maximum and minimum distances within cluster k. Training data is then sampled from $\mathcal{D}$ with replacement according to the following temperature-controlled softmax probabilities

$$P(x_i, \tau) = \frac{\exp(-\hat{d}_i/\tau)}{\sum_j \exp(-\hat{d}_j/\tau)}, \tag{4}$$

where τ is the temperature parameter that controls the sharpness of the distribution. In the early stages of training, τ is low to concentrate the sampling distribution around the cluster centers and prioritize prototypical examples. As training progresses, τ is gradually increased, allowing the model to encounter more diverse examples while building on its foundational understanding of prototypical patterns.

3.2.3. Annealing Temperature by Scheduling the Effective Data Size

While temperature-controlled sampling offers a flexible mechanism to adjust the curriculum over time, defining appropriate temperature values (τ) can be challenging due to their strong dependence on dataset characteristics. A more intuitive approach is to directly specify "effective dataset sizes", *i.e.*, what portion of the dataset the model should be exposed to at each training stage.

We define the effective dataset size $|\mathcal{D}_\tau|$ as the number of unique samples drawn from $\mathcal{D}$ over one epoch when sampling with replacement according to the dis-

tribution $P(\cdot, \tau)$, which can be expressed as

$$|\mathcal{D}_\tau| = \sum_{i=1}^{|\mathcal{D}|} \left[1 - (1 - P(x_i, \tau))^{|\mathcal{D}|}\right]. \quad (5)$$

The curriculum is implemented by gradually increasing $|\mathcal{D}_\tau|$ during training using a cosine schedule. This schedule spans from a pre-specified initial value to its theoretical limit

$$|\mathcal{D}| \times \lim_{\substack{\tau\to\infty \\ |\mathcal{D}|\to\infty}} \frac{|\mathcal{D}_\tau|}{|\mathcal{D}|} = |\mathcal{D}| \times (1 - 1/e). \quad (6)$$

For any desired $|\mathcal{D}_\tau|$, we can determine the corresponding temperature τ by solving Eq. 5. Although this equation lacks a closed-form solution, we can efficiently solve for τ using binary search, taking advantage of the monotonic relationship between $|\mathcal{D}_\tau|$ and τ. Fig. 3 illustrates this temperature annealing process.

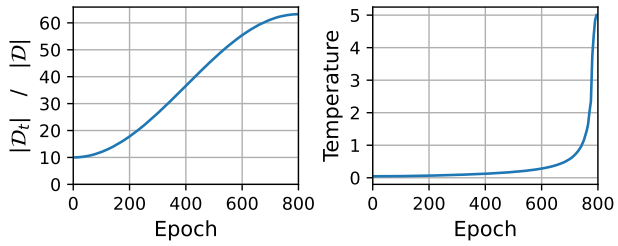

Figure 3. Temperature annealing schedule and its effect on training distribution. The temperature parameter τ gradually increases over training epochs, controlling the breadth of the sampled distribution. The effective dataset size, measured as $|\mathcal{D}_\tau|/|\mathcal{D}|$, shows the corresponding fraction of the training data that is sampled in a single epoch at each temperature level. As the temperature increases the effective dataset size approaches the theoretical limit of $1 - 1/e$.

4. Experiments

4.1. Experimental Setup

Pre-training All experiments are conducted on the ImageNet-1K [11] dataset, using a ViT-Base as the backbone architecture. Following the official MAE [17] implementation, images are randomly cropped and resized to 224×224 pixels, with random horizontal flipping for data augmentation. The mask ratio is set to 75%, and the temperature τ is annealed from 0.07 to 0.6 over the course of training. Models are optimized using the AdamW optimizer with a batch size of 4096, base learning rate of 1.5e-4, weight decay of 0.05. Linear learning rate warm-up is applied over the first 40 epochs, followed by a cosine learning rate decay schedule.

Curriculum	Epochs	NN	LP	FT
None (MAE) [17]	800	28.73	64.25	83.08
Hard-samples [40]	800	24.63	62.09	82.95
Prototypical (Ours)	800	47.40	68.84	83.31

Table 1. Comparison of MAE training strategies. Curriculum based on "Hard-samples" is adapted from [40]. MAE baseline uses uniform sampling of the training set. Results show NN, LP, and FT accuracy (%) on ImageNet-1K.

Downstream Evaluation We assess the quality of the learned representations through three complementary evaluation protocols. First, following standard practice in self-supervised learning [5, 6, 15, 16, 25, 47], we evaluate linear probing (LP) performance by training a supervised linear classifier on frozen features extracted from the `[CLS]` token. This provides a direct measure of the linear separability of the learned representations. Second, we perform full fine-tuning (FT) of the pre-trained model, allowing all parameters to be optimized end-to-end for the downstream classification task. Finally, we evaluate nearest neighbor (NN) classification accuracy using cosine similarity in the feature space, which provides a non-parametric assessment of the learned representations' ability to capture semantic similarities between samples without any additional training.

4.2. Main Results

We first evaluate our prototype-driven curriculum strategy against alternative approaches for MIM training (Table 1). We compare our method against two baselines. The first is a hard-sample curriculum adapted from [40], which prioritizes non-prototypical images distant from cluster centroids. As shown in [40], focusing on non-prototypical examples is beneficial for dataset distillation, enabling supervised training on smaller datasets while minimizing performance degradation. We also compare against the standard MAE training regime without any curriculum. All experiments use a ViT-B/16 architecture trained on ImageNet-1k.

The results demonstrate the clear advantages of a prototype-driven curriculum. Most notably, our approach achieves 47.40% nearest neighbor classification accuracy with 800 epochs of training, substantially outperforming the 800-epoch (30.25%) MAE baseline.

Interestingly, while [40] found that non-prototypical (hard) samples need to be preserved when reducing dataset size in supervised learning, our results reveal this strategy to be suboptimal for self-supervised MIM training. The significant performance gap between hard-sample curriculum (62.09% LP, 82.95% FT) and our

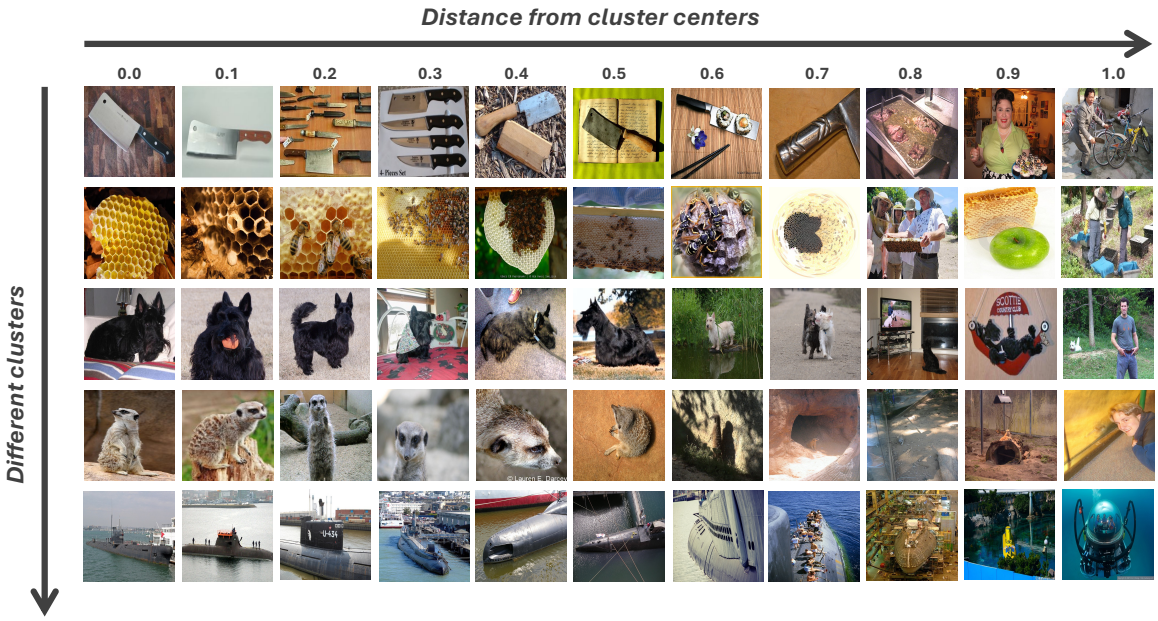

Figure 4. Visualization of cluster samples arranged by their distances from their respective cluster centers, ranging from nearest to farthest. Each row represents images from the same cluster, while each column indicates the distance of images from their cluster's centroid.

Curriculum	Epoch	NN	LP	FT
None (MAE [17])	200	24.88	59.31	82.46
	400	28.73	62.08	82.91
	800	30.25	64.25	83.08
Prototypical (Ours)	200	34.92	63.74	82.75
	400	38.09	66.34	83.09
	800	**47.40**	**68.84**	**83.31**

Table 2. Training efficiency comparison. Performance evaluation of prototype-driven curriculum against standard MAE across different training durations (200 to 800 epochs). The proposed method uses temperature annealing for curriculum learning, while the baseline follows standard uniform sampling. Results show NN, LP, and FT accuracy (%) on ImageNet-1K.

prototypical approach (68.84% LP, 83.31% FT) suggests that early exposure to prototypical examples plays a crucial role in developing robust visual representations. This aligns with our hypothesis that building from prototypical to complex examples creates a more stable optimization trajectory in the challenging MIM setting.

Training Efficiency To further investigate the efficiency of our approach, we conducted analyzed model performance across different training durations (Table 2). We compare our prototypical curriculum against the standard MAE training regime across epochs ranging from 200 to 800, using a ViT-B/16 architecture.

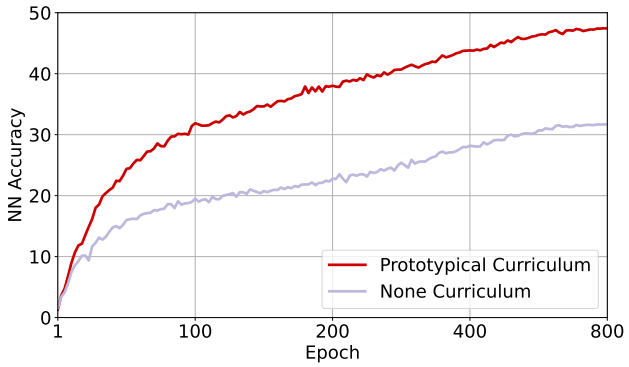

Figure 5. Training curve. Nearest-neighbor accuracy throughout training for an MAE model trained with and without a prototype-driven curriculum.

The results reveal that our prototypical curriculum consistently accelerates the learning process. At just 200 epochs, our method achieves 34.92% nearest neighbor accuracy and 63.74% linear probing performance, already surpassing the 800-epoch baseline (30.25% NN, 64.25% LP). This early advantage becomes more pronounced with additional training: at 400 epochs, our approach (38.09% NN, 66.34% LP) already exceeds the performance of the 800-epoch baseline (30.25% NN, 67.77% LP).

Overall, prototype-driven curriculum learning enables more efficient and effective representation learning in the MIM framework.

Training Dynamics We also analyzed the training dynamics using nearest-neighbor accuracy as a proxy

for representation quality. We train two MAE models with identical ViT-Base architectures on ImageNet-1k for 800 epochs: one with our proposed curriculum and one with the standard training procedure.

Our experiments demonstrate that incorporating a prototype-driven curriculum significantly improves the training dynamics of MAE models. As shown in Fig. 5, the model trained with our curriculum exhibits notably faster learning. The benefits become apparent very early in training, with curriculum training achieving the same level of performance of the baseline model after only 100 epochs (8× faster), and outperforming it by a large margin as the training progresses.

Few-shot Performance To further validate the quality and effectiveness of the representations learned by our accelerated pretraining method, we perform linear probing evaluations under a few-shot scenario. Specifically, we assess the performance of our pretrained models on eight fine-grained datasets, using a 5-shot setup where only five labeled examples per class are available for training the linear classifier. This evaluation tests the generalization capabilities and data efficiency of the learned representations. Results are summarized in Table 3, which clearly demonstrate that our method, even at significantly fewer training epochs (200 and 400 epochs), achieves comparable or superior performance compared to the baseline trained for 1600 epochs, highlighting both efficiency and effectiveness with up to an 8× speedup.

4.3. Impact of Temperature Annealing

Temperature annealing enables the model to transition from prototypical to diverse examples during training. To examine its role in our curriculum strategy, we compared it against fixed temperature τ schedules, with temperatures ranging from 0.01 to infinity in Table 4. Constant temperature effectively creates static training distributions, with lower temperatures prioritizing more prototypical examples.

The results demonstrate that static temperature choices face an inherent trade-off. At very low temperatures ($\tau = 0.01$), the model sees only the most prototypical examples but achieves poor performance (5.69% NN, 18.22% LP), likely due to limited exposure to variations in the data. As temperature increases, performance improves up to an optimal point around $\tau = 0.2$ (41.57% NN, 67.71% LP), before gradually declining at higher temperatures. Interestingly, the optimal static temperature of $\tau = 0.2$ outperforms the baseline MAE model which is trained under uniform sampling (*i.e.*, $\tau = \infty$). This demonstrates that even a static bias towards prototypical examples outperforms standard uniform sampling, suggesting that prototypical examples play a crucial role in MIM training.

However, the optimal static temperature still underperforms temperature annealing (47.40% NN, 68.80% LP). These findings show that a curriculum transitioning from prototypical to diverse examples is beneficial for effective MIM training. Rather than settling for a fixed compromise between prototypicality and diversity, temperature annealing allows the model to first establish robust representations from clear prototypes before gradually incorporating more challenging variations.

4.4. Impact of Prototype Identification

To study the impact of prototype identification on our curriculum learning approach, we conducted a comprehensive study examining three key aspects: the choice of feature space, backbone architecture, and clustering granularity. All experiments were conducted on ImageNet-1k using a ViT-B/16 backbone architecture.

First, we investigated the impact of different representation spaces (Table 4a). Prototypes selected from DiNO features yielded superior performance (NN: 40.15%, LP: 68.73%) compared to prototypes obtained from both an untrained model (NN: 32.63%, LP: 65.65%) and an MAE-pretrained model (NN: 34.14%, LP: 66.60%). Interestingly, DINO prototypes even slightly outperformed supervised pre-training (NN: 39.70%, LP: 68.09%), showing that supervised pretraining is not required for effective prototype identification. In fact, even curricula derived from clustering in a low-level feature space, such as SIFT, already performed well (NN: 36.85%, LP: 69.17%). Large pre-trained models, which might be expensive to train, are not required to establish the curriculum.

Table 4b shows that the size of the backbone architecture is not a critical factor in prototype identification. In fact, slightly better downstream performance is achieved with the smaller ViT-Small model (NN: 41.05%, LP: 69.23%) compared to the larger ViT-Base/16 model (NN: 40.15%, LP: 68.73%).

Finally, our study of clustering granularity (Table 4c) reveals an optimal range for the number of centroids. Performance peaks at 1000 clusters (NN: 41.05%, LP: 69.23%), with degradation observed at both lower (500 clusters) and higher (2000 clusters) granularities. Notably, the data-driven approach to determining cluster count through the Davies-Bouldin index (978 clusters) achieves comparable performance to the manually tuned optimum. Interestingly, comparing our self-supervised clustering approach to an oracle baseline using ground truth labels for centroid definition shows that our method actually outperforms the oracle in nearest neighbor classification (41.05% vs 37.83%). This surprising result suggests that our prototype identifica-

	CUB-200 [45]	SUN397 [48]	OxfordPets [33]	Caltech101 [13]	UCF101 [39]	DTD [8]	Flowers102 [30]	Food101 [4]	AVG
Baseline (1600 epoch)	14.42 ±0.38	30.20 ±0.61	39.76 ±2.43	74.81 ±2.11	49.06 ±0.67	44.47 ±1.89	64.48 ±0.85	16.97 ±0.42	41.77
Ours (200 epoch)	14.35 ±0.41	30.15 ±1.03	43.79 ±1.75	77.42 ±1.51	48.15 ±1.05	42.81 ±1.79	64.13 ±0.71	16.64 ±0.05	42.18
Ours (400 epoch)	**18.46** ±0.40	**34.33** ±1.24	**56.97** ±2.20	**82.02** ±1.86	**52.35** ±1.47	**45.59** ±1.20	**68.99** ±0.53	**19.11** ±0.36	**47.23**

Table 3. 5-shot linear probing accuracy (%) on eight fine-grained datasets. Mean and standard deviation reported over three runs. Our accelerated method (200 and 400 epochs) achieves comparable or improved performance compared to the baseline method (1600 epochs), highlighting a training speedup of up to 8×.

τ	$\lvert\mathcal{D}_\tau\rvert/\lvert\mathcal{D}\rvert$	NN	LP
0.01	0.004	5.69	18.22
0.05	0.122	29.51	46.38
0.08	0.236	39.67	61.86
0.2	0.495	41.57	67.71
0.4	0.592	36.86	67.15
0.6	0.614	35.17	66.45
inf	0.632	31.70	64.94
Annealed		47.40	68.84

Table 4. Impact of temperature on training. Comparison between fixed and annealed temperature sampling. Lower temperatures (τ) favor prototypical examples; $\tau = \infty$ represents uniform sampling. $\lvert\mathcal{D}_\tau\rvert/\lvert\mathcal{D}\rvert$ denotes effective dataset size. Annealed sampling adjusts τ from 0.07 to 0.6.

Representation Space	NN	LP
SIFT [10]	36.85	69.17
VIT-B/16 Untrained	32.63	65.65
VIT-B/16 MAE [16]	34.14	66.60
VIT-B/16 DiNO [5]	40.15	68.73
VIT-B/16 Supervised [12]	39.70	68.09

(a) Feature space.

Architecture	NN	LP
ViT B/16	40.15	68.73
ViT S/16	41.10	69.00
ViT S/8	41.05	69.23

(b) DiNO Backbone.

# Clusters	NN	LP
100	38.64	68.54
500	40.76	68.51
1000	41.05	69.23
1500	39.65	68.35
2000	38.90	68.27
978	41.05	69.00
Oracle (1000)	37.83	68.49

(c) Number of clusters. 978 is computed via DB 3.2.1.

Table 5. Empirical study on prototype identification. We evaluate the impact of (a) feature space, (b) model architectures, and (c) number of clusters. We further compare self-supervised DiNO clusters against "Oracle" clusters which use ground truth labels to define the centroids. Unless otherwise specified, experiments use DiNO features from a ViT-Base/16 backbone with 1000 clusters. All models are trained for 800 epochs on ImageNet-1K.

tion method captures underlying visual patterns that may be more fundamental than human-defined categorical boundaries, potentially offering a more nuanced representation of visual similarities.

These ablation studies collectively demonstrate the robustness of our approach and provide practical insights for implementation. The results support our hypothesis that prototype-driven curriculum learning benefits from (1) self-supervised feature spaces that capture rich semantic relationships, (2) architectures that preserve fine-grained spatial information, and (3) careful optimization of prototype granularity.

5. Conclusion

In this work, we introduced a prototype-driven curriculum learning approach for Masked Image Modeling that structures the learning process to progress from prototypical to more complex examples. As the computational demands of training large vision models continue to grow, improving training efficiency becomes increasingly critical for sustainable AI development. Through extensive experiments, we demonstrated that our method substantially outperforms standard MAE training and alternative curriculum strategies with significantly reduced training time. These improvements stem from our carefully designed temperature annealing strategy, which enables the model to first establish robust representations from clear prototypes before gradually incorporating more challenging variations.

While we focused our evaluation on ImageNet-1K, our findings have promising implications for scaling MAE methods to larger and more complex datasets. The demonstrated efficiency gains suggest that prototype-driven curricula could be particularly valuable in scenarios where computational resources are constrained or dataset complexity poses training challenges. Future work could explore the application of our approach to massive-scale datasets, where the benefits of structured learning from prototypical examples may be even more pronounced. Additionally, investigating how prototype selection strategies could be adapted for diverse visual domains could further broaden the impact of this approach.

References

[1] Guillaume Alain, Alex Lamb, Chinnadhurai Sankar, Aaron Courville, and Yoshua Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015. 2

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 348–367. Springer, 2022. 2

[3] Vighnesh Birodkar, Hossein Mobahi, and Samy Bengio. Semantic redundancies in image-classification datasets: The 10% you don't need. *arXiv preprint arXiv:1901.11409*, 2019. 2

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 8

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4, 5, 8

[6] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. 5

[7] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 108–124. Springer, 2022. 2

[8] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 8

[9] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019. 2

[10] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979. 3, 8

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[13] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 8

[14] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019. 2

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 5

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 1, 5, 8

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 4, 5, 6

[18] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 2

[19] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[20] Siddharth Joshi and Baharan Mirzasoleiman. Data-efficient contrastive self-supervised learning: Most beneficial examples for supervised learning contribute the least. In *International conference on machine learning*, pages 15356–15370. PMLR, 2023. 2

[21] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[22] Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pages 2525–2534. PMLR, 2018. 2

[23] Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021. 2, 3

[24] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14290–14302, 2022. 2

[25] Jinhong Lin, Cheng-En Wu, Yibing Wei, and Pedro

Morgado. Accelerating augmentation invariance pretraining. *arXiv preprint arXiv:2410.22364*, 2024. 5

[26] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015. 2

[27] Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pages 15630–15649. PMLR, 2022. 2

[28] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020. 2, 3

[29] Shentong Mo and Pedro Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27186–27196, 2024. 2

[30] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 8

[31] Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33:9960–9971, 2020. 1

[32] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 604–621. Springer, 2022. 2

[33] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 8

[34] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 1, 2

[35] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021. 2

[36] Omead Pooladzandi, David Davini, and Baharan Mirzasoleiman. Adaptive second order coresets for data-efficient machine learning. In *International Conference on Machine Learning*, pages 17848–17869. PMLR, 2022. 2

[37] Tom Schaul. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015. 2

[38] Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 8

[40] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35: 19523–19536, 2022. 2, 5

[41] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018. 2

[42] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093, 2022. 2

[43] Aditay Tripathi, Pradeep Shenoy, and Anirban Chakraborty. Dynamic data selection for efficient ssl via coarse-to-fine refinement. 2

[44] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12), 2010. 1, 2

[45] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 8

[46] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023. 2

[47] Yibing Wei, Abhinav Gupta, and Pedro Morgado. Towards latent masked image modeling for self-supervised visual representation learning. In *European Conference on Computer Vision*, pages 1–17. Springer, 2025. 5

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 8

[49] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 2

[50] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li.

Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27061–27074, 2022. 2