How Should Corruption Be Used in SSL? Empirical Insights for Effective Pretraining

Anonymous authors
Paper under double-blind review

Abstract

We study how corruption design—masking and additive noise—affects self-supervised pretraining of vision models. Although denoising diffusion models succeed in generation, noise-driven extensions of masked image modeling (MIM) achieve only marginal gains on recognition tasks, including fine-grained benchmarks. We thus investigate why this would be the case, seeking effective ways to combine masking and noising within the corruption-to-reconstruction (C2R) paradigm. We begin by analyzing prior noise-based MIM approaches, categorizing them into Substitutive Corruption (masked tokens replaced by noised ones) and Conjunctive Corruption (masked and noised tokens coexist), and further into Encoder- or Decoder-style depending on where corruption and restoration occur. Our study reveals that the literature trends toward a Decoder-style design. In contrast, we evaluate an Encoder-style alternative with a focus on transfer. Building on these analyses, we propose three principles for effective C2R pretraining: corruption and restoration should occur within the encoder, noise is most effective when injected at the feature level, and mask reconstruction and de-noising must be explicitly disentangled to avoid interference. By implementing these findings, we propose a framework that captures a broader frequency spectrum of representations and improves transferability, surpassing MIM by up to 8.1% and recent noise-driven pretraining methods by 8.0% across diverse recognition benchmarks. **Code** is available in the Supplementary Material.

1 Introduction

Self-supervised learning (SSL) has emerged as a key paradigm in computer vision, enabling the pretraining of large-scale models (Dosovitskiy et al., 2020; Liu et al., 2021) on massive unlabeled datasets and transferring them to diverse downstream tasks (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Bao et al., 2021; He et al., 2022; Xie et al., 2022). By removing the reliance on costly annotations, SSL has alleviated the data-hungry nature of foundational models and driven progress in image classification, semantic segmentation, object detection, and fine-grained recognition (Carion et al., 2020; Ranftl et al., 2021; Zhu et al., 2020; Zheng et al., 2021; Chen et al., 2021), establishing itself as a cornerstone of representation learning.

A dominant line of SSL research follows the **corruption-to-reconstruction (C2R)** paradigm, where inputs are intentionally corrupted and the model is trained to reconstruct the original data. Masked image modeling (MIM) exemplifies this strategy (Bao et al., 2021; He et al., 2022; Xie et al., 2022), masking large portions of input patches to encourage spatial reasoning and semantic understanding. These methods have shown remarkable effectiveness and scalability, achieving state-of-the-art results on various vision benchmarks (Deng et al., 2009; Zhou et al., 2017; Lin et al., 2014).

Meanwhile, motivated by the success of generative models such as denoising diffusion models (Ho et al., 2020; Rombach et al., 2022; Ramesh et al., 2021; Saharia et al., 2022), recent works have explored noise-based C2R pretraining. DiffMAE (Wei et al., 2023) and MaskDiT (Zheng et al., 2023) extend masking-based pretraining by introducing noise in different ways. DiffMAE (Wei et al., 2023) replaces masked tokens with noised ones, and MaskDiT (Zheng et al., 2023) further leverages both; collectively, they illustrate the integration of masking and noising within a unified pretraining framework. While diffusion models excel in high-fidelity image generation (Dhariwal & Nichol, 2021), these noise-driven approaches (Wei et al., 2023; Zheng et al., 2023) do not provide

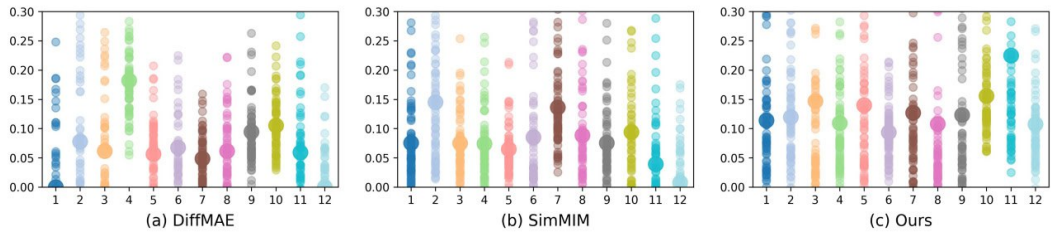

Figure 1: We plot the KL divergence of attention distributions across heads (small dots) and their layer-wise means (large dots) for (a) a recent noise-based MIM method (Wei et al., 2023), (b) a representative MIM method (Xie et al., 2022), and (c) ours. Higher KL divergence indicates broader frequency coverage. Our method achieves greater diversity than MIM and noise-based baselines, accounting for its strong performance on recognition tasks, including fine-grained settings.

Figure 2: We visualized the self-attention maps for the image classification token in the final layer of our model on a fine-grained visual categorization benchmark. The proposed method captures a range of frequencies by focusing on both key features and fine details within complex scenes.

notable gains over MIM on *recognition benchmarks* (Wah et al., 2011; Van Horn et al., 2015; 2017; 2018; Krause et al., 2013; Maji et al., 2013; Deng et al., 2009; Zhou et al., 2017; Lin et al., 2014), spanning both fine-grained recognition and general vision tasks. This indicates that although noise introduces high-frequency variations, such information is not effectively encoded into transferable representations for recognition tasks.

In Sec. 3, we systematically analyze why noise-based C2R pretraining yields limited gains. We first dissect their design choices, categorizing them into two paradigms: *Substitutive Corruption*, which replaces masked tokens with noised ones, and *Conjunctive Corruption*, where masked and noised tokens coexist. We then classify these paradigms into *Encoder-* and *Decoder-styles* depending on where corruption and reconstruction occur, noting that prior methods are largely Decoder-style. Building on these analyses, we propose three design principles for effectively unifying masking and noising: (1) **corruption and restoration should occur within the encoder**, as the encoder is what is ultimately transferred to downstream tasks; (2) **noise is most effective when injected at the feature level**, particularly in lower encoder layers, where high-frequency details are present; and (3) **masked token reconstruction and de-noising must be explicitly disentangled to avoid interference**, which we enforce by suppressing attention between the two token types.

With these findings, we design a novel pretraining setup that effectively utilizes both masking and noising. Our approach captures a richer frequency spectrum of image representations as shown in Figs. 1 and 2, enhancing transferability across a variety of downstream tasks and recognition benchmarks; CUB-200-2011 (Wah et al., 2011), NABirds (Van Horn et al., 2015), iNaturalist 2017/2018 (Van Horn et al., 2017; 2018), Stanford Cars (Krause et al., 2013), Aircraft (Maji et al., 2013), ImageNet (Deng et al., 2009), ADE20K (Zhou et al., 2017), and COCO (Lin et al., 2014). Our method achieves up to 8.1% performance gain over MIM baselines and 8.0% over recent noise-driven pretraining methods, validating the effectiveness of our design.

To summarize, our contributions are.

- We provide a thorough empirical study on why current noising-based pretraining approaches (Wei et al., 2023; Zheng et al., 2023) do not provide noticeable gains for recognition tasks.
- We provide guidelines from our detailed study on how to use corruptions within pretraining.

- With our findings, we propose a novel pretraining method that outperforms the state-of-the-art on a wide range of recognition tasks, including fine-grained tasks.

2 Preliminary and Related Works

Since the intuitions and findings of our work build upon masked image modeling (MIM) and denoising diffusion models, we first revisit these foundations for completeness.

2.1 Masked Image Modeling (MIM)

The core idea of MIM is to randomly mask a subset of image tokens and train the model to reconstruct the missing content in a self-supervised manner.

Random masking. Formally, let $X \in \mathbb{R}^{N\times L\times D}$ denote the input sequence of image tokens, where N is the batch size, L the number of tokens per image, and D the token dimension. We define the mask generation process as $M = \Phi_{\text{M}}(X, \gamma)$, where γ is the masking ratio and $M \in \{0,1\}^{N\times L}$ indicates a mask map. The masking operation generates a corrupted signal X_{cor} as

$$X_{cor} = X \odot M + \theta \odot (1 - M), \tag{1}$$

where $\odot$ denotes the Hadamard product, and θ is a learnable parameter for masked tokens.

Reconstruction. To learn to reconstruct the original tokens X back from only the visible ones X_{vis}, training of MIMs typically relies on mean squared error (MSE). With the token predictions $\hat{X}$ from MIM framework that takes as input the masked visible tokens X_{masked}, we minimize the loss:

$$\mathcal{L}_{\text{MIM}} = \frac{1}{\sum_{k,l}(1 - M_{k,l})} \sum_{k=1}^{N}\sum_{l=1}^{L}(1 - M_{k,l})\|\hat{X}_{k,l} - X_{k,l}\|^2. \tag{2}$$

Recent works. MIM approaches (He et al., 2022; Xie et al., 2022; Choi et al., 2024; 2025; Bao et al., 2021; Yi et al., 2022; Dong et al., 2022; Chen et al., 2024a; Assran et al., 2023) adapt the concept of Masked Language Modeling (MLM) from NLP. BEiT (Bao et al., 2021) applies MLM-like pretraining to images using discrete visual tokens generated by a pre-trained dVAE. MAE (He et al., 2022) focuses only on visible patches in the encoder, predicting masked pixel values through a decoder. SimMIM (Xie et al., 2022) uses both visible and masked patches in the encoder and predicts original pixels directly. Recent advances (Choi et al., 2024; 2025) focus on masked tokens for fast convergence and performance improvement.

2.2 Denoising Diffusion Model

Denoising diffusion models are trained by progressively corrupting inputs with Gaussian noise and learning to denoise them. This is in a similar spirit to MIMs, but unlike MIMs, the theoretical foundations allow the model to generate new data, hence they are generative (Song et al., 2020).

Forward diffusion. Forward diffusion iteratively adds noise to an input image sequence $X \in \mathbb{R}^{N\times L\times D}$ over T time steps. At each time step t, a noise schedule $\beta^t \in \mathbb{R}$ controls the amount of noise added, where β^t is a scalar that determines the noise level at time step t. The corrupted representation X^t at step t is then defined as:

$$X^t = \sqrt{1 - \beta^t} \cdot X^{t-1} + \sqrt{\beta^t} \cdot \epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the Gaussian noise with a zero matrix $\mathbf{0}$ and an identity covariance matrix $\mathbf{I}$. This iterative process gradually *diffuses* the data towards Gaussian noise as t approaches T.

Denoising. With the corrupted signal, the denoiser then learns to undo this corruption, effectively allowing the model to traverse back through the diffusion process, that is, transform Gaussian noise to clean data that follows the data distribution. Specifically, starting from X^T, the model learns to predict the clean image X^0 by estimating the intermediate states through a denoising function

Φ_{denoise}, which is typically parameterized as a neural network. The denoising step at time t can be represented as:

$$\hat{X}^{t-1} = \Phi_{\text{denoise}}(X^t, t), \tag{4}$$

where $\hat{X}^{t-1}$ represents the denoised estimate at time $t-1$. Training of $\Phi_{\text{denoise}}(X^t, t)$ is performed through various variations of the original DDIM (Song et al., 2020) and DDPM (Ho et al., 2020) methods, including recent family of Rectified Flow models (Liu et al., 2022), but these approaches are all essentially focusing on obtaining $\hat{X}^{t-1}$ estimates in some form that will accurately lead toward X^0 through various solvers (Lu et al., 2022; Karras et al., 2022).

Recent works. Denoising diffusion models (Ho et al., 2020; Nichol & Dhariwal, 2021; Rombach et al., 2022; Ramesh et al., 2021; Saharia et al., 2022) have gained prominence in generative tasks for their ability to produce high-quality, detailed images. DDPM (Ho et al., 2020) introduced the foundational framework, where Gaussian noise is gradually added to an image and then removed in a reverse process, Improved DDPM (Nichol & Dhariwal, 2021) enhanced this approach with modifications in noise scheduling and model architecture. LDM (Rombach et al., 2022) further improved efficiency by operating in a compressed latent space rather than pixel space, allowing various practical applications.

2.3 Pretraining via denoising

With preliminaries on MIMs and denoising diffusion models, we now review two representative works that aim to marry the two schools of thought into a single pretraining framework.

DiffMAE. DiffMAE (Wei et al., 2023) combines diffusion-based modeling with MIM. Instead of replacing masked patches with a learnable token, DiffMAE corrupts them by progressively injecting Gaussian noise following a predefined schedule $\alpha_t \in \mathbb{R}$. Given a binary mask M, Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is progressively added to the selected tokens over T steps, formulated as:

$$X_{cor} = X_v + X_n^t = X \odot M + \left(\sqrt{\alpha_t} \cdot X + \sqrt{1-\alpha_t} \cdot \epsilon\right) \odot (1 - M). \tag{5}$$

For clarity, we denote the *visible tokens* as X_v and the *noised tokens* as X_n^t. The model then reconstructs X by denoising the corrupted tokens X_n through an iterative reverse process conditioned on the visible tokens X_v:

$$\hat{X}_n^{t-1} = \Phi_{\text{denoise}}(X_n^t, X_v, t), \tag{6}$$

where Φ_{denoise} denotes the denoising function.

MaskDiT. MaskDiT (Zheng et al., 2023) introduces a noise-based pretraining approach that leverages masked transformers for faster training of diffusion models. Unlike DiffMAE, MaskDiT generates a corrupted input using both *noised tokens* and *masked tokens*:

$$X_{cor} = X_n^t + X_m = \left(\sqrt{\alpha_t} \cdot X + \sqrt{1-\alpha_t} \cdot \epsilon\right) \odot M \; + \; \theta \odot (1 - M), \tag{7}$$

where θ denotes a learnable parameter for the masked tokens. Similarly, as before, we denote the *noised tokens* X_n^t and the *masked tokens* X_m. The model then learns to reconstruct both the noised tokens X_n^t and the masked tokens X_m via denoising and reconstruction function Φ:

$$(\hat{X}_n, \hat{X}_m) = \Phi(X_n^t, X_m, t). \tag{8}$$

3 An Analysis of Prior pretraining Methods

Recent noise-based C2R methods (Wei et al., 2023; Zheng et al., 2023) augment masked image modeling (MIM) by injecting additive Gaussian noise to capture fine-grained detail. Yet, consistent with the results reported in the prior study (Zheng et al., 2023), our experiments show no notable gains over MIM baselines (Xie et al., 2022; He et al., 2022) on recognition tasks (Fig. 7). Under matched pretraining and identical fine-tuning, both variants perform on par with MIM on ImageNet (Deng et al., 2009) and underperform on FGVC (Wah et al., 2011; Van Horn et al., 2015; 2017; 2018; Krause et al., 2013; Maji et al., 2013), where fine detail matters most. Simply adding a denoising stage to MIM does not improve representation quality for recognition. We therefore examine *how* masking and noising are combined and *where* corruption is applied.

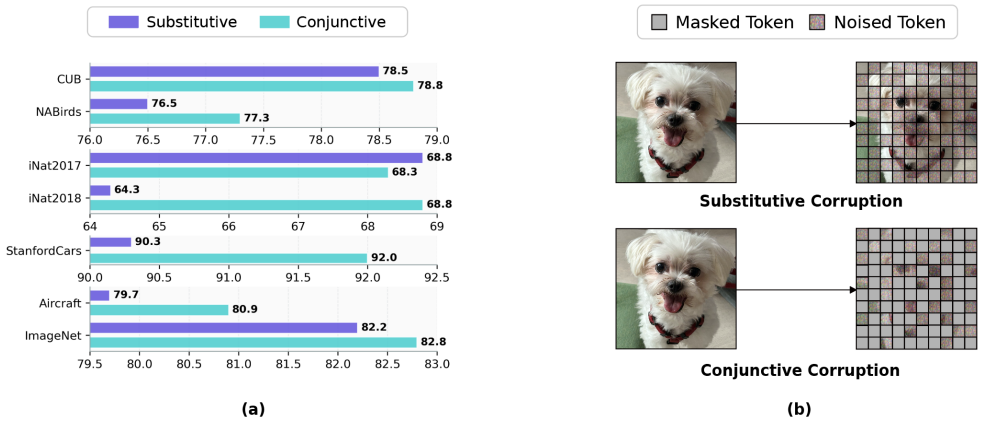

Figure 3: **(a)** Conjunctive Corruption achieved better performance across datasets, as it consistently retains a fully masked portion that enhances semantic discriminability. In contrast, Substitutive Corruption relies on tokens with random noise intensities, which may limit its effectiveness. **(b)** Illustration of two corruption paradigms: Substitutive Corruption, where masked tokens are replaced by noised ones; and Conjunctive Corruption, where masked and noised tokens coexist.

3.1 How should masking and noising be combined?

We first examine how recent methods integrate noising with masking to investigate why they provide limited gains on recognition tasks. Specifically, we look into the two representative cases of integrations, DiffMAE (Wei et al., 2023) and MaskDiT (Zheng et al., 2023).

- **Substitutive Corruption**: masked tokens are *replaced* with noised tokens;
- **Conjunctive Corruption**: masked tokens are *retained* while visible tokens are additionally noised.

Substitutive Corruption (Wei et al., 2023) employs a noised token alongside a clean visible token, as specified in (5), and the model focuses on denoising (6). On the other hand, *Conjunctive Corruption* (Zheng et al., 2023) utilizes both a masked token and a noised token, as described in (7), and the model performs both denoising and reconstruction (8). Fig. 3 (b) illustrates these two alternatives.

We evaluated the two corruption methods used in recent baselines (Wei et al., 2023; Zheng et al., 2023). Fig. 3 (a) presents the transfer learning performance measured after pretraining on ImageNet-1K (Deng et al., 2009) and fine-tuning across recognition benchmarks (Wah et al., 2011; Van Horn et al., 2015; 2017; 2018; Krause et al., 2013; Maji et al., 2013), confirming that Conjunctive Corruption consistently outperforms Substitutive Corruption. We attribute this to the limitation of Substitutive Corruption, which relies solely on noise as a corruption. Since random time sampling often produces nearly clean inputs, the pretraining task may become trivial and fail to encourage meaningful semantic learning. In contrast, Conjunctive Corruption always retains masked regions, compelling the model to jointly solve denoising and reconstruction, encouraging richer feature representations.

3.2 Where should corruption be applied?

Beyond the strategy to combine masking and noising, a further key question is *where* corruption should be applied. **We first categorize MIM paradigms into two types** based on the placement of masked tokens, as illustrated in Fig. 4 (a):

- **Encoder-style** (Xie et al., 2022; Bao et al., 2021; Yi et al., 2022): masked tokens are injected *into the encoder*, and reconstruction is performed across the encoder–decoder.
- **Decoder-style** (He et al., 2022; Chen et al., 2024a; Dong et al., 2022): masked tokens are processed only *in the decoder*, while the encoder learns solely from visible tokens.

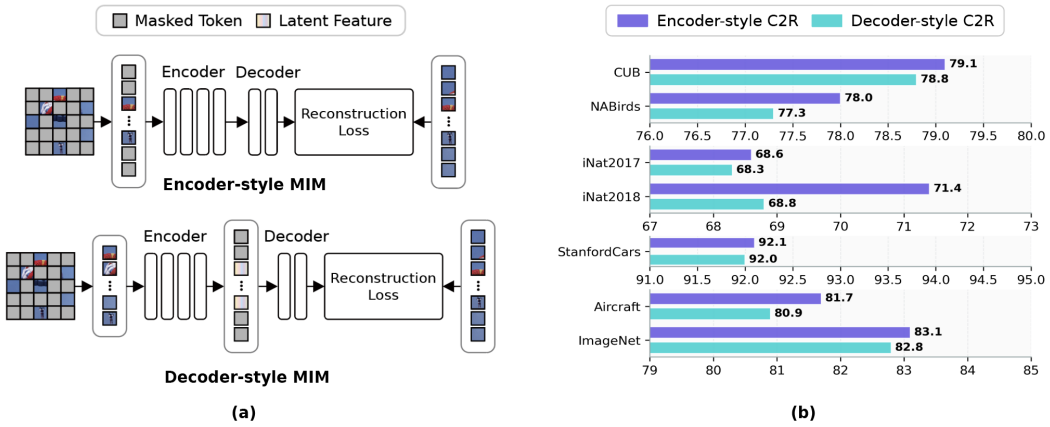

Figure 4: **(a)** The study of MIM is broadly segmented into two types based on masked token placement: Encoder-style, which reconstructs masked regions within the encoder (Xie et al., 2022; Bao et al., 2021; Yi et al., 2022; Choi et al., 2024), and Decoder-style, where reconstruction occurs solely in the decoder (He et al., 2022; Chen et al., 2024a; Dong et al., 2022). The recent noise-based C2R baselines (Wei et al., 2023; Zheng et al., 2023) build on MAE (He et al., 2022), can be seen as Decoder-style. **(b)** We implemented naive Encoder- and Decoder-style C2R frameworks, both with Conjunctive Corruption. Consistent with our hypothesis, the Encoder-style performs better across standard and fine-grained tasks, though the gains are modest, suggesting that naive implementations fall short. We thus further examine how Encoder-style design can more fully exploit its advantages in Sec. 4.1.

Recent noise-based pretraining approaches (Wei et al., 2023; Zheng et al., 2023) primarily adopt a Decoder-style design built on MAE (He et al., 2022).

However, it is important to note that only the encoder is transferred for downstream fine-tuning. This suggests that the placement of corruption and reconstruction may matter and has motivated MIM baselines to explore Encoder-style designs; accordingly, recent works (Xie et al., 2022; Bao et al., 2021; Yi et al., 2022; Choi et al., 2024) adopts an Encoder-style design. We therefore study an Encoder-style variant that applies corruption and reconstruction inside the encoder, such that noising directly shapes the learned transferable representations. The next section (Sec. 4.1) details this design and compares it head-to-head with matched Decoder-style baselines.

4 Proposed Method

Building on the analysis of prior methods, which revealed strengths and limitations of existing designs, we move beyond them and identify three novel design principles to effectively unify masking and noising. These principles, detailed in the following subsections, are as follows:

- **Encoder-style**: corruption and restoration should occur within the encoder;
- **Feature-level noise**: noise is most effective when injected at the feature level; and
- **Task disentanglement**: masked token reconstruction and de-noising must be explicitly disentangled.

4.1 Corruption and Restoration Should Occur Within the Encoder

In most transfer learning pipelines, the encoder is transferred and fine-tuned for downstream tasks, while the decoder is typically discarded. Consequently, when corruption is applied only at the latent level and restoration is confined to the decoder, the encoder neither learns to handle corrupted signals nor explicitly engages in restoration, limiting the relevance of its learned features. In contrast, introducing corruption and enforcing restoration within the encoder directly couples representation learning with corruption handling, which, in principle, should promote richer and more transferable

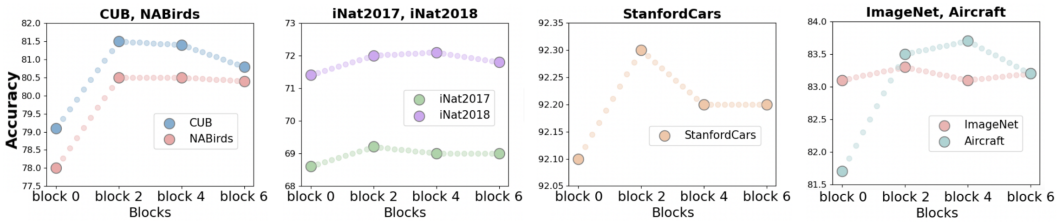

Figure 5: We introduced noise at encoder blocks 0, 2, 4, and 6. Feature-space injection (blocks 2, 4, and 6) outperformed pixel-space (block 0) on recognition tasks, with optimal performance at block 2, where high-frequency details are captured.

features. This theoretical motivation forms the basis of our hypothesis that Encoder-style corruption design offers clear advantages for pretraining. We then evaluated their performance on transfer learning by pretraining both models on ImageNet-1K (Deng et al., 2009) and fine-tuning them on the range of recognition tasks and datasets.

Fig. 4 (b) reports the transfer learning performance of each design. We implemented two naive noise-based frameworks featuring Conjunctive Corruption strategies that differ only in their placement of corruption. We kept all other factors identical. Consistent with our hypothesis, the Encoder-style structure performs better than the Decoder-style design across both standard recognition tasks and fine-grained benchmarks. However, the margin of improvement was smaller than anticipated, suggesting that a *naive implementation alone does not fully reveal its potential*. In the following subsection, we delve deeper into why this is the case and outline how the Encoder-style paradigm can better realize its advantages.

4.2 Noise is most effective when injected at the feature level

Much of the success of denoising diffusion models is rooted in the application of noise at the latent (feature) space (Rombach et al., 2022; Chen et al., 2024b). While pixel-space diffusion exists (Hoogeboom et al., 2024), they need careful strategies on how noise should be applied. As such, we also suspect this to be the case for pretraining. However, in the naive implementation that we consider in Sec. 4.1, the Decoder-style method adds noise at the latent space immediately before the decoder, whereas the Encoder-style approach injects noise in pixel-space, prior to the encoder. We thus evaluate the Encoder-style setting by adding noise to various blocks within the encoder to investigate the impact of different noise-addition locations.

In Fig. 5, we conducted experiments by varying the stage at which noise is introduced within the encoder, specifically at different encoder blocks (blocks 0, 2, 4, and 6). The transfer learning performance results for recognition tasks verify that adding noise in feature-space (blocks 2, 4, and 6) is more effective than in pixel-space (block 0). Additionally, the highest performance observed at 'encoder block 2' suggests that noise addition is particularly effective when applied in the lower layers of the encoder, where high-frequency details are captured. This result reveals that the injecting noise at the feature level is crucial for maximizing the transfer learning potential of the model.

4.3 Masked token reconstruction and de-noising should be explicitly disentangled

Referring to recent studies of MIM (Choi et al., 2024; He et al., 2022; Dong et al., 2022), Encoder-style approaches have shown mixed outcomes compared to Decoder-style. A plausible contributor is that masked tokens are optimized along directions weakly aligned with those of visible tokens (Choi et al., 2024), which can interfere with updates to visible token representations. Since Conjunctive Corruption uses masked tokens alongside noisy visible tokens, we hypothesize a similar risk in noise-based C2R pretraining, where masked tokens may interact undesirably with the encoding of noisy visible tokens.

To address this, we propose an explicit objective that disentangles the masked token reconstruction from the de-noising strategy. We introduce disruption loss, a variant of masked token optimization

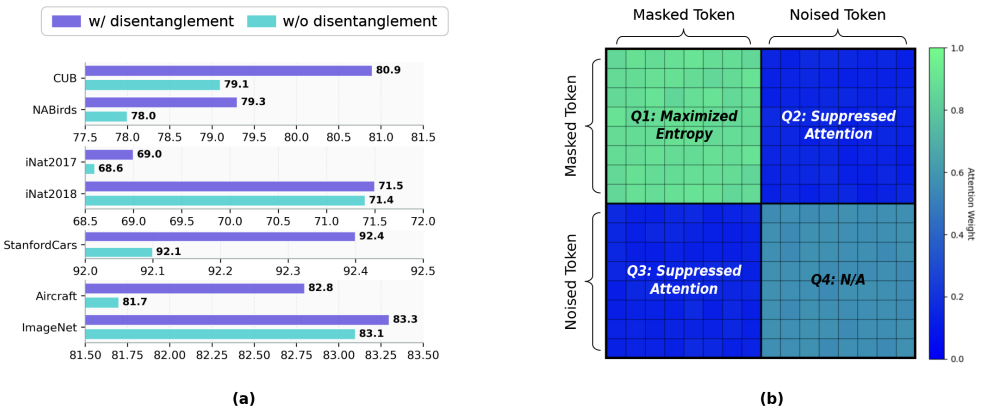

Figure 6: **(a)** We propose an explicit objective to disentangle the de-masking task from the de-noising task, fully harnessing both reconstructions within the encoder. The disruption loss adjusts the weight distribution of the affinity map, minimizing the influence of masked tokens on noisy visible tokens, and enhancing performance across both fine-grained and standard recognition tasks. **(b)** Disruption loss adjusts the affinity matrix to suppress masked-noised interactions, enforcing disentanglement between de-noising and mask reconstruction within the encoder.

proposed in MTO (Choi et al., 2024), designed to suppress attention between the two different token types. Disruption loss leverages per-row sparsities within the affinity matrix, which consists of four quadrants in Fig. 6 (b) as follows:

$$A = \begin{bmatrix} A_{mm} & A_{mn} \\ A_{nm} & A_{nn} \end{bmatrix}, \tag{9}$$

where A_{mm} represents weights between masked-masked tokens, A_{mn} and A_{mn} represents affinities between masked-noised tokens, and A_{nn} represents weights among noised tokens. The disruption loss $\mathcal{L}_d$ *suppresses the attention of the masked-noised tokens* by increasing the entropy of the attention between masked-masked token in the row unit of the affinity matrix. Thus, $\mathcal{L}_d$ recalibrates the weight distribution of A, minimizing the impact of masked tokens x_m on noisy visible tokens x_n^t:

$$\mathcal{L}_d = -\sum_{i \in \mathcal{N}} \sum_j \tilde{p}_{i,j} \log \tilde{p}_{i,j} \tag{10}$$

where $\mathcal{N}$ denotes the index set of masked tokens x_n^t, and $\tilde{p}$ is the row-wise softmax of the affinity entries in A, satisfying $0 < \tilde{p}_{i,j} < 1$ and $\sum_j \tilde{p}_{i,j} = 1$. The application of $\mathcal{L}_d$ reduces interference between different token types and thus ensures effective task disentanglement between masked token reconstruction and denoising within the encoder.

The experimental results in Fig. 6 (a) exhibit a performance improvement from this weight adjustment on both fine-grained and standard recognition tasks. This demonstrates that explicitly separating de-masking and de-noising objectives maximizes the transferability of the Encoder-style approach.

5 Experiments

5.1 Implementation Details

All experiments in this manuscript were conducted under precisely identical conditions to ensure accurate analysis. For consistency, we evaluated all methods using official implementations (except DiffMAE, which we reimplemented due to lack of release) under our hardware configuration (4 × A100 GPUs). Since baselines have been trained in large cluster resources that are not available to everyone, reproduced results may differ from original papers. Please note that we ensured all comparisons followed the same setup, with code provided for verification in the Supplementary Material.

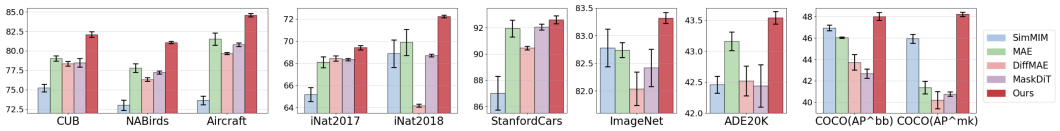

Figure 7: To ensure statistical significance, we conducted 5 trials with different random seeds, reporting the mean (bars) and standard deviation (lines) for each method. The proposed method (Ours) consistently outperforms representative MIM (Xie et al., 2022; He et al., 2022) and noise-based methods (Wei et al., 2023; Zheng et al., 2023) across a wide range of recognition tasks, capturing diverse frequency details as shown in Fig. 1 and Fig. 2, that improve accuracy in FGVC, image classification, semantic segmentation, object detection, and instance segmentation tasks.

All experiments used ViT-B (Dosovitskiy et al., 2020) as the backbone architecture applying a unified 400-epoch training schedule. pretraining was performed on the ImageNet-1K (Deng et al., 2009) classification dataset, followed by fine-tuning on respective downstream task datasets.

5.2 Main Result

In Fig. 7, we evaluate the proposed method on diverse tasks, including fine-grained visual categorization (FGVC), image classification, semantic segmentation, object detection, and instance segmentation, each with task-specific datasets. FGVC datasets (CUB-200-2011 (Wah et al., 2011), NABirds (Van Horn et al., 2015), iNaturalist 2017 (Van Horn et al., 2017), iNaturalist 2018 (Van Horn et al., 2018), Stanford Cars (Krause et al., 2013), Aircraft (Maji et al., 2013)) demand detailed, fine-grained feature learning to distinguish visually similar classes. In contrast, standard recognition tasks (ImageNet (Deng et al., 2009)), semantic segmentation (ADE20K (Zhou et al., 2017)), object detection and instance segmentation (COCO (Lin et al., 2014)) emphasize broader spatial details at different levels of granularity.

To ensure the *statistical significance* of the results, we conducted 5 trials with different random seeds and included the mean and standard deviation for each method in the figure. In the graph, the bars represent the mean performance, while the lines indicate the standard deviation.

The proposed method (Ours) consistently outperforms representative MIM (Xie et al., 2022; He et al., 2022) and noise-based methods (Wei et al., 2023; Zheng et al., 2023) across tasks. In FGVC tasks, our method effectively captures high-frequency, localized features, as also shown in Fig. 1 and Fig. 2, surpassing comparison methods in accuracy. Even in standard recognition tasks, where spatial detail is key, our method shows favourable gains, highlighting our proposed design enhance transfer potential and capture diverse frequency information, as demonstrated in Fig. 1. These statistically significant improvements strongly validate the effectiveness and robustness of our approach over comparison methods.

Ablation studies in the Appendix. Comprehensive ablation studies are provided in the Appendix below and should be consulted for a complete understanding of our method. They cover various decoupling frameworks, time-embedding placement, longer pre-training schedules, evaluation on denoising task, and extended comparisons with related works.

6 Conclusion

We have investigated why existing noise-based C2R pretraining yields only limited gains on recognition tasks. Through systematic analysis, we proposed architectural guidelines that advocate encoder-style corruption, feature-level noise injection, and explicit disentanglement of masking and noising objectives. Our framework following these principles captures a broader frequency spectrum and achieves consistent improvements, surpassing both MIM and prior noise-based methods by significant margins across standard and fine-grained benchmarks. We believe that these findings highlight the importance of corruption design in self-supervised pretraining and open new directions to exploit generative principles in representation learning. Nevertheless, our analysis is currently limited to recognition tasks. It would be interesting to broaden the scope of our study to other applications.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12299–12310, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024a.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024b.

Hyesong Choi, Hunsang Lee, Seyoung Joung, Hyejin Park, Jiyeong Kim, and Dongbo Min. Emerging property of masked token for effective pre-training. *arXiv preprint arXiv:2404.08330*, 2024.

Hyesong Choi, Hyejin Park, Kwang Moo Yi, Sungmin Cha, and Dongbo Min. Salience-based adaptive masking: revisiting token dynamics for enhanced pre-training. In *European Conference on Computer Vision*, pp. 343–359. Springer, 2025.

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*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *European Conference on Computer Vision*, pp. 247–264. Springer, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23164–23173, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pp. 37–55. Springer, 2024.

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*, pp. 9729–9738, 2020.

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*, pp. 16000–16009, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. *arXiv preprint arXiv:2410.19324*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 595–604, 2015.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist challenge 2017 dataset. *arXiv preprint arXiv:1707.06642*, 1 (2):4, 2017.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

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

Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16284–16294, 2023.

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*, pp. 9653–9663, 2022.

Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. *arXiv preprint arXiv:2205.09616*, 2022.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Appendix

A Ablation Study on Decoupling Frameworks

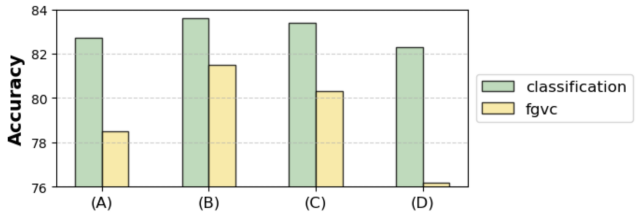

Figure 8: Comparison of four decoupling frameworks for combining de-noising (dn) and de-masking (dm) across encoder/decoder: (A) Enc-dn, Dec-dm; (B) Enc-dn+dm with disentanglement (proposed); (C) Enc-dm, Dec-dn; (D) Dec-dn+dm with disentanglement. Under matched pre-training and fine-tuning on ImageNet and FGVC, results suggest B $>$ C $>$ A $>$ D, indicating a role for task placement. Encoder-centric placement with explicit separation tends to reduce cross-objective interference and yield more transferable features, while fully decoder-based training is weaker on average.

One key finding of this study is the critical role of disentangling de-noising and de-masking tasks within the encoder, as it is the component transferred during downstream fine-tuning. To maximize transfer learning potential, our framework explicitly separates these tasks through a disentanglement objective while keeping them within the encoder. To further investigate this, we evaluated *four decoupling frameworks* on fine-grained visual categorization (FGVC) (Wah et al., 2011) and image classification (Deng et al., 2009) tasks, varying the placement of de-noising and de-masking tasks across the encoder and decoder.

Specifically, we implemented and analyzed the following four configurations.

- **(A)** Encoder de-noises, decoder de-masks (Zheng et al., 2023).
- **(B)** Proposed framework: Encoder de-noises and de-masks with disentanglement loss, ensuring task separation.
- **(C)** Encoder de-masks, decoder de-noises.
- **(D)** Decoder de-noises and de-masks with disentanglement loss, fully shifting both tasks to the decoder.

The results in Fig. 8 suggest an ordering **(B)** $>$ **(C)** $>$ **(A)** $>$ **(D)** under our protocol, pointing to a role for task placement. Framework (B) attains the highest mean score, indicating potential benefits when both objectives reside in the encoder with an explicit separation of responsibilities; a plausible explanation is reduced cross-objective interference while letting de-noising and mask reconstruction shape transferable features. Comparing (A) and (C) hints that de-masking inside the encoder can be more helpful than de-noising in our setup, possibly because mask reconstruction pressures features to encode part-level and semantic cues, whereas denoising can emphasize lower-level statistics. Framework (D) trails on most benchmarks, consistent with encoder supervision being more indirect when both objectives sit in the decoder. Overall, we read these trends as supportive of an encoder-centric placement with explicit separation, yielding semantically meaningful and transferable features across downstream tasks.

B Ablation Study on Time-Embedding

In denoising diffusion models, time-embedding plays a crucial role that encodes the temporal information associated with the noise levels introduced at various steps. Injecting time-embedding at appropriate points that align with the temporal dynamics of the noise is essential, as it enables the model to more effectively capture the relationship between the noise levels and the input features. Thus, for pre-training, we introduce time-embedding just before the block where noise is

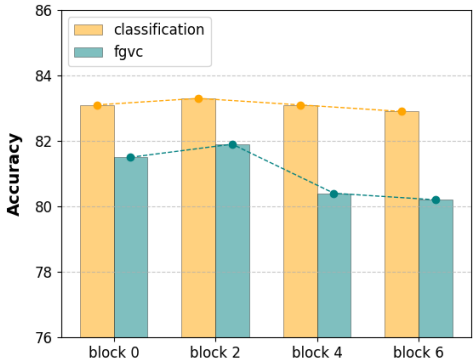

Figure 9: We evaluated the placement of time-embedding during fine-tuning on FGVC (Wah et al., 2011) and image classification tasks (Deng et al., 2009). Results show that the placement at block 2, consistent with pre-training, achieves the best performance by aligning temporal encoding with noise dynamics to retain fine-grained features.

added (block 2) to ensure the optimal alignment between the noise addition process and temporal information.

To evaluate how well the model utilizes fine-grained features aligned with temporal information for downstream tasks, we adjusted the placement of time-embedding during fine-tuning and assess its performance on FGVC (Wah et al., 2011) and image classification (Deng et al., 2009) tasks. Specifically, we tested time-embedding at four locations: initial embedding (block 0), immediately prior to noise addition (block 2), and post-noise addition blocks (block 4 and block 6).

In Fig. 9, the results show that placing time-embedding at block 2 achieves the best performance, followed by block 0, block 4, and block 6. This indicates that using the same time-embedding placement during fine-tuning as the pre-training (block 2) yields optimal results. The alignment of temporal encoding with noise dynamics is crucial for retaining the fine-grained features learned during pre-training.

The placement at block 0 achieves the second-highest performance, as injecting time-embedding at the initial stage allows the model to incorporate temporal information from the very beginning, allowing it to guide the extraction of features that are consistent with the progression of noise levels across the diffusion process. This observation aligns with findings reported in prior diffusion model studies (Ho et al., 2020; Nichol & Dhariwal, 2021), where early time-embedding helps initialize representations that remain consistent throughout the network. Blocks 4 and 6 perform worse as they occur after the noise has already been added and partially processed, making the temporal information less relevant to the feature refinement for downstream tasks. Late-stage time-embedding can introduce redundancy by repeating temporal information already captured in earlier layers, or fail to impact the already-learned representations effectively.

C Ablation Study on Longer Pre-Training Schedule

All experiments in the manuscript use a 400-epoch pre-training schedule. To assess schedule length effects, we also pre-train for 800 epochs and evaluate on FGVC (Wah et al., 2011). Figure 10 reports our method alongside representative MIM baselines—MAE (He et al., 2022) and SimMIM (Xie et al., 2022)—under both 400 and 800 epochs. Across methods, extending to 800 epochs yields consistently higher accuracy, but the improvements are modest relative to the additional compute. Notably, the relative ranking among methods is largely preserved, with no systematic cross-overs, suggesting that longer schedules primarily refine existing representations rather than alter inductive biases. Taken together, these results indicate that while longer schedules are beneficial, the cost–benefit trade-off is weak in this regime and does not change our main conclusions.

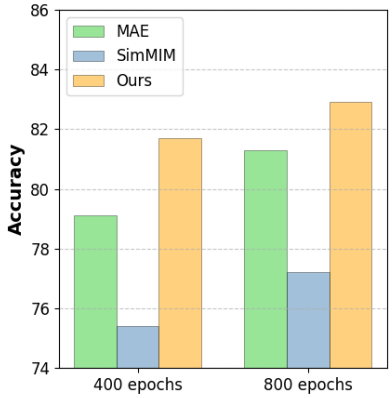

Figure 10: Effect of pre-training schedule length (400 vs. 800 epochs) on FGVC (Wah et al., 2011). We report our method and MIM baselines (MAE (He et al., 2022), SimMIM (Xie et al., 2022)). Accuracy increases at 800 epochs across methods, and rankings are largely preserved. Longer schedules appear to refine rather than change representations, yielding limited cost–benefit and leaving our conclusions unchanged.

D Evaluation on Denoising Task

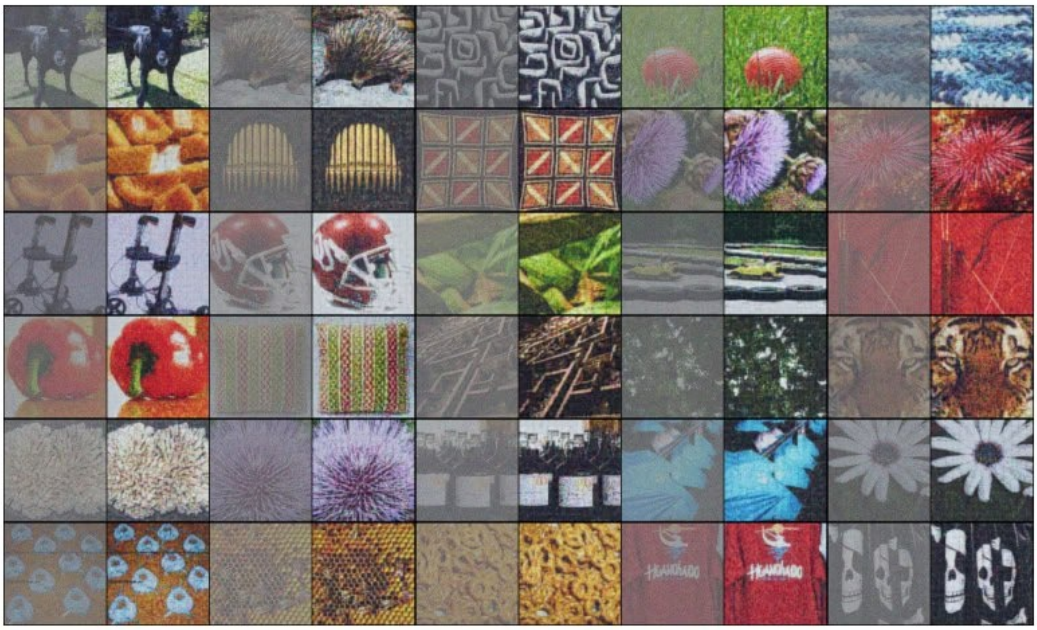

Figure 11: We qualitatively evaluate our method on denoising tasks, demonstrating its ability to accurately reconstruct corrupted scenes by capturing fine-grained details and high-level semantics through effective de-noising and de-masking strategy. (Left: input images with applied noise. Right: Denoised outputs produced by the proposed model.)

The main manuscript evaluates a broad set of recognition tasks including FGVC, image classification, semantic segmentation, object detection, and instance segmentation. Since our approach builds

on denoising diffusion models, we hypothesize potential benefits on denoising-style downstreams as well.

To validate this, we provide qualitative results on denoising tasks over ImageNet validation benchmarks (Deng et al., 2009), as shown in Fig. 11. Our method demonstrates fair predictions even under heavy noise on the task that requires capturing both fine-grained textures and high-level semantics. This highlights how our model leverages noise-based learning to understand holistic scene representations, enabling it to also excel in generative tasks. The combined de-noising and de-masking framework further promotes the learning of semantic discriminability, enabling precise reconstruction of corrupted scenes.

E Comparison with Other Related Works

E.1 Comparison with Masked Image Modeling Methods

Dataset	iBOT (Zhou et al., 2021)	Ours
CUB (Wah et al., 2011)	64.1	81.7
NABirds (Van Horn et al., 2015)	65.9	80.9

Table 1: Our comparisons focus on representative MIM baselines (SimMIM and MAE), demonstrating MIM's tendency to underperform on FGVC tasks. While a full comparison is unnecessary, we include iBOT for reference, where our method achieves significantly higher accuracy.

Our main comparisons focus on representative MIM baselines (SimMIM and MAE) to highlight the overall *tendency* of MIM to underperform on FGVC tasks. While a complete comparison with all MIM methods is not essential, we include a comparison of iBOT (Zhou et al., 2021) for reference. Results are provided in Tab. 1. The results clearly demonstrate that our proposed framework significantly outperforms representative MIM methods (Xie et al., 2022; He et al., 2022; Zhou et al., 2021) in fine-grained visual categorization (FGVC) tasks. Compared to iBOT, a strong representative of masked image modeling (MIM) approaches, our method achieves remarkably higher accuracy (17.6% on CUB (Wah et al., 2011), 15.0% on NABirds (Van Horn et al., 2015)), highlighting its superiority in FGVC. These datasets require capturing subtle differences between visually similar categories, and our framework excels at learning richer, more transferable representations that better preserve fine-grained details.

E.2 Comparison with Denoising-Based Methods

Our work focuses on pre-training methods specifically designed for recognition tasks, such as Diff-MAE (Wei et al., 2023) and MaskDiT (Zheng et al., 2023). While other denoising-based methods (e.g., DiT (Peebles & Xie, 2023), MDT (Gao et al., 2023), DiffiT (Hatamizadeh et al., 2024)) exist, they are not tailored for recognition tasks and showed significantly lower performance in preliminary evaluations. Since our work aims to advance recognition-specific pre-training, not general denoising techniques, including such comparisons would be misleading rather than informative. In addition, l-DAE (Chen et al., 2024b) is designed for recognition tasks but lacks publicly available code, making direct comparison infeasible.

F The Use of LLMs

LLMs were used only for minor language improvements. They were not involved in the conception of the research, experiments, analysis, interpretation, or drafting.