ECC: Encoder-Centric Corruption for Fine-Grained Vision in VLMs

Hyesong Choi[1], Daeun Kim[2], Sungmin Cha[3], Kwang Moo Yi[4], and Dongbo Min[2†]

[1] Soongsil University, South Korea
[2] Division of Al and Software, Ewha W. University, South Korea
[3] Meta, USA
[4] University of British Columbia, Canada

Abstract. Modern Vision-Language Models (VLMs) often suffer from *granularity mismatch*, failing to encode high-density visual details not explicitly described in abstract text prompts, which results in low-frequency bias and object hallucination. While additive noise naturally introduces high-frequency variations, suggesting corruption-to-reconstruction (C2R) as a potential remedy, existing noise-based extensions of masked image modeling (MIM) surprisingly yield little improvement on recognition tasks. We identify the cause as a structural limitation: recent trends toward decoder-style designs prevent noise-induced signals from being deeply integrated into the encoder representations. To address this issue, we propose **Encoder-Centric Corruption (ECC)**, a framework guided by three principles: (i) **Encoder-Centricity**, performing restoration within the encoder to shape transferable features; (ii) **Feature-Level Noising**, injecting noise at intermediate layers to capture fine-grained textures; and (iii) **Task Disentanglement**, separating mask reconstruction and denoising via a disruption loss. ECC surpasses prior MIM and noise-based methods by up to 8.1%. Moreover, integrating ECC into large-scale frameworks such as CLIP and AIM-v2 consistently improves performance, demonstrating its effectiveness as a scalable pretraining strategy for next-generation foundation models. Code is available at `https://github.com/doihye/ecc`.

1 Introduction

The emergence of Vision-Language Models (VLMs) [1, 11, 19, 20, 22, 27, 32, 41] has fundamentally shifted the paradigm of visual representation learning toward global semantic alignment. By leveraging massive image-text pairs, these models have achieved remarkable zero-shot capabilities. However, a critical limitation persists: while VLMs excel at capturing global context, they often suffer from *granularity mismatch*, where the model fails to encode the high-density visual details of an image that are not explicitly described in the brief, abstract text prompts [1,19,27,32]. This lack of fine-grained understanding frequently leads to

† Corresponding author.

object hallucination [13, 22] and a low-frequency bias [12, 37], as the contrastive objective encourages the model to ignore subtle textures and small objects in favor of coarse semantic shapes.

To address these fundamental deficits, we revisit the **corruption-to-reconstruction (C2R)** paradigm. We argue that for a model to be truly effective, its vision backbone must possess a robust base capability to capture the full frequency spectrum of an image—details that text labels alone cannot provide. Theoretically, both image masking [3,14,39] and the addition of Gaussian noise [15,28–30] are ideal candidates for this task. While masking large portions of input patches compels the model to infer missing structures and enhances its sensitivity to spatial locality, the injection of additive noise introduces high-frequency variations that force the model to recover even finer structures and textures. By combining these two complementary corruptions, C2R pretraining [38, 42] holds the potential to bridge the gap between global semantic alignment and local visual precision, effectively compensating for the low-frequency bias of VLMs.

However, we observe a paradoxical gap: recent noise-driven extensions of Masked Image Modeling (MIM) [38, 42] have surprisingly yielded only marginal gains on recognition benchmarks [8, 18, 21, 25, 33–36, 43]. This paradox suggests that although these corruptions provide essential high-frequency and local structural information, current architectures fail to effectively translate these signals into transferable representations for downstream recognition.

In this paper, we systematically investigate why noise-based C2R pretraining has struggled and how it can be repurposed to elevate the performance of both general-purpose vision backbones and the visual foundations of VLMs. In Sec. 3, 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.

Based on these insights, we establish a novel framework, **Encoder-Centric Corruption (ECC)**, built on three key principles: (i) **Encoder-Centricity**, where corruption and restoration occur within the encoder to directly shape transferable representations; (ii) **Feature-Level Noising**, where noise injected at intermediate encoder layers captures high-frequency visual details that pixel-space perturbations fail to model; (iii) **Task Disentanglement**, where reconstruction and denoising are explicitly separated through a disruption loss to prevent task interference.

We first validate the Base Capability of our framework on standard vision backbones [10]. Our method captures a broader frequency spectrum as shown in Figs. 1 and 2, significantly improving accuracy across a variety of downstream tasks and recognition benchmarks; CUB-200-2011 [36], NABirds [33], iNaturalist 2017/2018 [34, 35], Stanford Cars [18], Aircraft [25], ImageNet [8], ADE20K [43], and COCO [21]. Furthermore, we demonstrate the systemic utility of our approach by showing that these enhanced backbones lead to superior mul-

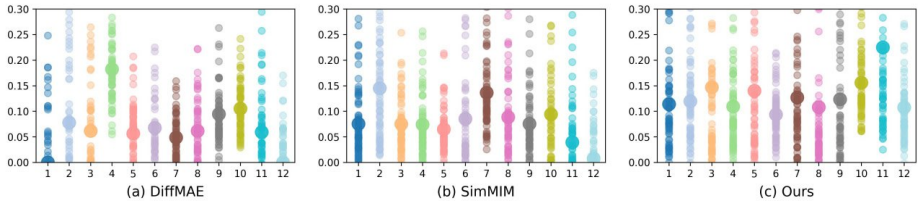

Fig. 1: We plot the KL divergence of attention distributions across heads (small dots) and their layer-wise means (large dots) for (a) a recent noise-based MIM method [38], (b) a representative MIM method [39], and (c) ours. Higher KL divergence indicates broader frequency coverage. Our method achieves greater diversity than MIM and noise-based baselines, accounting for its strong performance on recognition tasks, including fine-grained settings.

Fig. 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.

timodal alignment when integrated into vision-language frameworks [11, 22, 27]. Our method achieves up to 8.1% gain over prior C2R baselines, validating the effectiveness of our design for next-generation vision and multimodal foundation models.

To summarize, our contributions are as follows:

- **Empirical Analysis of the Fine-Grained Deficit**: We provide a thorough empirical study revealing that current noise-based C2R approaches fail to provide noticeable gains for recognition tasks due to their designs, which overlook high-frequency visual details.
- **General Principles for Encoder-Centric Corruption**: We establish three key design guidelines—encoder-centricity, feature-level noising, and task disentanglement—that serve as a foundation for enhancing the *Base Capability* of vision backbones to capture a broader frequency spectrum.
- **A Unified Framework for Vision and VLMs**: We propose a novel method, **ECC**, that effectively captures rich latent representations. Our framework not only sets a new state-of-the-art on diverse vision-only benchmarks but also demonstrates significant systemic utility by consistently improving the multimodal alignment and zero-shot accuracy of large-scale VLMs.

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 $M = \Phi_{\mathrm{M}}(X, \gamma)$, 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:

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

Recent works. MIM approaches [2–4,6,7,9,14,39,40] adapt the concept of Masked Language Modeling (MLM) from NLP. BEiT [3] applies MLM-like pretraining to images using discrete visual tokens generated by a pre-trained dVAE. MAE [14] focuses only on visible patches in the encoder, predicting masked pixel values through a decoder. SimMIM [39] uses both visible and masked patches in the encoder and predicts original pixels directly. Recent advances [6, 7] 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 [31].

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 [31] and DDPM [15] methods, including recent family of Rectified Flow models [23], 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 [17, 24].

Recent works. Denoising diffusion models [15, 26, 28–30] have gained prominence in generative tasks for their ability to produce high-quality, detailed images. DDPM [15] introduced the foundational framework, where Gaussian noise is gradually added to an image and then removed in a reverse process, Improved DDPM [26] enhanced this approach with modifications in noise scheduling and model architecture. LDM [29] 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 [38] 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:

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

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:

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

where Φ_{denoise} denotes the denoising function.

MaskDiT. MaskDiT [42] 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*:

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

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:

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

3 An Analysis of Prior pretraining Methods

Recent noise-based C2R methods [38, 42] 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 [42], our experiments show no notable gains over MIM baselines [14,39] on recognition tasks (Tab. 1). Under matched pretraining and identical fine-tuning, both variants perform on par with MIM on ImageNet [8] and underperform on FGVC [18, 25, 33–36], 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.

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 [38] and MaskDiT [42].

– **Substitutive Corruption**: masked tokens are *replaced* with noised tokens;

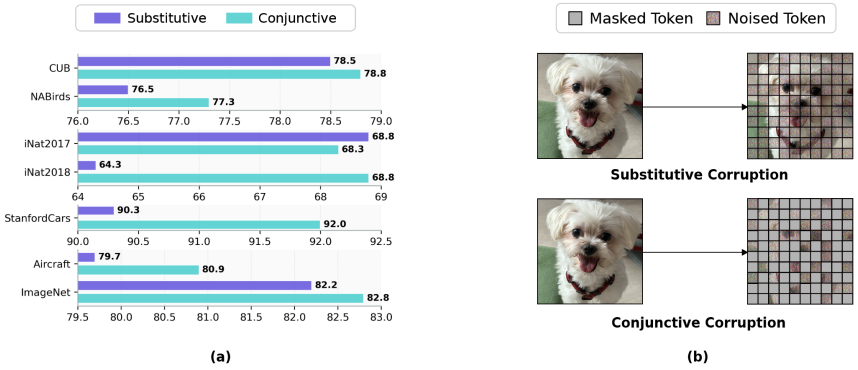

Fig. 3: (a) Conjunctive Corruption achieved better performance across datasets, as it consistently retains a fully masked portion that enhances semantic discriminability. In contrast, Substitutive Corruption relies on tokens with random noise intensities, which may limit its effectiveness. **(b)** Illustration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

– **Conjunctive Corruption**: masked tokens are *retained* while visible tokens are additionally noised.

Substitutive Corruption [38] 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* [42] 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 [38, 42]. Fig. 3 (a) presents the transfer learning performance measured after pretraining on ImageNet-1K [8] and fine-tuning across recognition benchmarks [18,25,33–36], 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** [3, 39, 40]: masked tokens are injected *into the encoder*, and reconstruction is performed across the encoder–decoder.

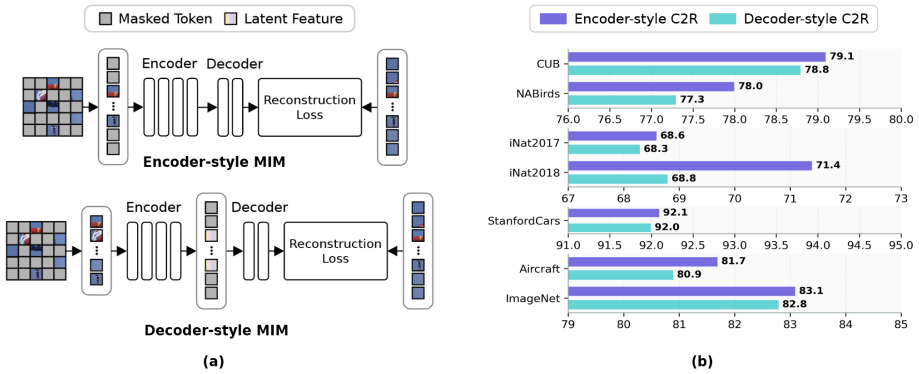

Fig. 4: (a) The study of MIM is broadly segmented into two types based on masked token placement: Encoder-style, which reconstructs masked regions within the encoder [3, 6, 39, 40], and Decoder-style, where reconstruction occurs solely in the decoder [4, 9, 14]. The recent noise-based C2R baselines [38, 42] build on MAE [14], 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.

– **Decoder-style** [4, 9, 14]: masked tokens are processed only *in the decoder*, while the encoder learns solely from visible tokens.

Recent noise-based pretraining approaches [38,42] primarily adopt a Decoder-style design built on MAE [14].

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 [3, 6, 39, 40] 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-Centricity**: corruption and restoration occur within the encoder to directly shape transferable representations;

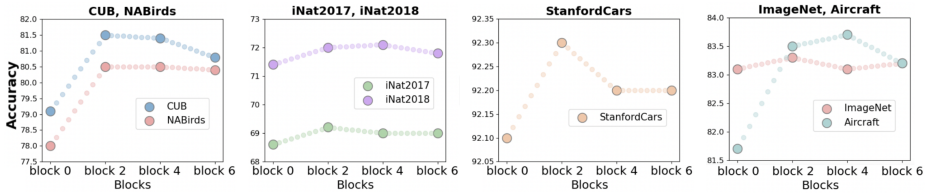

Fig. 5: We introduced noise at encoder blocks 0, 2, 4, and 6. Feature-space injection (blocks 2, 4, and 6) outperformed pixel-space (block 0) on recognition tasks, with optimal performance at **block 2**, where high-frequency details are captured.

– **Feature-Level Noising**: noise injected at intermediate encoder layers captures high-frequency visual details that pixel-space perturbations fail to model; and
– **Task Disentanglement**: reconstruction and denoising are explicitly separated through a disruption loss to prevent task interference.

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 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 [8] 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.

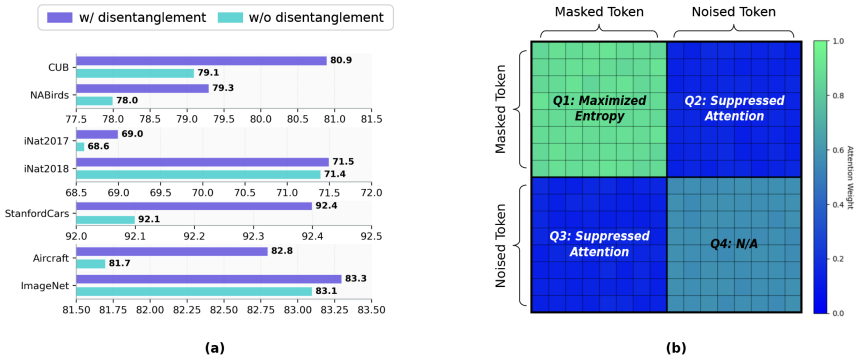

Fig. 6: (a) We propose an explicit objective to disentangle the de-masking task from the de-noising task, fully harnessing both reconstructions within the encoder. The disruption loss adjusts the weight distribution of the affinity map, minimizing the influence of masked tokens on noisy visible tokens, 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.

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 [5, 29]. While pixel-space diffusion exists [16], 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 [6, 9, 14], 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 [6], 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 proposed in MTO [6], 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 on Vision Models

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.

Table 1: To ensure statistical significance, we conduct 5 trials with different random seeds and report the mean (bars) and standard deviation (lines). Our method consistently outperforms representative MIM [14, 39] and noise-based methods [38, 42] across diverse recognition tasks.

Method	Fine-Grained Visual Categorization (Top-1 Acc %)					
	CUB	NABirds	iNat2017	iNat2018	StanfordCars	Aircraft
SimMIM [39]	75.40±0.27	73.08±0.25	65.50±0.37	69.46±0.19	87.94±0.24	73.80±0.26
MAE [14]	79.14±0.34	77.86±0.26	68.18±0.27	70.84±0.22	92.20±0.27	81.56±0.24
DiffMAE [38]	78.50±0.32	76.22±0.22	68.66±0.23	64.08±0.14	90.46±0.27	79.74±0.18
MaskDiT [42]	78.80±0.29	77.40±0.25	68.46±0.07	68.74±0.12	92.00±0.05	80.80±0.24
Ours (ECC)	**81.76**±0.36	**81.16**±0.24	**69.58**±0.15	**72.26**±0.11	**92.60**±0.24	**84.42**±0.17

Method	ImageNet	ADE20K	COCO (AP^{bb})	COCO (AP^{mk})
SimMIM [39]	83.02±0.09	42.46±0.15	47.06±0.40	46.88±0.44
MAE [14]	82.74±0.15	43.14±0.08	45.86±0.12	41.74±0.15
DiffMAE [38]	82.16±0.28	42.66±0.10	44.18±0.24	38.64±0.21
MaskDiT [42]	82.44±0.15	42.58±0.31	43.16±0.12	40.82±0.18
Ours (ECC)	**83.41**±0.13	**43.54**±0.14	**48.44**±0.38	**47.90**±0.30

All experiments used ViT-B [10] as the backbone architecture applying a unified 400-epoch training schedule. pretraining was performed on the ImageNet-1K [8] classification dataset, followed by fine-tuning on respective downstream task datasets.

5.2 Computational Cost and Training Efficiency

To provide a comprehensive evaluation of the proposed ECC framework, we analyze its computational complexity and empirical training efficiency compared to representative MIM baselines. Our method adopts an Encoder-style architecture, which processes all input tokens—including masked and noised ones—within the encoder to facilitate deep feature-level restoration. Consequently, ECC exhibits a theoretical complexity of 16.94 GFLOPs for a standard 224×224 input, which is higher than the Decoder-style MAE at 11.67 GFLOPs but remains highly comparable to SimMIM at 16.72 GFLOPs. Despite this increase in theoretical GFLOPs, the actual wall-clock training time for a 400-epoch schedule on our 4×A100 GPU configuration is remarkably efficient; ECC requires 55 hours, which is nearly identical to SimMIM (54h) and only marginally higher than MAE (51h). These relatively modest differences in practice suggest that the unified encoder design benefits from hardware-level optimization for dense tensor operations, making the significant performance gains in both fine-grained vision tasks and VLM alignment a justified and scalable trade-off.

5.3 Main Result

In Tab. 1, we evaluate the proposed method on diverse tasks, including fine-grained visual categorization (FGVC), image classification, semantic segmenta-

tion, object detection, and instance segmentation, each with task-specific datasets. FGVC datasets (CUB-200-2011 [36], NABirds [33], iNaturalist 2017 [35], iNaturalist 2018 [34], Stanford Cars [18], Aircraft [25]) demand detailed, fine-grained feature learning to distinguish visually similar classes. In contrast, standard recognition tasks (ImageNet [8]), semantic segmentation (ADE20K [43]), object detection and instance segmentation (COCO [21]) 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 [14, 39] and noise-based methods [38, 42] 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 Experiments on Vision-Language Models

To demonstrate the Systemic Utility and scalability of our proposed principles, we integrate ECC into three multimodal frameworks: CLIP [27], AIM-v2 [11], and LLaVA [22]. To ensure fair comparison, all experiments follow the official settings of each framework. Training hyperparameters also follow the original implementations. All experiments are conducted on 4×A100 GPUs. While our core analysis was conducted on vision-only backbones to establish a robust Base Capability, we hypothesize that the enhanced ability to capture fine-grained details will directly translate into superior multimodal alignment and reasoning.

6.1 Experimental Results

As summarized in Tab. 2, integrating ECC design principles leads to consistent performance gains across diverse architectures and training objectives. These results confirm that the benefits of our optimized C2R framework persist and amplify at scale. The consistent gains suggest that ECC provides a more *visually rich* foundation that allows the multimodal head to distinguish subtle intra-class variations that standard global alignment often overlooks.

Table 2: Effect of ECC on Vision-Language Models. Applying ECC to the vision encoder consistently improves fine-tuning performance across representative VLM architectures.

Model	Backbone	Baseline	+ Ours (ECC)
CLIP [27]	ViT-B/16	83.35	**84.83**
AIM-v2 [11]	ViT-L/14	86.23	**87.11**
LLaVA [22]	ViT-L/14	84.76	**85.42**

Table 3: Comparison of visual grounding sharpness measured by attention entropy (lower is better).

Method	Attention Entropy (bits) ↓
CLIP [27]	4.2
CLIP [27] + Ours (ECC)	**3.5**

6.2 Analysis on Visual Grounding

The performance gains observed in Tab. 2 raise a fundamental question: Why does ECC's corruption design improve multimodal alignment so effectively? To answer this, we investigate how the model's visual lens is 'sharpened' through our encoder-centric principles.

Metric: Attention Entropy We introduce Attention Entropy as a quantitative measure of visual grounding certainty. For a given text prompt, we compute the spatial entropy of the cross-modal attention maps where text tokens attend to L image patches. The entropy H is defined as $H = -\sum_{i=1}^{L} p_i \log p_i$, where p_i is the normalized attention weight for patch i. A lower entropy value indicates a more *focused* attention distribution on discriminative regions.

Results and Interpretation. As illustrated in Tab. 3, our analysis reveals a significant reduction in grounding uncertainty when ECC is integrated: (i) **Sharpness of Grounding**: Baseline CLIP [27] exhibits an average entropy of 4.2 bits, suggesting a diffuse attention pattern. In contrast, CLIP + ECC achieves a much lower entropy of 3.5 bits. This reduction confirms that our principles effectively force the model to ground abstract semantic concepts in high-frequency visual details that are otherwise ignored by standard objectives. (ii) **Mitigating Hallucination**: This enhanced grounding directly correlates with reduced object hallucination. By providing precise visual evidence, ECC ensures that the model's predictions are grounded in actual high-frequency features rather than coarse semantic priors. To directly substantiate the hallucination claim, we evaluate ECC on LLaVA-1.5 with object-hallucination (POPE, CHAIR) and cross-modal retrieval (Flickr30K) metrics (Tab. 4). Built atop an already-strong CLIP ViT-L/14 backbone, ECC lowers both CHAIR_s (48.4→46.5) and CHAIR_i (14.0→12.9) while improving POPE-F1 (85.9→86.6) and Flickr R@1

Table 4: Direct hallucination and cross-modal alignment on LLaVA-1.5. Applying ECC to the vision encoder reduces object hallucination (CHAIR) and improves grounding (POPE-F1) and image–text retrieval (Flickr30K R@1).

Model	POPE-F1 ↑	CHAIR_i ↓	CHAIR_s ↓	Flickr R@1 ↑
LLaVA-1.5	85.9	14.0	48.4	69.1
LLaVA-1.5 + Ours (ECC)	86.6	12.9	46.5	70.2

(69.1→70.2). These results confirm that the sharpened, high-frequency grounding induced by ECC translates into measurably reduced object hallucination and stronger image–text alignment, beyond vision-only recognition.

7 Conclusion

We address the granularity mismatch in modern VLMs, where global alignment often incurs a low-frequency bias and object hallucination. While image masking and additive noise theoretically provide structural locality and high-frequency textures, we identify that prior noise-driven extensions fail to yield transferable representations due to suboptimal Decoder-style designs. To bridge this gap, we propose Encoder-Centric Corruption (ECC), a framework governed by three principles: (i) Encoder-Centricity for direct feature shaping , (ii) Feature-Level Noising to capture fine structures, and (iii) Task Disentanglement via a disruption loss to mitigate interference. ECC surpasses prior MIM and noise-based methods across diverse benchmarks. Furthermore, integrating ECC into large-scale frameworks—including CLIP, AIM-v2, and LLaVA—consistently enhances multimodal alignment and fine-tuning accuracy. These results validate ECC as an efficient, scalable pretraining strategy for next-generation foundation models.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2025-24803204 and RS-2026-25498839); by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2026-RS-2026-25546026, Leading Generative AI Human Resources Development); and by the IITP-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2026-RS-2020-II201602, 20%).

References

1. Alayrac, J.B., Donahue, J., Luc, P., et al.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: 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)
3. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: International Conference on Learning Representations (2021)
4. Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J.: Context autoencoder for self-supervised representation learning. International Journal of Computer Vision **132**(1), 208–223 (2024)
5. Chen, X., Liu, Z., Xie, S., He, K.: Deconstructing denoising diffusion models for self-supervised learning. arXiv preprint arXiv:2401.14404 (2024)
6. Choi, H., Lee, H., Joung, S., Park, H., Kim, J., Min, D.: Emerging property of masked token for effective pre-training. arXiv preprint arXiv:2404.08330 (2024)
7. Choi, H., Park, H., Yi, K.M., Cha, S., Min, D.: Salience-based adaptive masking: revisiting token dynamics for enhanced pre-training. In: European Conference on Computer Vision. pp. 343–359. Springer (2025)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Bootstrapped masked autoencoders for vision bert pretraining. In: European Conference on Computer Vision. pp. 247–264. Springer (2022)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Fini, E., Shukor, M., Li, X., Dufter, P., Klein, M., Haldimann, D., Aitharaju, S., da Costa, V.G.T., Béthune, L., Gan, Z., et al.: Multimodal autoregressive pre-training of large vision encoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9641–9654 (2025)
12. Geirhos, R., et al.: Imagenet-trained cnns are biased towards texture. In: ICLR (2019)
13. Goyal, S., et al.: Object hallucination in vision-language models. arXiv (2023)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Hoogeboom, E., Mensink, T., Heek, J., Lamerigts, K., Gao, R., Salimans, T.: Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. arXiv preprint arXiv:2410.19324 (2024)
17. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems **35**, 26565–26577 (2022)

18. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (ICML) (2023)
20. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (ICML) (2022)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: 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)
22. Liu, H., et al.: Visual instruction tuning. In: NeurIPS (2023)
23. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
24. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: 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)
25. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
26. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021)
27. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
28. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
30. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
31. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
32. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
33. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: 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)
34. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)

35. Van Horn, G., Mac Aodha, O., Song, Y., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist challenge 2017 dataset. arXiv preprint arXiv:1707.06642 **1**(2), 4 (2017)
36. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
37. Wang, H., et al.: High-frequency components help explain the generalization of convolutional neural networks. In: CVPR (2020)
38. Wei, C., Mangalam, K., Huang, P.Y., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A., Feichtenhofer, C.: Diffusion models as masked autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16284–16294 (2023)
39. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: 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)
40. Yi, K., Ge, Y., Li, X., Yang, S., Li, D., Wu, J., Shan, Y., Qie, X.: Masked image modeling with denoising contrast. arXiv preprint arXiv:2205.09616 (2022)
41. Yu, J., Wang, X., Zhang, W., et al.: Coca: Contrastive captioners are image-text foundation models. In: International Conference on Machine Learning (ICML) (2022)
42. Zheng, H., Nie, W., Vahdat, A., Anandkumar, A.: Fast training of diffusion models with masked transformers. arXiv preprint arXiv:2306.09305 (2023)
43. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)