Exposure Bias Can Alleviate Itself via Directional and Frequency Rectification in Flow Matching

Guanbo Huang[1,2*,‡], Jingjia Mao[1,2*,‡], Fanding Huang[1*], Fengkai Liu[1], Xiangyang Luo[1], Yaoyuan Liang[1], Jiasheng Lu[2], Xiaoe Wang[2], Pei Liu[2], Ruiliu Fu[2†], Ruqi Huang[1†], and Shao-Lun Huang[1†]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2] Central Media Technology Institute, Huawei, China

Abstract. Flow Matching (FM) has achieved remarkable generative performance, yet it suffers from exposure bias due to discrepancies between training and inference. Existing mitigation strategies typically rely on static constraints or external heuristics. In this work, we propose that exposure bias itself inherently contains dynamic signals that can guide its own rectification. To leverage this, we introduce **DEFAR** (**D**ir**E**ctional-**F**requency **A**daptive **R**ectification). This framework simulates the single-step inference process during training to identify exposure bias. It utilizes the directional and frequency adaptive feedback signals within bias itself to enhance the bias tolerance of the model. It consists of two key components: (1) **Anti-Drift Rectification (ADR)**. ADR treats inference-time drift as a signal to learn the direction to steer deviated states back toward the target. ADR endows the model with intrinsic active self-rectification capabilities; (2) **Frequency Compensation (FC)**. Empirically, we observe that accumulated bias often stems from a lack of low-frequency components in high-noise stages and exposure bias carries the missing frequency information. FC leverages the bias itself as a self-feedback weighting factor to reinforce the missing frequency components. Experiments on CIFAR-10, CelebA-64, and ImageNet-256/512 show that DEFAR outperforms prior baselines and further demonstrates favorable scalability, compatibility, and inference robustness. Code will be made available in `https://github.com/wuliwuliy/DEFAR`.

Keywords: Exposure bias · Flow matching · Adaptive rectification

1 Introduction

Recently, Flow Matching (FM) [1,33,34] has emerged as a promising framework for generative modeling. By modeling continuous-time flows, FM offers a more direct and efficient training paradigm compared to traditional diffusion-based approaches [16,48,49]. Its strong empirical performance and theoretical simplicity

* Equal contributions.
† Corresponding authors: Ruiliu Fu, Ruqi Huang and Shao-Lun Huang.
‡ This work was conducted during the internship at Central Media Technology Institute, Huawei.

have established FM as a foundational method across diverse generation tasks, including image, video, and audio generation [4, 9, 20, 25, 31, 38, 51, 54, 59, 69].

However, FM models remain inherently vulnerable to exposure bias, a fundamental challenge stemming from the discrepancies between training and inference inputs. As illustrated in Fig. 1, during training, the model is conditioned on perturbed input, namely a mixture of Gaussian noise and ground-truth data. The model effectively learns to predict the vector field without predicted error accumulation. In contrast, during multi-step inference, the model must exclusively rely on its own previous predictions. This propagates and accumulates the potential predicted error at each inference step. This forces the model to operate in the unexplored data space compared to the training stage where its predictive capability significantly degrades, ultimately leading to substantial deviation from the target distribution.

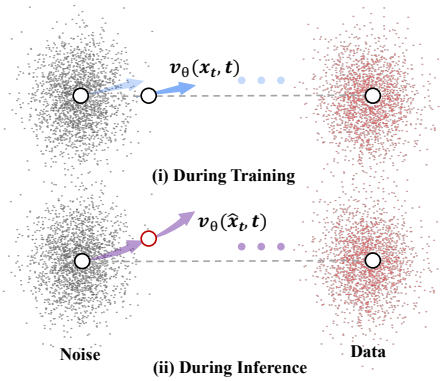

Fig. 1: Illustration of exposure bias in Flow Matching. (i) During training, the model is conditioned on perturbed inputs sampled strictly along the ideal linear path (blue arrows). (ii) During inference, the input suffers from accumulated errors generated by previous steps, causing training-inference mismatch (purple arrows).

While prior research has explored this issue within the DDPM framework [11, 27, 29, 40, 62], existing solutions predominantly rely on static alignment strategies or exogenous noise injection. Some approaches introduce inference-time regularization without altering the training process [27, 40, 62], while others attempt to mitigate bias during training by simulating potentially imperfect inference. For instance, IP [41] injects fixed perturbations into training input data, and SS [7] directly replaces perturbed samples with predicted samples. These methods generally enforce a static alignment with the original ground truth or rely on externally fixed noise augmentation to improve robustness. Notably, [43] proposes Multi-step Denoising Scheduled Sampling (MDSS) and Single-step Denoising Scheduled Sampling (SDSS), which focus on passively enhancing the model's robustness to input deviations. By exposing the model to perturbed states, MDSS encourages stability against accumulated errors but generally enforces a fixed alignment target regardless of the deviation's severity. In this work, we advance beyond passive robustness to establish a dynamic correction mechanism. We empower the model to learn adaptive correction signals that are modulated by the severity of exposure bias.

Specifically, we investigate the intrinsic properties of exposure bias from both directional and frequency perspectives. **First**, we simulate the inference process during training to expose drift. After inference, the exposure bias is inherently carried in the inferred state. This simulation enables us to explicitly character-

ize the deviation from the state to the target endpoint. Building on this, we introduce a directional regularizer. This term acts as a steering force, guiding the model to rectify the deviated state back towards the target endpoint. This allows us to actively adjust the rectification magnitude according to the severity of the exposure bias itself. **Second**, identifying that exposure bias stems from prediction imperfection, we conduct a frequency analysis and observe a distinct deficiency in low-frequency components during the early denoising outputs from the model. Since these components encode critical structural information, their absence exacerbates error propagation. To quantify this, we introduce two metrics: Predicted Frequency Ratio (PFR), which measures the ratio of low-to-high frequency energy in prediction outputs, and Frequency Emphasis of Loss (FEL), which evaluates the relative focus of the loss function on different frequency bands. Intriguingly, we observe that exposure bias itself exhibits a complementary characteristic in relatively early high-noise timesteps: it naturally highlights neglected low-frequency regions. This motivates us to use the bias as a dynamic reweighting signal to mitigate itself.

Building on these insights, we propose **DEFAR** (**D**ir**E**ctional-**F**requency **A**daptive **R**ectification), a flexible framework that leverages the dynamic nature of exposure bias itself to adaptively rectify FM models. DEFAR comprises two core components: (1) Anti-Drift Rectification (ADR): Functioning as a regularizer to the original training objective, ADR constructs a direction that explicitly steers the model from the deviated state back towards the target data distribution, actively correcting the prediction drift based on the severity of exposure bias itself. (2) Frequency Compensation (FC): Applied to the original training objective, FC utilizes the exposure bias itself as a reweighting signal to enhance the learning of missing low-frequency components via a negative feedback loop. Our main contributions are summarized as follows:

- We propose Anti-Drift Rectification (ADR), a method that actively provides directional correction, endowing flow matching with superior resistance to bias compared to static alignment methods.
- We identify the connection between exposure bias and missing low-frequency information, proposing Frequency Compensation (FC) to dynamically reinforce structural learning by leveraging the bias signal itself.
- We integrate ADR and FC into a unified framework, DEFAR. Extensive experiments on CIFAR-10, ImageNet-256/512, and CelebA-64 demonstrate that DEFAR significantly outperforms existing baselines, while exhibiting scalability, architectural compatibility, and inference robustness.

2 Related Work

Exposure Bias. Exposure bias arises from the mismatch between training and inference inputs in sequential models [3, 42, 46, 55, 67, 68]. Among early approaches, DAD [53] combines ground truth and predicted tokens during training, and SS [7] samples biased inputs to better approximate inference. In diffusion

models, DDPM-IP [41] first formalizes exposure bias and perturbs training samples to simulate it. EP-DDPM [29] studies error propagation and introduces a cumulative-error regularizer. AE-DDPM [65] mitigates exposure bias via prompt learning, and MCDO [61] imposes manifold constraints. Inspired by distribution adaptation methods [18,19,66], training-free techniques include a time-shift sampler (TS-DDPM [27]), noise scaling (ES-DDPM [40]), training-distribution leak signal exploitation [11] and frequency-domain analysis with wavelet-based regularization [62]. Among prior works, MDSS/SDSS [43] is most relevant to our approach. It simulates multi- or single-step inference trajectories during training but persists in aligning predictions with the static ground-truth target. Consequently, this strategy merely bolsters the model's passive robustness against input perturbations, failing to equip it with the capability to actively rectify accumulated errors. In contrast, we investigate the intrinsic signal properties of exposure bias in Flow Matching and introduce a novel paradigm that leverages the bias itself to drive active self-rectification.

Reweighting objectives for emphasizing informative signals. Several works explore reweighting strategies in losses or objectives to emphasize salient or informative signals, guiding models to focus on the most influential components. In decomposable or multi-stage generation, Decomposable Flow Matching [14] applies spatially varying masks to highlight critical regions on outputs. In video generation, Proteus-ID [64] and MotiF [56] adopt motion-aware weighting to prioritize frequently moving regions. MotionCharacter [12] emphasizes motion-sensitive areas, while MOCA [5] models object consistency for 3D camera control. Beyond spatial cues, frequency or modality modeling, representation alignment, retrieval or ranking guidance, and reward signaling redirect learning toward informative components [10, 26, 28, 35, 47, 60].

3 Preliminaries

3.1 Flow Matching

We adopt the standard Flow Matching formulation [33], where time $t \in [0, 1]$ drives the flow from the source noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at $t = 0$ to the target data $\mathbf{x}_* \sim p_{\text{data}}(\mathbf{x})$ at $t = 1$. A sample at time t is constructed as

$$\mathbf{x}_t = a_t \mathbf{x}_* + b_t \boldsymbol{\epsilon}, \tag{1}$$

where a_t and b_t are predefined schedules. The corresponding conditional velocity is given by the time derivative of the path: $\mathbf{v}_t = a'_t \mathbf{x}_* + b'_t \boldsymbol{\epsilon}$, where $a'_t = \frac{da_t}{dt}$ and $b'_t = \frac{db_t}{dt}$. FM trains a neural network $\mathbf{v}_\theta(\mathbf{x}, t)$ to approximate this conditional velocity by minimizing

$$\mathcal{L}_{\text{FM}}(\theta) := \mathbb{E}_{\mathbf{x}_*, \boldsymbol{\epsilon}, t}\left[\left\| \mathbf{v}_\theta(\mathbf{x}_t, t) - (a'_t \mathbf{x}_* + b'_t \boldsymbol{\epsilon}) \right\|^2 \right]. \tag{2}$$

This objective encourages the model to learn a time-dependent velocity field that consistently transports the intermediate distribution towards the true data

distribution. For instance, to ensure the flow follows Optimal Transport (OT), we can employ the straight-line schedule $a_t = t$ and $b_t = 1 - t$, which simplifies the target velocity to a constant $\mathbf{x}_* - \boldsymbol{\epsilon}$ along the linear path.

3.2 Fourier Frequency Analysis

The Fourier transform provides a way to represent spatial-domain signals in the frequency domain, decomposing them into sinusoidal components of different frequencies and amplitudes. Since the predicted velocity field $\mathbf{v} \in \mathbb{R}^{H \times W}$ shares a similar spatial structure with images, we can directly apply frequency-domain analysis to it. The 2D Discrete Fourier Transform (DFT) of $\mathbf{v}$ is defined as:

$$\mathbf{V}(u, v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \mathbf{v}(x, y)\, e^{-i2\pi\left(\frac{ux}{H} + \frac{vy}{W}\right)}, \tag{3}$$

where (u, v) denotes the frequency coordinates, and $i = \sqrt{-1}$ is the imaginary unit. The resulting $\mathbf{V}(u, v)$ consists of a real part $R(\mathbf{V})$ and an imaginary part $I(\mathbf{V})$, from which the amplitude and phase spectra can be derived as:

$$\begin{aligned} |\mathbf{V}(u, v)| &= \sqrt{R(\mathbf{V})^2 + I(\mathbf{V})^2}, \\ \angle\mathbf{V}(u, v) &= \arctan\left(\frac{I(\mathbf{V})}{R(\mathbf{V})}\right). \end{aligned} \tag{4}$$

The inverse Fourier transform reconstructs the spatial-domain signal from its frequency representation:

$$\mathbf{v}(x, y) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \mathbf{V}(u, v)\, e^{i2\pi\left(\frac{ux}{H} + \frac{vy}{W}\right)}. \tag{5}$$

In practice, we adopt the Fast Fourier Transform (FFT) and its Inverse (IFFT) for efficient computation.

3.3 Exposure Bias Problem Formulation

Exposure bias is a well-documented phenomenon in generative sequence modeling [3, 42] and has been recently investigated within diffusion-based frameworks [41, 43]. In the context of Flow Matching (FM), this issue manifests as a critical discrepancy between the training and inference phases. Specifically, during training, the model $\mathbf{v}_\theta(\mathbf{x}_t, t)$ relies on inputs $\mathbf{x}_t$ derived from a linear interpolation between the true data distribution $\mathbf{x}_*$ and Gaussian noise $\boldsymbol{\epsilon}$ (Eq. 1). Conversely, during inference, the input $\hat{\mathbf{x}}_t$ is generated recursively from previous predictions, inevitably accumulating errors. This mismatch induces a distribution drift between the training state $\mathbf{v}_\theta(\mathbf{x}_t, t)$ and the inference state $\mathbf{v}_\theta(\hat{\mathbf{x}}_t, t)$, which propagates through timesteps and progressively degrades sample quality.

To explicitly model this effect during the training phase, we simulate the generation process to capture the resulting drift. As illustrated in Fig. 2, we

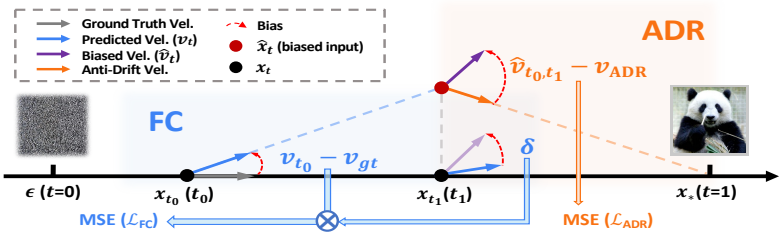

Fig. 2: Overview of DEFAR. (a) **Anti-Drift Rectification**: Introduces a learning target that actively guides the model from the drift-affected distribution back toward the data distribution. (b) **Frequency Compensation**: Reweights the original objective using exposure bias as a negative-feedback signal to mitigate low-frequency deficiency at relatively high-noise timesteps (e.g., t_0). Together, DEFAR adaptively rectifies exposure bias across both directional and frequency dimensions by leveraging the bias signal itself.

incorporate a one-step inference simulation in the training loop. At time t_0, given the ideal perturbed input $\mathbf{x}_{t_0}$, the model predicts the velocity $\mathbf{v}_\theta(\mathbf{x}_{t_0}, t_0)$. Due to unavoidable prediction errors, the estimated velocity naturally deviates from the ground-truth direction. Proceeding to the next timestep t_1, the resulting state is computed as:

$$\hat{\mathbf{x}}_{t_0,t_1} = \mathbf{x}_{t_0} + (t_1 - t_0) \cdot \mathbf{v}_\theta(\mathbf{x}_{t_0}, t_0). \tag{6}$$

This estimated state carries the prediction drift. Consequently, the subsequent velocity $\mathbf{v}_\theta(\hat{\mathbf{x}}_{t_0,t_1}, t_1)$ exacerbates the deviation from the previous step.

Based on this simulation, we operationalize exposure bias at a single timestep to guide our training. We formulate the bias for any timestep interval t_0, t_1 as:

$$\boldsymbol{\delta}_{t_0,t_1} = \mathbf{v}_\theta(\hat{\mathbf{x}}_{t_0,t_1}, t_1) - \mathbf{v}_\theta(\mathbf{x}_{t_1}, t_1). \tag{7}$$

This formulation quantifies the variation in predicted velocity caused solely by replacing the ideal input $\mathbf{x}_{t_1}$ with the drift-affected input $\hat{\mathbf{x}}_{t_0,t_1}$, serving as a tractable proxy for the actual inference discrepancy.

4 DirEctional-Frequency Adaptive Rectification

This section presents our method, **DEFAR**, a framework designed to adaptively mitigate exposure bias by leveraging the intrinsic directional and frequency signals of the bias itself. As illustrated in Fig. 2, our approach comprises two core components: *Anti-Drift Rectification* (ADR) and *Frequency Compensation* (FC), detailed in Sec. 4.1 and Sec. 4.3 respectively, with Sec. 4.2 providing the metrics used for the subsequent frequency analysis. ADR operates in the directional dimension, actively learning a restorative direction to steer drifted states back toward the target endpoint based on the severity of the bias. In parallel, FC functions in the frequency dimension, leveraging the frequency signature of the bias as a self-feedback cue to adaptively compensate for deficient low-frequency components during high-noise timesteps. By integrating them into the unified **DEFAR** framework, these two mechanisms complement each other, bolstering

the bias tolerance of the model to training-inference mismatch from both directional and frequency dimensions. The learning objective combines the two mechanisms as follows:

$$\mathcal{L} := \beta_1 \mathcal{L}_{\text{ADR}} + \beta_2 \mathcal{L}_{\text{FC}}, \tag{8}$$

where β_1 and β_2 are weighting coefficients for ADR and FC, respectively. $\mathcal{L}_{\text{FC}}$ represents the $\mathcal{L}_{\text{FM}}$ reweighted by exposure bias signal. DEFAR thus offers an effective solution to alleviate exposure bias by exploiting the informative signals inherent in the bias itself. The training procedure is summarized in Algorithm 1.

4.1 Anti-Drift Rectification (ADR)

As formulated in Sec. 3.3, exposure bias emerges from the recursive accumulation of prediction errors, integrating drift into $\hat{\mathbf{x}}_{t_0,t_1}$. When existing Flow Matching frameworks process such biased states, they typically impose a static optimization objective [43], which lacks sensitivity to the dynamically varying severities of prediction drift. To empower the model with active resistance, we introduce Anti-Drift Rectification (ADR), a mechanism that turns the exposure bias into an endogenous control signal to mitigate itself.

Our fundamental motivation is grounded in the invariance of the target data distribution. Regardless of the severity of the accumulated exposure bias, the ground-truth endpoint $\mathbf{x}_*$ remains deterministic. This invariance allows us to establish a dynamic anchor: the anti-drift target velocity $\mathbf{v}_{\text{ADR}} = a'_{t_1}\mathbf{x}_* + b'_{t_1}\hat{\mathbf{x}}_{t_0,t_1}$. This vector acts as a restorative trajectory, originating strictly from the current biased state and guiding toward the target distribution. By leveraging this anchor, the model is empowered to actively recognize the precise corrective direction required to navigate back to the target manifold, adapting seamlessly to varying degrees of exposure bias.

To enable the model to learn this rectifying direction, we introduce ADR as a regularization term appended to the standard FM objective (which will be further upgraded to our FC objective in Sec. 4.3). Notably, when no exposure bias occurs, $\mathbf{v}_{\text{ADR}}$ naturally degenerates to align parallel to $a'_{t_1}\mathbf{x}_* + b'_{t_1}\boldsymbol{\epsilon}$. This guarantees consistency with the original probability flow, ensuring that the regularization only actively penalizes trajectory drift without perturbing optimal predictions. The ADR regularization term is formulated as follows:

$$\mathcal{L}_{\text{ADR}} = \mathbb{E}_{\mathbf{x}_*,\hat{\mathbf{x}}_{t_0,t_1},t_0,t_1}\left[\left\| \frac{\mathbf{v}_\theta(\hat{\mathbf{x}}_{t_0,t_1},t_1)}{\|\mathbf{v}_\theta(\hat{\mathbf{x}}_{t_0,t_1},t_1)\|_2} - \frac{a'_{t_1}\mathbf{x}_* + b'_{t_1}\hat{\mathbf{x}}_{t_0,t_1}}{\|a'_{t_1}\mathbf{x}_* + b'_{t_1}\hat{\mathbf{x}}_{t_0,t_1}\|_2} \right\|^2\right]. \tag{9}$$

This design serves as a dynamic self-rectifying mechanism: the regularization gradient is intrinsically proportional to the angle of deviation. As the exposure bias exacerbates the drift, the angular discrepancy widens, automatically amplifying the rectification signal. Through this dynamic modulation, ADR forces the model to actively learn a restorative direction, thereby alleviating sampling error and dismantling the cycle of recursive error propagation, thus enabling the model to leverage the exposure bias signal to mitigate the bias itself.

4.2 Frequency Analysis Metrics

Recent studies [27,40,62] attribute exposure bias to the distributional mismatch between training and inference predictions. While ADR effectively rectifies the *geometric direction* of this drift, the *internal structure* of exposure bias, specifically its frequency characteristics, remains underexplored. In this section, we analyze this pixel-level frequency mismatch and propose a fine-grained compensation strategy derived from the exposure bias itself.

Inspired by the frequency analyses in [32, 50], we introduce two metrics to quantify the frequency distribution of velocity and loss for further analysis: the Predicted Frequency Ratio (PFR), which measures the ratio of low-to-high frequency energy in prediction, and Frequency Emphasis of Loss (FEL), which evaluates the relative focus of the loss function on different frequency bands.

We first introduce the computation of PFR. Specifically, given the predicted velocity $\mathbf{v} \in \mathbb{R}^{H\times W}$ produced by the FM model, we perform a 2D FFT on it:

$$\mathbf{V}(u,v) = \mathrm{FFT}(\mathbf{v}(i,j)). \tag{10}$$

To investigate the proportion of low- and high-frequency bands, we apply Low-Pass Filter ($\mathbf{F}_{LP}$) and High-Pass Filter ($\mathbf{F}_{HP}$) defined as:

$$\mathbf{F}_{\mathrm{LP}}(u,v) = \mathbb{I}\left[D(u,v) \le D_{\mathrm{cutoff}}\right], \quad \mathbf{F}_{\mathrm{HP}}(u,v) = \mathbf{1} - \mathbf{F}_{\mathrm{LP}}(u,v), \tag{11}$$

where $D(u,v) = \sqrt{u^2+v^2}$ is the frequency magnitude and D_{cutoff} is the cutoff frequency. We compute the energy ratio between the low- and high-frequency components and define PFR as follows:

$$\mathrm{PFR} := \frac{\sum_{u,v} ||\mathbf{V}(u,v)||^2 \cdot \mathbf{F}_{\mathrm{LP}}(u,v)}{\sum_{u,v} ||\mathbf{V}(u,v)||^2 \cdot \mathbf{F}_{\mathrm{HP}}(u,v)}. \tag{12}$$

We then define the FEL. Here, we consider the target velocity map $\mathbf{v}_{\mathrm{target}} \in \mathbb{R}^{H\times W}$ (typically the standard FM target velocity) and the MSE loss map $\boldsymbol{\mathcal{L}} \in \mathbb{R}^{H\times W}$. To compute it, we divide $\mathbf{v}_{\mathrm{target}}$ into high- and low-frequency regions. The magnitude of the loss map $\boldsymbol{\mathcal{L}}$ in each region reflects the relative emphasis of loss. Thus, we separately sum the loss values within the low- and high-frequency regions to quantify how the loss distributes its emphasis across different frequency components. This metric further serves as an indicator of the capability of loss to compensate for frequency discrepancies.

We first apply Eq. 10 and 11 to perform the FFT of $\mathbf{v}_{\mathrm{target}}$ and compute the corresponding low- and high-frequency masks in the frequency domain. These masks are then used to separate $\mathbf{v}_{\mathrm{target}}$ into its low- and high-frequency components, which are subsequently transformed back to the spatial domain via IFFT. Finally, we compute the salient low- and high-frequency region masks in the spatial domain by thresholding the corresponding values according to their percentile distributions.

$$\begin{aligned} \mathbf{v}_{\mathrm{low}} &= \mathrm{IFFT}(\mathbf{V}\cdot\mathbf{F}_{\mathrm{LP}}), \quad \mathbf{v}_{\mathrm{high}} = \mathrm{IFFT}(\mathbf{V}\cdot\mathbf{F}_{\mathrm{HP}}), \\ \mathbf{M}_{\mathrm{LFR}}(i,j) &= \mathbb{I}\big[\mathbf{v}_{\mathrm{low}} > p(\mathbf{v}_{\mathrm{low}})\big], \quad \mathbf{M}_{\mathrm{HFR}}(i,j) = \mathbb{I}\big[\mathbf{v}_{\mathrm{high}} > p(\mathbf{v}_{\mathrm{high}})\big], \end{aligned} \tag{13}$$

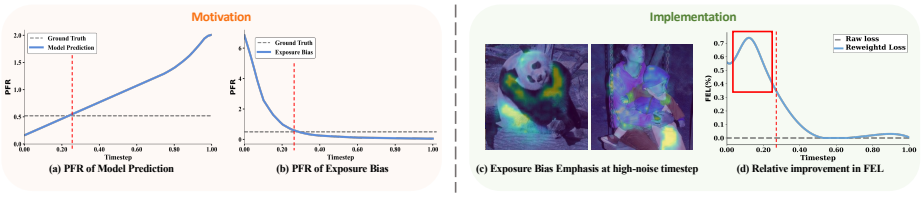

Fig. 3: Motivation and Verification of FC. (a) illustrates frequency trends derived from forward-perturbed inputs, while (b) reveals contrasting trends using inputs generated via single-step inference. In (b), for any starting timestep t_0, the exposure bias PFR is averaged over all subsequent sampled timesteps $t_1 \in (t_0, 1]$, capturing the cumulative impact across varying inference intervals. (c) visualizes the heatmap of exposure bias at high-noise timesteps, focusing on low-frequency semantics. (d) shows that the reweighted loss effectively emphasizes low-frequency regions during high-noise timesteps. The continuous time horizon $[0, 1]$ is discretized into 50 uniform timesteps. The metrics are averaged across $50,000$ samples from ImageNet-256 on SiT-B/4.

where $p(\cdot)$ denotes a percentile threshold function, $\mathbf{v}_{\text{low}}$ and $\mathbf{v}_{\text{high}}$ represent the low- and high-frequency components of $\mathbf{v}_{\text{target}}$. The resulting masks, $\mathbf{M}_{\text{LFR}}$ and $\mathbf{M}_{\text{HFR}}$, indicate the low- and high-frequency region masks, respectively. Finally, we compute FEL as follows:

$$\text{FEL} := \frac{\sum \tilde{\boldsymbol{\mathcal{L}}} \cdot \mathbf{M}_{\text{LFR}}}{\sum \tilde{\boldsymbol{\mathcal{L}}} \cdot \mathbf{M}_{\text{HFR}}}, \quad \text{where} \quad \tilde{\boldsymbol{\mathcal{L}}}^{(i,j)} = \frac{\boldsymbol{\mathcal{L}}^{(i,j)}}{\sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{\mathcal{L}}^{(i,j)}}. \tag{14}$$

The trend of FEL reveals how the loss distributes attention across low- and high-frequency regions during training.

4.3 Frequency Compensation (FC)

Leveraging the PFR metric proposed in Sec. 4.2, we analyze the forward perturbation process and identify that prediction drift is primarily attributed to a frequency deficiency, which subsequently induces exposure bias.

Exposure bias intrinsically compensates for the low-frequency deficiency during high-noise timesteps. As visualized in Fig. 3 (a), the model prediction exhibits a distinct frequency shift over time. During the early high-noise steps ($t \to 0$), the prediction is characterized by a lack of low-frequency content (indicated by low PFR values) compared to the target $\mathbf{v}_{\text{target}}$. Conversely, Fig. 3 (b) reveals that the exposure bias follows an inverse trend, exhibiting high PFR values in this phase. This uncovers a complementary relationship: the exposure bias encapsulates the low-frequency structural components that the model inherently struggles to capture initially.

Exposure bias serves as a dynamic and self-decaying corrective signal. When approaching low-noise ($t \to 1$), the model progressively reconstructs sufficient low-frequency content. Concurrently, the low-frequency dominance within the exposure bias autonomously diminishes. This prevents excessive structural compensation, transforming the exposure bias from a static error into a dynamically weighted variable for self-correction.

Algorithm 1 DEFAR: Training

Input: model f to predict velocity with parameters θ, dataset $\mathcal{D}$, learning rate η, ADR loss weight β_1, FC loss weight β_2, normalization coefficient α, path interpolation coefficients a_t, b_t (with derivatives a'_t, b'_t) and stability constant ξ.
While θ not converged **do**
  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, $\mathbf{x}_* \sim \mathcal{D}$, $t_0, t_1 \in [0, 1]$, and $t_1 > t_0$
  $\mathbf{x}_{t_0} \leftarrow t_0 \mathbf{x}_* + (1 - t_0)\boldsymbol{\epsilon}$, $\mathbf{x}_{t_1} \leftarrow t_1 \mathbf{x}_* + (1 - t_1)\boldsymbol{\epsilon}$
  $\mathbf{v}_{t_0} \leftarrow f_\theta(\mathbf{x}_{t_0}, t_0), \mathbf{v}_{t_1} \leftarrow \text{StopGradient}(f_\theta(\mathbf{x}_{t_1}, t_1))$
  $\hat{\mathbf{x}}_{t_0,t_1} \leftarrow \mathbf{x}_{t_0} + (t_1 - t_0)\mathbf{v}_{t_0}$ // Single-step inference simulation
  $\hat{\mathbf{v}}_{t_0,t_1} \leftarrow f_\theta(\hat{\mathbf{x}}_{t_0,t_1}, t_1)$
  $\mathbf{v}_{\text{ADR}} \leftarrow a'_{t_1} \mathbf{x}_* + b'_{t_1} \hat{\mathbf{x}}_{t_0,t_1}$ // ADR target velocity
  $\mathcal{L}_{\text{ADR}} \leftarrow \mathbb{E}\left[\left\| \frac{\hat{\mathbf{v}}_{t_0,t_1}}{\|\hat{\mathbf{v}}_{t_0,t_1}\|_2} - \frac{\mathbf{v}_{\text{ADR}}}{\|\mathbf{v}_{\text{ADR}}\|_2} \right\|^2\right]$ // ADR regularization loss
  $\boldsymbol{\delta}_{t_0,t_1} \leftarrow \hat{\mathbf{v}}_{t_0,t_1} - \mathbf{v}_{t_1}$
  $\mathbf{W}^{(i,j)}_{t_0,t_1} \leftarrow 1 + \alpha \frac{\|\boldsymbol{\delta}^{(i,j)}_{t_0,t_1}\|^2}{\sum_{i=1}^{H}\sum_{j=1}^{W} \|\boldsymbol{\delta}^{(i,j)}_{t_0,t_1}\|^2 + \xi}$ // Spatial frequency weights
  $\mathcal{L}_{\text{FC}} \leftarrow \mathbb{E}\left[\left\|\mathbf{W}_{t_0,t_1} \cdot \left(\mathbf{v}_{t_0} - (a'_{t_0}\mathbf{x}_* + b'_{t_0}\boldsymbol{\epsilon})\right)\right\|^2\right]$ // FC-reweighted FM loss
  $\mathcal{L} \leftarrow \beta_1 \mathcal{L}_{\text{ADR}} + \beta_2 \mathcal{L}_{\text{FC}}$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$

The low-frequency content of exposure bias is semantically grounded in the target data. To verify that this complementary signal is not noise, we visualize the spatial distribution of the exposure bias in Fig. 3 (c). The heatmaps confirm that during the high-noise timesteps, the bias is concentrated on the low-frequency structural regions of the original images (More samples in Appendix).

These observations motivate our core insight: the dynamic frequency signatures of exposure bias constitute an endogenous feedback loop. This empowers the model to rectify low-frequency deficiencies using the bias signal, thereby reducing prediction errors and ultimately mitigating the bias itself.

Inspired by [12, 47, 56, 64], which demonstrate that reweighting the loss can enhance saliency alignment, we design a negative-feedback weight mask for the original loss based on exposure bias. The weight is formally defined as:

$$\mathbf{W}^{(i,j)}_{t_0,t_1} = 1 + \alpha \frac{\|\boldsymbol{\delta}^{(i,j)}_{t_0,t_1}\|^2}{\sum_{i=1}^{H}\sum_{j=1}^{W} \|\boldsymbol{\delta}^{(i,j)}_{t_0,t_1}\|^2 + \xi}, \tag{15}$$

where α is a scaling coefficient, and ξ is a small constant introduced to ensure numerical stability. We then reweight the FM learning objective as follows:

$$\mathcal{L}_{\text{FC}} := \mathbb{E}_{\mathbf{x}_*, \boldsymbol{\epsilon}, t_0, t_1}\left[\left\|\mathbf{W}_{t_0,t_1} \cdot \left(\mathbf{v}_\theta(\mathbf{x}_{t_0}, t_0) - (a'_{t_0}\mathbf{x}_* + b'_{t_0}\boldsymbol{\epsilon})\right)\right\|^2\right]. \tag{16}$$

This weighting strategy adaptively adjusts the loss distribution according to the exposure bias, emphasizing the missing frequency components.

Exposure bias reweighting prioritizes low-frequency learning during high-noise timesteps. To quantitatively evaluate the impact of $\mathcal{L}_{\text{FC}}$, we analyze the FEL. Prior studies [2, 17, 22, 30, 37, 58, 70] indicate that the early generation phase, especially high-noise timesteps, is critical for establishing the global

structure of generated samples, which directly influences the final generation quality. Consistent with these findings, as illustrated in Fig. 3 (d), the exposure-bias-reweighted loss exhibits a relative increase over the raw baseline, directing the learning focus toward low-frequency components during the high-noise phase ($t \to 0$). By utilizing the bias signal to compensate for frequency deficiencies, FC translates the exposure bias into a corrective signal, thereby leveraging the bias to mitigate its underlying cause.

5 Experiments

5.1 Experiment Setting

Main Implementation. We evaluate our methods on conditional image generation tasks on ImageNet-256/512 [6] and CIFAR-10 [23], as well as unconditional tasks on CIFAR-10 and CelebA-64 [36]. Following prior works [13, 41], we generate 50K samples using 50 Number of Function Evaluations (NFE). CIFAR-10 and CelebA-64 are processed in pixel space, while ImageNet-256 is encoded in the latent space of a pre-trained VAE-ft-EMA tokenizer [44]. All models are trained from scratch on Ascend 910B NPUs, except that REPA, DDT and OT-CFM experiments are conducted on A100 GPUs. Based on grid search results (detailed in Appendix), we set $\beta_1 = 10$, $\beta_2 = 1$ and $\alpha = 1$ in all experiments. During training, all variants of our methods randomly select a pair of timesteps (t_0, t_1) with $t_1 > t_0$ in each iteration. For ablation studies, we adopt SiT-B/4 as the default backbone and $a_t = t, b_t = 1 - t$ unless stated otherwise. For Discriminator Guidance (DG), we follow [21] and train a shallow U-Net discriminator for the SiT-B/4 score model with an ADM classifier. For mini-batch OT coupling, we follow [52] and report five-seed statistics. More details in Appendix.

Evaluation Metrics. We report the training budget in NPU/GPU hours (GPU for REPA, DDT, and mini-batch OT coupling experiments), denoted as "Hour". We follow [8] for the computation of all evaluation metrics. Fréchet Inception Distance (FID) [15] is reported as our primary metric, alongside sFID [39], Inception Score (IS) [45], Precision and Recall [24]. Following prior works [8,27,43], we utilize the full training sets of CIFAR-10 and CelebA-64, and the standard testset of ImageNet, as our reference distributions. Details in Appendix.

Frequency Analysis Details. To maximize the separation between low- and high-frequency regions, the cutoff frequency is set to $\min(H, W)//8$ in Eq. 11, with 20% and 25% thresholds respectively in Eq. 13.

Baselines. We compare our method with four baselines: SiT [38], IP [41], SDSS [43], and MDSS [43]. For IP, SDSS, and MDSS, which were originally developed under the DDPM framework, we adapt their core ideas to the flow matching setting. Specifically, IP perturbs the input with additional noise to simulate inference, SDSS aligns predictions with the static target during single-step inference in training, and MDSS performs multi-step inference (we use 4 steps as reported to achieve the best performance [43]). We further evaluate two complementary exposure-bias mitigation paradigms. DG [21] corrects generated states at inference time with an external classifier and discriminator. Mini-batch

Table 1: Quantitative comparison on conditional datasets. Class-conditional (without/with CFG following SiT [38]) generation on ImageNet-256 and CIFAR-10. All entries are reported using 50 sampling steps (NFE). Red highlights improvements by our method, and green indicates degradations compared to gray baseline. Under a comparable cost (e.g., 144h vs. 166h on ImageNet), DEFAR outperforms the baselines.

Method	Backbone	Params	ImageNet-256 (conditional)								CIFAR-10 (conditional)						
			Iters	Hour	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑	Iters	Hour	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
SiT (original) [38]	SiT-B/4	131M	500k	166	61.64/31.64	12.20/8.44	24.67/55.65	0.387/0.542	0.581/0.524	250k	11	14.09/9.36	6.28/5.73	8.22/8.90	0.617/0.676	0.538/0.500
					+0.0/+0.0	+0.0/+0.0	+0.0/+0.0	+0.0/+0.0	+0.0/+0.0			+0.0/+0.0	+0.0/+0.0	+0.0/+0.0	+0.0/+0.0	+0.0/+0.0
IP [41]	SiT-B/4	131M	500k	166	60.45/29.84	12.49/8.80	26.05/60.20	**0.404**/**0.563**	0.578/0.523	250k	11	13.63/8.76	6.43/5.58	8.18/8.97	0.624/0.685	0.539/0.500
SDSS [43]	SiT-B/4	131M	500k	228	61.25/30.29	13.23/8.95	25.81/58.69	0.387/0.549	0.588/0.534	250k	15	13.27/8.27	5.97/5.41	8.21/8.91	0.612/0.679	0.550/0.506
MDSS (4steps) [43]	SiT-B/4	131M	500k	228	63.66/31.90	14.38/10.05	24.72/56.89	0.382/0.544	0.583/0.527	250k	15	12.73/8.45	6.29/5.93	8.25/9.06	**0.631**/**0.689**	0.529/0.484
DEFAR w/o FC	SiT-B/4	131M	500k	228	59.77/29.69	12.77/9.00	25.81/59.25	0.377/0.539	**0.598**/**0.549**	200k	12	11.61/7.62	6.07/6.21	8.30/9.05	0.599/0.669	0.556/0.507
DEFAR w/o ADR	SiT-B/4	131M	500k	289	57.17/28.71	10.65/8.25	26.55/**60.43**	0.398/0.557	0.590/0.532	200k	16	11.25/**7.53**	5.71/5.53	8.30/**9.10**	0.601/0.677	**0.557**/0.506
DEFAR	SiT-B/4	131M	250k	144	60.40/29.81	**9.05**/**7.35**	26.21/58.81	0.375/0.531	0.586/0.532	120k	10	12.01/7.73	5.46/**5.13**	8.27/9.01	0.605/0.667	0.555/**0.514**
					-1.24/ -1.83	-3.15/ -1.09	+1.54/+3.16	-0.012/-0.011	+0.005/+0.008			-2.08/-1.63	-0.82/-0.60	+0.05/+0.11	-0.012/-0.009	+0.017/+0.014
DEFAR	SiT-B/4	131M	500k	289	**56.39**/**28.34**	9.50/7.84	**26.62**/58.94	0.386/0.551	0.594/0.524	180k	15	**11.22**/7.56	**5.32**/5.21	**8.31**/9.06	0.618/0.682	0.553/0.500

Table 2: Unconditional generation and complementary mitigation strategies. Left: Quantitative results on unconditional CIFAR-10 and CelebA-64 (50 NFE), where DEFAR (120k) surpasses baselines under a comparable training budget. **Right:** Complementarity with two exposure-bias mitigation paradigms, DG on conditional ImageNet-256 and mini-batch OT coupling on unconditional CIFAR-10.

Method	Backbone	Params	Iters	CIFAR-10 Hour	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑	CelebA-64 Hour	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
SiT (original) [38]	SiT-B/4	131M	250k	12	17.41	6.19	7.82	0.586	0.539	36	7.04	6.88	2.76	0.626	0.521
					+0.0	+0.0	+0.0	+0.0	+0.0		+0.0	+0.0	+0.0	+0.0	+0.0
IP [41]	SiT-B/4	131M	250k	12	15.95	6.65	7.97	0.583	0.556	36	5.93	6.57	2.72	0.640	0.543
SDSS [43]	SiT-B/4	131M	250k	17	16.02	6.37	7.70	0.590	0.542	60	6.12	8.01	2.71	0.671	0.476
MDSS(4steps) [43]	SiT-B/4	131M	250k	17	16.10	6.59	7.91	**0.594**	0.547	60	6.10	8.29	2.91	0.623	0.527
DEFAR w/o FC	SiT-B/4	131M	250k	17	15.92	6.45	7.83	0.585	0.554	60	6.83	6.47	2.67	**0.676**	0.498
DEFAR w/o ADR	SiT-B/4	131M	250k	23	15.99	5.99	**8.07**	0.580	0.562	80	4.96	6.26	**2.94**	0.641	0.541
DEFAR	SiT-B/4	131M	120k	11	15.93	5.99	7.86	0.567	**0.567**	36	5.91	**6.13**	2.77	0.634	0.521
					-1.48	-0.20	+0.04	-0.019	+0.028		-1.13	-0.75	+0.01	+0.008	+0.000
DEFAR	SiT-B/4	131M	250k	23	**14.82**	**5.88**	7.79	0.570	0.563	80	**4.53**	6.41	2.92	0.629	**0.551**

Discriminator guidance [21] (conditional ImageNet-256 with SiT-B/4):

Method	Cls./Disc.	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑	ΔFID	ΔIS
SiT	-	61.64	12.20	24.67	0.387	0.581	+0.00	+0.00
DEFAR	-	60.40	9.05	26.21	0.375	0.586	-1.24	+1.54
SiT-G++	ADM/U-Net$_{\text{shallow}}$	60.92	11.10	25.08	**0.388**	0.579	-0.72	+0.41
DEFAR-G++	ADM/U-Net$_{\text{shallow}}$	**59.03**	**8.84**	**26.49**	0.374	0.590	-2.61	+1.82

Mini-batch OT coupling [52] (unconditional CIFAR-10):

Method	Backbone	Hour	Iters	FID$_{100}$ ↓	FID$_{1000}$ ↓	Avg. FID↓	ΔAvg. FID
OT-FM	U-Net	20	400k	4.643±0.035	3.824±0.037	4.233±0.027	-0.000±0.000
OT-FM + DEFAR	U-Net	20	200k	4.397±0.029	3.716±0.017	4.056±0.018	-0.177±0.023
OT-CFM	U-Net	20	400k	4.446±0.037	3.744±0.029	4.095±0.022	-0.138±0.033
OT-CFM + DEFAR	U-Net	20	200k	**4.260**±0.047	**3.660**±0.037	**3.960**±0.034	-0.273±0.029

OT coupling (OT-CFM) [52] changes the noise-data coupling during training to reduce transport-path variance. These two paradigms allow us to test whether DEFAR remains effective when exposure bias is also mitigated by inference-time correction or coupling-based path straightening.

5.2 Quantitative Experiments

To account for differences in computational environments, we reproduce the key baselines and report the results in Tab. 1 and Tab. 2. We highlight five primary findings: *(i)* Under comparable training budgets, DEFAR consistently improves upon the SiT baseline, reducing FID by **1.24**/**1.83** on ImageNet-256 without/with Classifier-Free Guidance (CFG), **2.08**/**1.63** on conditional CIFAR-10, and **1.48** and **1.13** on unconditional CIFAR-10 and CelebA-64, respectively. This confirms its effectiveness in enhancing generation quality. *(ii)* ADR outperforms MDSS and SDSS [43]. Unlike these baselines, which rely on static alignment targets, ADR learns an active anti-drift target, validating the advantage of dynamic directional rectification. *(iii)* Both ADR and FC independently yield performance gains, while their integration within the unified DEFAR framework achieves better results. *(iv)* While prior methods passively mitigate exposure bias via fixed input perturbation (IP) or static target alignment (SDSS, MDSS), DEFAR actively rectifies prediction drift and compen-

Fig. 4: Qualitative Comparison. DEFAR produces the most realistic samples on ImageNet-256 and CelebA-64. Red boxes highlight the blurriness or distorted details.

sates for deficient frequency components, achieving consistent empirical gains over these approaches. *(v)* In Tab. 2 (right), the DG and mini-batch OT comparisons further verify complementarity. For DG, DEFAR improves FID by **0.52** over DG alone, and combining DEFAR with DG reduces FID by **1.89**, suggesting that DEFAR strengthens the base model during training while DG provides complementary inference-time guidance. For mini-batch OT coupling, combining DEFAR with OT-CFM achieves the best Avg. FID **3.960**$\pm$0.034, outperforming OT-CFM alone by **0.135** and DEFAR alone by **0.096** in terms of the mean value. This suggests that OT-CFM reduces the transport-path difficulty and thereby provides a better basis for DEFAR to perform accurate fine-grained rectification according to the magnitude of exposure bias, leading to further gains.

5.3 Qualitative Experiments

All methods use the same random seed and 50 NFE for a fair comparison. As shown in Fig. 4, baselines exhibit two clear visual artifacts: *(i)* Rows 1, 2,

Table 3: Scalability of DEFAR on ImageNet-256. 50-NFE generation (without/with CFG). DEFAR consistently outperforms SiT across different scales.

Method	Backbone	Params	Iters	Hour	ImageNet-256				
					FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
SiT [38]	SiT-B/4	131M	500k	166	61.64/31.64	12.20/8.44	24.67/55.65	**0.387**/0.542	0.581/0.524
DEFAR	SiT-B/4	131M	250k	144	60.40/29.81	**9.05**/**7.35**	26.21/58.81	0.375/0.531	0.586/**0.532**
					-1.24/-1.83	-3.15/-1.09	+1.54/+3.16	-0.012/-0.011	+0.005/+0.008
DEFAR	SiT-B/4	131M	500k	289	**56.39**/**28.34**	9.50/7.84	**26.62**/**58.94**	0.386/**0.551**	**0.594**/0.524
SiT [38]	SiT-M/2	308M	700k	311	30.57/8.23	10.90/6.13	58.66/154.71	0.547/0.750	0.654/**0.531**
DEFAR	SiT-M/2	308M	350k	330	29.46/7.85	**5.60**/**4.64**	59.01/156.27	**0.549**/0.751	0.639/0.518
					-1.11/-0.38	-5.30/-1.49	+0.35/+1.56	+0.002/+0.001	-0.015/-0.013
DEFAR	SiT-M/2	308M	700k	661	**27.58**/**7.30**	7.35/4.96	**62.09**/**159.76**	0.546/**0.752**	**0.658**/**0.531**
SiT [38]	SiT-XL/2	675M	700k	2260	25.85/5.88	7.63/4.73	68.93/183.58	0.570/0.762	**0.659**/**0.541**
DEFAR	SiT-XL/2	675M	350k	2300	24.93/5.74	**5.71**/**4.71**	71.63/188.75	0.580/0.775	0.651/0.532
					-0.92/-0.14	-1.92/-0.02	+2.70/+5.17	+0.010/+0.013	-0.008/-0.009
DEFAR	SiT-XL/2	675M	700k	4640	**18.45**/**4.14**	6.34/4.95	**77.80**/**199.59**	**0.621**/**0.810**	0.644/0.529

Table 4: Architectural compatibility and sampling robustness. Left: Integration into REPA and DDT. 250-NFE generation under CFG, with and without DEFAR. Models are fine-tuned on ImageNet-512 from official ImageNet-256 checkpoints. **Right:** Generation across varying NFEs without CFG on ImageNet-256.

Model	Params	Iters	Hour	ImageNet-512				
				FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
REPA-XL/2 [63]	675M	100k	400	3.99	11.76	248.60	0.796	**0.595**
+ **DEFAR**	675M	45k	410	**3.46**	**9.12**	**255.26**	**0.798**	0.582
				-0.53	-2.64	+6.66	+0.002	-0.013
DDT-XL/2 [57]	675M	100k	570	1.90	4.33	281.25	0.793	**0.601**
+ **DEFAR**	675M	45k	580	**1.82**	**4.21**	**285.42**	**0.799**	0.596
				-0.08	-0.12	+4.17	+0.006	-0.005

Steps	SiT		IP		SDSS		MDSS		DEFAR(250k)		DEFAR(500k)	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑
30	63.63	24.70	61.25	**26.21**	63.41	25.80	65.93	24.78	61.12	25.62	**57.20**	25.84
50	61.64	24.67	60.45	26.05	61.25	25.81	63.66	24.72	60.40	26.21	**56.39**	**26.62**
100	60.42	24.48	58.26	26.01	59.92	25.78	61.33	24.64	58.11	26.09	**55.89**	**26.51**
250	59.84	24.30	57.75	25.83	59.23	25.64	61.33	24.48	57.56	25.91	**55.66**	**26.33**
500	59.70	24.24	57.58	25.76	59.06	25.56	61.10	24.41	57.32	25.79	**55.60**	**26.30**

and 5 exhibit severe structural disintegration and missing details. *(ii)* Rows 3, 4, and 6 show blurred object-background boundaries. These issues arise from the exposure bias accumulating during high-noise timesteps, causing confusion between the primary object and the surrounding context. In contrast, DEFAR generates images with more coherent structures and consistent textures.

6 Discussion

In this section, we conduct a multifaceted evaluation of DEFAR, assessing its scalability across model sizes, frequency restoration capability, and generalizability to diverse architectures and inference configurations.

Scalability across Model Sizes. Tab. 3 reports FID alongside four metrics of DEFAR across different model scales (B/4, M/2 and XL/2). DEFAR consistently outperforms the SiT models and exhibits scalability in generative quality.

Frequency Restoration Analysis. As shown in Fig. 5, we compare the PFR of predicted velocities between DEFAR and the SiT baseline when given forward-perturbed inputs. DEFAR effectively recovers low-frequency components during high-noise timesteps, alleviating the low-frequency deficiency.

Universality across Architectures. We integrate our method into two strong and distinct diffusion transformers: REPA [63] (featuring representation align-

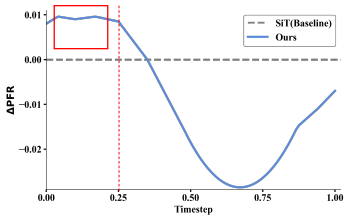

Fig. 5: Low-frequency Restoration. After comparable training, DEFAR can compensate for missing low-frequency components during high-noise timesteps (Red box).

Table 5: Generalization across path interpolants and sampling methods. FID on unconditional CIFAR-10.

Interpolant	Model	Hour	ODE	SDE
Linear	SiT-B/4	12	17.41	16.81
Linear	DEFAR-B/4	11	**15.93**	**15.31**
SBDM-VP	SiT-B/4	12	19.89	18.89
SBDM-VP	DEFAR-B/4	11	**18.33**	**17.35**
GVP	SiT-B/4	12	17.31	16.42
GVP	DEFAR-B/4	11	**15.84**	**14.91**

ment) and DDT [57] (featuring structural decoupling). We fine-tune their default REPA-XL/2 and DDT-XL/2 backbones on ImageNet-512 from the official ImageNet-256 pre-trained checkpoints. This demonstrates that our method is compatible with these advanced paradigms. As shown in Tab. 4 (left), under a comparable training budget, our approach achieves superior performance.

Adaptability to Path Interpolants and Sampling Methods. As shown in Tab. 5, we examine three interpolant types: Linear, SBDM-VP, and GVP [38]. The results indicate that DEFAR is adaptable to the choice of the forward process. Under both Ordinary Differential Equation (ODE) and Stochastic Differential Equation (SDE) sampling methods, DEFAR yields performance gains.

Robustness to Inference Steps. Potential prediction bias can accumulate during inference. Increasing inference steps often amplifies exposure bias. To assess this effect, we evaluate all methods under various step settings. DEFAR consistently achieves the lowest FID across nearly all settings, maintaining stable performance as the sampling step increases in Tab. 4 (right).

7 Conclusion

This work introduces DEFAR, a framework designed to adaptively mitigate exposure bias by exploiting its intrinsic properties rather than relying on passive robustness. We demonstrate that exposure bias serves as a valuable signal, providing both directional guidance and frequency weighting cues. By integrating Anti-Drift Rectification (ADR) and Frequency Compensation (FC), DEFAR empowers the model to perform dynamic self-rectification based on its own deviations. Extensive experiments verify the superior performance, scalability, compatibility, and inference robustness of DEFAR across various benchmarks.

Acknowledgements

The research of Shao-Lun Huang is supported in part by National Key R&D Program of China under Grant 2021YFA0715202, the National Natural Science Foundation of China under Grants 62571296 and Huawei.

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: The Eleventh International Conference on Learning Representations (2023)
2. Bai, C., Li, Y., Zhao, Z., Chen, J., Jia, P., She, Q., Lu, M., Zhang, S.: Fastinit: Fast noise initialization for temporally consistent video generation. arXiv preprint arXiv:2506.16119 (2025)
3. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. Advances in neural information processing systems **28** (2015)
4. Chen, Z., Seetharaman, P., Russell, B., Nieto, O., Bourgin, D., Owens, A., Salamon, J.: Video-guided foley sound generation with multimodal controls. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18770–18781 (2025)
5. Cheng, Z., Zhang, X., Di, D., Wei, C., Li, H., Yang, X.: Moca: Modeling object consistency for 3d camera control in video generation. In: The Fourteenth International Conference on Learning Representations (2026)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
7. Deng, Y., Kojima, N., Rush, A.M.: Markup-to-image diffusion models with scheduled sampling. In: The Eleventh International Conference on Learning Representations (2022)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
9. Ding, D., Ju, Z., Leng, Y., Liu, S., Liu, T., Shang, Z., Shen, K., Song, W., Tan, X., Tang, H., et al.: Kimi-audio technical report. arXiv preprint arXiv:2504.18425 (2025)
10. Du, S., Zou, Y., Wang, Z., Li, X., Li, Y., Shang, C., Shen, Q.: Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. International Journal of Computer Vision **134**, 152 (2026)
11. Everaert, M.N., Fitsios, A., Bocchio, M., Arpa, S., Süsstrunk, S., Achanta, R.: Exploiting the signal-leak bias in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4025–4034 (2024)
12. Fang, H., Qiu, D., Mao, B., Yan, P., Tang, H.: Motioncharacter: Identity-preserving and motion controllable human video generation. arXiv e-prints pp. arXiv–2411 (2024)
13. Geng, Z., Deng, M., Bai, X., Kolter, Z., He, K.: Mean flows for one-step generative modeling. Advances in Neural Information Processing Systems **38**, 75460–75482 (2026)
14. Haji-Ali, M., Menapace, W., Skorokhodov, I., Sahni, A., Tulyakov, S., Ordonez, V., Siarohin, A.: Improving progressive generation with decomposable flow matching. Advances in Neural Information Processing Systems **38**, 163844–163885 (2026)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Huang, F., Huang, G., Fan, X., He, Y., Liang, X., Chen, X., Jiang, Q., Khan, F.N., Jiang, J., Wang, Z.: Semantic-space exploration and exploitation in RLVR for LLM

reasoning. In: Liakata, M., Moreira, V.P., Zhang, J., Jurgens, D. (eds.) Findings of the Association for Computational Linguistics: ACL 2026. pp. 38402–38449. Association for Computational Linguistics, San Diego, California, United States (Jul 2026), https://aclanthology.org/2026.findings-acl.1915/, accessed: 30 June 2026

18. Huang, F., Jiang, J., Jiang, Q., Li, H., Khan, F.N., Wang, Z.: Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9772–9781 (2025)
19. Huang, F., Yao, Z., Zhou, W.: Dtbs: Dual-teacher bi-directional self-training for domain adaptation in nighttime semantic segmentation. In: ECAI 2023, pp. 1084–1091. IOS Press (2023)
20. Huang, X., Chen, Z., Shen, W., Zhang, X.P.: Learnibridge: Learnable calibration of feature caching for diffusion models acceleration. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=8sD74Krbw7, accessed: 30 June 2026
21. Kim, D., Kim, Y., Kwon, S.J., Kang, W., Moon, I.C.: Refining generative process with discriminator guidance in score-based diffusion models. In: International Conference on Machine Learning. pp. 16567–16598. PMLR (2023)
22. Kim, K., Kim, S.: Model already knows the best noise: Bayesian active noise selection via attention in video diffusion model. arXiv preprint arXiv:2505.17561 (2025)
23. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009)
24. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in neural information processing systems **32** (2019)
25. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025)
26. Li, H., Wang, Y., Hao, W., Zhang, P., Wang, D., Lu, H.: Ragtrack: Language-aware rgbt tracking with retrieval-augmented generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28179–28189 (2026)
27. Li, M., Qu, T., Yao, R., Sun, W., Moens, M.F.: Alleviating exposure bias in diffusion models through sampling with shifted time steps. In: International Conference on Learning Representations. vol. 2024, pp. 16816–16838 (2024)
28. Li, R., Liang, Y., Ni, Z., Huang, H., Zhang, C., Li, X.: Rethinking reward signals in video grpo: When scores become targets (2026)
29. Li, Y., van der Schaar, M.: On error propagation of diffusion models. In: International Conference on Learning Representations. vol. 2024, pp. 32791–32807 (2024)
30. Liang, Y., Cai, Z., Xu, J., Huang, G., Wang, Y., Liang, X., Liu, J., Li, Z., Wang, J., Huang, S.L.: Unleashing region understanding in intermediate layers for mllm-based referring expression generation. Advances in Neural Information Processing Systems **37**, 120578–120601 (2024)
31. Lin, H., Qin, C., Liu, Z., Pei, Q., Li, Y., Zhong, Z., Gao, X., Wang, Y., He, C., Wu, L.: Scientific image synthesis: Benchmarking, methodologies, and downstream utility. arXiv preprint arXiv:2601.17027 (2026)

32. Lin, Z., Gao, Y., Yang, Y., Sang, J.: Revisiting visual model robustness: A frequency long-tailed distribution view. Advances in Neural Information Processing Systems **36**, 59239–59251 (2023)
33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
34. Liu, X., Gong, C., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
35. Liu, Y., Fu, P., Li, H., Qi, Y., Jiang, C., Fu, J., Liu, Z., Qin, B., Luo, Z., Luan, J., et al.: Elva: Exploring ranking-driven universal multimodal retrieval. arXiv preprint arXiv:2606.20280 (2026)
36. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
37. Luo, X., Li, Q., Li, Y., Huang, G., Zhu, Y., Qin, W., Wang, M., Wan, P., Huang, S.L.: Beyond the golden data: Resolving the motion-vision quality dilemma via timestep selective training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 43440–43449 (2026)
38. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV. pp. 23–40. Springer (2024)
39. Nash, C., Menick, J., Dieleman, S., Battaglia, P.: Generating images with sparse representations. In: International Conference on Machine Learning. pp. 7958–7968. PMLR (2021)
40. Ning, M., Li, M., Su, J., Salah, A.A., Onal Ertugrul, I.: Elucidating the exposure bias in diffusion models. In: International Conference on Learning Representations. vol. 2024, pp. 15167–15189 (2024)
41. Ning, M., Sangineto, E., Porrello, A., Calderara, S., Cucchiara, R.: Input perturbation reduces exposure bias in diffusion models. In: International Conference on Machine Learning. pp. 26245–26265. PMLR (2023)
42. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732 (2015)
43. Ren, Z., Zhan, Y., Ding, L., Wang, G., Wang, C., Fan, Z., Tao, D.: Multi-step denoising scheduled sampling: Towards alleviating exposure bias for diffusion models. In: AAAI. vol. 38, pp. 4667–4675 (2024)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
45. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
46. Schmidt, F.: Generalization in generation: A closer look at exposure bias. In: Proceedings of the 3rd Workshop on Neural Generation and Translation. vol. 19, pp. 157–167. Association for Computational Linguistics (2019)
47. Sigillo, L., He, S., Comminiello, D.: Latent wavelet diffusion: Enabling 4k image synthesis for free. arXiv e-prints pp. arXiv–2506 (2025)
48. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning(ICML). pp. 2256–2265. pmlr (2015)
49. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)

50. Tang, J., Li, J., Gao, Z., Li, J.: Rethinking graph neural networks for anomaly detection. In: International conference on machine learning. pp. 21076–21089. PMLR (2022)
51. Team, M.L., Cai, X., Huang, Q., Kang, Z., Li, H., Liang, S., Ma, L., Ren, S., Wei, X., Xie, R., et al.: Longcat-video technical report. arXiv preprint arXiv:2510.22200 (2025)
52. Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., Bengio, Y.: Improving and generalizing flow-based generative models with minibatch optimal transport. Transactions on Machine Learning Research pp. 1–34 (2024)
53. Venkatraman, A., Boots, B., Hebert, M., Bagnell, J.A.: Data as demonstrator with applications to system identification. In: ALR Workshop, NIPS (2014)
54. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
55. Wang, C., Sennrich, R.: On exposure bias, hallucination and domain shift in neural machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3544–3552 (2020)
56. Wang, S., Azadi, S., Girdhar, R., Rambhatla, S., Sun, C., Yin, X.: Motif: Making text count in image animation with motion focal loss. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7773–7783 (2025)
57. Wang, S., Tian, Z., Huang, W., Wang, L.: Ddt: Decoupled diffusion transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 40633–40642 (2026)
58. Wu, T., Si, C., Jiang, Y., Huang, Z., Liu, Z.: Freeinit: Bridging initialization gap in video diffusion models. In: European conference on computer vision. pp. 378–394. Springer (2024)
59. Xia, Z., Li, Y., Tang, S., Fan, Z., Fang, X., Tao, H., Cai, X., Ke, G., Zhang, L., Hong, Y., et al.: Uniem-3m: A universal electron micrograph dataset for microstructural segmentation and generation. arXiv preprint arXiv:2508.16239 (2025)
60. Xie, X., Liu, J., Lin, Z., Fan, H., Han, Z., Tang, Y., Qu, L.: Unleashing the potential of large language models for text-to-image generation through autoregressive representation alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11105–11113 (2026)
61. Yao, Y., Chen, J., Huang, Z., Lin, H., Wang, M., Dai, G., Wang, J.: Manifold constraint reduces exposure bias in accelerated diffusion sampling. In: International Conference on Learning Representations. vol. 2025, pp. 96580–96616 (2025)
62. Yu, M., Zhan, K.: Frequency regulation for exposure bias mitigation in diffusion models. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10370–10378 (2025)
63. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: The Thirteenth International Conference on Learning Representations (2025)
64. Zhang, G., Shi, C., Jiang, Z., Xiang, X., Qian, J., Shi, S., Jiang, L.: Proteus-id: Id-consistent and motion-coherent video customization. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
65. Zhang, J., Liu, D., Park, E., Zhang, S., Xu, C.: Anti-exposure bias in diffusion models. In: The Thirteenth International Conference on Learning Representations (2025)

66. Zhang, Q., Fu, H., Huang, G., Liang, Y., Chu, C., Peng, T., Wu, Y., Li, Q., Li, Y., Huang, S.L.: A high-dimensional statistical method for optimizing transfer quantities in multi-source transfer learning. Advances in Neural Information Processing Systems **38**, 25528–25563 (2026)
67. Zhang, W., Feng, Y., Meng, F., You, D., Liu, Q.: Bridging the gap between training and inference for neural machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4334–4343 (2019)
68. Zhang, Y., Gu, J., Wu, Z., Zhai, S., Susskind, J., Jaitly, N.: Planner: Generating diversified paragraph via latent language diffusion model. Advances in Neural Information Processing Systems **36**, 80178–80190 (2023)
69. Zhao, C., Chen, J., Li, H., Kang, Z., Lu, S., Wei, X., Zhang, K., Yang, J., Tai, Y.: LUVE : Latent-cascaded ultra-high-resolution video generation with dual frequency experts. In: Forty-third International Conference on Machine Learning (2026)
70. Zhao, M., Zhu, H., Xiang, C., Zheng, K., Li, C., Zhu, J.: Identifying and solving conditional image leakage in image-to-video diffusion model. Advances in Neural Information Processing Systems **37**, 30300–30326 (2024)