Hierarchical Koopman Diffusion: Fast Generation with Interpretable Diffusion Trajectory

Hanru Bai[1] Weiyang Ding[1] Difan Zou[2]

[1]Fudan University

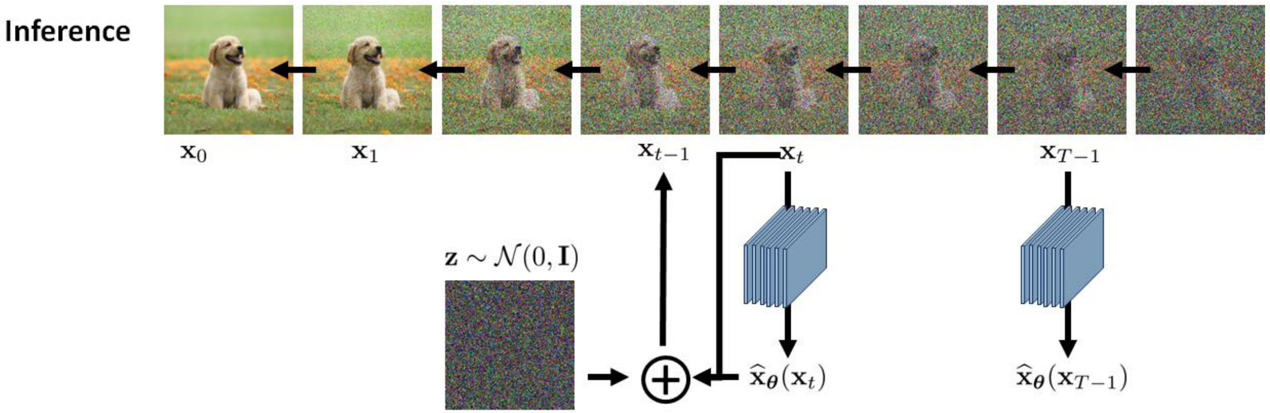
[2]The University of Hong Kong

NeurIPS 2025

Motivation

- The sampling process of diffusion models requires thousands of denoising steps

  → Accelerate sampling in diffusion models: one-step generation


Inference
$\mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{x}_{t-1}$
$\mathbf{x}_t$
$\mathbf{x}_{T-1}$
$\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
$\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{x}_t)$
$\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{x}_{T-1})$

- Existing one-step generation methods learn direct noise-to-image mappings

  → limited interpretability of the generation process

  → limit the ability to intervene at specific stages along the generative trajectory, which enables controllable image synthesis at inference time.

Develop an explicitly interpretable one-step generation framework

Method: Hierarchical Koopman Diffusion

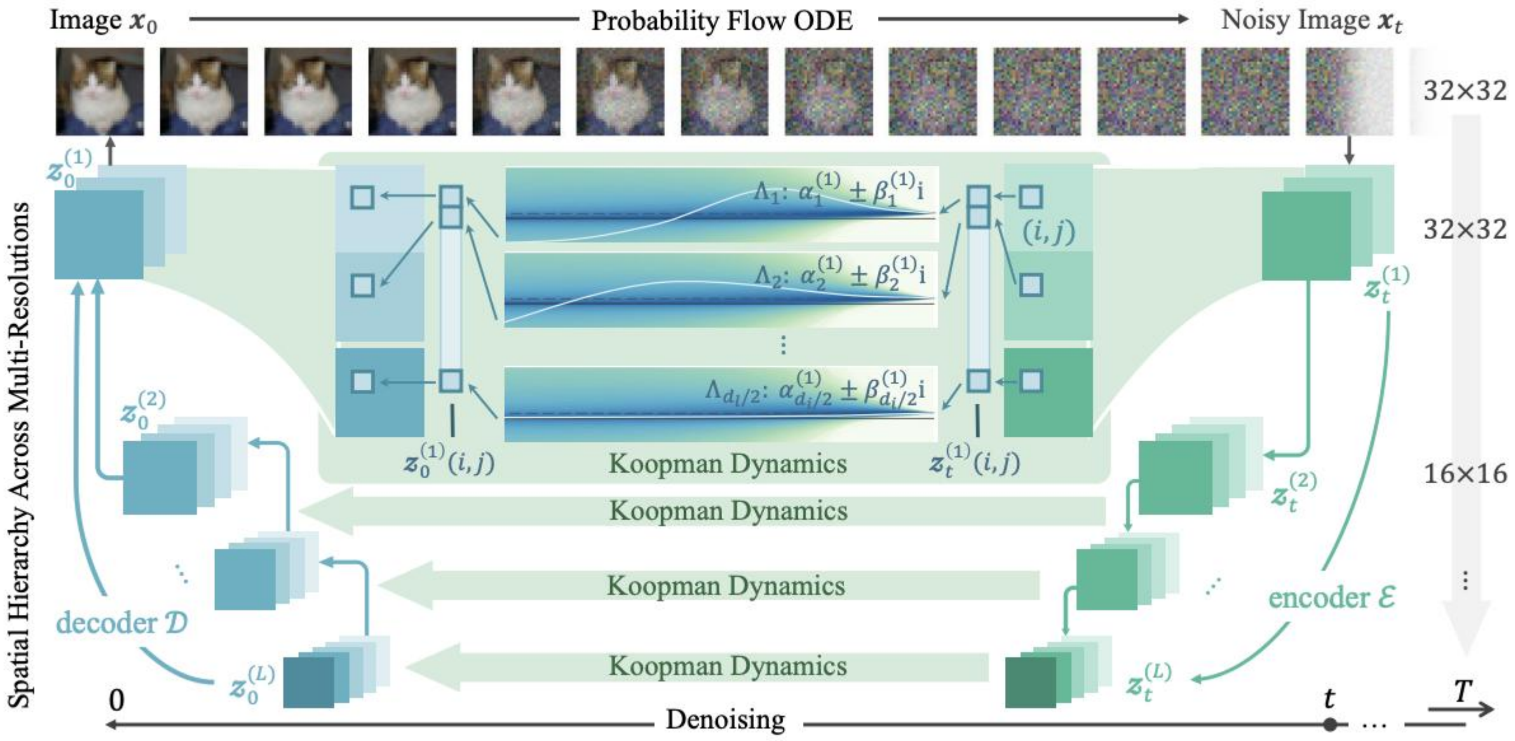


Figure 1: The framework of the proposed method. The HKD model first hierarchically extracts different-level features from the given noisy image at any time t by encoder $\mathcal{E}$. Secondly, the Koopman dynamics model is applied for each level to the skips and the bottleneck. Last, a uniform decoder $\mathcal{D}$ performs the mapping from the Koopman spaces back to the image space.

Main idea: map the whole deterministic diffusion trajectory into the **Koopman space**, where the dynamics become linear with closed-form ODE solutions.

- This explicit formulation makes every intermediate state analytically accessible, yielding *interpretability*, *extra supervision* along trajectories, and *fine-grained control* that implicit one-step methods lack.

Method: Hierarchical Koopman Diffusion

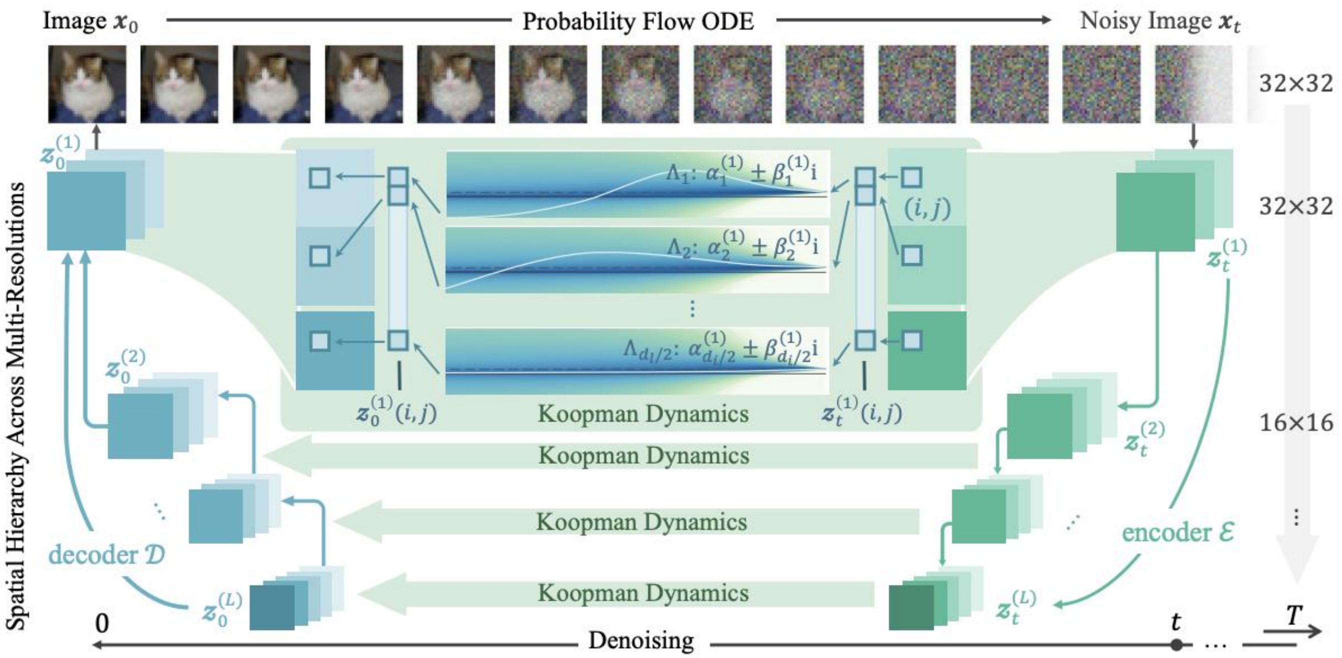


Figure 1: The framework of the proposed method. The HKD model first hierarchically extracts different-level features from the given noisy image at any time t by encoder $\mathcal{E}$. Secondly, the Koopman dynamics model is applied for each level to the skips and the bottleneck. Last, a uniform decoder $\mathcal{D}$ performs the mapping from the Koopman spaces back to the image space.

Encoder:

$$\boldsymbol{\mathcal{E}_\theta} : \boldsymbol{x}_t \in \mathbb{R}^{C\times H\times W} \mapsto \{\boldsymbol{z}_t^{(l)} \in \mathbb{R}^{d_l\times h_l\times w_l}\}_{l=1}^{L}.$$

Hierarchical Koopman Dynamics:

$$\frac{\mathrm{d}\boldsymbol{z}_t^{(l)}(i,j)}{\mathrm{d}t} = \boldsymbol{A}^{(l)}(i,j)\, \boldsymbol{z}_t^{(l)}(i,j),\ \forall(i,j),\ t\in[\epsilon,T].$$

Decoder:

$$\boldsymbol{\mathcal{D}_\phi} : \{\boldsymbol{z}_\epsilon^{(l)}\}_{l=1}^{L} \mapsto \boldsymbol{x}_\epsilon \in \mathbb{R}^{C\times H\times W}.$$

Experiments

Compare with Prior One-step Generation Methods

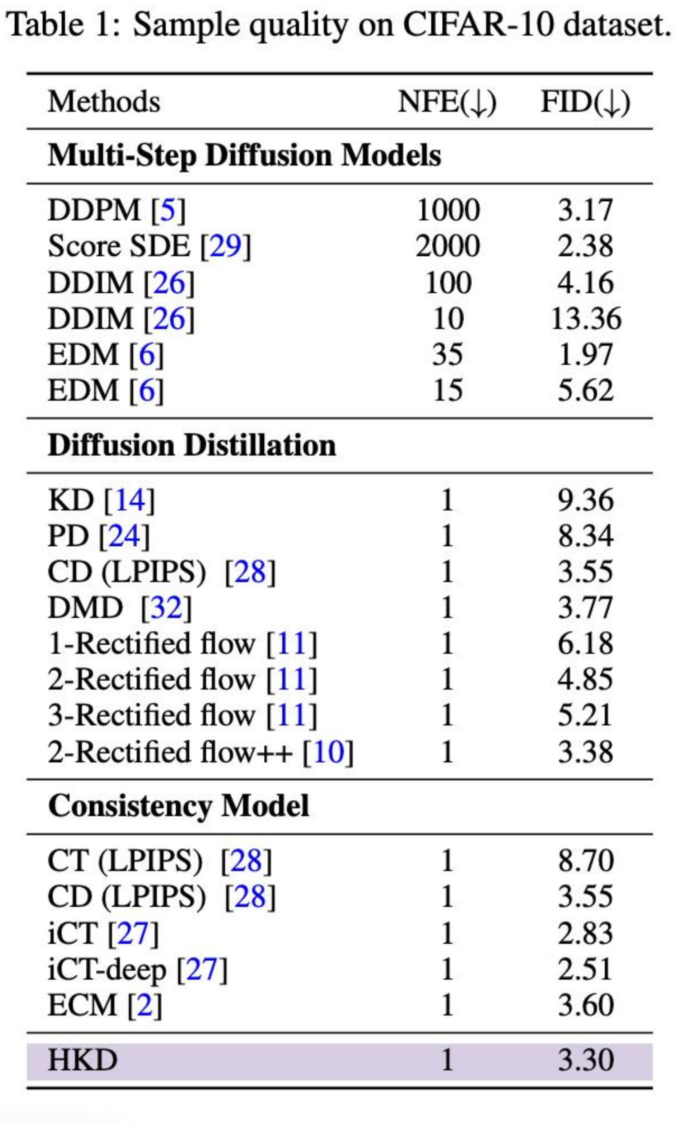


Table 1: Sample quality on CIFAR-10 dataset.

Methods	NFE(↓)	FID(↓)
Multi-Step Diffusion Models		
DDPM [5]	1000	3.17
Score SDE [29]	2000	2.38
DDIM [26]	100	4.16
DDIM [26]	10	13.36
EDM [6]	35	1.97
EDM [6]	15	5.62
Diffusion Distillation		
KD [14]	1	9.36
PD [24]	1	8.34
CD (LPIPS) [28]	1	3.55
DMD [32]	1	3.77
1-Rectified flow [11]	1	6.18
2-Rectified flow [11]	1	4.85
3-Rectified flow [11]	1	5.21
2-Rectified flow++ [10]	1	3.38
Consistency Model		
CT (LPIPS) [28]	1	8.70
CD (LPIPS) [28]	1	3.55
iCT [27]	1	2.83
iCT-deep [27]	1	2.51
ECM [2]	1	3.60
HKD	1	3.30

- Datasets: CIFAR-10 and FFHQ
- Evaluation Metric: Fréchet Inception Distance (FID)
- Results: While enhanced methods such as iCT-deep have pushed performance boundaries, they suffer from high sensitivity tohyperparameters. Moreover, state-of-the-art consistency training typically requires nearly a week of training on 8 GPUs and remains unstable in practice.

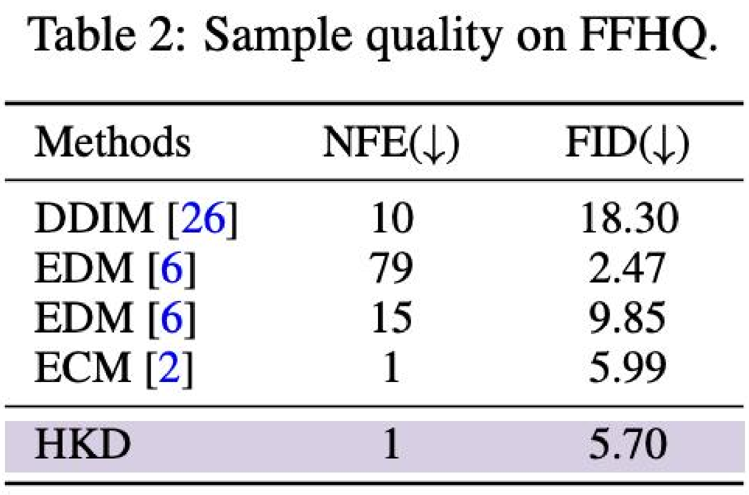


Table 2: Sample quality on FFHQ.

Methods	NFE(↓)	FID(↓)
DDIM [26]	10	18.30
EDM [6]	79	2.47
EDM [6]	15	9.85
ECM [2]	1	5.99
HKD	1	5.70

Koopman Spectral Analysis for Generative Process

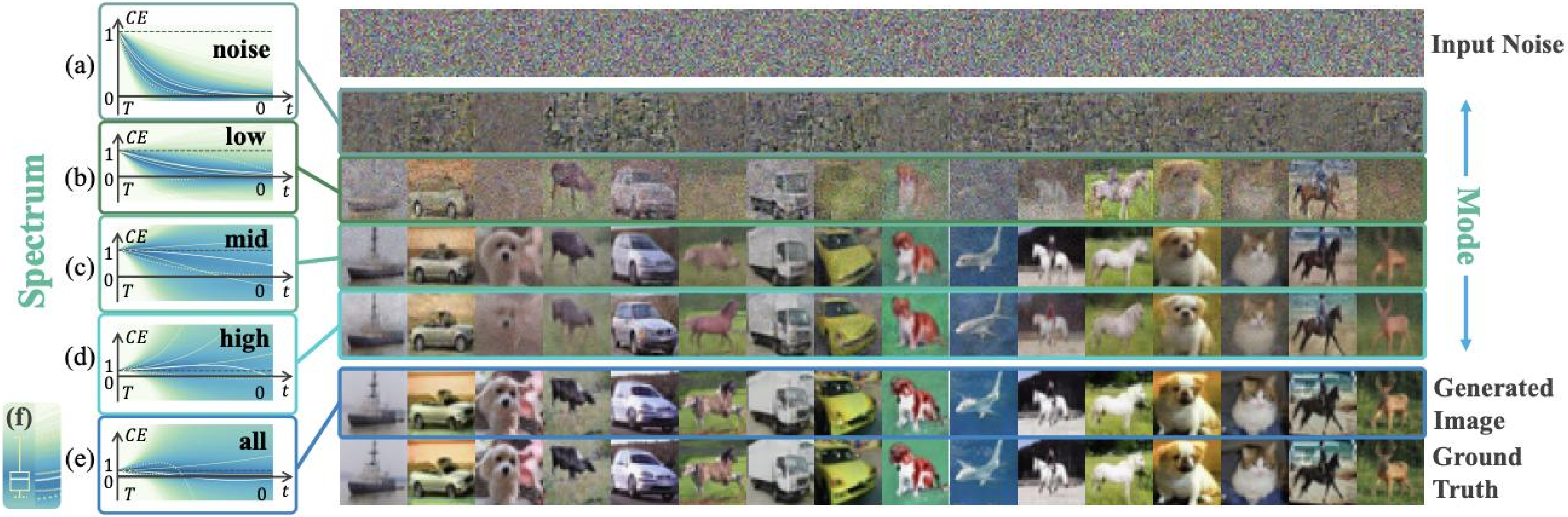


Figure 2: Visualization of Spectral Contributions in Generated Images Across Koopman Modes. (a) visualizes the contribution of Koopman spectrum with the smallest real-part magnitudes to the reconstructed state over time. It corresponds to the noisy part of the image. (b), (c) and (d) illustrate progressively larger spectral components and their reconstructed image components.

One-Step Image Editing: A Case of Model Interpretability

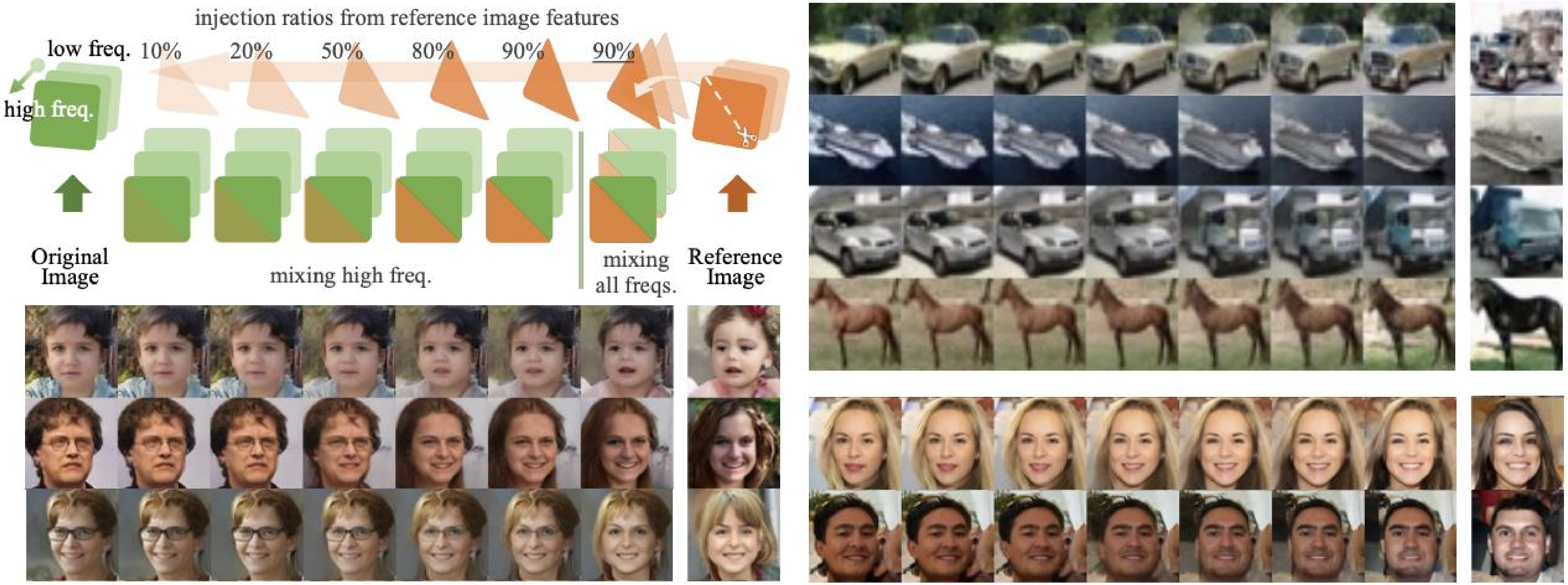


Figure 3: One-step image editing via frequency-aware interventions along the diffusion trajectory. We controlled the image generation by the high-frequency features from a reference image through injecting them into the lower-left half of the generating image at different mixing ratios (10%, 20%, 50%, 80%, 90%). The modifications were performed at the midpoint of the Koopman trajectory. Columns 2-6 showcase the frequency-aware editing, where only high-frequency components were mixed, preserving the low-frequency structure of the original image. Column 7 is for frequency agnostic editing, where all-frequency features of reference images are mixed with all frequency bands of the original image at a mixing ratio of 90%. We exhibit the results from both datasets.