A-SelecT: Automatic Timestep Selection for Diffusion Transformer Representation Learning

Supplementary Material

The supplementary is organized as follows:

- §S1 details the **Datasets** utilized in the study, including statistical descriptions of each dataset.
- §S2 offers **Additional Preliminary Results**.
- §S3 provides more **Implementation Details**.
- §S4 offers **Additional Visualizations** of feature decomposition.
- §S5 presents new **Discussions on Feature Fusion**.
- §S6 presents **Discussions on Feature Selection between Query and Value**.
- §S7 provides **Additional Results on Impact of Feature and Block Selection**.
- §S8 offers **Additional Feature Visualization Examples**.
- §S9 presents **Visualization Feature Selection Results**.
- §S10 presents **Discussions on Impact of Resolution on HFR**.
- §S11 provides **DDAE Classification Results on FGVC**.
- §S12 provides **Additional Details on HFR**.
- §S13 provides **Additional Details on Fisher Score**.
- §S14 provides **Additional HFR Results across multiple DiT Models**.
- §S15 gathers **Additional Discussions** on license, reproducibility, technical contributions, social impact and limitations, and future work.

S1. Datasets

In Table S1, we provide statistical information of FGVC benchmark (*i.e.*, Caltech-UCSD Birds (CUB) [83], Aircraft [59], Stanford Cars [46], NABirds [82], Stanford Dogs [43], Oxford Flowers [66]), ImageNet [20], and ADE20K [98].

Dataset	Class Number	Training Number	Test Number
Aircraft	100	6,667	3,333
Stanford Cars	196	8,144	8,041
CUB	200	5,994	5,794
Stanford Dogs	120	12,000	8,580
Oxford Flowers	102	2,040	6,149
NABirds	555	23,929	24,633
ImageNet	1,000	1.28M	50,000
ADE20K	150	2,0210	2000

Table S1. **Datasets Statistical Details.**

S2. Additional Preliminary Results

We present additional preliminary results illustrating the relationship between HFR and classification accuracy on the Stanford Cars and Aircraft datasets, as shown in Fig. S1.

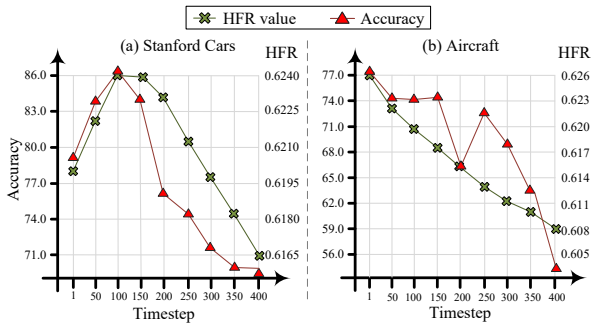

Figure S1. **More Preliminary Results** on the impact of the High-Frequency Ratio (HFR) with classification performance on Stanford Cars (a) and Aircraft (b).

The results indicate a clear positive correlation: classification accuracy is consistently highest at the timestep where HFR reaches its maximum. Similar trends are observed on CUB and Oxford Flowers, further supporting HFR as a reliable and robust indicator of discriminative feature quality.

S3. Additional Implementation Details

Here we provide more implementation details for our experiments. SD 3.5 [24] released Large version model with 8 billion weights, while Medium version with 2 billion weights. In our experiment, we use the SD 3.5 Medium, comprising 24 MMDiT blocks and features a backward process with 1,000 time steps for denoising. Given that SD 3.5 operates as a text-to-image model, we standardized the text condition to an empty string, ensuring uniformity across all features extraction. For text encoding, we exclusively employed the CLIP-G/14 encoder [71], electing not to utilize the CLIP-L/14 [71] or T5 XXL text encoders [72]. For a fair comparison, we include SDXL [63]) as a baseline, utilizing its U-Net backbone with 2.6 billion weights. This model is trained with textual conditioning, consistent with our SD 3.5 backbone. For the classification task, we report SDXL results under the same setting as our method. We also use an empty string as the text condition for SDXL. For the ImageNet classification task, we do not report results for ResNet-50, as publicly available pretrained ResNet-50 checkpoints are mostly trained on ImageNet. Utilizing these models for evaluation will lead to an unfair comparison.

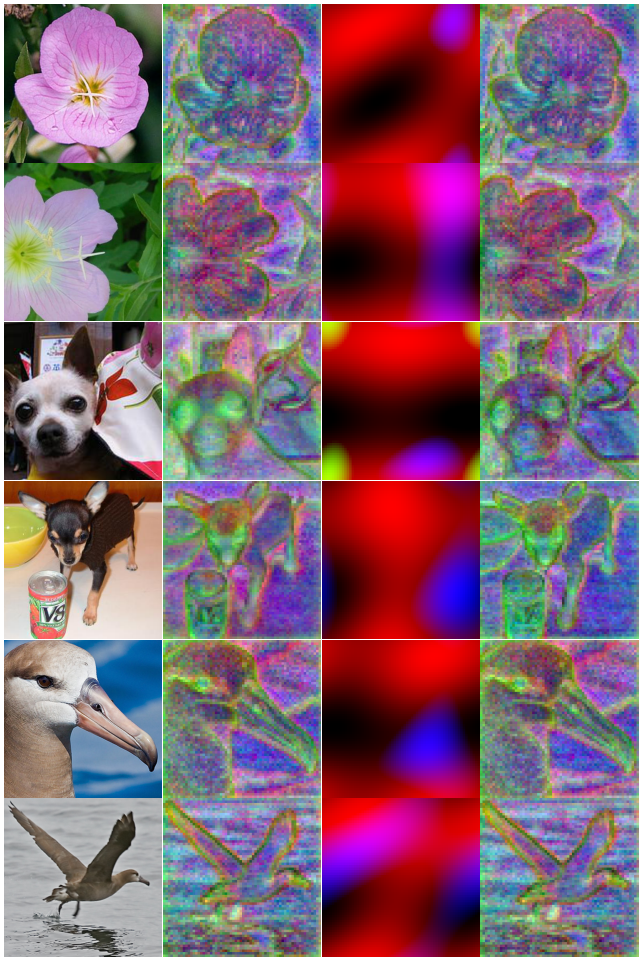

Figure S2. **More Visualization of Feature Decomposition Examples from CUB, Oxford Flowers and Stanford Dogs**. We present six sets of decomposition of the original features extracted from SD 3.5 into components that exclusively contain high-frequency and low-frequency information.

S4. Additional Results on High-Frequency Components

In Fig. S2, we provide more visualization results by decomposing of the original extracted features into components that exclusively contain high-frequency and low-frequency information. As seen, high-frequency features turn to contain more discriminative information, which is consistent with our observation in the main paper §3.3.

S5. Discussions on Feature Fusion

One major research question that may arise is whether we can achieve better performance when utilizing features from multiple blocks or different timesteps. We thus conduct an experiment on CUB dataset in Table S2. The results indicate that incorporating additional features does not enhance accuracy.

Block	Timestep	CUB [83]
9-th + 6-th	50	71.9%
9-th + 12-th	50	72.5%
9-th	50 + 1	76.1%
9-th	50 + 100	78.0%
9-th	50	**78.6%**

Table S2. **Impact of Feature Fusion**. We extract features from different blocks at the same timestep and from the same block at various timesteps. The results indicate that the addition of more features does not lead to an improvement in performance; rather, it may actually decrease performance.

	Aircraft			CUB		
Timestep	Query	Value	Key	Query	Value	Key
1	77.5%	76.6%	72.1%	72.3%	71.3%	47.5%
50	74.3%	73.4%	69.7%	78.6%	76.1%	68.2%
100	74.2%	74.0%	70.7%	77.9%	73.1%	64.2%
150	74.6%	74.9%	62.0%	75.0%	71.0%	56.7%
200	66.5%	72.0%	48.7%	72.2%	68.7%	62.0%
250	72.5%	55.6%	59.6%	67.3%	66.0%	63.3%
300	68.9%	47.2%	59.8%	63.9%	63.8%	56.7%
350	63.8%	62.6%	56.8%	64.8%	43.8%	58.3%
400	53.5%	54.0%	37.8%	63.6%	52.0%	41.8%
450	43.8%	55.2%	56.1%	57.7%	17.3%	35.5%
500	45.9%	49.5%	28.9%	37.2%	40.5%	35.8%
550	30.1%	21.3%	30.0%	27.2%	33.5%	26.3%
600	35.7%	17.5%	23.5%	27.5%	25.2%	24.5%
650	25.8%	20.5%	10.3%	19.8%	16.5%	17.2%
700	15.3%	9.7%	6.3%	15.5%	13.0%	7.2%
750	12.0%	9.9%	8.1%	12.3%	7.8%	5.3%
800	6.6%	5.1%	6.7%	8.0%	3.7%	2.1%
850	5.8%	3.0%	4.8%	4.2%	3.5%	3.5%
900	3.4%	3.3%	3.7%	2.0%	3.5%	1.9%
950	2.2%	2.5%	2.7%	1.3%	1.7%	1.2%
1000	1.3%	1.5%	1.6%	0.8%	0.8%	0.6%

Table S3. **Accuracies of Query, Value, and Key Features Across Timesteps**, respectively. We report top-1 classification accuracy using query, value and key features at various diffusion timesteps on the Aircraft and CUB datasets. While query and value features show similar performance overall, query features outperform in a greater number of cases.

Even worse, the additional operations deteriorate our model's performance. This decrease in performance can be attributed to the increased complexity introduced to the model. We propose that the augmented complexity burdens the downstream classifier, thereby impairing its effectiveness.

S6. Discussions on Feature Selection

Table S3 presents the classification accuracies of query, value, and key features on the CUB and Aircraft across different timesteps, respectively. We observe that both query and value features achieve comparable performance, substantially outperforming the key feature. Notably, the query feature generally outperforms the value feature in most cases. Based on this observation, we adopt the query feature for DiT represen-

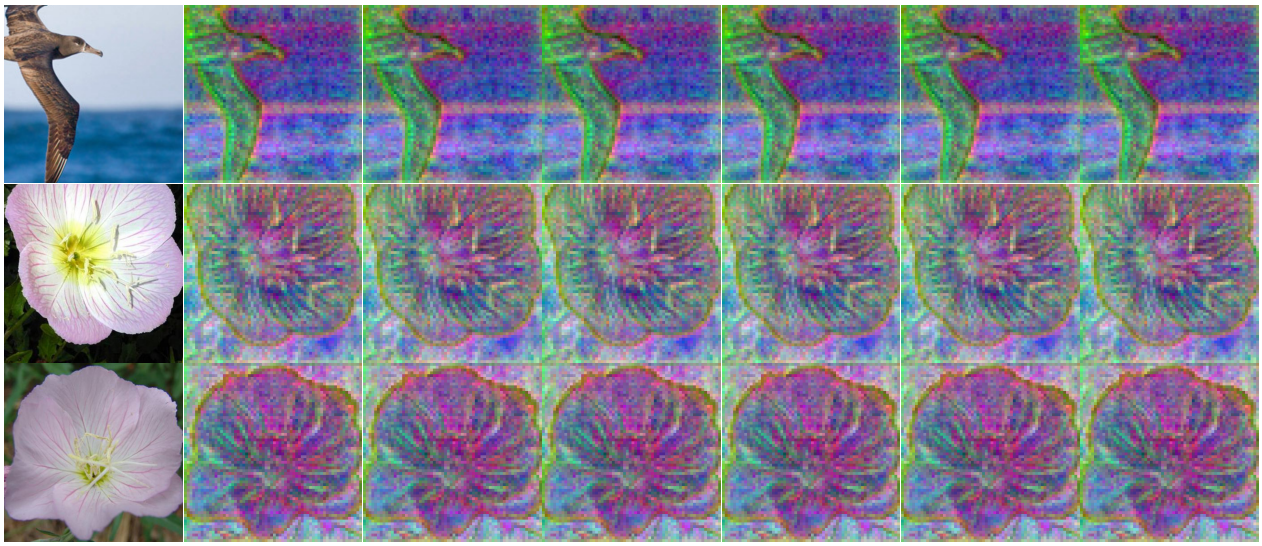

Figure S3. **More Feature Visualization Examples from CUB and Oxford Flowers.** We present visualizations of three feature sets extracted from the 9-th block of SD 3.5. The columns represent features extracted at different timesteps.

tation learning for all experiments.

S7. Additional Results on Impact of Feature and Block Selection

In Fig. S4, we provide additional results analyzing the impact of block and feature selection across multiple datasets, including CUB, Aircraft, and Stanford Cars. Consistent to the results shown in our main paper, the block and feature influence the performance. Consistent with the main paper (§4.3), both factors strongly influence representation's discriminative performance. Among transformer block components, Q features achieve the highest accuracy, followed by V, while K, A, and O perform worse. For block selection, features from middle layers consistently outperform those from early or late layers, as they capture a balanced mix of coarse and fine-grained representations. These findings confirm that optimal feature and block choices are crucial for maximizing discrimination.

S8. Additional Feature Visualization Examples

In Fig. S3, we provide more visualization examples of features extracted at different timesteps. Consistent to the results shown in our main paper, manual selections of discriminative features at different timesteps are impractical and ambiguous.

S9. Visualization Feature Selection Results

We exam the visualization feature selection method from [63] further for DiT block selection on the CUB and Oxford Flowers datasets, visualizing features at timestep 50 for CUB and timestep 100 for Oxford Flowers in Fig. S5. Based on visualization, we manually identify the most informative features from blocks 7, 0, 20, and 12. These selected features are subsequently combined and evaluated on downstream tasks. However, Table S4 shows that these selected features underperform compared to the single feature from block 9. This suggests that the visualization selection method is not effective for the DiT model.

Block	CUB	Oxford Flowers
7-th	66.2%	52.0%
7-th + 0-th	69.7%	84.3%
7-th + 0-th + 20-th	66.2%	67.2%
7-th + 0-th + 20-th + 12-th	58.5%	73.0%

Table S4. **Visualization Feature Selection Results on DiT**.

S10. Discussions on Impact of Resolution on HFR

In Table S5, we report the classification accuracies of query features on the CUB dataset across different input resolutions (*i.e.*, 256, 512, and $1,024$) and timesteps (*i.e.*, 1000, 950, 900, 850, 800, 750, 700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 200, 150, 100, 50, 1). The results indicate that the optimal timestep yielding the highest classification performance varies with the input resolution. Notably, these optimal timesteps consistently correspond to the highest HFR values, suggesting that HFR remains robust to changes in resolution. This implies that HFR effectively identifies the optimal timestep regardless of the input resolution.

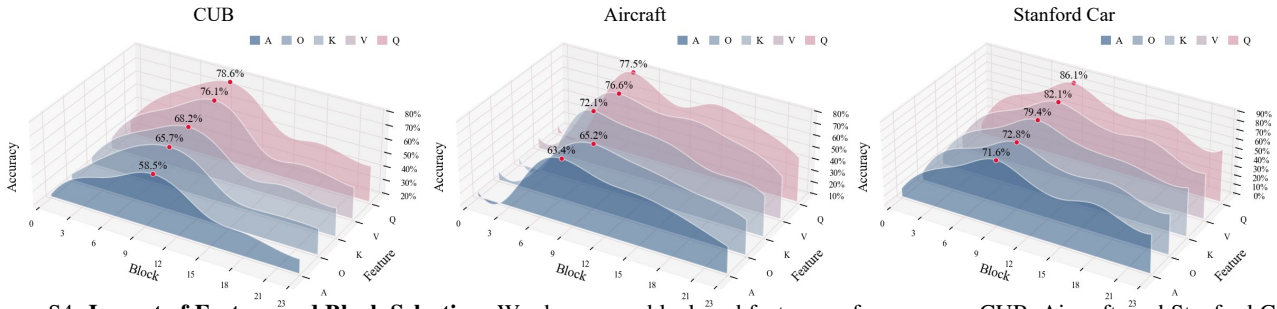

Figure S4. **Impact of Feature and Block Selection.** We show more block and feature performance on CUB, Aircraft, and Stanford Car. The figure shows the consistent result with Oxford Flower dataset.

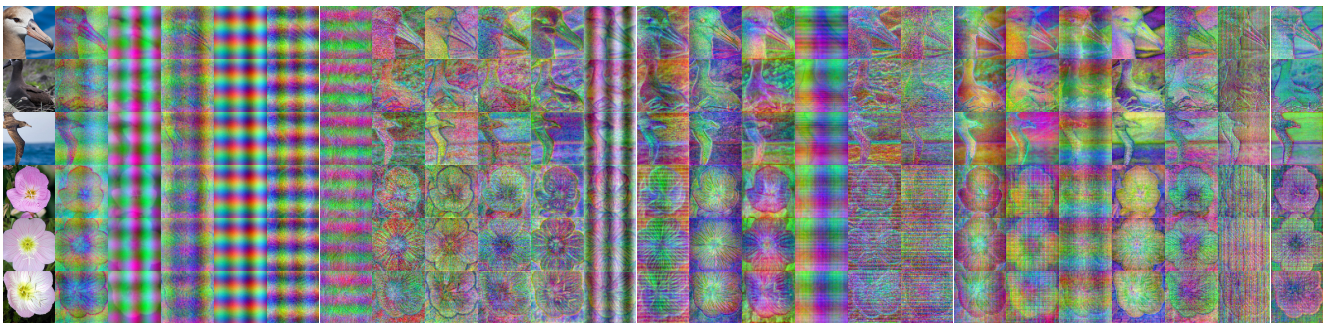

Figure S5. **Feature Visualization Across Blocks on CUB and Oxford Flowers.** The first column shows original input images and subsequent columns presents feature visualizations extracted from the 0-th to 23-th block of SD 3.5. The top three rows correspond to samples from CUB and the bottom three rows are from Oxford Flowers.

	256		512		1024	
Timestep	Acc.	HFR	Acc.	HFR	Acc.	HFR
1	**71.5%**	**0.6163**	72.3%	0.6199	65.2%	0.6180
50	69.3%	0.6131	**78.6%**	**0.6223**	66.2%	0.6195
100	68.2%	0.6110	77.9%	0.6221	67.8%	0.6208
150	62.0%	0.6092	75.0%	0.6217	**74.8%**	**0.6234**
200	61.3%	0.6077	72.2%	0.6212	69.7%	0.6164
250	57.8%	0.6059	67.3%	0.6200	72.4%	0.6144
300	46.5%	0.6045	63.9%	0.6201	70.0%	0.6128
350	34.0%	0.6031	64.8%	0.6186	63.3%	0.6111
400	27.5%	0.6020	63.6%	0.6171	68.2%	0.6096
450	26.2%	0.6011	57.7%	0.6155	53.3%	0.6082
500	14.3%	0.5998	37.2%	0.6138	52.0%	0.6067
550	14.5%	0.5987	27.2%	0.6123	51.7%	0.6055
600	11.7%	0.5977	27.5%	0.6112	46.2%	0.6048
650	7.3%	0.5962	19.8%	0.6100	36.3%	0.6039
700	5.7%	0.5943	15.5%	0.6090	38.7%	0.6034
750	4.3%	0.5916	12.3%	0.6075	24.4%	0.6025
800	3.5%	0.5889	8.0%	0.6063	16.8%	0.6019
850	2.0%	0.5857	4.2%	0.6042	11.1%	0.6000
900	1.7%	0.5824	2.0%	0.6013	8.2%	0.5959
950	1.2%	0.5711	1.3%	0.5946	2.7%	0.5851
1000	1.2%	0.5720	0.8%	0.5870	0.8%	0.5789

Table S5. **Accuracies and HFR across Input Resolutions and Timesteps on CUB.** The highest classification accuracy consistently corresponds to the highest HFR value across input resolutions.

S11. DDAE Classification Results on FGVC

Denoising Diffusion Autoencoders (DDAE) [89] extracts layer-wise output features from Diffusion Transformer [70]. We evaluate DDAE on FGVC datasets (*i.e.*, Aircraft, Stanford Cars, CUB, Stanford Dogs, Oxford Flowers and NABirds) and compare it against our method. The results, in Table S6, demonstrate that our method significantly outperforms DDAE across all datasets. For example, our method achieves **86.1%** accuracy on Stanford Cars and **78.4%** on NABirds, compared to 20.0% and 17.1% with DDAE, respectively. These consistent gains across datasets demonstrate the effectiveness of our approach.

Dataset	DDAE	Ours
Aircraft	19.3%	**77.5%**
Stanford Cars	20.0%	**86.1%**
CUB	25.4%	**78.6%**
Stanford Dogs	49.2%	**83.5%**
Oxford Flowers	73.4%	**90.6%**
NABirds	17.1%	**78.4%**
Mean	34.1%	**82.5%**

Table S6. **DDAE Classification Results on FGVC.**

S12. Additional Details on HFR

We compute HFR on the test dataset to ensure that it captures discriminative information from unseen data. The Gaussian high-pass filter threshold is set to 30.

S13. Additional Details on Fisher Score

We compute the Fisher score on the test dataset same as HFR. To obtain a one-dimensional embedding for each sample, we apply mean pooling over the token dimension of the two-dimensional feature representations. More results about relationship between Fisher Score and HFR are shown in Fig. S6

S14. Additional HFR Results across multiple DiT Models

To further examine the generalization of our proposed HFR, we evaluate it on different DiT models, including Vanilla DiT [70] and SiT [58], using Oxford Flowers dataset under the same experimental settings as in §4.1. As shown in Fig. S7, HFR values exhibit strong alignment with classification accuracy across both models. The highest accuracy consistently appears at the timestep where HFR reaches its maximum, demonstrating the robustness and generalization of HFR across DiT models.

S15. Discussion

S15.1. Asset License and Consent

Stable Diffusion 3.5 is licensed under https://huggingface.co/stabilityai/stable-diffusion-3.5-large/blob main/LICENSE.md

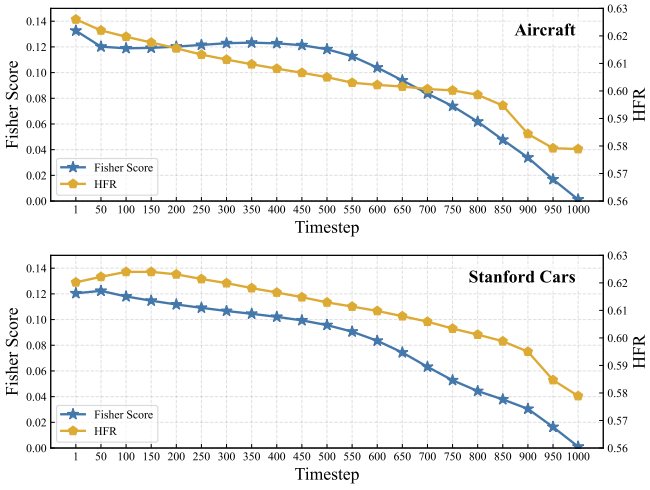

Figure S6. **More Comparison of HFR and Fisher Score** across timesteps on Aircraft (top) and Stanford Cars (bottom). The results show that HFR and Fisher Score exhibit consistent trends.

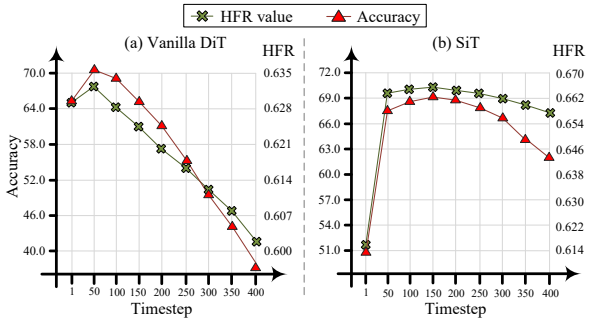

Figure S7. **Comparison of HFR and Classification Accuracy across multiple DiT Models** on Oxford Flowers. The alignment of peak accuracy with maximum HFR values demonstrates the consistent generalization of HFR.

Algorithm 1 Pseudo-code of A-SelecT in a PyTorch-like style.

```
# timesteps: timesteps used for computing HFR
# epochs: number of training epochs
# dif_model_path: diffusion model path

def A-SelecT(timesteps, epochs, dif_model_path):
    pipe = StableDiffusion3Pipeline.
        from_pretrained(dif_model_path)
    downstream_network = setup_head()
    optimizer = AdamW(downstream_network.
        named_parameters())
    HFR_list = []

    #get optimal timestep
    for t in timesteps:
       HFR_t = compute_HFR(t, train_dataloader,
           pipe)
       HFR_list.append(HFR_t)
    optimal_t, highest_HFR = get_optimal_t(
        HFR_list)

    #train downstream network
    for epoch in range(epochs):
       for step, batch in enumerate(train_loader
           ):
          features = pipe.transformer.extract(
              batch, optimal_t)
          output = downstream_network(features)
          loss = loss_fn(output, batch)
          loss.backward()
          optimizer.step()
          downstream_network.zero_grad()
```

S15.2. Reproducibility

To guarantee reproducibility, our full implementation shall be publicly released upon paper acceptance. We provide the pseudo code of our proposed A-SelecT in Algorithm 1.

S15.3. Technical Contributions

Our study presents three principal technical contributions: **First**, the inspiration for this research derives from the observation that high-frequency details, such as edges, textures, and corners, typically harbor more discriminative information. This insight has led to the development of the High-Frequency Ratio (HFR) metric. **Second**, a significant chal-

lenge in using diffusion models for extracting features is the selection of the most informative timestep from the extensive denoising trajectory. Traditional methods depend on exhaustive brute force searching or subjective manual selection, both of which are inefficient and potentially inaccurate. Our implementation of the HFR addresses this issue by providing a reliable and computationally efficient method for identifying the optimal timestep. **Third**, this paper is pioneering in its analysis of Diffusion Transformer (DiT)-based models for feature extraction. Through comprehensive experiments, we demonstrate that our approach not only overcomes the limitations of existing methods but also achieves state-of-the-art performance, substantiating the efficacy of DiT-based models as robust tools in representation learning.

S15.4. Limitations

Although our HFR performs effectively for selecting a single feature within the DiT model, it remains unclear whether this approach is equally viable for simultaneously selecting multiple features. In §S5, we find that under the pipeline of A-SelecT, additional features from different blocks or timesteps do not lead to better performance. However, other quantitative metric might be suitable under such the scenarios. In this sense, further investigation is needed to ascertain the applicability and effectiveness of the HFR metric in involving multi-feature extraction from DiT models.

S15.5. Future Work

As discussed in §S15.4, although our ablation study in §S5 demonstrates that incorporating more features leads to diminished downstream performance, it is plausible that additional features could provide more discriminative information. The effective utilization of this increased information warrants further investigation. Furthermore, incorporating additional features introduces new challenges, including training efficiency and the precise identification of multiple discriminative feature candidates.