USP: Unified Self-Supervised Pretraining for Image Generation and Understanding

Supplementary Material

A. More Ablation Studies

Incorporating Noise into the Pretraining Stage. We investigate the effect of introducing noise into the latent space during pretraining by adopting the formulation $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, with the objective of reconstructing the unmasked clean target. Despite the similarity of this setting to the data distribution used in generative tasks, it yields a FID of 32.20 and an Inception Score (IS) of 43.36 without Classifier-Free Guidance (CFG), which is inferior to our baseline performance. This suggests that masking modeling already serves as a highly effective form of data augmentation, and the introduction of additional, stronger noise significantly increases the difficulty of the learning task.

High-Resolution Results Table 10 demonstrates the effectiveness of USP at a higher resolution of 512×512. We initialize the downstream image generation task using weights obtained from pretraining with USP at 256 resolution and directly transfer them to the 512 resolution setting, adjusting the positional encodings via bilinear interpolation. The results confirm that USP maintains strong performance and transferability under higher-resolution settings.

Model	Params	Steps	FID (↓)	IS (↑)
SiT-B/2	130M	400K	42.80	37.37
USP	130M	400K	33.89	45.03

Table 10. Results at 512×512 Resolution.

Image Normalization (IN). Image normalization is a standard transformation in the community, and the default mean ([0.485, 0.456, 0.406]) and std ([0.229, 0.224, 0.225]) are widely used. However, VAE of DiT [66] has a different setting (mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5]). We perform pretraining using these two groups and report the downstream results in Table 11. Although the default setting of ImageNet has a lower loss (0.375), it doesn't bring in higher performance. Therefore, we utilize the settings of SD-VAE [71].

Method	$Loss_{pretrain}$	Acc_{SFT}	Acc_{LP}	FID
USP	0.465	82.6%	62.8%	28.26
ImageNet IN	0.375	82.0%	60.8%	78.23

Table 11. All results are reported using the same VAE as [66].

AdaLN-Zero or Skip Connection. AdaLN-zero initializes the attention and MLP branches with zero weights, effectively disabling them at the beginning of training. This approach alleviates the difficulties associated with the training of deep transformers [66]. We explored an alternative initialization strategy where the attention and MLP blocks are activated from the start by calibrating the gate bias to 1. This setting achieved similar performance to the zero-initialized approach. However, considering the minimal modification required for DiT and the established effectiveness of AdaLN-Zero in stabilizing training, we opted to retain the original AdaLN-Zero initialization scheme.

Comparison with UMD. UMD [35] integrates the diffusion loss and MAE loss through a weighted sum approach, aiming to achieve robust performance in both understanding and generation tasks. However, it still significantly underperforms compared to its single-task counterparts in each task. We attribute this shortfall to inherent conflicts between the MAE and diffusion models arising from their coupling.

In terms of performance and efficiency, our method achieves a substantial reduction in training cost, requiring only 15% of the computational resources compared to the DiT baseline to match its performance (see Table 5). This highlights the superior efficiency of our approach. In contrast, UMD [35] significantly underperforms the DiT baseline under the same computational constraints, further corroborating the effectiveness of our method. For image recognition tasks, our method achieves performance that is on par with, and in some cases surpasses, the strong baseline MAE. By contrast, UMD falls significantly short, demonstrating a clear performance gap. The results reported in Table 3 of the UMD study are not reliable and significantly deviate from the commonly reproduced and published results in the community. For instance, the FID score for DiT-L/2 with 400 epochs is reported as 9.6, which is consistent with widely accepted results. In contrast, it is known that the DiT-XL/2 architecture requires approximately 1400 epochs to achieve a similar FID score.

B. More Related Work

Generative Models with Auxiliary Task. MaskDiT [101] introduces an asymmetric encoder–decoder architecture based on the DiT framework, leveraging masked reconstruction to reduce the training cost of diffusion models. However, this approach involves substantial modifications to the original DiTs, resulting in limited transferability, and because the encoder always receives noisy inputs, it cannot be applied to downstream recognition tasks. Similarly,

MDT [30] employs an additional decoder for mask token modeling to enhance semantic contextual learning. Unlike [101], it performs noise prediction on all tokens rather than just on the unmasked ones. Although this improves generation quality, it also introduces significant computational overhead. Essentially, these methods incorporate an extra masked reconstruction task alongside noise prediction, which compromises architectural flexibility and, due to the input mismatch, restricts their applicability to understanding tasks. MAGE [50] proposes a unified framework for image generation and self-supervised representation learning, simultaneously conducting generation and representation learning through a variable mask ratio and an additional contrastive loss. MAGE utilizes VQ-GAN [28] encoder and quantizer to tokenize the input images and focuses solely on class-unconditional generation, whereas our approach operates in continuous space, aiming to enhance the generation performance of diffusion models while maintaining strong representation. In contrast, our method introduces minimal modifications to the original DiT/SiT architecture, ensuring excellent transferability and scalability. Moreover, by employing a single masked token reconstruction task, we decouple the heterogeneous optimization objectives between pretraining and downstream tasks.

MLLMs for Unified Understanding and Generation. Multimodal Large Language Models (MLLMs) have recently drawn extensive attention from both academia and industry. MLLMs [2, 16, 60] enable visual question answering in multimodal understanding tasks by aligning image embeddings with textual embeddings and jointly feeding them into a large language model, ultimately yielding text token IDs.

Several works address unified multimodal understanding and generation tasks, which can be broadly divided into two categories. One line of research [61, 78, 87, 91] employs VQ-GAN [28] or VQ-VAE [81] to tokenize images into discrete token IDs that are then fed into MLLMs for autoregressive image generation, thus aligning with the discrete input format of large language models. To mitigate potential performance degradation on understanding tasks due to discretization, [87] proposes utilizing discretized image token IDs only during the image generation stage. Another line of research [26, 31, 34, 63, 102] does not require converting images into discrete token IDs consistent with text. Instead, it leverages the image tokens output by the LLM as conditions for external generative models (e.g., Stable Diffusion [70]) to produce images. Our approach is a purely vision-based pre-training method, providing a robust weight initialization for subsequent fine-tuning on downstream understanding and generation tasks.

C. Visualization

C.1. Image Restoration

We visualize image reconstruction results using a ViT-Large model (encoder + decoder) pretrained with MAE and our method (see Figure 6). We randomly mask 25% of the patches and infer the restored images. Our method achieves much better performance.

config	value
optimizer	LARS [96]
base learning rate	0.1
weight decay	0
optimizer momentum	0.9
batch size	16384
learning rate schedule	cosine decay
warm-up epochs	10
training epochs	90(B), 50(L)
augmentation	RandomResizedCrop

Table 12. **Linear probing setting.**

C.2. Fully Tuning on ImageNet

The hyperparameter setting for fine-tuning in ImageNet is shown in Table 13.

C.3. Image Generation on ImageNet

The hyperparameter setting for generation in ImageNet is shown in Table 14.

config	value
optimizer	AdamW
base learning rate	1e-3
weight decay	0.05
optimizer momentum	β_1, β_2=0.9, 0.999
layer-wise lr decay [3]	0.75
batch size	1024
learning rate schedule	cosine decay
warmup epochs	5
training epochs	100 (B), 50 (L)
augmentation	RandAug (9, 0.5) [23]
label smoothing [77]	0.1
mixup [100]	0.8
cutmix [98]	1.0
drop path	0.1

Table 13. **Fine-tuning the whole neural network.**

C.4. Image Generation

We visualize more image generation results from DiT-XL/2 and SiT-XL/2, as shown in Figure 7 and Figure 8, respec-

Figure 6. Reconstruction results using ViT-Large on the ImageNet validation set. For each group of samples, we present the ground-truth image (left), MAE [39] (middle) reconstructed image and USP reconstructed image (right). The masking ratio is set to 75%.

tively. All results are generated with a CFG scale of 4.0.

D. HyperParameters

D.1. Linear Probe on ImageNet

We follow the setting of [19, 39] and show the details in Table 12.

E. Pretraining Code

We provide the code for the pre-training stage bundled with the supplementary materials. The pre-trained weights can be conveniently transferred to downstream understanding and generation tasks.

model	config	value
DiTs	optimizer	AdamW
	constant learning rate	1e-4
	weight decay	0.
	optimizer momentum	β_1, β_2=0.9, 0.999
	batch size	256
	augmentation	RandomHorizontalFlip
SiTs	optimizer	AdamW
	constant learning rate	1e-4
	weight decay	0.
	optimizer momentum	β_1, β_2=0.9, 0.999
	batch size	256
	augmentation	RandomHorizontalFlip
	path type	Linear
	prediction	velocity

Table 14. **Image generation on ImageNet.**

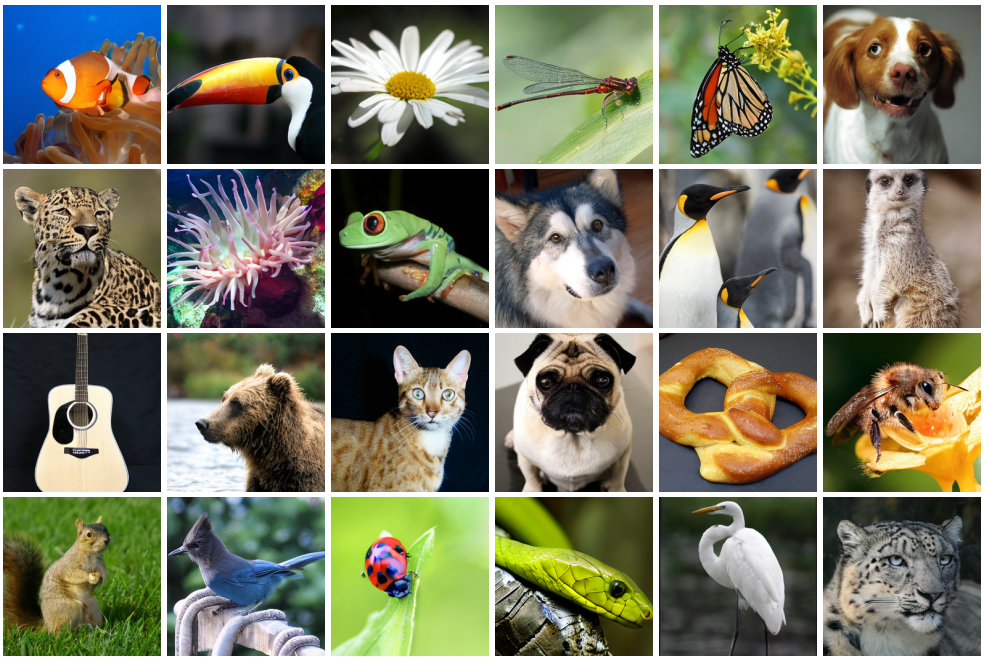

Figure 7. 256×256 generation samples: DiT-XL/2 (1.2M steps) with CFG=4.0.

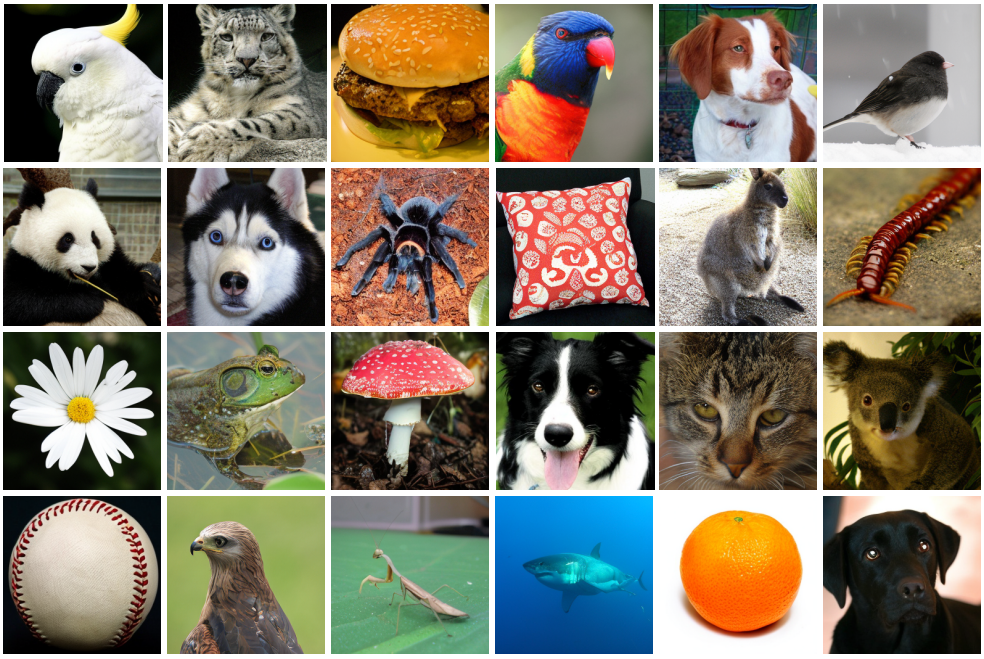

Figure 8. 256×256 generation samples: SiT-XL/2 (800K steps) with CFG=4.0.