Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think

Sihyun Yu[1] **Sangkyung Kwak**[1] **Huiwon Jang**[1] **Jongheon Jeong**[2]
Jonathan Huang[3] **Jinwoo Shin**[1*] **Saining Xie**[4*]
[1]KAIST [2]Korea University [3]Scaled Foundations [4]New York University

2024.10.16
Jaihoon Kim

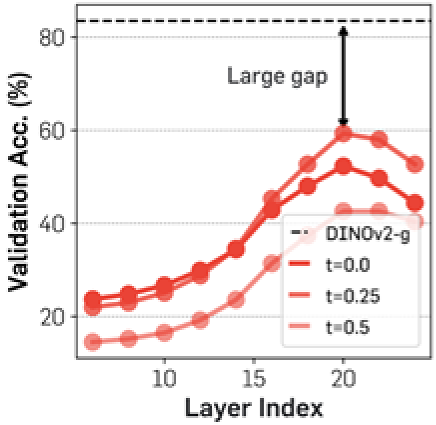
KAIST Visual AI Group

Self-Supervised Learning Visual Models

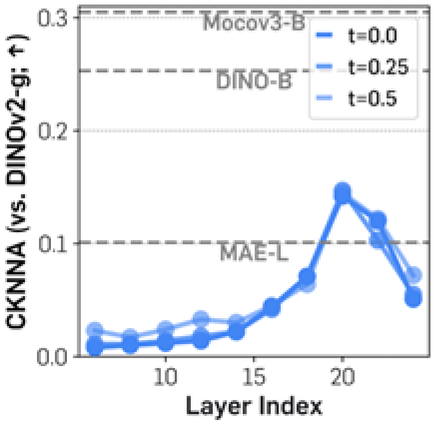
SotA self-supervised model such as DINOv2 show rich feature representations, showing its general applicability.

Motivation

(a): Representations of DMs exhibit a **significant semantic gap** compared to SotA self-supervised models (e.g., DINOv2) that show rich representations.



(a) Semantic gap: Linear probing



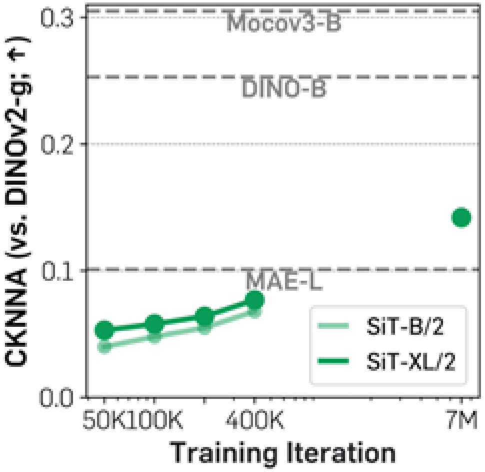
(b) Alignment to DINOv2-g

(b): The representations of the two models are **weakly** aligned.

Key Ideas

(c) While additional training slightly improves the alignment, it is not an efficient way.

→ The reconstruction task may not be ideal for learning effective representations as it does not incentivize the model for removing unnecessary details in **x**.

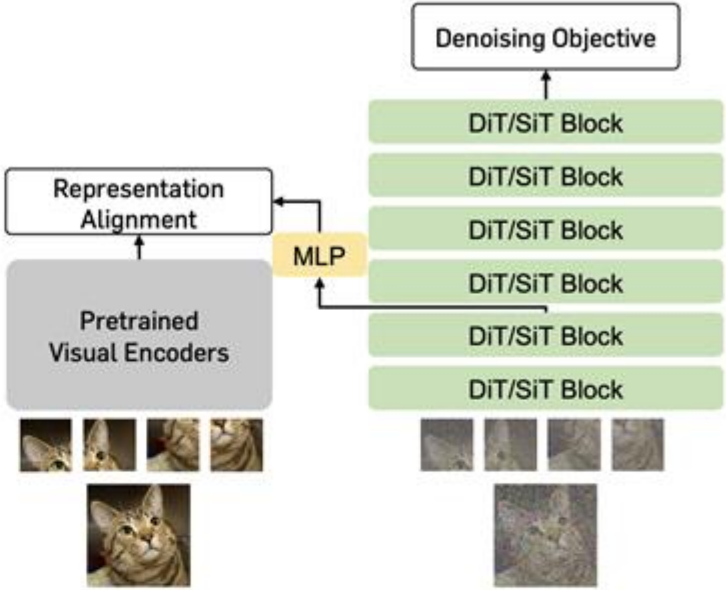


(c) Alignment progression

Meaningful representations can lead to efficient training of a diffusion model.

Method - REPresentation Alignment (REPA)

REPA aligns patch-wise projections of the diffusion model hidden states (noisy images) with pre-trained self-supervised visual representations (clean images).

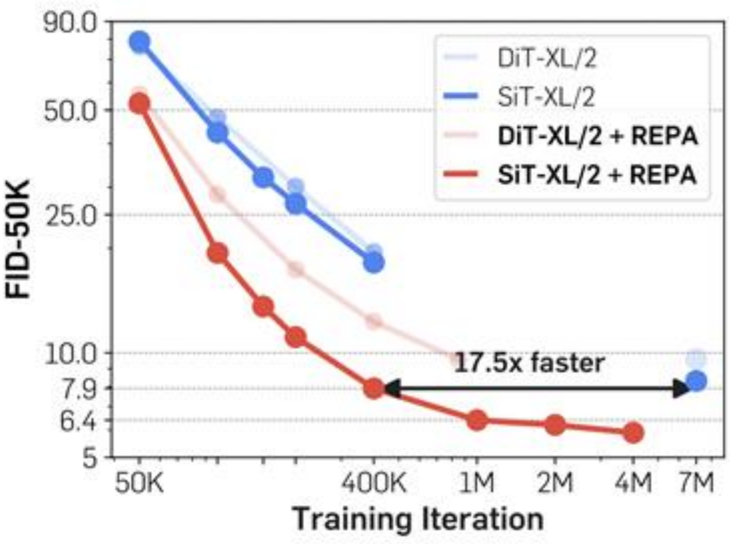


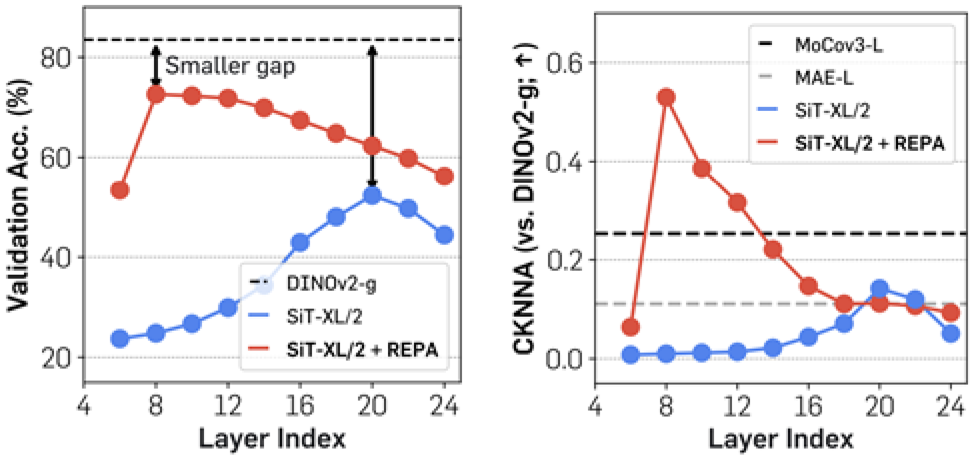
$$\mathcal{L}_{\text{REPA}}(\theta, \phi) := -\mathbb{E}_{\mathbf{x}_*, \epsilon, t}\left[\frac{1}{N}\sum_{n=1}^{N} \text{sim}(\mathbf{y}_*^{[n]}, h_\phi(\mathbf{h}_t^{([n]}))\right]$$

$$\mathcal{L} := \mathcal{L}_{\text{velocity}} + \lambda \mathcal{L}_{\text{REPA}}$$

Experiments

REPA accelerates the trainig process:
reaching the performance of 7M steps in less than 400K steps.





(a) Semantic gap: Linear probing

(b) Alignment to DINOv2-g

Experiments

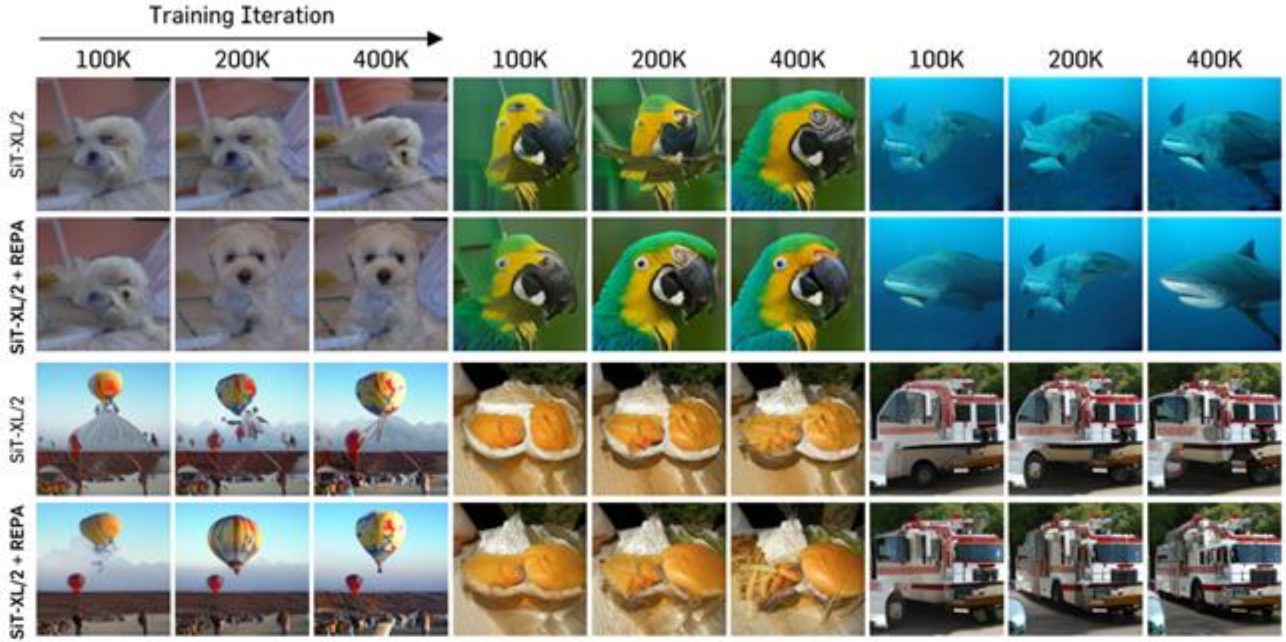


Table 3: **FID comparisons with vanilla DiTs and SiTs** on ImageNet 256×256. We do not use classifier-free guidance (CFG). ↓ denotes lower values are better. Iter. indicates the training iteration.

Model	#Params	Iter.	FID↓
DiT-L/2	458M	400K	23.3
+ REPA (ours)	458M	**400K**	**15.6**
DiT-XL/2	675M	400K	19.5
+ REPA (ours)	675M	**400K**	**12.3**
DiT-XL/2	675M	**7M**	9.6
+ REPA (ours)	675M	**850K**	**9.6**
SiT-B/2	130M	400K	33.0
+ REPA (ours)	130M	**400K**	**24.4**
SiT-L/2	458M	400K	18.8
+ REPA (ours)	458M	**400K**	**9.7**
+ REPA (ours)	458M	**700K**	**8.4**
SiT-XL/2	675M	400K	17.2
+ REPA (ours)	675M	**150K**	**13.6**
SiT-XL/2	675M	7M	8.3
+ REPA (ours)	675M	**400K**	**7.9**
+ REPA (ours)	675M	**1M**	**6.4**
+ REPA (ours)	675M	**4M**	**5.9**