ICLR

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







Sihyun Yu Sangkyung Kwak Huiwon Jang Jongheon Jeong Jonathan Huang Jinwoo Shin Saining Xie

KAIST NYU KOREA UNIVERSITY SCALED FOUNDATIONS

Introduction: Diffusion/Flow Models

Show state-of-the-art results in recent image/video generation

- Sora, SD3, Flux, DreamMachine, etc.

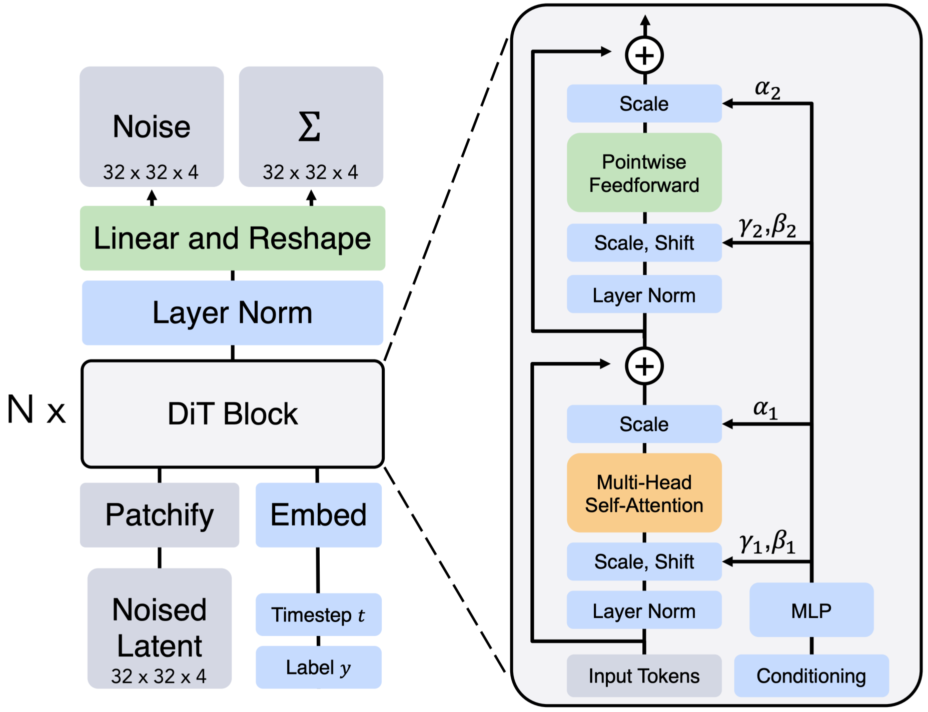
Images/Videos from https://stability.ai/news/introducing-stable-diffusion-3-5, https://openai.com/sora/

Diffusion Transformer (DiT) Training Is Too Slow

DiT: A recent scalable architecture for diffusion models

Issue: Extremely high training cost

- e.g.) Requires 1400 epochs on ImageNet to achieve reasonable FIDs



Latent Diffusion Transformer

DiT Block with adaLN-Zero

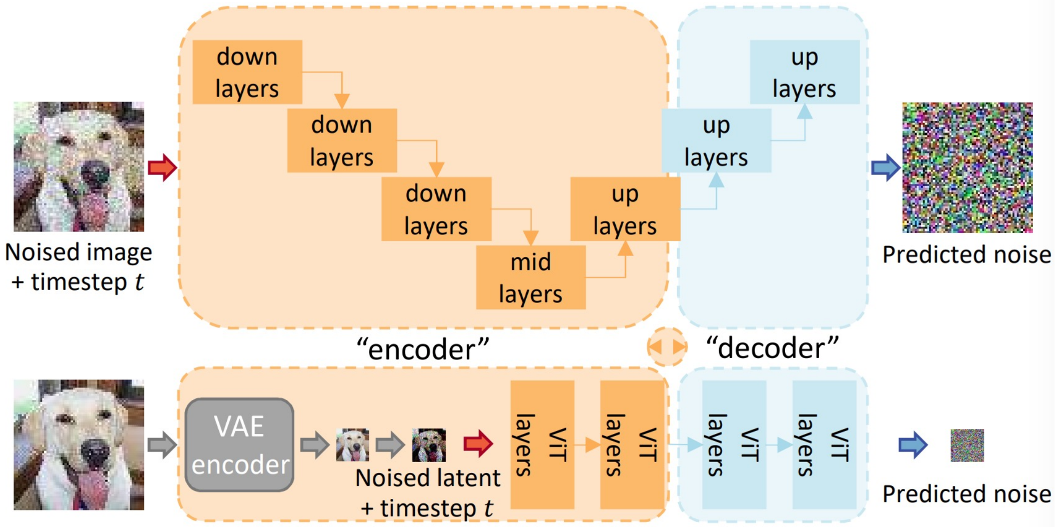
Model	Params(M)	Training Steps	FID ↓
DiT-S	33	400K	68.4
SiT-S	33	400K	**57.6**
DiT-B	130	400K	43.5
SiT-B	130	400K	**33.0**
DiT-L	458	400K	23.3
SiT-L	458	400K	**18.8**
DiT-XL	675	400K	19.5
SiT-XL	675	400K	**17.2**
DiT-XL	675	7M	9.6
SiT-XL	675	7M	**8.3**
DiT-XL (cfg=1.5)	675	7M	2.27
SiT-XL (cfg=1.5)	675	7M	**2.06**

Images from DiT [https://arxiv.org/abs/2212.09748] and SiT [https://arxiv.org/abs/2401.08740]

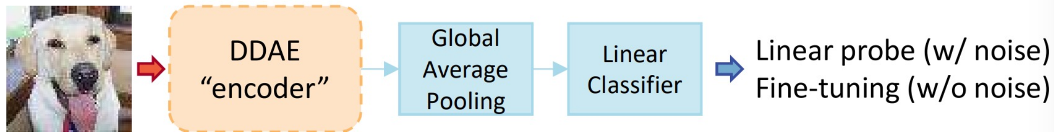
Generation for Representation Learning

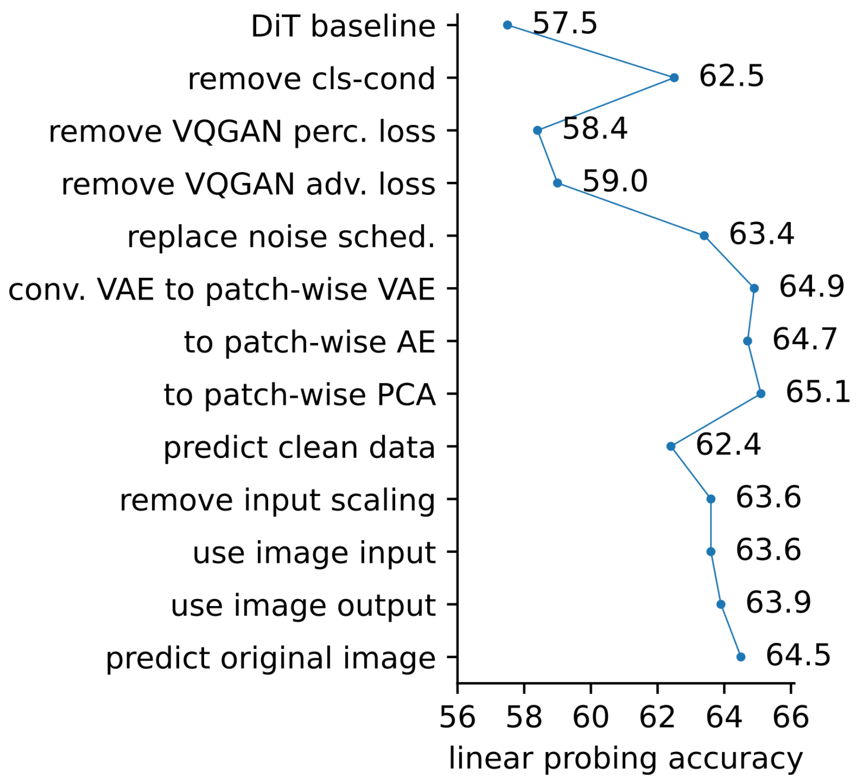
Recent work: Diffusion models learn acceptable representations

- e.g., DDAE [Xiang et al., 2023], l-DAE [Chen et al., 2024]
- But they still leg behind recent state-of-the-art SSL representations



(a) Denoising networks in pixel-space and latent-space diffusion models.





Images from DDAE [https://arxiv.org/pdf/2303.09769] and l-DAE [https://arxiv.org/pdf/2401.14404]

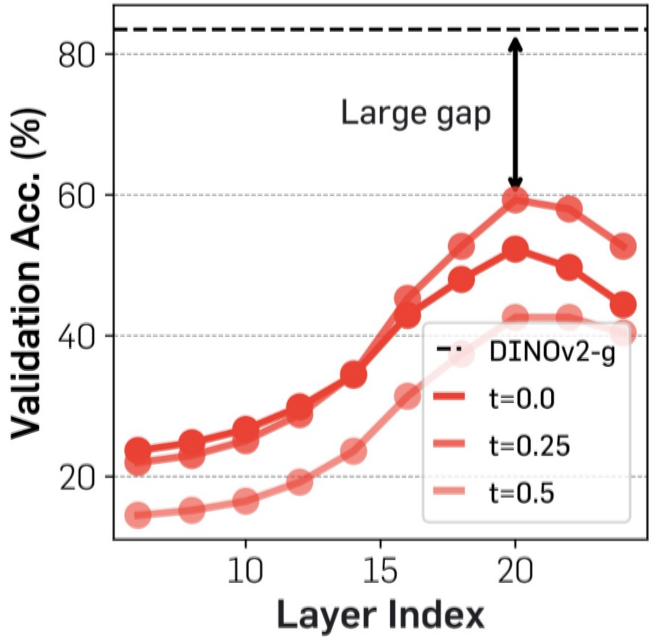
Our focus: Representation for Better Generation

Question: Can good representation improve training efficiency and generation quality of diffusion models?

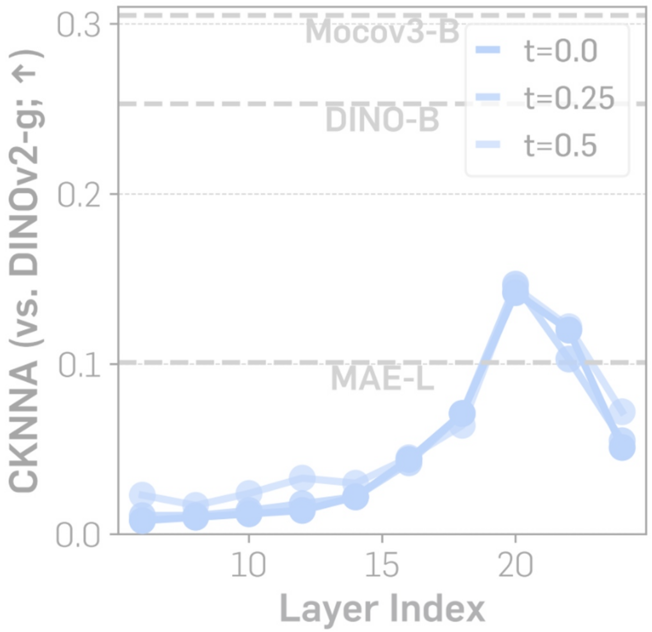
Observations from Pretrained SiT-XL/2 Representations

Three main observations from pretrained SiT-XL/2 representations:

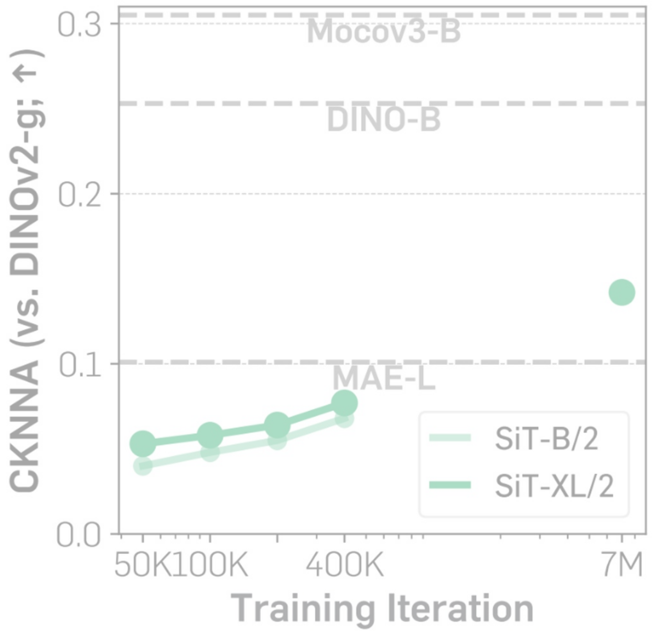
- The model learns reasonably (discriminative) representations
- Representation are partially aligned with state-of-the-art visual encoders
- Alignment improves slowly and inefficiently with increased training/model size



(a) Semantic gap: Linear probing



(b) Alignment to DINOv2-g

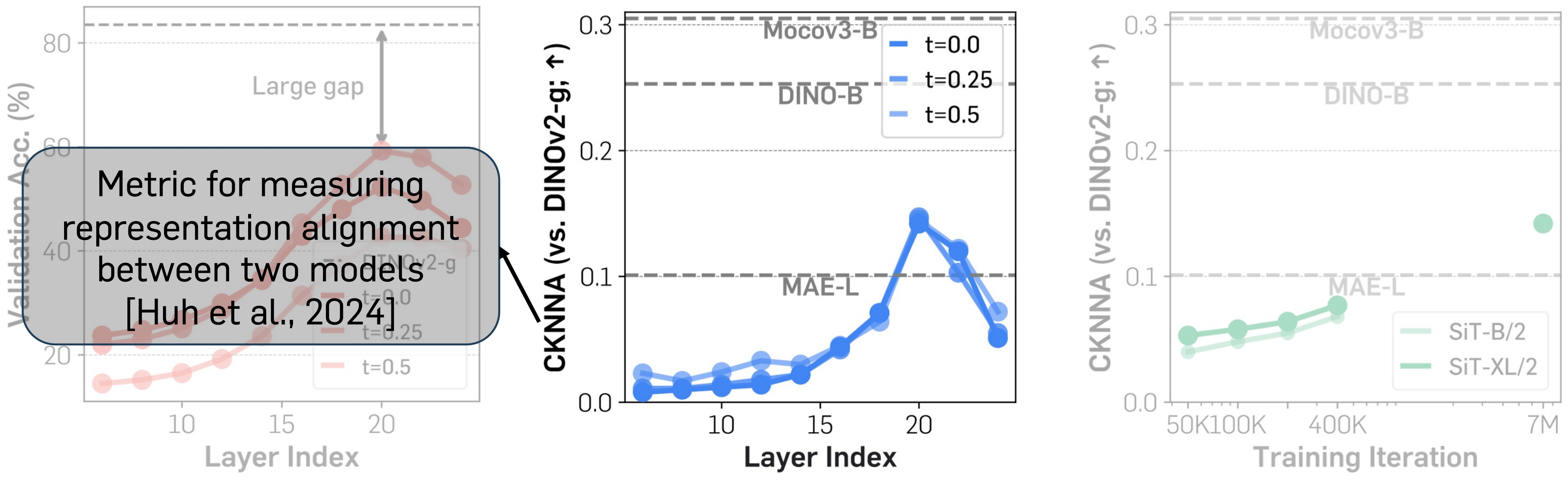


(c) Alignment progression

Observations from Pretrained SiT-XL/2 Representations

Three main observations from pretrained SiT-XL/2 representations:

- The model learns reasonably (discriminative) representations
- Representation are partially aligned with state-of-the-art visual encoders
- Alignment improves slowly and inefficiently with increased training/model size

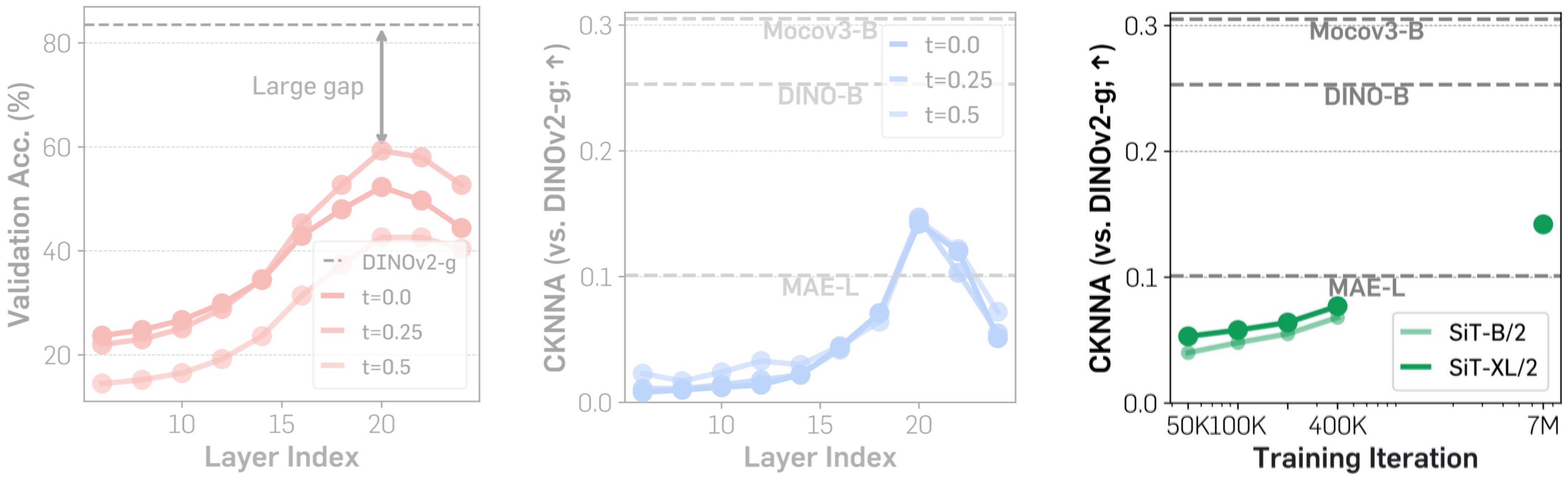


(a) Semantic gap: Linear probing (b) Alignment to DINOv2-g (c) Alignment progression

Observations from Pretrained SiT-XL/2 Representations

Three main observations from pretrained SiT-XL/2 representations:

- The model learns reasonably (discriminative) representations
- Representation are partially aligned with state-of-the-art visual encoders
- Alignment improves but inefficiently with increased training/model size



(a) Semantic gap: Linear probing

(b) Alignment to DINOv2-g

(c) Alignment progression

Representation for Better Generation

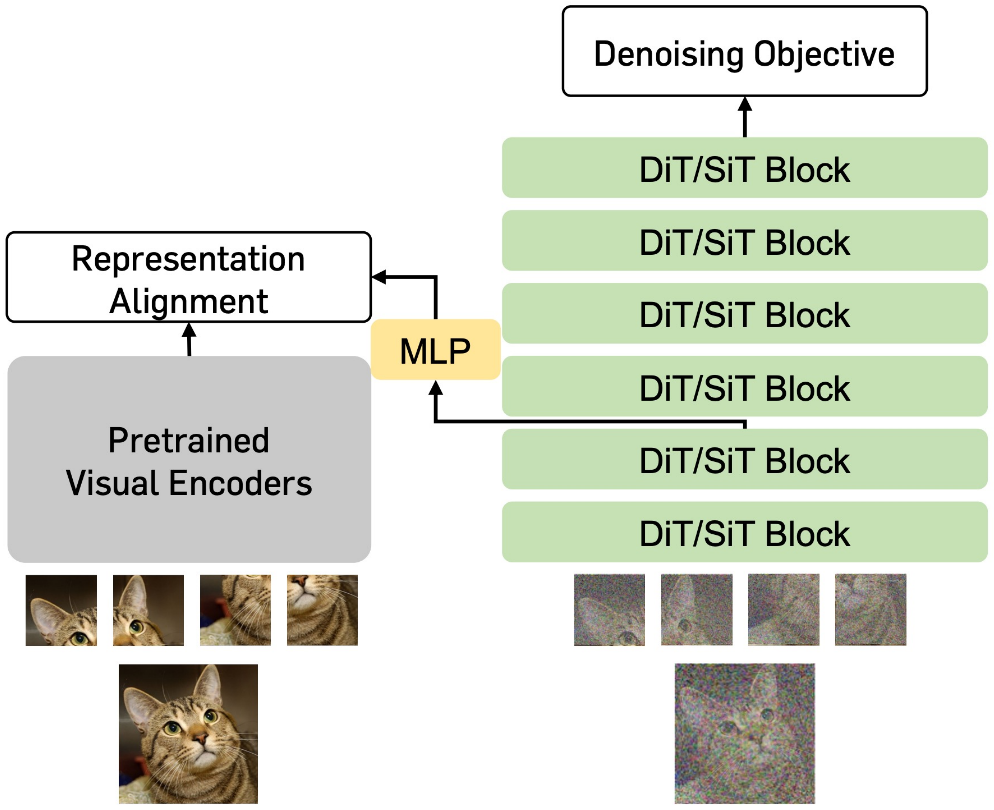
Hypothesis: Model should first learn good “representations” before focusing on “reconstructing” pixel-wise details

- The denoising objective alone might be insufficient to achieve this
- If we can guide representation learning of DiTs, then training becomes much easier

REPA: A Simple Regularization

We guide representation learning via a simple regularization

- **REPA:** Distills pretrained SSL representations into diffusion representations



Alignment between the target representation and the projected hidden state

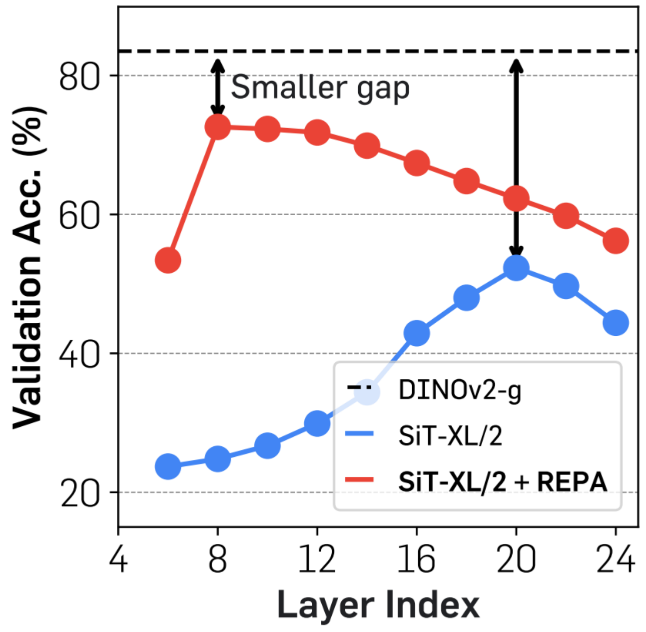
Patch Index

$$-\mathbb{E}_{\mathbf{x}_*,\epsilon,t}\left[\frac{1}{N}\sum_{n=1}^{N}\mathrm{sim}(\underbrace{\mathbf{y}_*^{[n]}}_{\text{Target}}, \underbrace{h_\phi}_{\text{MLP}}(\underbrace{\mathbf{h}_t^{[n]}}_{\text{Hidden state}}))\right]$$

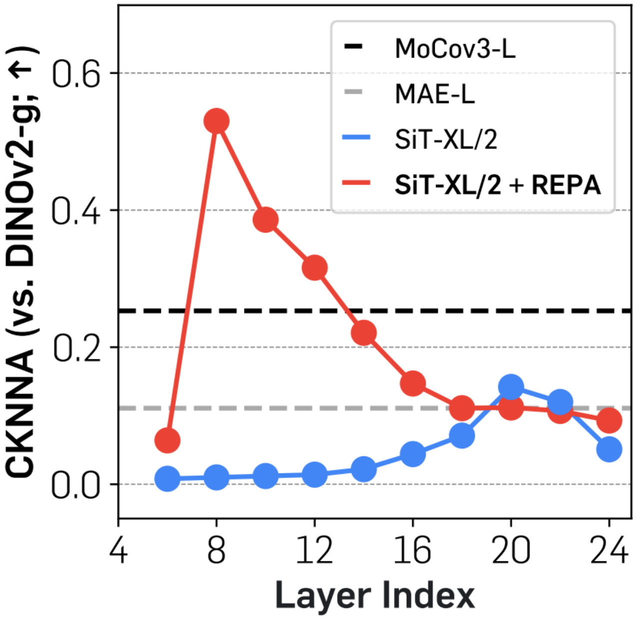
Bridging the Representation Gap

With REPA, the model shows

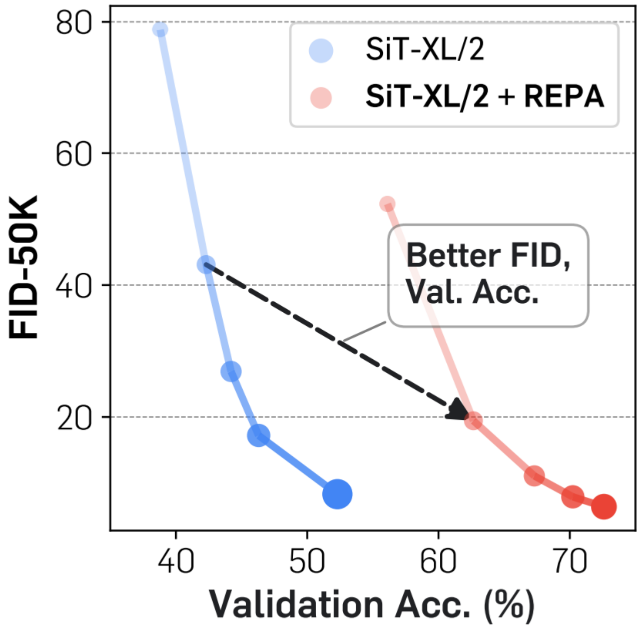
- **Reduced semantic gap:** Improved linear probing performance
- **Stronger alignment:** Higher CKNNA values
- **Improved training dynamics:** Better FID and validation accuracy



(a) Semantic gap: Linear probing



(b) Alignment to DINOv2-g

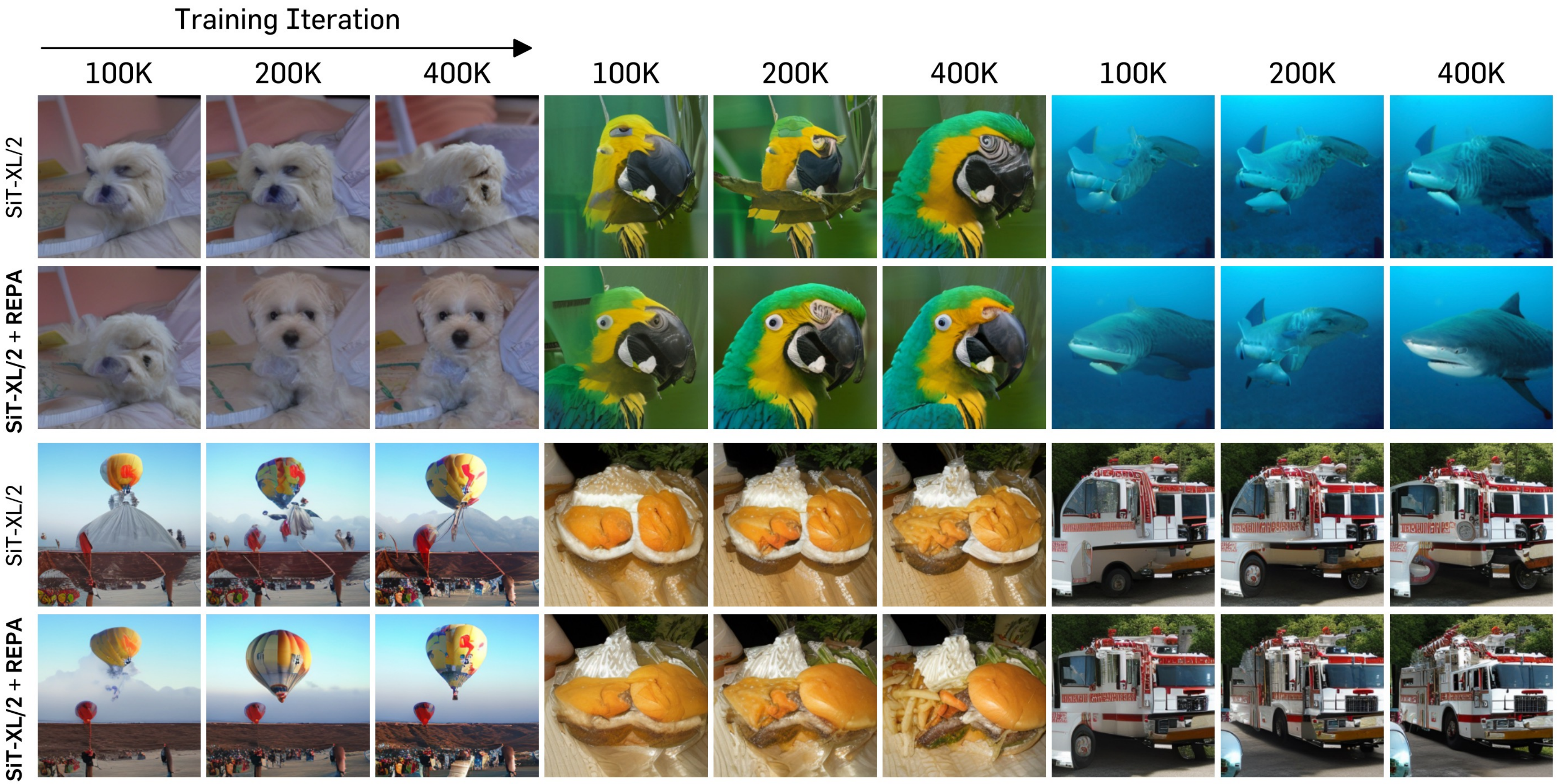


(c) Acc. and FID progression

REPA Improves Visual Scaling

REPA enables much better visual scaling

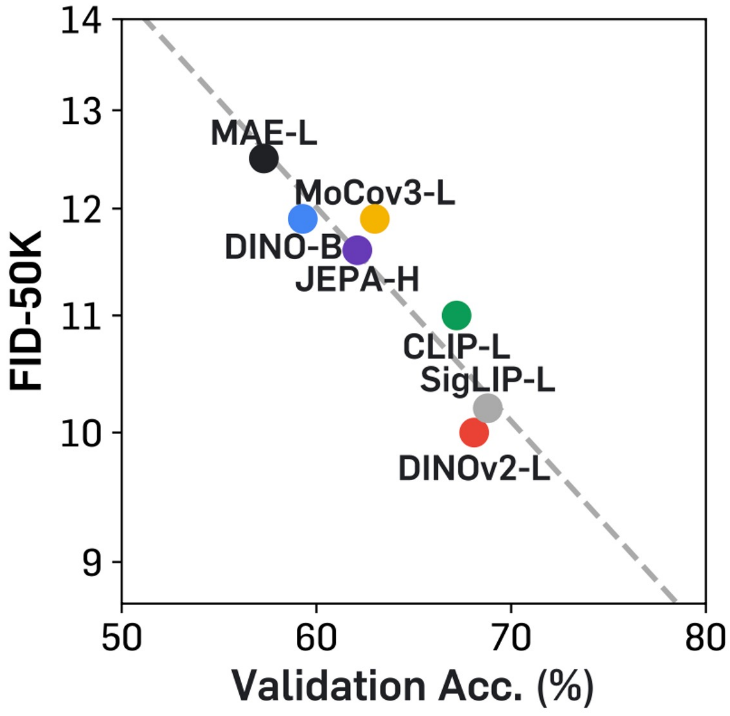
- The model produces significantly better generation at the same training iteration



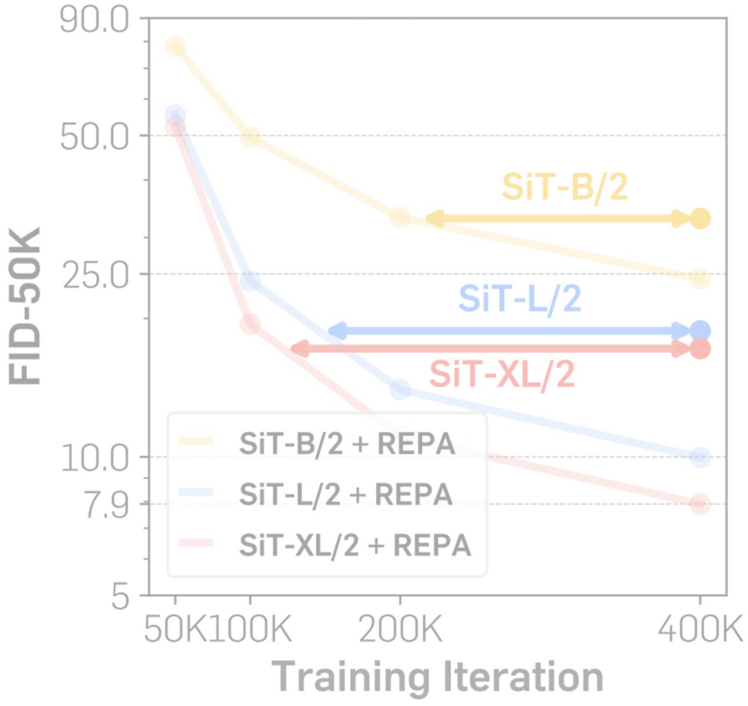
Analysis: Scalability (ImageNet 256x256)

For different target representations:

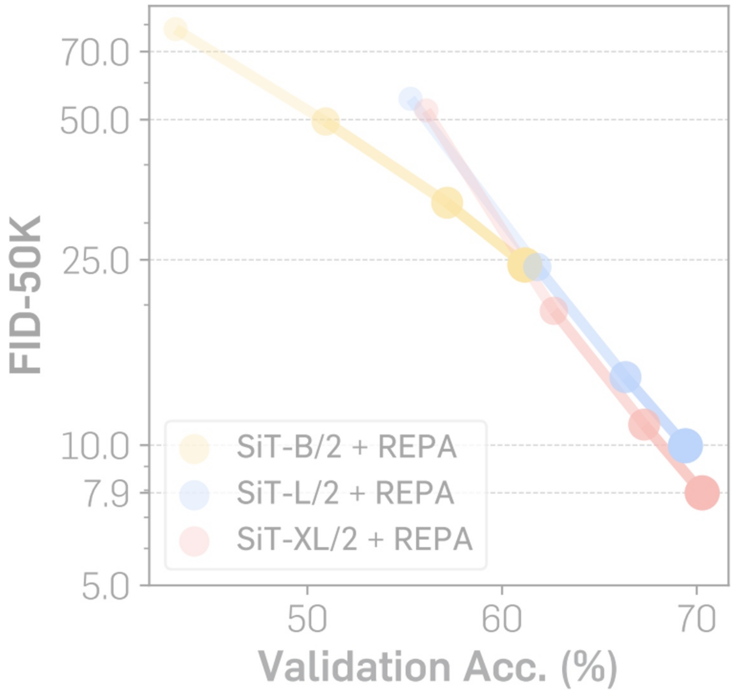
- Higher-quality representations lead to better linear probing results/generation quality



(a) Different visual encoders



(b) Relative convergence

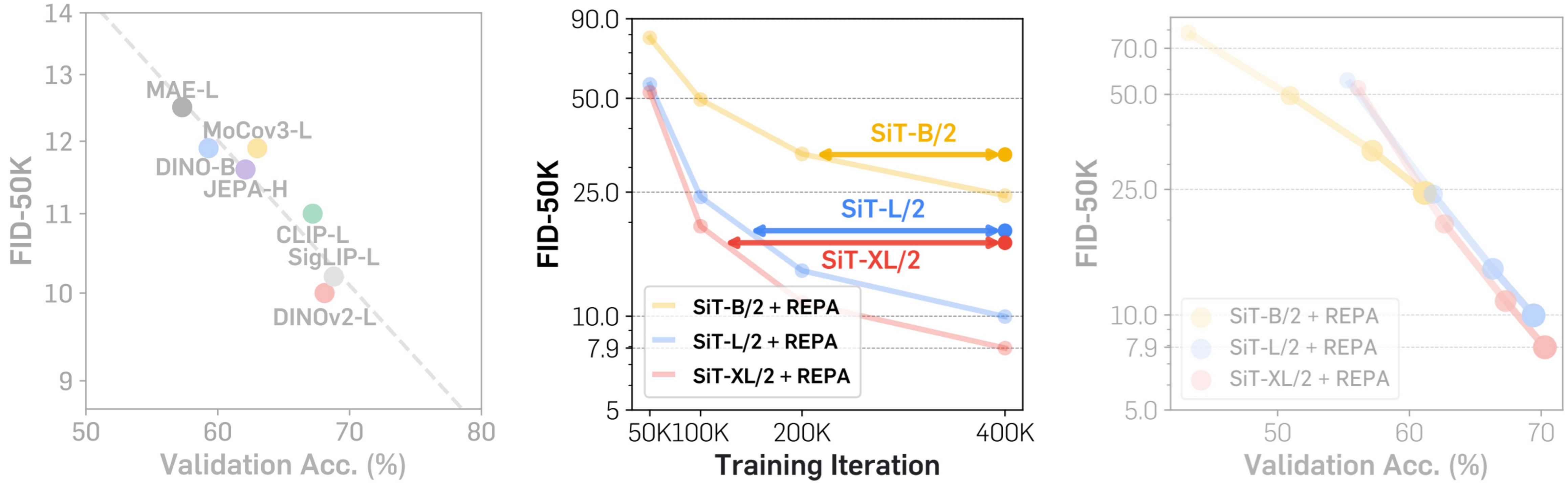


(c) Validation acc. vs. FID

Analysis: Scalability (ImageNet 256x256)

For different model sizes:

- Larger model with REPA reaches the same FID level faster as model size increases

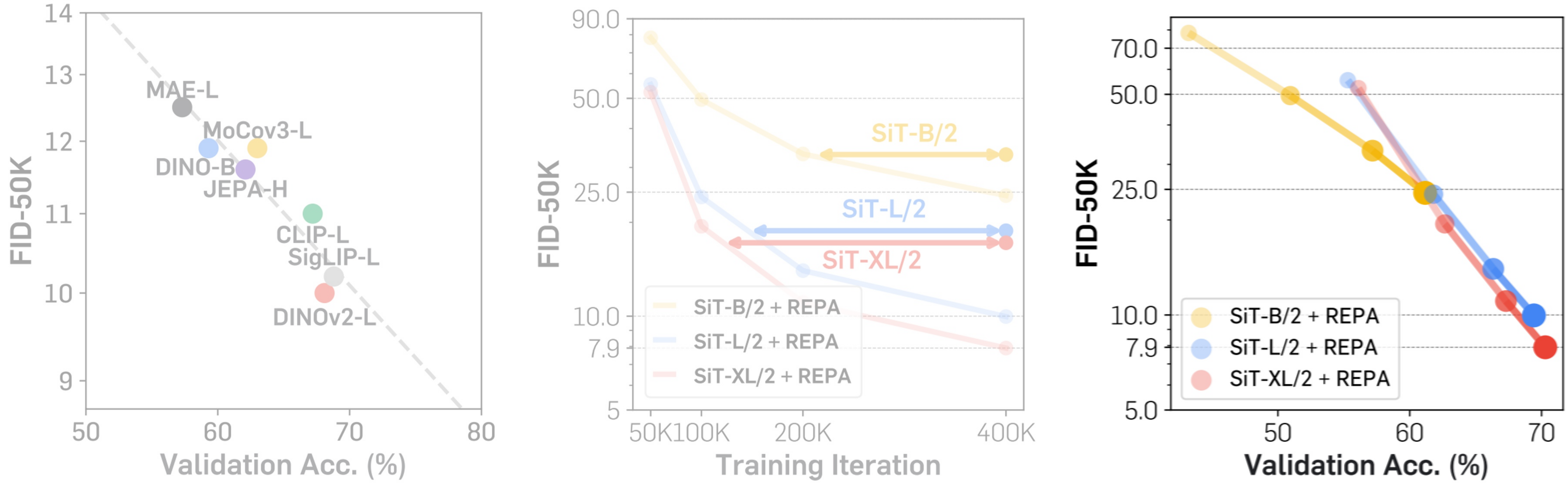


(a) Different visual encoders (b) Relative convergence (c) Validation acc. vs. FID

Analysis: Scalability (ImageNet 256x256)

For different model sizes:

- Larger models show steeper performance gain with REPA

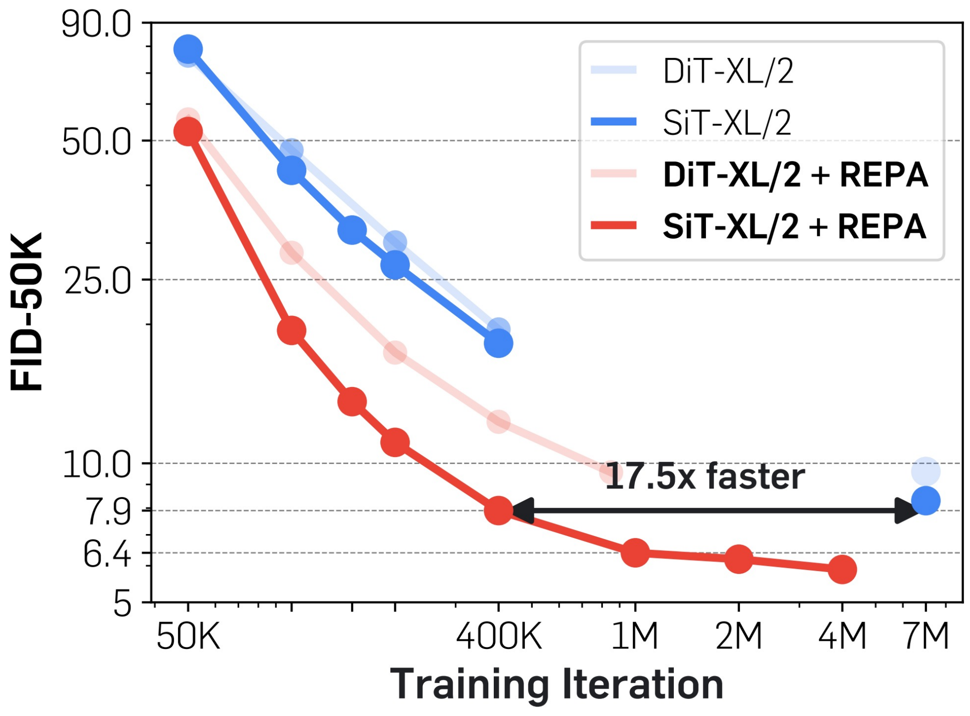


(a) Different visual encoders (b) Relative convergence (c) Validation acc. vs. FID

System-level Comparison: ImageNet 256x256

Results on ImageNet 256x256

- Accelerates training by over 17.5×
- Achieves state-of-the-art performance
- With guidance interval, FID=1.42



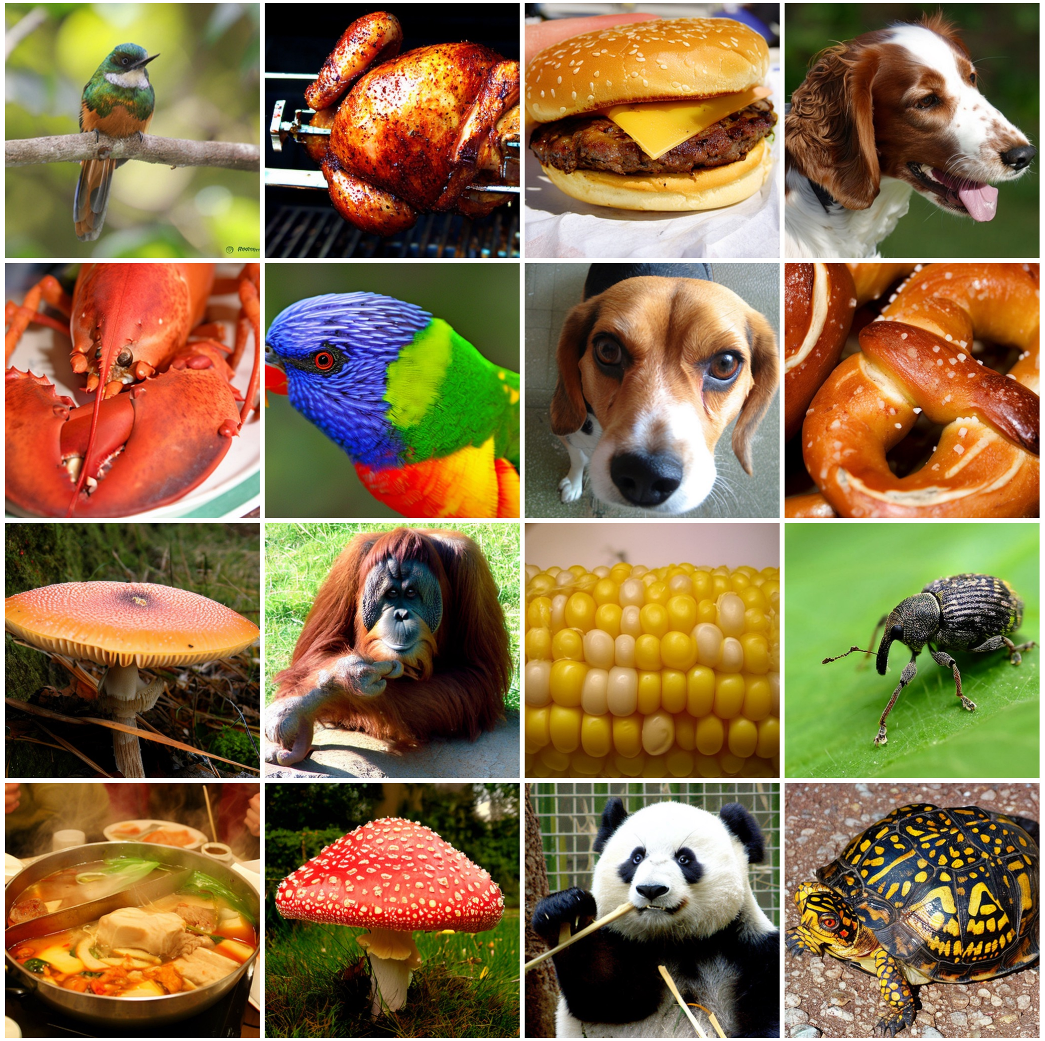
Model	Epochs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
Pixel diffusion						
ADM-U	400	3.94	6.14	186.7	0.82	0.52
VDM++	560	2.40	-	225.3	-	-
Simple diffusion	800	2.77	-	211.8	-	-
CDM	2160	4.88	-	158.7	-	-
Latent diffusion, U-Net						
LDM-4	200	3.60	-	247.7	0.87	0.48
Latent diffusion, Transformer + U-Net hybrid						
U-ViT-H/2	240	2.29	5.68	263.9	0.82	0.57
DiffiT*	-	1.73	-	276.5	0.80	0.62
MDTv2-XL/2*	1080	1.58	4.52	314.7	0.79	0.65
Latent diffusion, Transformer						
MaskDiT	1600	2.28	5.67	276.6	0.80	0.61
SD-DiT	480	3.23	-	-	-	-
DiT-XL/2	1400	2.27	4.60	278.2	**0.83**	0.57
SiT-XL/2	1400	2.06	4.50	270.3	0.82	0.59
+ REPA (ours)	200	1.96	**4.49**	264.0	0.82	0.60
+ REPA (ours)	800	1.80	4.50	284.0	0.81	0.61
+ REPA (ours)*	**800**	**1.42**	4.70	**305.7**	0.80	**0.65**

System-level Comparison: ImageNet 512x512

Results on a higher-resolution dataset

- Also shows significant improvements
- REPA exceeds the vanilla model's FID >7.5x faster

Model	Epochs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
Pixel diffusion						
VDM++	-	2.65	-	278.1	-	-
ADM-G, ADM-U	400	2.85	5.86	221.7	0.84	0.53
Simple diffusion (U-Net)	800	4.28	-	171.0	-	-
Simple diffusion (U-ViT, L)	800	4.53	-	205.3	-	-
Latent diffusion, Transformer						
MaskDiT	800	2.50	5.10	256.3	0.83	0.56
DiT-XL/2	600	3.04	5.02	240.8	0.84	0.54
SiT-XL/2	600	2.62	4.18	252.2	0.84	0.57
+ REPA (ours)	80	2.44	4.21	247.3	0.84	0.56
+ REPA (ours)	100	2.32	**4.16**	255.7	**0.84**	0.56
+ REPA (ours)	200	**2.08**	4.19	**274.6**	0.83	**0.58**



System-level Comparison: Text-to-Image Generation

Results on MS-COCO

- Shows better image quality
- Improves image-text alignment

Method	Type	FID
AttnGAN (Xu et al., 2018)	GAN	35.49
DM-GAN (Zhu et al., 2019)	GAN	32.64
VQ-Diffusion (Gu et al., 2022)	Discrete Diffusion	19.75
DF-GAN (Tao et al., 2022)	GAN	19.32
XMC-GAN (Zhang et al., 2021)	GAN	9.33
Frido (Fan et al., 2023)	Diffusion	8.97
LAFITE (Zhou et al., 2021)	GAN	8.12
U-Net (Bao et al., 2023)	Diffusion	7.32
U-ViT-S/2 (Bao et al., 2023)	Diffusion	5.95
U-ViT-S/2 (Deep) (Bao et al., 2023)	Diffusion	5.48
MMDiT (ODE; NFE=50)	Diffusion	6.05
MMDiT+REPA (ODE; NFE=50)	Diffusion	**4.73**
MMDiT (SDE; NFE=250)	Diffusion	5.30
MMDiT+REPA (SDE; NFE=250)	Diffusion	**4.14**

REPA: Summary & Conclusion

Summary: Representation alignment significantly improves DiT/SiT training

We propose REPA = REPresentation Alignment

1. **Hypothesis:** "Good representation" is a key for diffusion transformer training
2. Shows great scalability in terms of target representation, model size, etc.
3. State-of-the-art FID on ImageNet 256x256 (FID=1.42)
4. Significant improvements in higher resolution datasets or text-to-image generation

Paper | Project page | Code