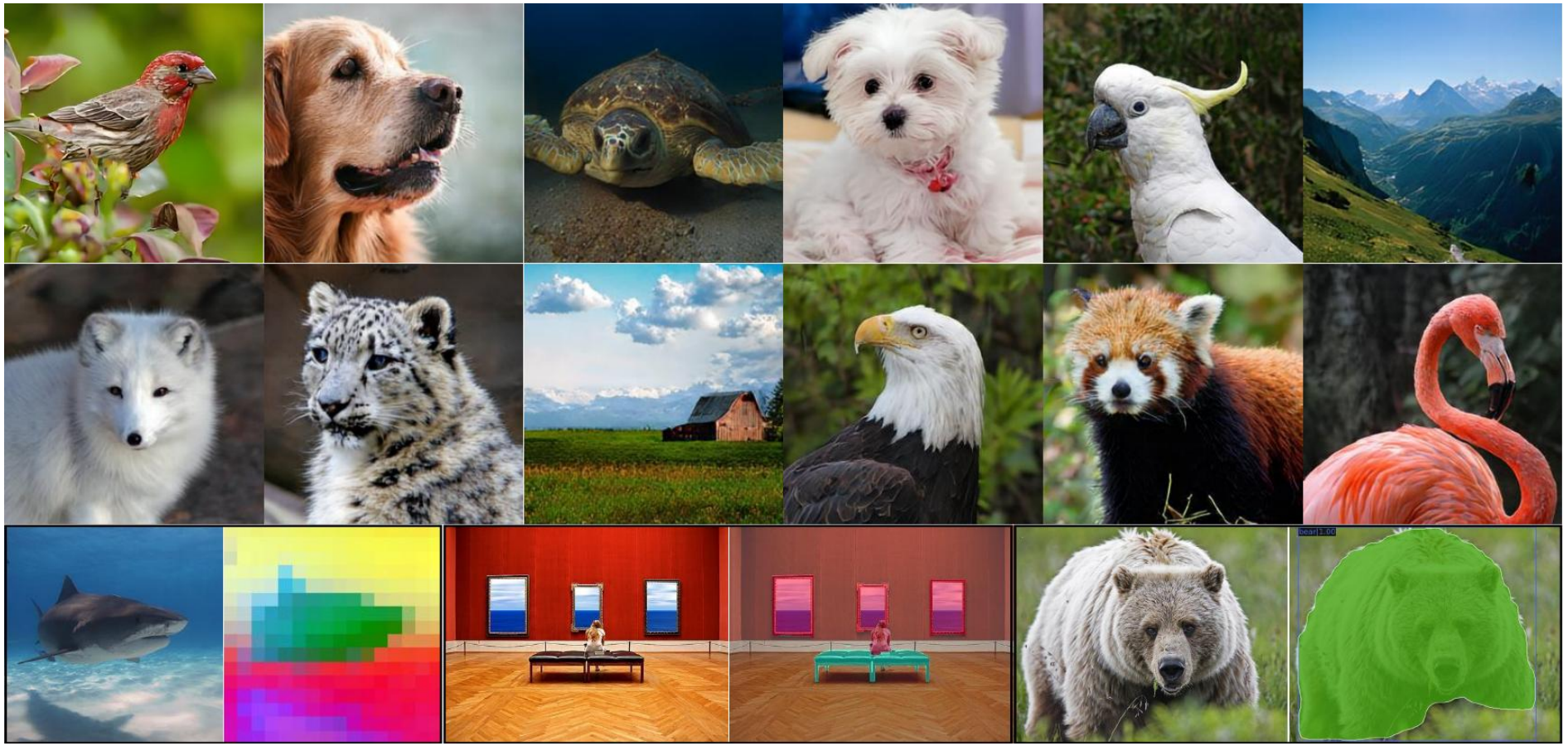
VUM

Visual Unified Model for Image Generation and Perception

Zidong Wang* Yiyuan Zhang* Xiaoyu Yue Xiangyuan Xue
Zhangquan Chen Manyuan Zhang† Wanli Ouyang†

The Chinese University of Hong Kong
The University of Sydney · Tsinghua University

One model, two visual capabilities



Can one pixel-native model learn image generation and transferable visual representations?

Direct fusion causes severe degradation

ImageNet-1K baselines, 160 training epochs

Model	FID ↓	Linear probe (%) ↑
MAE-B	—	44.85
JiT-B, 24 layers	9.16	6.97
JiT-B + 75% masking	**316.9**	**6.57**

Heavy masking compounds the information loss from noise

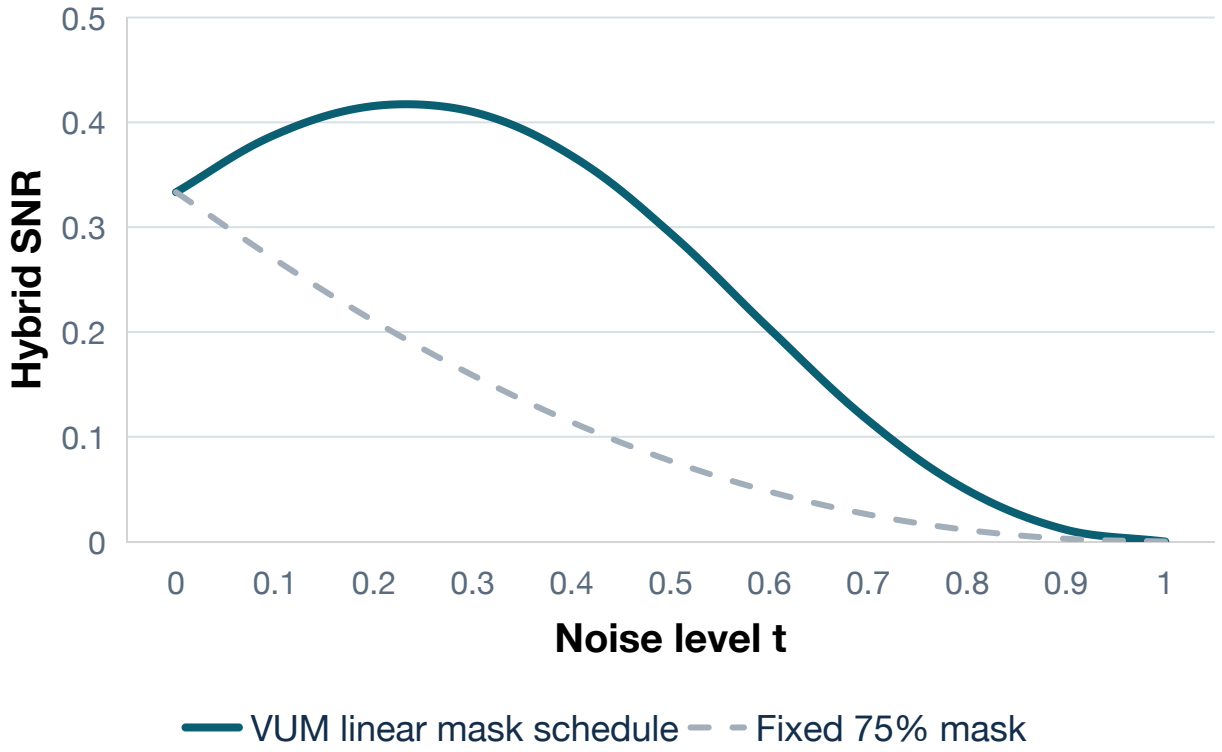
Noise-dependent masking

$$x = (1 - t)x + t\varepsilon$$

$$r(t) = 0.75(1 - t)$$

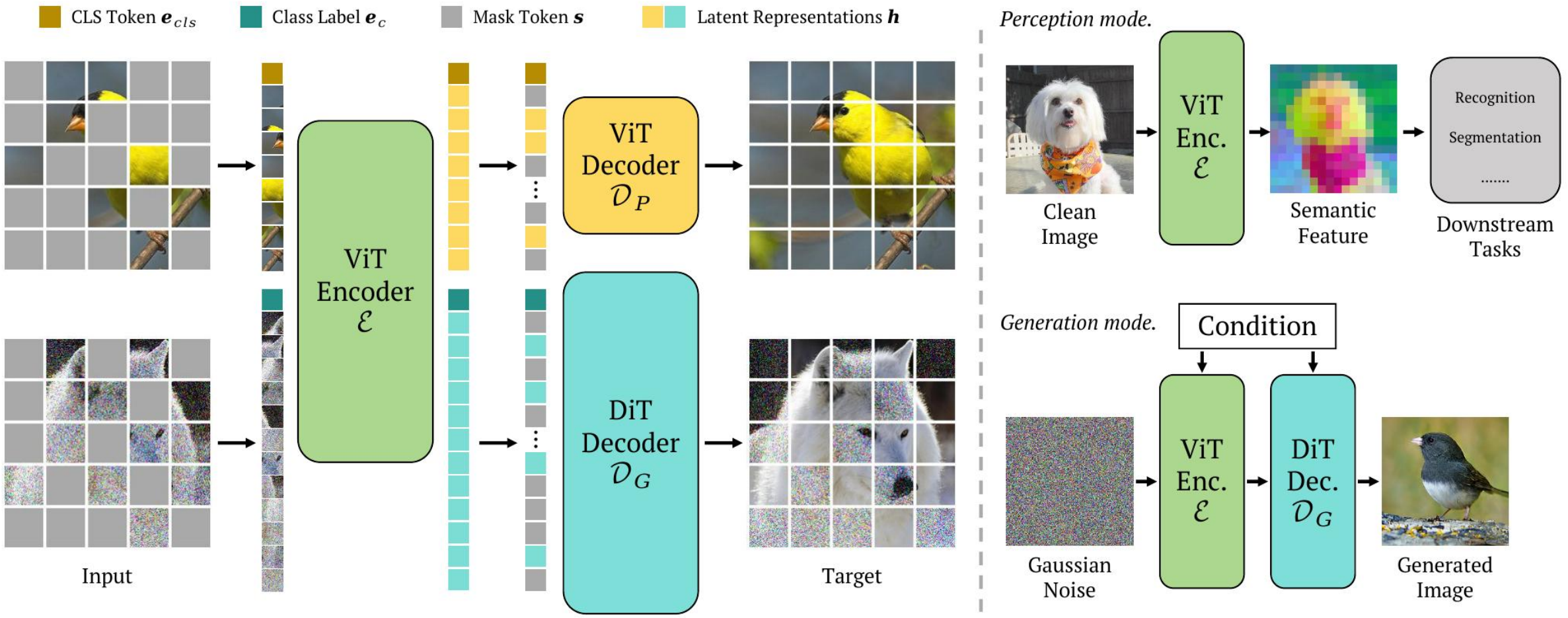
Low noise: recover missing patches
High noise: retain spatial context

t is diffusion time within each sample



$$HSNR(r, t) = (1 - r)(1 - t)^2 / [(1 - r)t^2 + r]$$

VUM architecture



(a) Overview of Visual Unified Training Pipeline.

(b) Dual-Task Inference.

Recovery targets and inference paths

Joint training

Clean state, t = 0

Auxiliary ViT recovers normalized clean pixels on masked patches

Noisy state, t > 0

DiT predicts clean visible pixels and normalized noisy masked pixels

$$L = \beta L_{} + L_G$$

$\beta = 1,\ \lambda = 0.1$

Dual-task inference

Perception

Shared ViT encoder

Generation

Shared encoder + DiT

Auxiliary decoder is removed at inference

Competitive ImageNet-256 generation

Model	Parameters	Epochs	FID ↓
DiT	675M + 86M	1400	2.27
PixNerd	700M	320	1.95
DeCo	682M	320	1.90
JiT	953M	600	1.86
VUM	**763M**	**800**	**1.91**

1.91 FID

One shared encoder also supports perception

All results use classifier-free guidance. Model sizes and training budgets differ.

The shared encoder transfers to perception

Metric	MAE ViT-L	VUM ViT-L
ImageNet fine-tuning (%)	86.0	84.4
ImageNet linear probe (%)	76.0	68.4
ADE20K mIoU	53.6	51.3
COCO box AP	53.3	53.1
COCO mask AP	47.2	46.3

COCO box AP: 53.1 vs. 53.3

VUM: 800 pretraining epochs. MAE: 1600 epochs.
Dense prediction uses downstream fine-tuning.

Components improve both generation and perception

VUM-B ablation, 160 epochs, CFG = 2.5

Configuration	FID ↓	Linear probe (%) ↑
Naive mask + noise	316.9	6.57
Encoder–decoder, static mask	159.6	21.25
Noise-dependent masking	**22.07**	**43.25**
Full VUM with auxiliary decoder	**9.77**	**48.70**

The schedule helps at fixed architecture

The full model adds auxiliary decoder capacity

Masked recovery controls the task balance

Masked-region weight λ	FID ↓	Linear probe (%) ↑
0	9.52	22.79
0.1	**9.77**	**48.70**
0.5	12.42	46.36
1.0	13.41	48.08

+25.91 points in linear probing

λ = 0 to 0.1 changes FID from 9.52 to 9.77

Remaining gaps and scope

7.6 points

Linear-probing gap to MAE
68.4% vs. 76.0%

Generation scope

Class-conditional ImageNet
at 256 × 256

Scaling questions

Pixel-space computation
Text conditioning
Web-scale learning

Representation quality still leaves room for improvement

Shared pixel recovery supports both capabilities

Noise-dependent masking

Shared encoding with specialized decoders

ImageNet generation

FID 1.91

COCO detection

53.1 box AP

github.com/WZDTHU/VUM