Understand Before You Generate:

Self-Guided Training for Autoregressive Image Generation

Presented by Xiaoyu Yue

THE UNIVERSITY OF SYDNEY

Celebrating 175 years

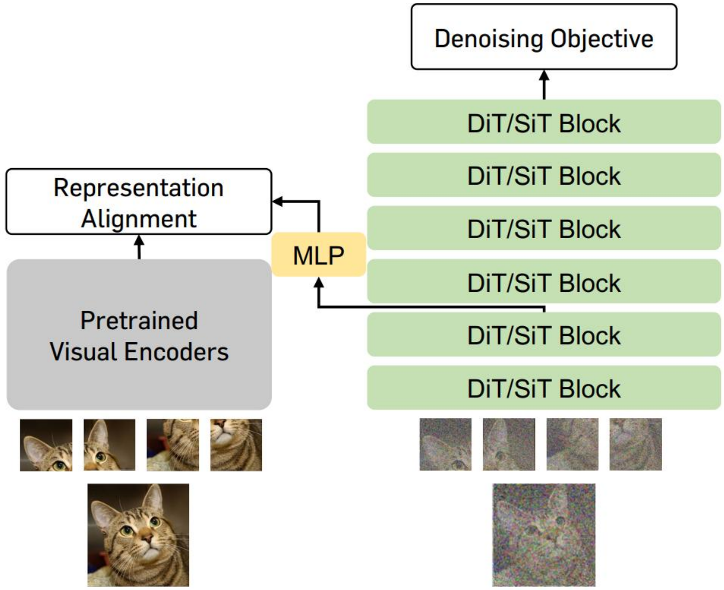
香 港 大 學
THE UNIVERSITY OF HONG KONG

香 港 中 文 大 學
The Chinese University of Hong Kong

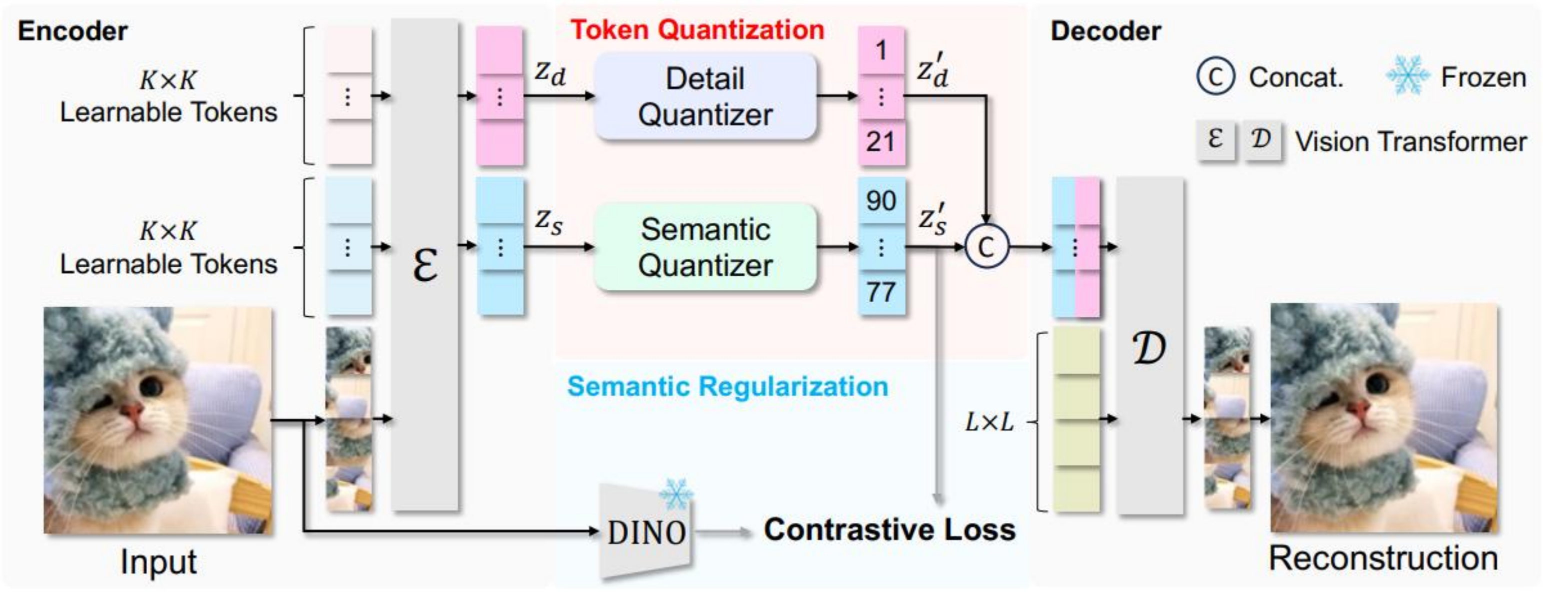
上 海 人 工 智 能 实 验 室
Shanghai Artificial Intelligence Laboratory

Representation Learning in Generative Models

Image Understanding Enhances Image Generation Performance



REPA [1]



ImageFolder [2]

[1] Representation alignment for generation: Training diffusion transformers is easier than you think.

[2] Imagefolder: Autoregressive image generation with folded tokens.

Representation Learning in Generative Models

Image Understanding in Autoregressive Generative Models

Autoregressive models can learn high-level semantics from text.

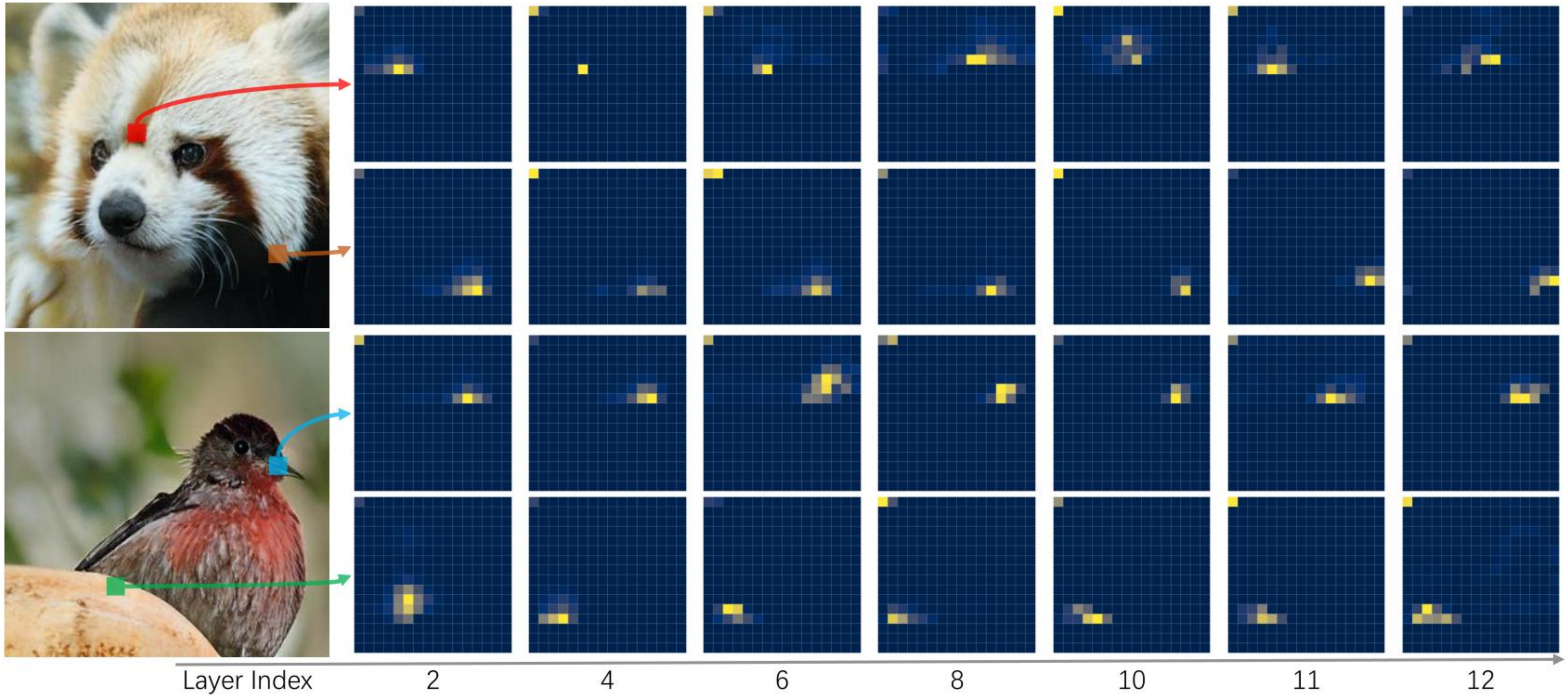
But what about images?

An orange cat hiding on the wheel of a red car.

Representation Learning in Generative Models

Image Understanding in Autoregressive Generative Models

Local and conditional dependence.

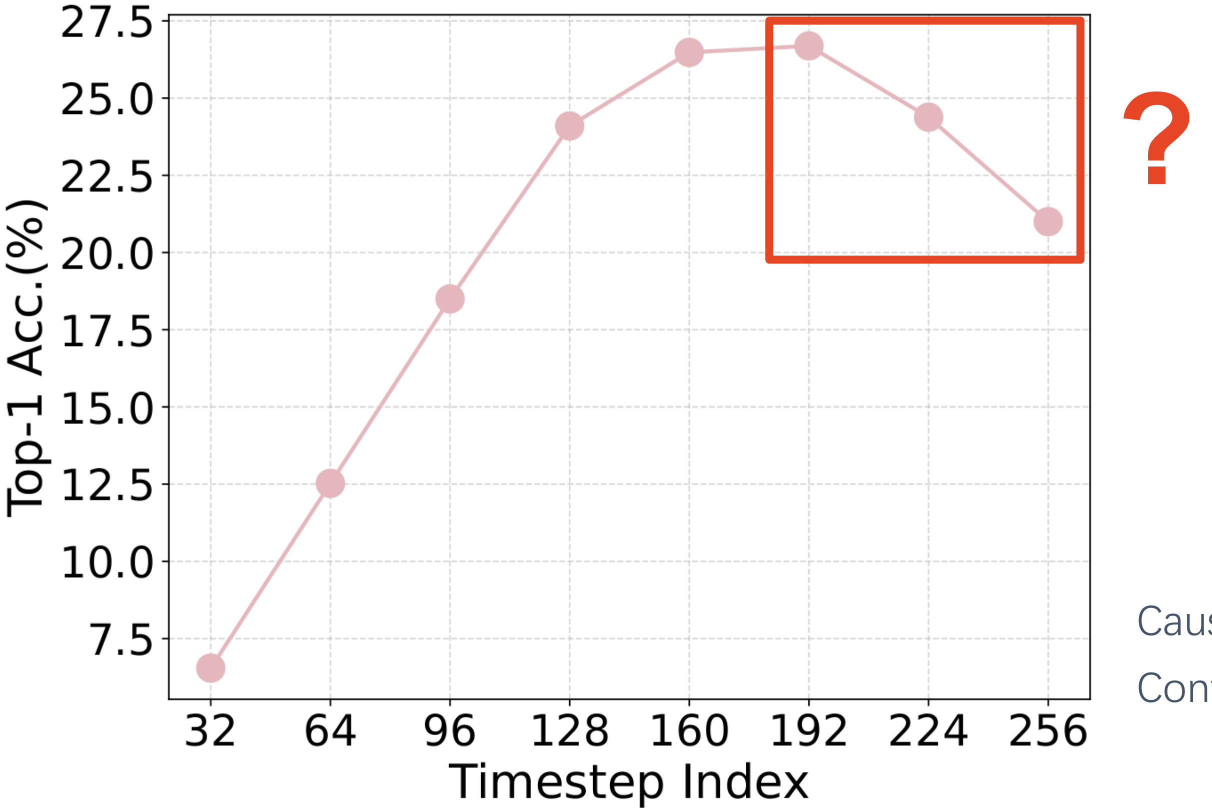


Autoregressive models primarily rely on local and conditional information.

Representation Learning in Generative Models

Image Understanding in Autoregressive Generative Models

Inter-step semantic inconsistency

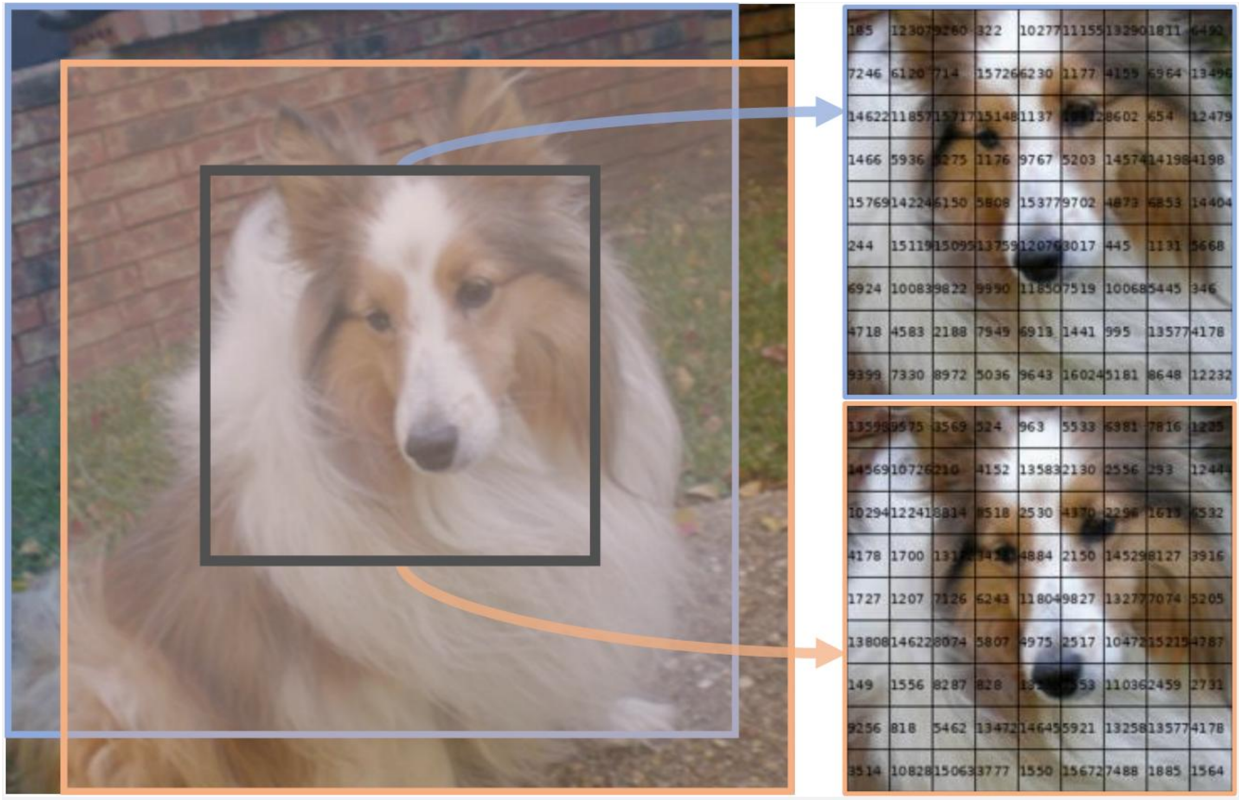


Causal Attention Challenges Bi-directional Image Context Modeling.

Representation Learning in Generative Models

Image Understanding in Autoregressive Generative Models

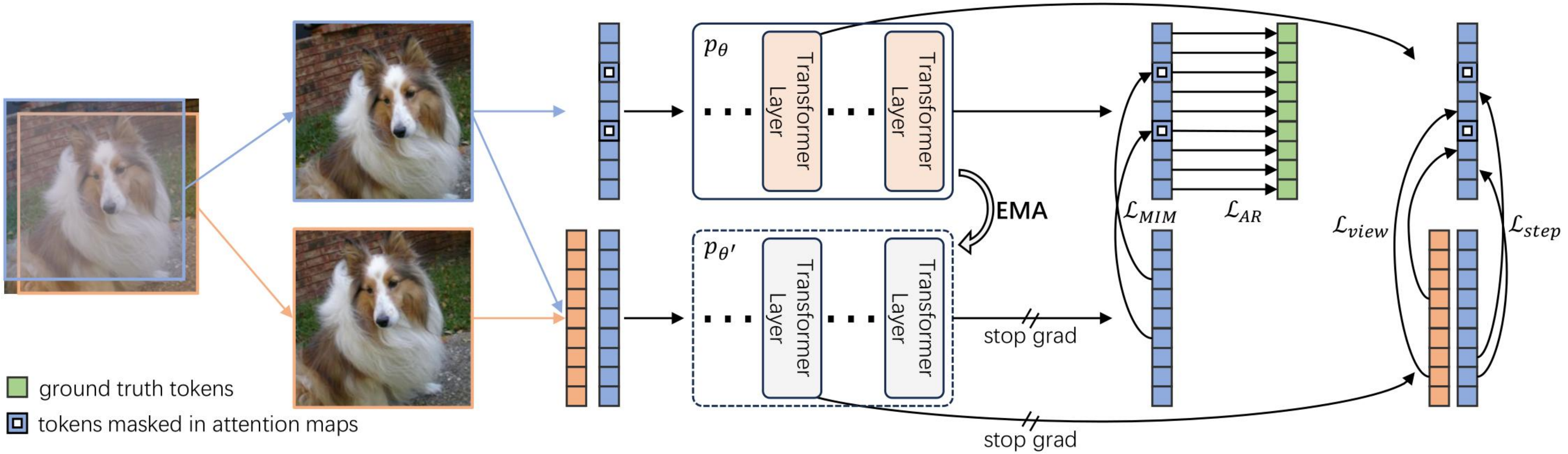
Spatial invariance deficiency



Visual tokens lack invariance.

Representation Learning in Generative Models

ST-AR: Self-guided Training for AutoRegressive models



Two Branches:

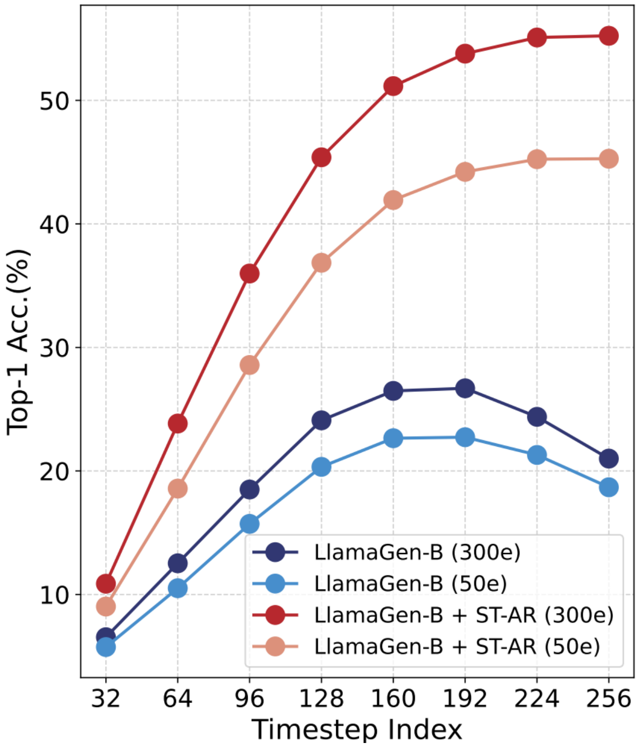
- Student Network
- Teacher Network

Three Objective Functions:

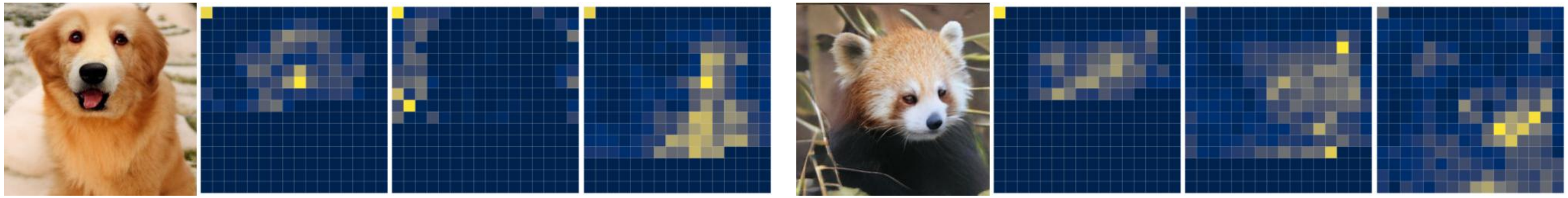
- Masked learning for longer contexts.
  - MIM Loss ($\mathcal{L}_{MIM}$)
- Contrastive learning for consistency.
  - Inter-step contrastive loss ($\mathcal{L}_{step}$)
  - Inter-view contrastive loss ($\mathcal{L}_{view}$)

Representation Learning in Generative Models

ST-AR: Experiments



Model	#Params	Epochs	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
LlamaGen-B	111M	50	31.35	8.75	39.58	0.57	0.61
+ ST-AR	111M	50	26.58	7.70	49.91	0.60	0.62
LlamaGen-B	111M	300	26.26	9.22	48.07	0.59	0.62
+ ST-AR	111M	300	18.44	6.71	66.18	0.64	0.62
LlamaGen-L	343M	50	21.81	8.77	59.18	0.62	0.64
+ ST-AR	343M	50	12.59	6.79	91.19	0.65	0.64
LlamaGen-L	343M	300	13.45	8.32	82.29	0.66	0.64
+ ST-AR	343M	300	9.38	6.64	112.71	0.70	0.65
LlamaGen-XL†	775M	300	15.55	7.05	79.16	0.62	**0.69**
LlamaGen-XXL†	1.4B	300	14.65	8.69	86.33	0.63	0.68
LlamaGen-3B†	3.1B	300	9.38	8.24	112.88	0.69	0.67
LlamaGen-XL	775M	50	19.42	8.91	66.20	0.61	0.67
+ ST-AR	775M	50	9.81	6.94	109.77	0.71	0.63
+ ST-AR	775M	300	**6.20**	**6.47**	**147.47**	**0.73**	0.65



Thank you!

THE UNIVERSITY OF
SYDNEY

Celebrating 175 years