Hierarchical Masked Autoregressive Models with Low-Resolution Token Pivots

Guangting Zheng

Joint work with Yehao Li, Yingwei Pan, Jiajun Deng, Ting Yao, Yanyong Zhang, Tao Mei (ICML 2025)

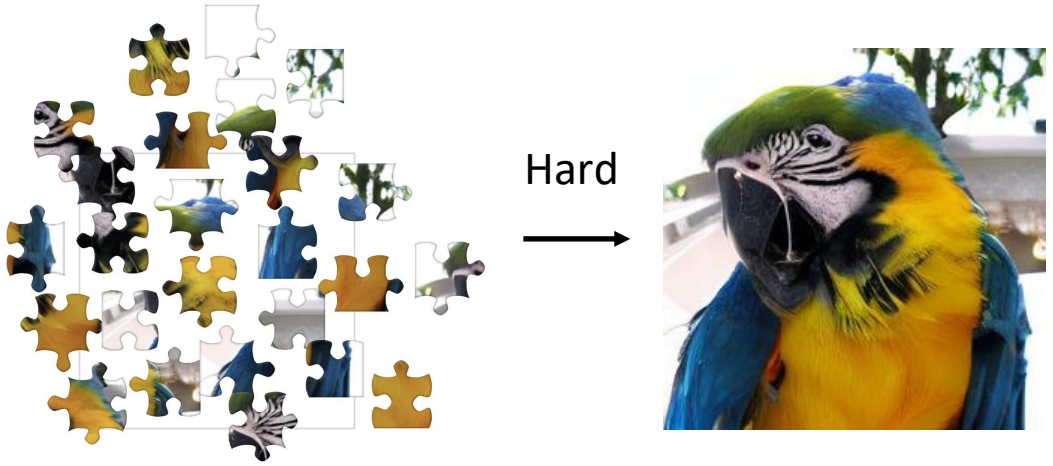
中国科学技术大学
University of Science and Technology of China

智象未来
HiDream.ai

Lack of global context in next-token visual generation

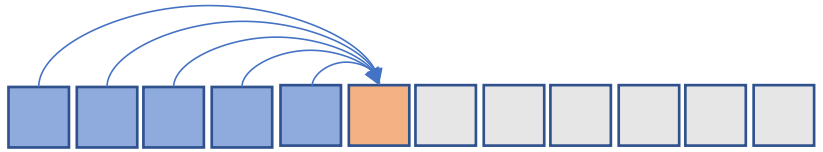
$$P(x_1, x_2, \ldots, x_N)$$

Next-Token Visual Generation Paradigm
(Analogy to Jigsaw Puzzles)



$$\prod_N P(x_i \mid x_1, x_2, \ldots, x_{i-1})$$

Incapable of utilizing global context

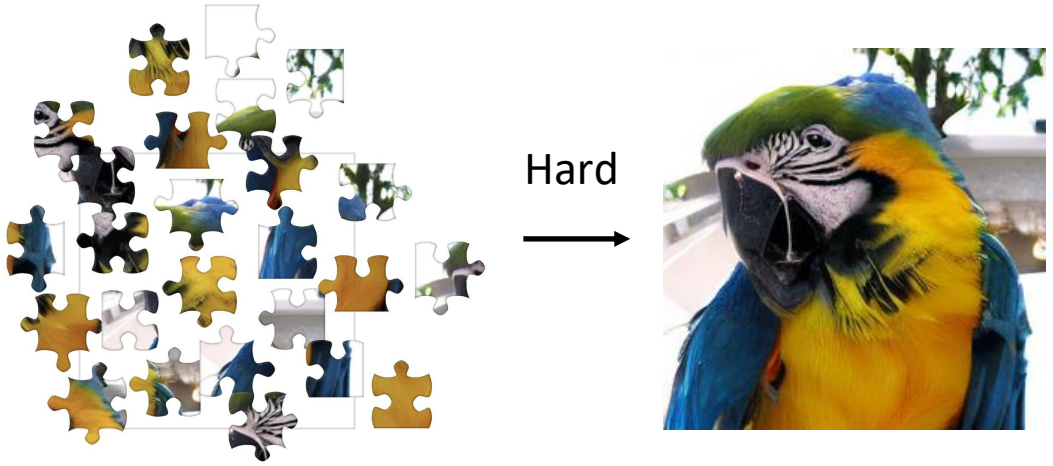


$$P(x_i \mid x_1, x_2, \ldots, x_{i-1})$$

Introducing global context by low-resolution token pivots

$$P(x_1, x_2, \ldots, x_N)$$

Next-Token Visual Generation Paradigm
(Analogy to Jigsaw Puzzles)

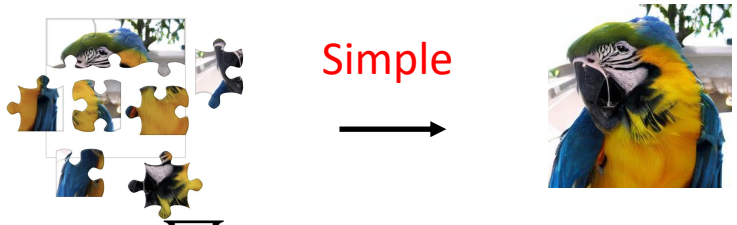


$$\prod_{N} P(x_i \mid x_1, x_2, \ldots, x_{i-1})$$

Incapable of utilizing global context

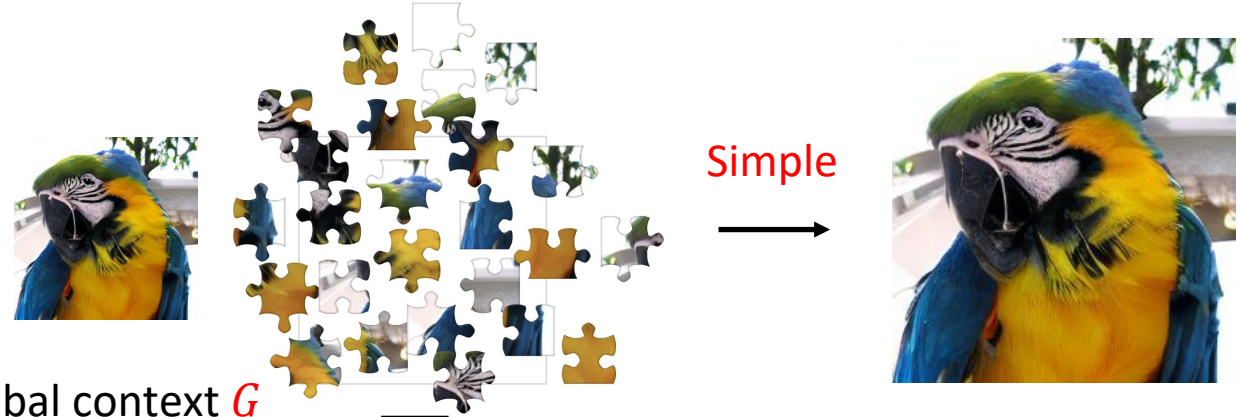
Hierarchical Masked Autoregressive Modeling (ours)

Phase 1:



$$\prod_{n} P(x_i \mid x_1, x_2, \ldots, x_{i-1})$$

Phase 2:



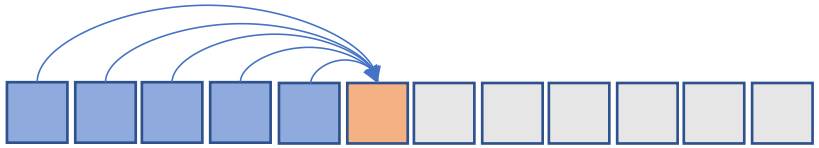
Global context G

$$\prod_{N} P(x_i \mid G, x_1, x_2, \ldots, x_{i-1})$$

Accuracy/speed trade-off in next-token visual generation

$$P(x_1, x_2, \ldots, x_N)$$

Next-Token Visual Generation Paradigm

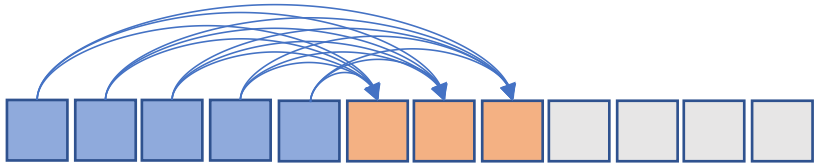


Speed ↓ Quality ↑

$$P(x_i \,|x_1, x_2, \ldots, x_{i-1})$$

Decrease inference steps

Accuracy/Speed Trade-off



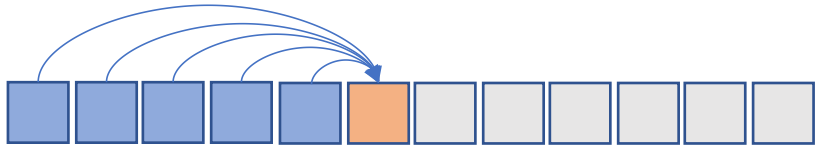
Speed ↑ Quality ↓

$$P(x_i, x_{i+1}, x_{i+2}|x_1, x_2, \ldots, x_{i-1})$$

Accuracy/speed trade-off in next-token visual generation

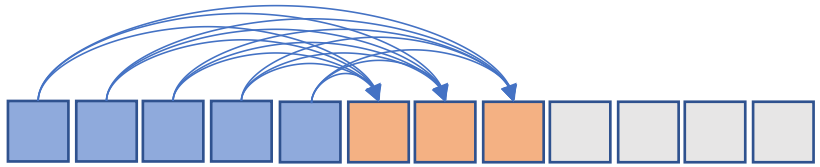
$$P(x_1, x_2, \ldots, x_N)$$

Next-Token Visual Generation Paradigm



$$P(x_i \,|x_1, x_2, \ldots, x_{i-1})$$

Decrease inference steps



$$P(x_i, x_{i+1}, x_{i+2}|x_1, x_2, \ldots, x_{i-1})$$

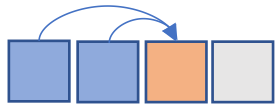
Speed ↓ Quality ↑

Accuracy/Speed Trade-off

Speed ↑ Quality ↓

Hierarchical Masked Autoregressive Modeling (ours)

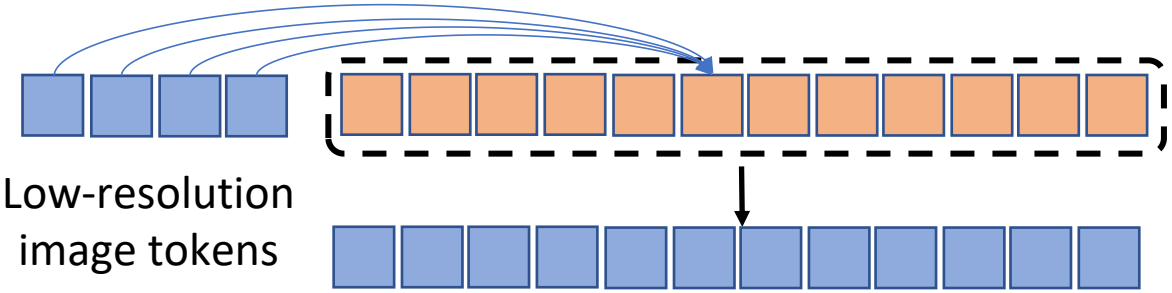
Phase 1:



Low-resolution image tokens

$$P(x_i \,|x_1, \ldots, x_{i-1})$$

Phase 2:



Low-resolution image tokens

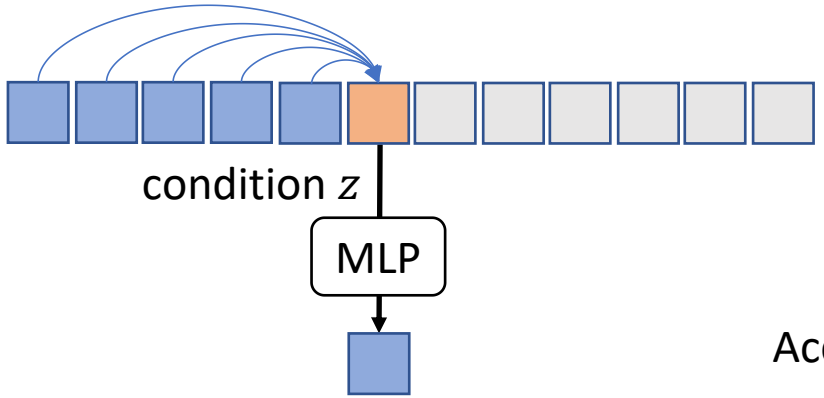
Only a few inference steps are needed, or even just one.

Speed ↑ Quality ↑

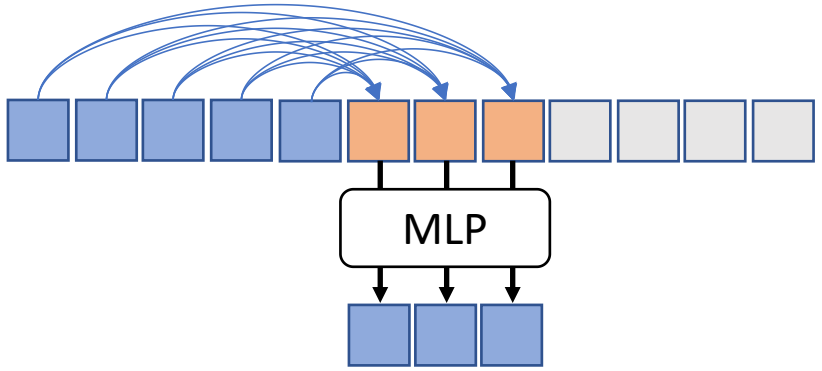
Lack of global context in MAR (continuous-valued token)

$$P(x_1, x_2, \ldots, x_N)$$

MAR



$$MLP(x_i \mid z_i)\, P(z_i \mid x_1, x_2, \ldots, x_{i-1})$$

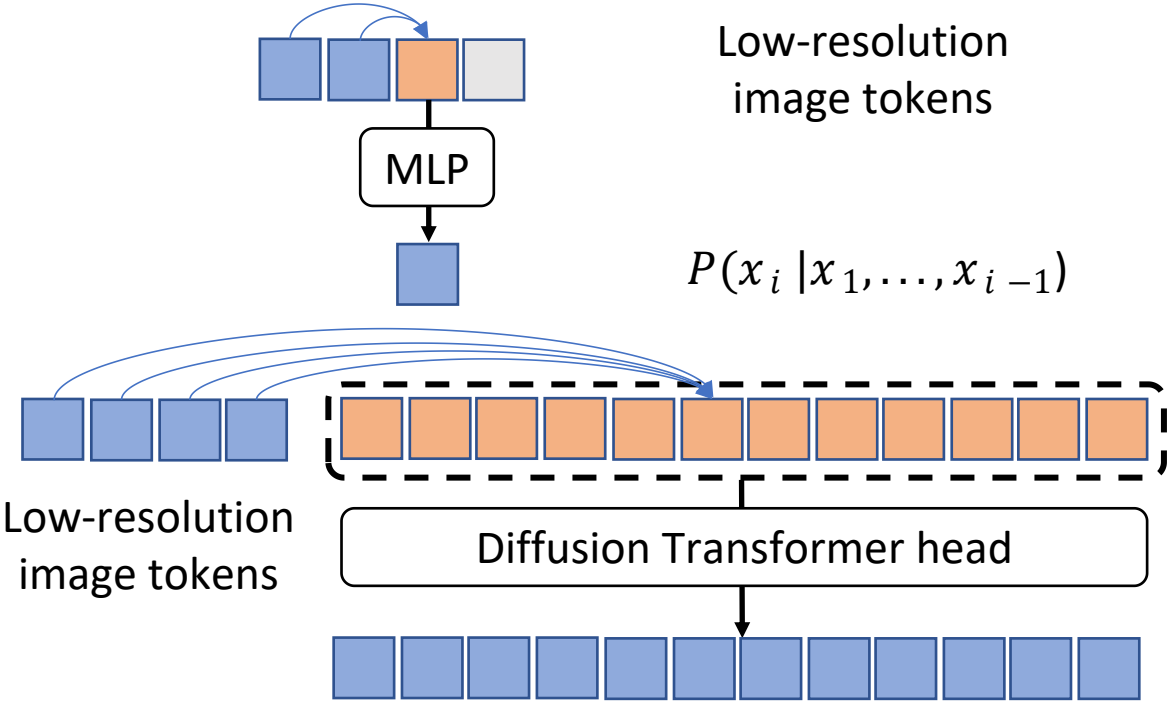


Speed ↓ Quality ↑

Accuracy/Speed Trade-off

Speed ↑ Quality ↓

Hi-MAR



$$Di\,T(x_1, \ldots, x_N \mid z_1, \ldots, z_N) P(z_1, \ldots, z_N \mid x_1, \ldots, x_n)$$

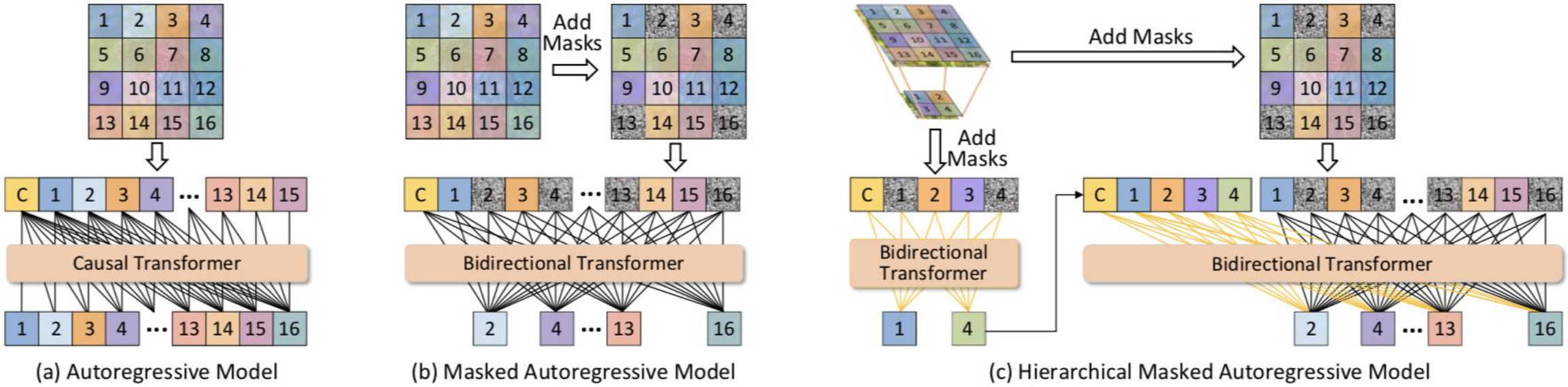
$MLP(x_i \mid z_i) MLP(x_{i+1} \mid z_{i+1}) MLP(x_{i+1} \mid z_{i+1})$ $P(z_i, z_{i+1}, z_{i+2} \mid x_1, x_2, \ldots, x_{i-1})$

Infer indepently, lack of global context

Modeling the global distribution $P(x_1, \ldots, x_N \mid z_1, \ldots, z_N)$

Overview of our method Hi-MAR

- **Hierarchical Masked Autoregressive Models (Hi-MAR):**
  - **Stage 1:** Mask autoregressive for $\prod_n P(x_i \mid x_1, x_2, \ldots, x_{i-1})$.
  - **Stage 2:** Mask autoregressive for $\prod_N P(x_i \mid G,\ x_1, x_2, \ldots, x_{i-1})$ with low-resolution toke pivots G.



(a) Autoregressive Model (b) Masked Autoregressive Model (c) Hierarchical Masked Autoregressive Model

Training-inference discrepancy in multi-scale AR

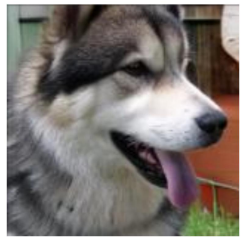
Previous multi-scale autoregressive model

Training:

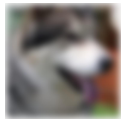
Ground-truth clean low-resolution tokens

Training-inference discrepancy → Generation quality ↓

Inference:

② 

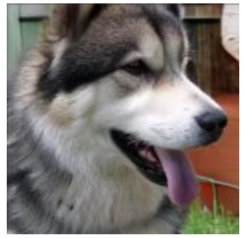
Predicted noisy low-resolution tokens

① 

Training-inference discrepancy in multi-scale AR

Previous multi-scale autoregressive model

Training:

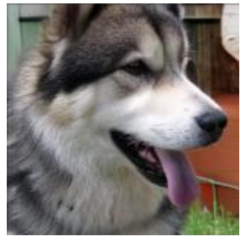
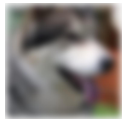


P(|  , Cond)

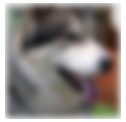
Ground-truth clean low-resolution tokens

Training-inference discrepancy

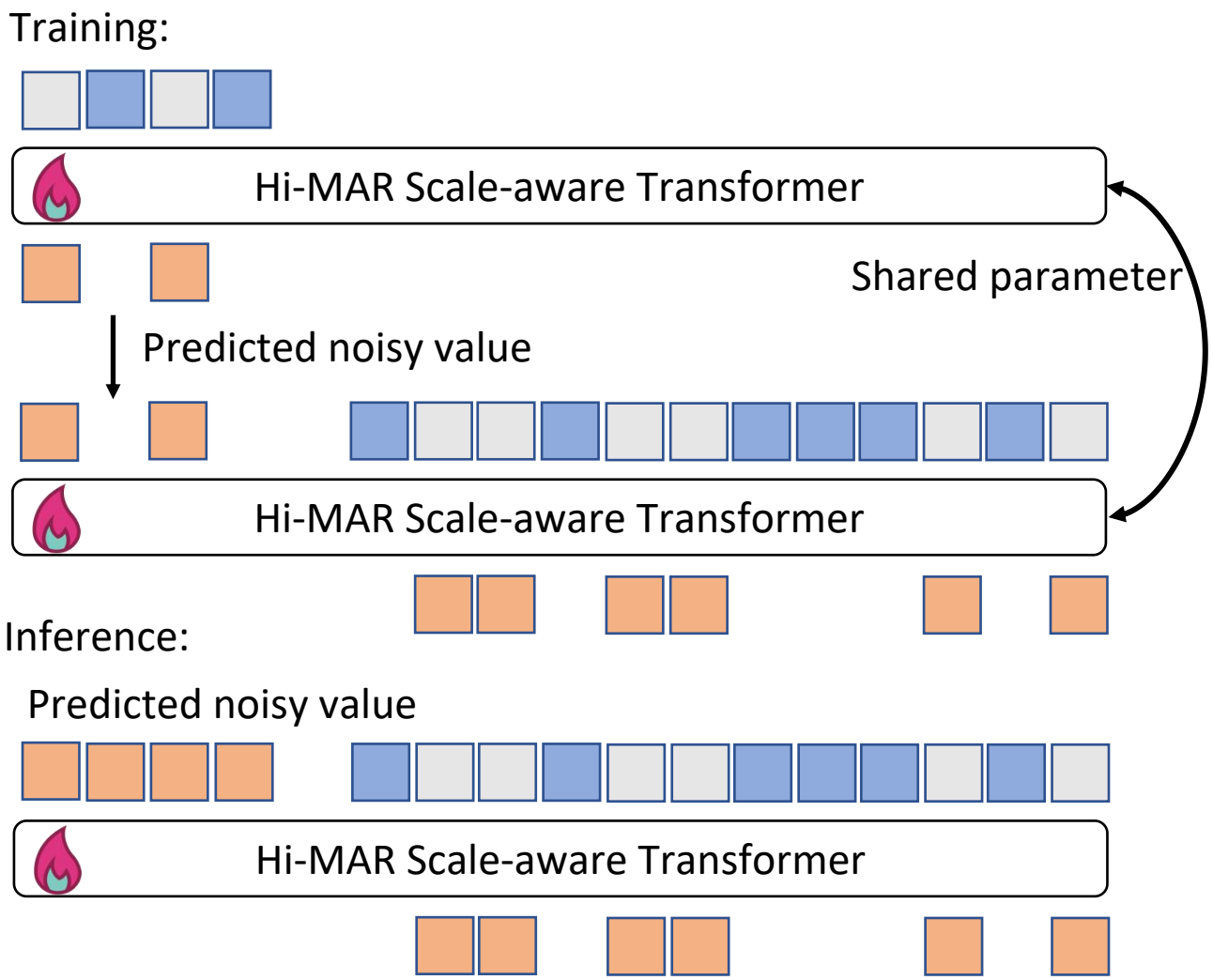
Inference:

② P( |  , Cond)

Predicted noisy low-resolution tokens

① P( | Cond)

Previous multi-scale autoregressive model

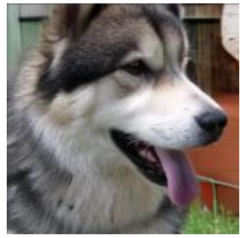


Eliminating the train-inference distribution gap significantly improves performance

Training-inference discrepancy in multi-scale AR

Previous multi-scale autoregressive model

Training:



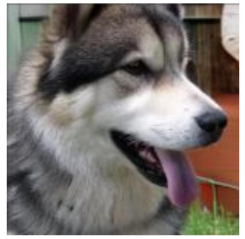
P(| , Cond)

Ground-truth clean low-resolution tokens

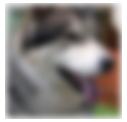
Training-inference discrepancy

Inference:

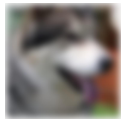
② P(| , Cond)



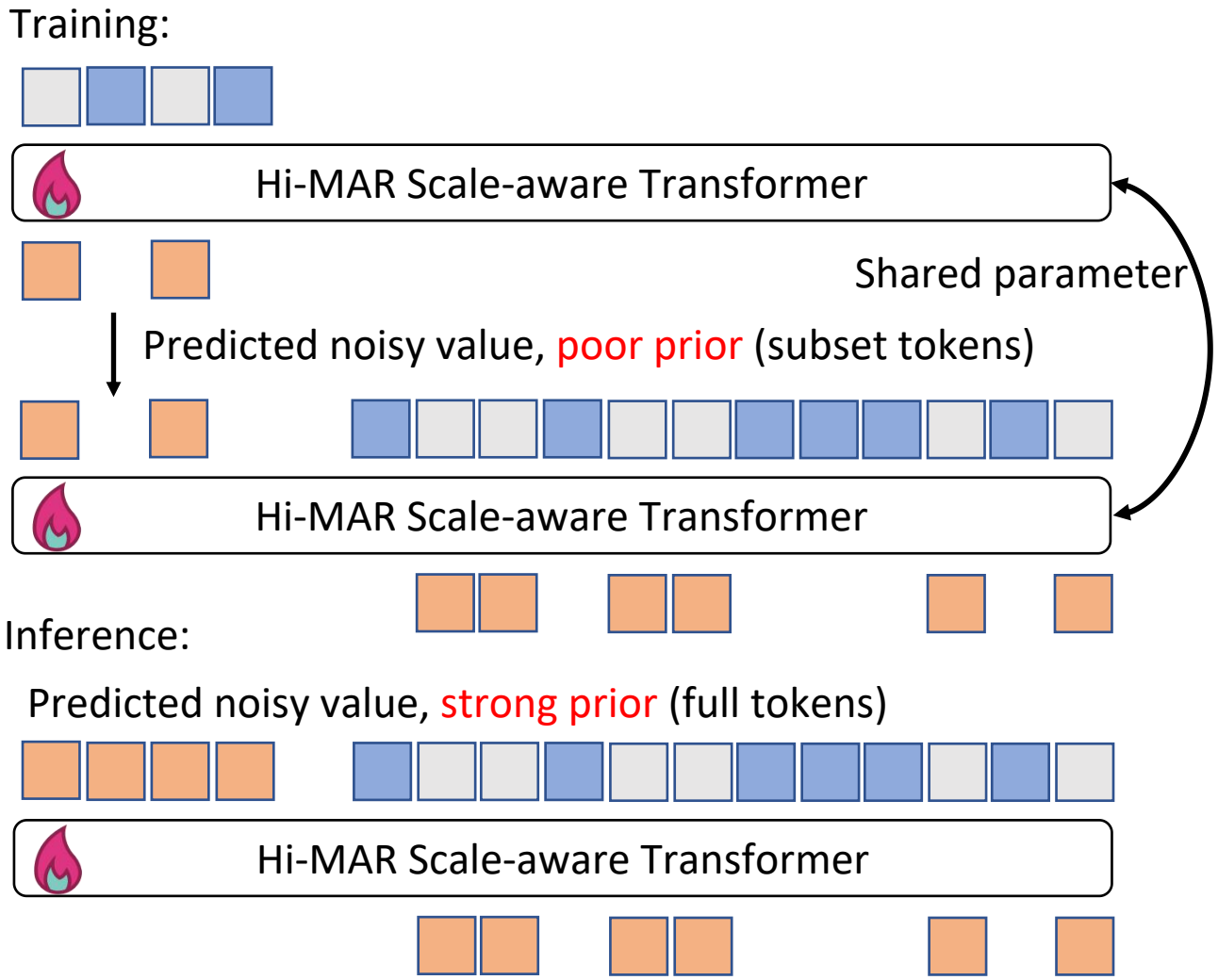
Predicted noisy low-resolution tokens



① P(| Cond)

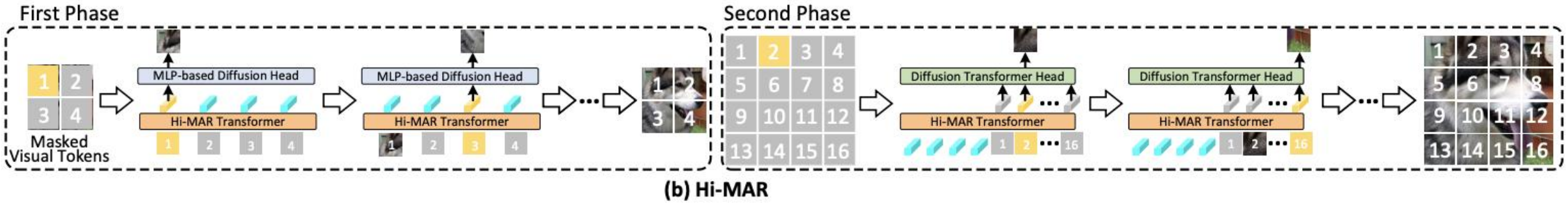


Previous multi-scale autoregressive model



Eliminating the train-inference distribution gap significantly improves performance

Scale-aware transformer backbone



System level comparison

Table 2. **Generative model family comparison on class-conditional ImageNet 256×256.** "↓" or "↑" indicate lower or higher values a better. Metrics include Fréchet inception distance (FID), inception score (IS), precision and recall. Models with the suffix "-re" use rejection sampling.

Type	Model	#Para.	w/o CFG FID↓	IS↑	Precision↑	Recall↑	w/ CFG FID↓	IS↑	Precision↑	Recall↑
GAN	BigGAN (Brock, 2018)	112M	6.95	224.5	**0.89**	0.38	–	–	–	–
	GigaGAN (Kang et al., 2023)	569M	3.45	225.5	0.84	0.61	–	–	–	–
	StyleGan-XL (Sauer et al., 2022)	166M	2.30	265.1	0.78	0.53	–	–	–	–
Diff.	ADM (Dhariwal & Nichol, 2021)	554M	10.94	101.0	0.69	0.63	4.59	186.7	0.82	0.52
	CDM (Ho et al., 2022)	–	–	–	–	–	4.88	158.7	–	–
	LDM-4-G (Rombach et al., 2022)	400M	10.56	103.5	0.71	0.62	3.60	247.7	**0.87**	0.48
	U-ViT-H/2 (Bao et al., 2023)	501M	–	–	–	–	2.29	263.88	0.82	0.57
	DiT-XL/2 (Peebles & Xie, 2023)	675M	9.62	121.5	0.67	0.67	2.27	278.2	0.83	0.57
AR	VQGAN (Esser et al., 2021)	227M	18.65	80.4	0.78	0.26	–	–	–	–
	VQGAN-re (Esser et al., 2021)	1.4B	5.20	280.3	–	–	–	–	–	–
	RQTran. (Lee et al., 2022)	3.8B	–	–	–	–	7.55	134.0	–	–
	RQTran.-re (Lee et al., 2022)	3.8B	–	–	–	–	3.80	**323.7**	–	–
	GIVT (Tschannen et al., 2025)	304M	5.67	–	0.75	0.59	3.35	–	0.84	0.53
	LlamaGen-L (Sun et al., 2024)	343M	19.07	64.3	0.61	0.67	3.07	256.06	0.83	0.52
	LlamaGen-XL (Sun et al., 2024)	775M	15.54	79.2	0.62	**0.69**	2.62	244.08	0.80	0.57
	LlamaGen-XXL (Sun et al., 2024)	1.4B	14.65	86.3	0.63	0.68	2.34	253.90	0.80	0.59
	VAR-*d*16 (Tian et al., 2024)	310M	–	–	–	–	3.30	274.4	0.84	0.51
	VAR-*d*20 (Tian et al., 2024)	600M	–	–	–	–	2.57	302.6	0.83	0.56
	VAR-*d*24 (Tian et al., 2024)	1.0B	–	–	–	–	2.09	312.9	0.82	0.59
Mask.	MaskGIT (Chang et al., 2022)	227M	6.18	182.1	0.80	0.51	–	–	–	–
	AutoNAT-L (Ni et al., 2024)	422M	–	–	–	–	2.68	278.8	–	–
	MAR-B (Li et al., 2024)	208M	3.48	192.4	0.78	0.58	2.31	281.7	0.82	0.57
	MAR-L (Li et al., 2024)	479M	2.60	221.4	0.79	0.60	1.78	296.0	0.81	0.60
	MAR-H (Li et al., 2024)	943M	2.35	227.8	0.79	0.62	1.55	303.7	0.81	0.62
Hi-MAR	Hi-MAR-B	244M	2.11	251.46	0.80	0.59	1.93	293.0	0.81	0.59
	Hi-MAR-L	529M	1.72	278.63	0.79	0.62	1.66	322.3	0.79	0.61
	Hi-MAR-H	1090M	**1.55**	**300.72**	0.80	0.63	**1.52**	322.78	0.80	**0.63**

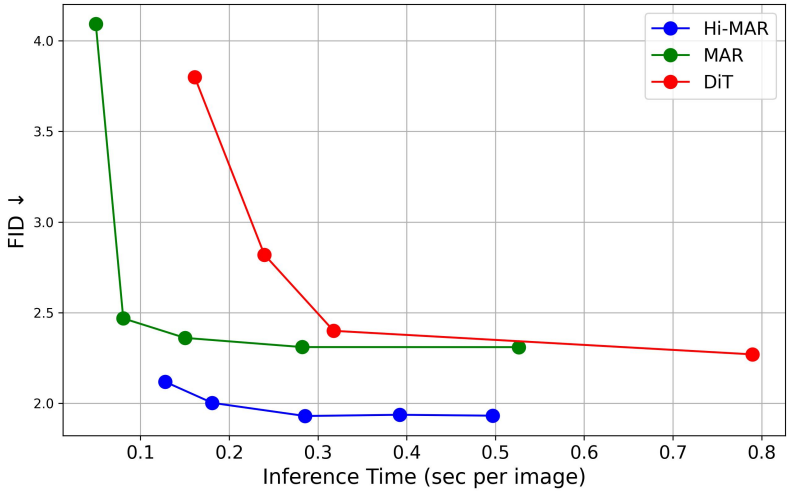


Table 3. **FID results of different models on MS-COCO 256×256 validation.** "↓" indicates lower values are better.

Type	Model	FID ↓
GAN	AttnGAN (Xu et al., 2018)	35.49
	DM-GAN (Zhu et al., 2019)	32.64
	DF-GAN (Tao et al., 2022)	19.32
	XMC-GAN (Zhang et al., 2021)	9.33
	LAFITE (Zhou et al., 2022)	8.12
Diffusion	VQ-Diffusion (Gu et al., 2022)	19.75
	Friro (Fan et al., 2023)	8.97
	U-ViT-S/2 (Deep)	5.48
Mask.	AutoNAT-S (Ni et al., 2024)	5.36
	MAR (Li et al., 2024)	6.36
Hi-MAR	Hi-MAR-S	**4.77**