ControlAR: Controllable Image Generation with Autoregressive Models

Zongming Li[1*], Tianheng Cheng[1*], Shoufa Chen[2], Peize Sun[2], Haocheng Shen[3], Longjing Ran[3], Xiaoxin Chen[3], Wenyu Liu[1] & Xinggang Wang[1†]

[1]School of EIC, Huazhong University of Science and Technology
[2]Department of Computer Science, The University of Hong Kong
[3]vivo AI Lab

Autoregressive Image Generation

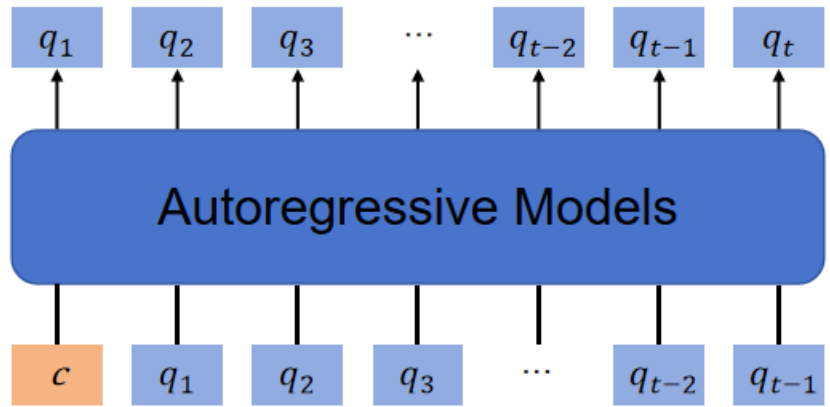
Next-token prediction:

$$p(x) = \prod_{i=1}^{n} p(x_i|x_1, x_2, \ldots, x_{i-1}) = \prod_{i=1}^{n} p(x_i|x_{<i})$$

For image generation:

- First divide image patches by fixed size
- Then convert compressed image patches into discrete image tokens q_i
- Predict the next image token based on the condition c and existing image tokens

$$p(q) = \prod_{i=1}^{n} p(q_t|q_{<t}, c)$$



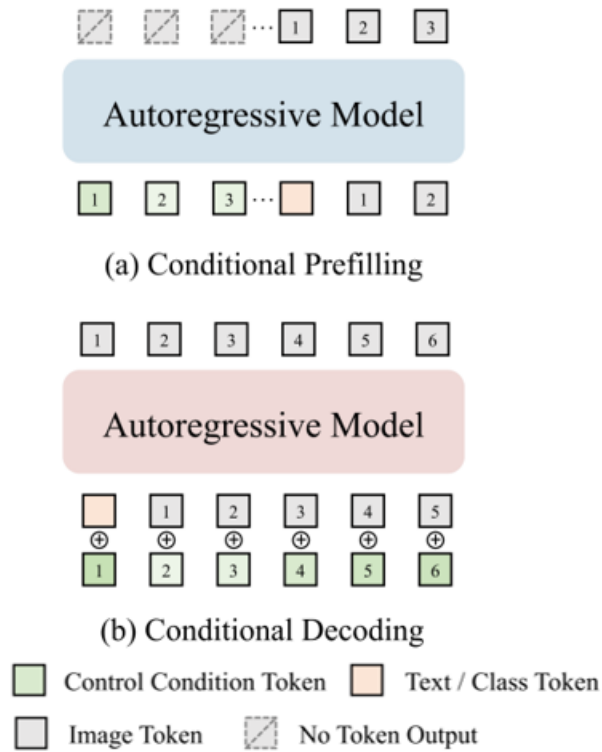
Add Condition to AR Model

Conditional Prefilling

- Concatenate conditional token before class/text token
- Sequence length: $L_{condition} + L_{text} + L_{image}$

Conditional Decoding

- Add the conditional token directly to the image token
- Sequence length: $L_{text} + L_{image}$



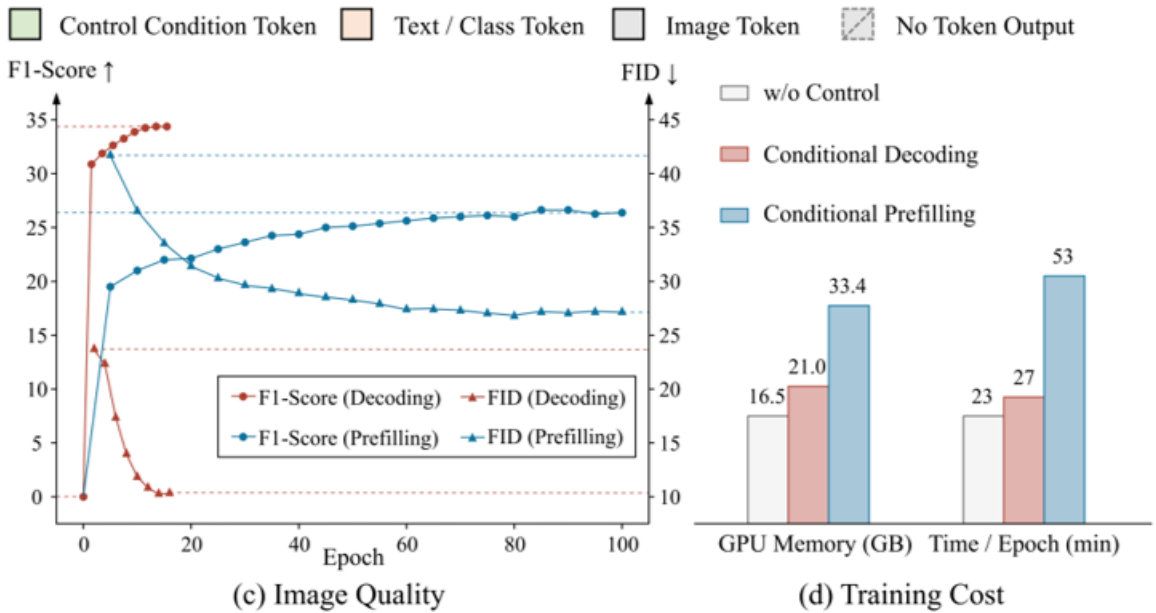
Performance Comparision

Conditional Prefilling

- Slow and poor convergence
- High training computing costs

Conditional Decoding

- Fast and strong convergence
- Low training computing costs



(c) Image Quality (d) Training Cost

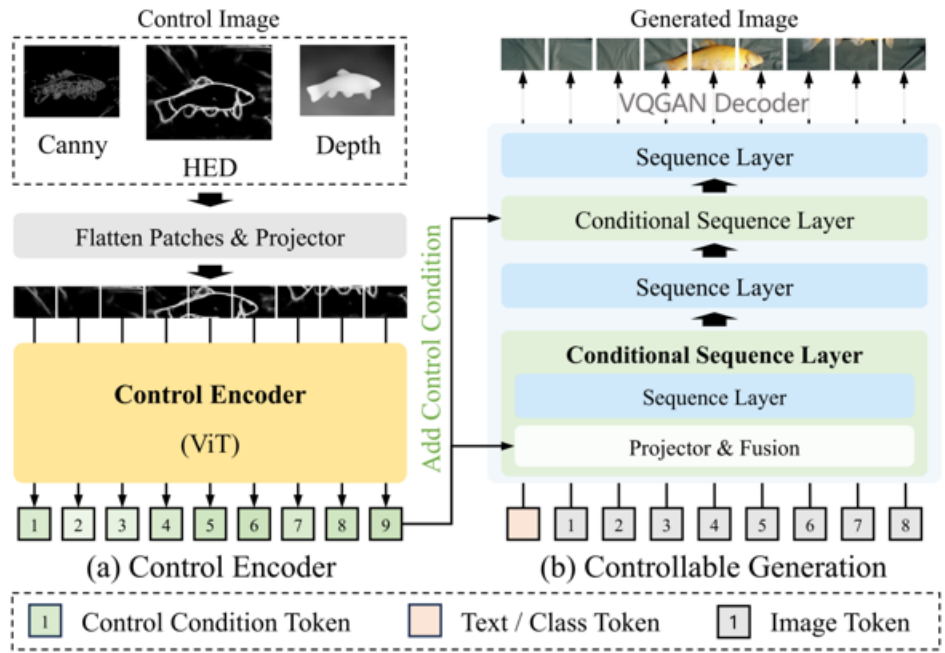
ControlAR

Control Encoder

- Encode 2D conditional image into 1D conditional sequence
- Lightweight ViT is enough

Conditional Sequence Layer

- Fusing control conditional token with image token
- $S_{out} = \mathcal{F}(S_{in} + \mathcal{P}(C)) = \mathcal{F}([c + C_1, I_1 + C_2, I_2 + C_3, ..., I_{i-1} + C_i])$

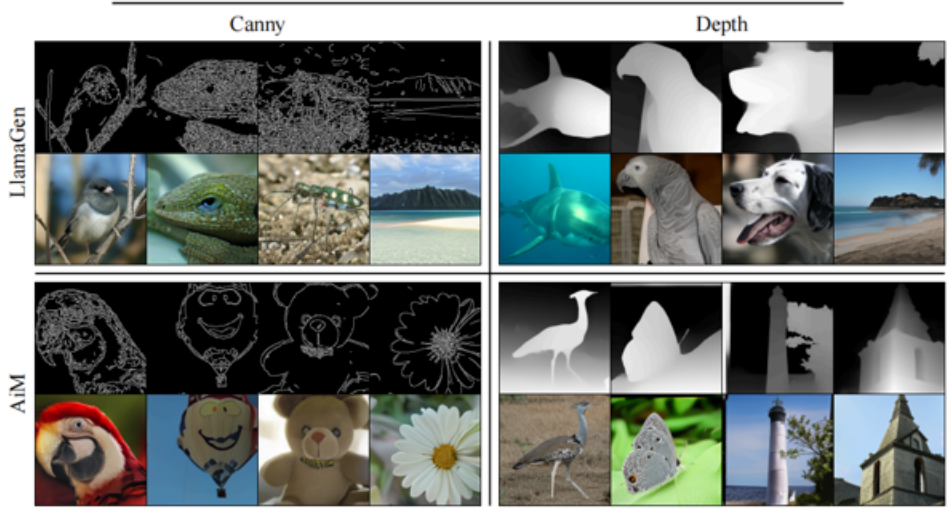


Class-to-Image Performance

- Conduct experiments in Transformer-based LlamaGen[1] and Mamba-based AiM[2]
- ControlAR achieves lower FID based on the LlamaGen-L, which only has 16.7% of parameters of VAR-d30
- Good visualization in ImageNet-1K validation data

Table 1: **C2I controllable generation.** Param. denotes the number of parameters of the C2I model. "↑" or "↓" indicate lower or higher values are better. "*" indicates that ControlVAR's FID values are estimated from its histograms (Li et al., 2024c). The results are conducted on 256 × 256 resolution.

Method	C2I Model	Param.	Canny		Depth	
			F1-Score ↑	FID ↓	RMSE ↓	FID ↓
ControlVAR*	VAR-d16	310M	-	16.20	-	13.80
	VAR-d20	600M	-	13.00	-	13.40
	VAR-d24	1.0B	-	15.70	-	12.50
	VAR-d30	2.0B	-	7.85	-	6.50
Ours	AiM-L	350M	30.36	9.66	35.01	7.39
	LlamaGen-B	111M	34.15	10.64	32.41	6.67
	LlamaGen-L	343M	34.91	7.69	31.11	4.19



[1] Sun P, Jiang Y, Chen S, et al. Autoregressive model beats diffusion: Llama for scalable image generation[J]. arXiv preprint arXiv:2406.06525, 2024.

[2] Li H, Yang J, Wang K, et al. Scalable autoregressive image generation with mamba[J]. arXiv preprint arXiv:2408.12245, 2024.

Text-to-Image Performance

ICLR

Compared with Diffusion Method

- Highly competitive condition consistency with diffusion method
- Better FID results

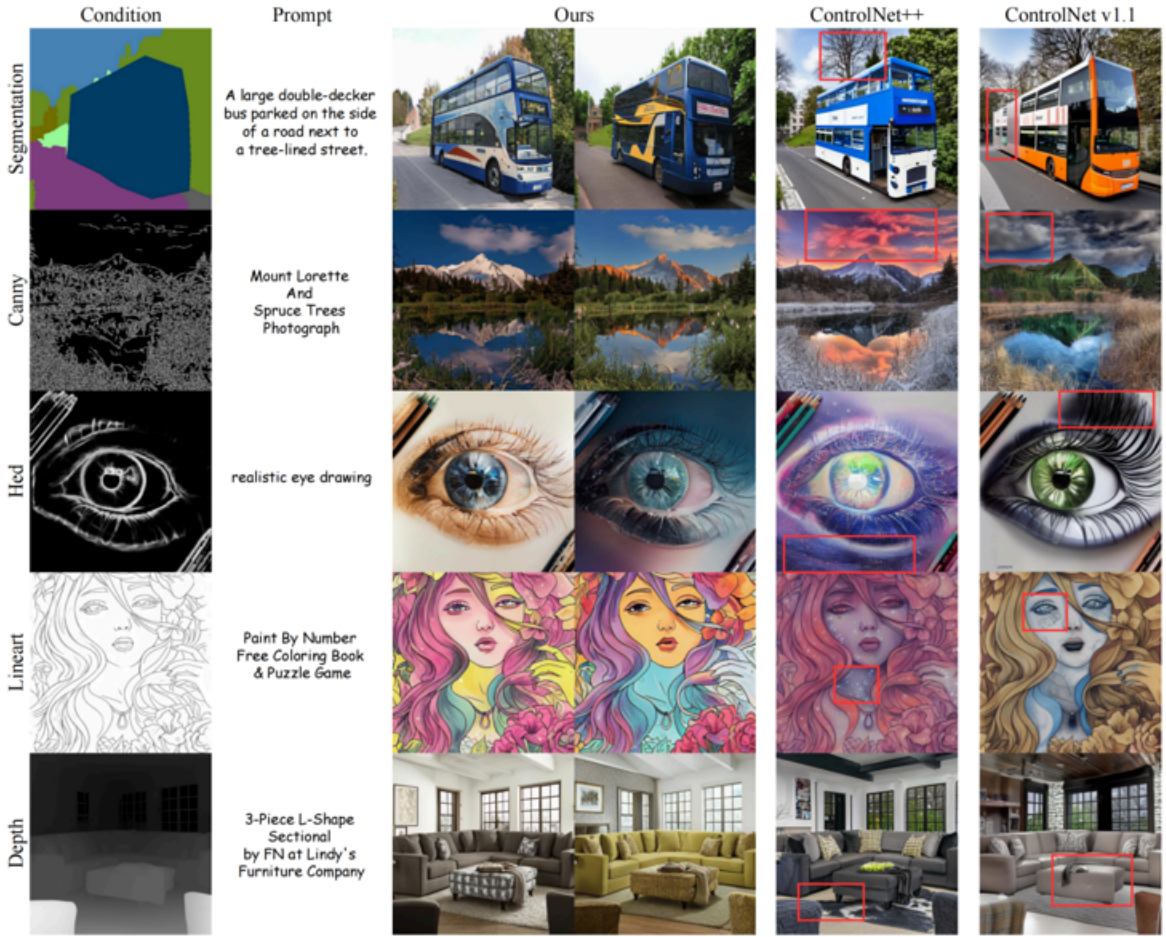
Table 2: **Conditional consistency of T2I controllable generation.** "↑" or "↓" indicate lower or higher values are better. "-" denotes that the method does not release a model for testing. The results are conducted on 512×512 resolution.

Method	Seg.		Canny	Hed	Lineart	Depth
	mIoU ↑	mIoU ↑	F1-Score ↑	SSIM ↑	SSIM ↑	RMSE ↓
	ADE20K	COCOStuff	MultiGen-20M	MultiGen-20M	MultiGen-20M	MultiGen-20M
GLIGEN	23.78	-	26.94	-	-	38.83
T2I-Adapter	12.61	-	23.65	-	-	48.40
Uni-ControlNet	19.39	-	27.32	69.10	-	40.65
UniControl	25.44	-	30.82	79.69	-	39.18
ControlNet	32.55	27.46	34.65	76.21	70.54	35.90
ControlNet++	**43.64**	<u>34.56</u>	<u>37.04</u>	<u>80.97</u>	**83.99**	**28.32**
Ours	<u>39.95</u>	**37.49**	**37.08**	**85.63**	<u>79.22</u>	<u>29.01</u>

Table 3: **FID of T2I controllable generation.** "-" denotes that the method does not release a model for testing. Our ControlAR achieves significant FID improvements.

Method	Seg.		Canny	Hed	Lineart	Depth
	ADE20K	COCOStuff	MultiGen-20M	MultiGen-20M	MultiGen-20M	MultiGen-20M
GLIGEN	33.02	-	18.89	-	-	18.36
T2I-Adapter	39.15	-	<u>15.96</u>	-	-	22.52
Uni-ControlNet	39.70	-	17.14	17.08	-	20.27
UniControl	46.34	-	19.94	15.99	-	18.66
ControlNet	33.28	21.33	**14.73**	15.41	17.44	17.76
ControlNet++	<u>29.49</u>	<u>19.29</u>	18.23	<u>15.01</u>	<u>13.88</u>	<u>16.66</u>
Ours	**27.15**	**14.51**	17.51	**10.53**	**12.41**	**14.61**

Text-to-Image Performance



Arbitrary-Resolution Generation

The resolution of the generated image is consistent with the input control image and is not limited to a fixed resolution

Arbitrary-Resolution Generation without Condition Image

- By adjusting the training strategy, we can freely control the resolution of the T2I generation

- Generate a grayscale map, and use conditional decoding strategies to tell the model where to make line breaks and where to end

A cowboy riding a horse across a river

A blue Porsche 356 parked in front of a yellow brick wall

A two-story modern villa by the sea

Ablations

ICLR

- ViT performs better in C2I task, DINOv2 performs better in T2I task
- Three conditional sequence layer work best
- Full fine-tuning gives better results

Table 4: **Ablations on the Control Encoder.** "↑" or "↓" indicate lower or higher values are better.

Control Encoder	Params	Canny (C2I)		Depth (C2I)		Hed (T2I)	
		F1-Score ↑	FID ↓	RMSE ↓	FID ↓	SSIM ↑	FID ↓
CNN (4× Res. Blocks)	21.8M	33.55	12.27	33.36	6.97	81.64	15.33
ViT-S	22.1M	**34.15**	10.64	32.41	6.64	82.37	14.59
DINOv2-S	22.1M	33.38	10.87	32.82	7.31	85.63	10.53
DINOv2-B	86.6M	34.07	**9.47**	**31.81**	**6.36**	**86.12**	**8.58**

Table 5: **Ablations on the control fusion strategy.** "↑" or "↓" indicate lower or higher values are better.

Fusion Strategy	#Layer	F1-Score ↑	FID ↓
Cross-Attention	1-th	30.86	15.34
Addition	1-th	34.01	11.02
Addition	1,5,9-th	34.15	10.64
Addition	1∼12-th	34.21	11.75

Table 6: **Ablations on the training strategy.** "↑" or "↓" indicate lower or higher values are better.

Training Strategy	F1-Score ↑	FID ↓
Freeze	30.62	13.67
LoRA	32.90	13.20
Full fine-tune	34.15	10.64

Conclusion

- We explore controllable autoregressive image generation and present ControlAR, which enables precise control and generates high-quality images.
- We exploit the properties of our proposed conditional decoding to extend the ability of the autoregressive model to generate arbitrary resolution.
- Our ControlAR demonstrates its highly competitive performance towards conditional consistency and image quality compared to state-of-the-art diffusion methods.

Thanks!

Paper: https://arxiv.org/abs/2410.02705

Project Page: https://github.com/hustvl/ControlAR