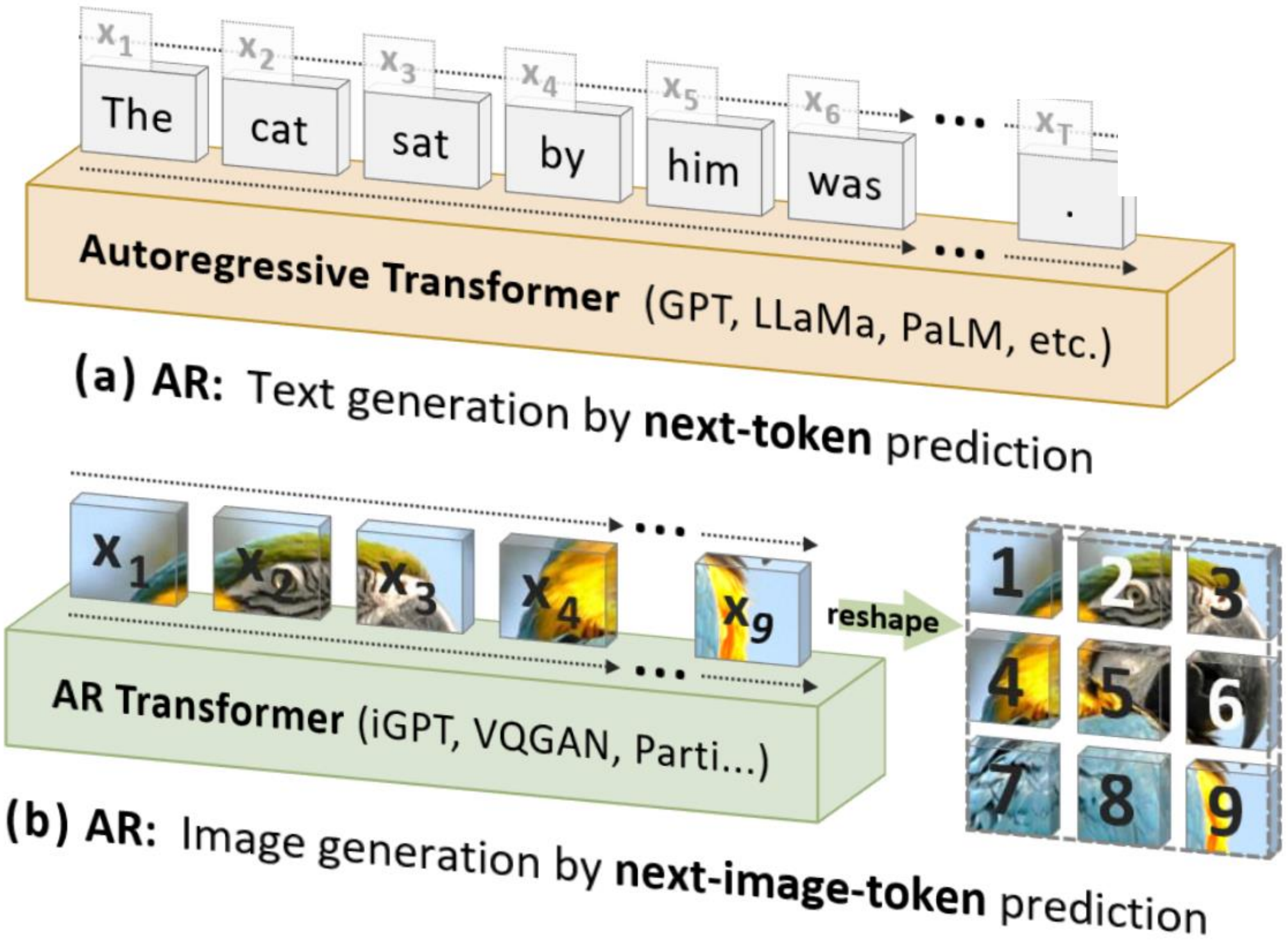
Toward Guidance-Free AR Visual Generation via Condition Contrastive Alignment

Huayu Chen, Hang Su, Peize Sun, Jun Zhu

AR Visual Generation



AR guided sampling (CFG)

Vanilla Decoding

$$\ell^c$$

Classifier-Free Guidance

$$\ell^{\text{sample}} = \ell^c + s(\ell^c - \ell^u)$$

AR guided sampling (CFG)

Poor Image-Condition Alignment

Vanilla Decoding

$$\ell^c$$

LlamaGen (w/o Guidance)

IS=64.7

Classifier-Free Guidance

$$\ell^{\text{sample}} = \ell^c + s(\ell^c - \ell^u)$$

AR guided sampling (CFG)

Poor Image-Condition Alignment

Vanilla Decoding

$$\ell^c$$

LlamaGen (w/o Guidance)
IS=64.7

Classifier-Free Guidance

$$\ell^{\text{sample}} = \ell^c + s(\ell^c - \ell^u)$$

LlamaGen (w/ CFG)
IS=404.0

Problems

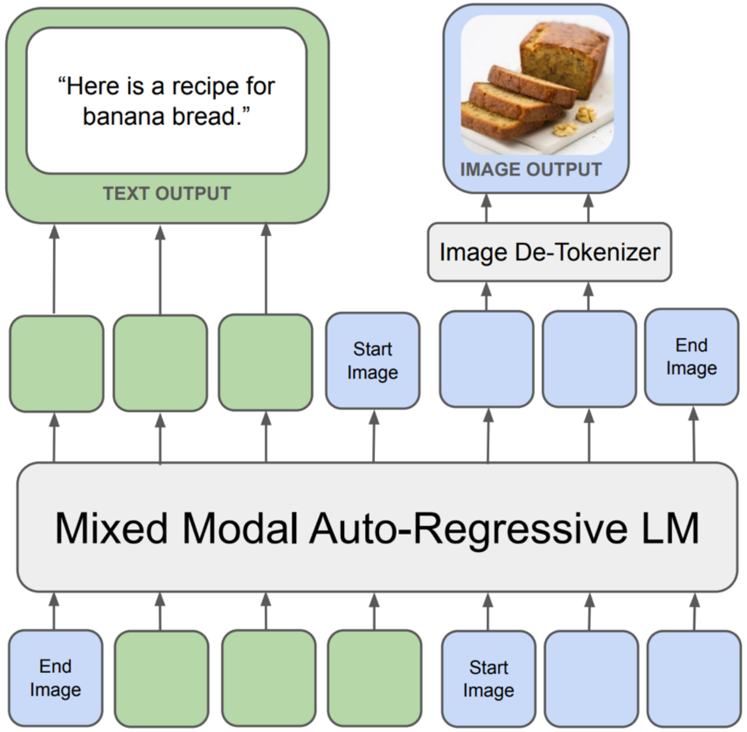
Is Classifier Free Guidance good enough for us?

Problems

- What's the ultimate goal for studying Visual AR?

Unified Mixed-Modal Modeling:

- Unified Representation.
- Unified Algorithm.
- Unified Decoding.



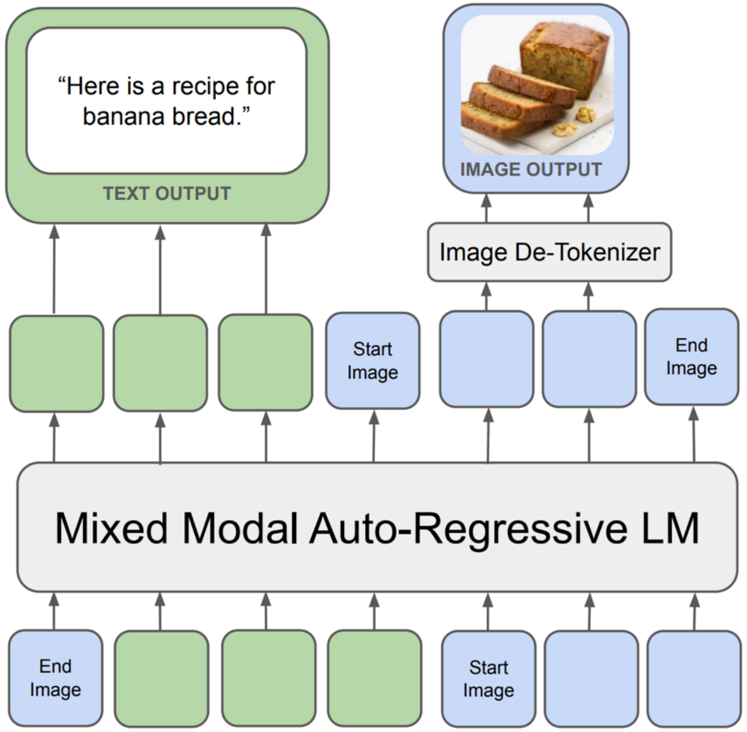
Mixed-modal training

Problems

- What's the ultimate goal for studying Visual AR?

  Unified Mixed-Modal Modeling:

  - Unified Representation.
  - Unified Algorithm.
  - Unified Decoding.



Mixed-modal training

CFG causes inconsistencies between language & vision

- Training → Randomly masking text conditions in loss
- Sampling → Inconsistent Decoding System + 2x inference times

Condition Contrastive Alignment

$$p(\boldsymbol{x}|\boldsymbol{c})$$

$$p^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c}) \propto p(\boldsymbol{x}|\boldsymbol{c}) \left[\frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})}\right]^s$$

$$\ell^{\text{sample}} = \ell^c + s(\ell^c - \ell^u)$$

LlamaGen (w/o Guidance)
IS=64.7

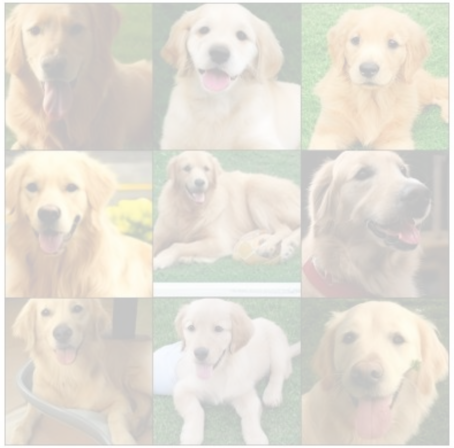
LlamaGen (w/ CFG)
IS=404.0

Condition Contrastive Alignment

$$p(\boldsymbol{x}|\boldsymbol{c}) \longrightarrow p^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c}) \propto p(\boldsymbol{x}|\boldsymbol{c}) \left[\frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})}\right]^s$$

LlamaGen (w/o Guidance)
IS=64.7

$\ell^{\text{sample}} = \ell^c + s(\ell^c - \ell^u)$



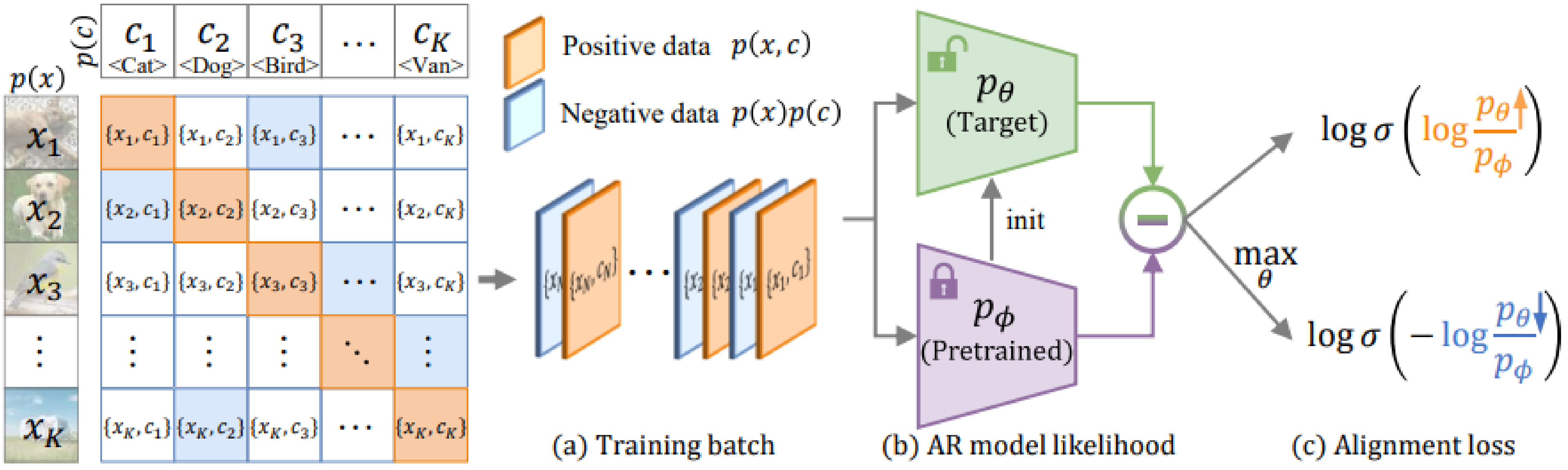
LlamaGen (w/ CFG)
IS=404.0

~~Unconditional Training~~

Condition Contrastive Alignment

LlamaGen + CCA (w/o G.)
IS=384.6

Implementation



(a) Training batch (b) AR model likelihood (c) Alignment loss

$$\mathcal{L}_\theta^{\text{CCA}}(\boldsymbol{x}_k, \boldsymbol{c}_k, \boldsymbol{c}_k^{\text{neg}}) = -\underbrace{\log \sigma \left[\beta \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}_k|\boldsymbol{c}_k)}{p_\phi(\boldsymbol{x}_k|\boldsymbol{c}_k)}\right]}_{\text{relative likelihood for positive conditions} \uparrow} \underbrace{-\lambda \log \sigma \left[-\beta \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}_k|\boldsymbol{c}_k^{\text{neg}})}{p_\phi(\boldsymbol{x}_k|\boldsymbol{c}_k^{\text{neg}})}\right]}_{\text{relative likelihood for negative conditions} \downarrow}$$

Primary Result

LlamaGen (w/o Guidance)
IS=64.7

LlamaGen + CCA (w/o G.)
IS=384.6

LlamaGen (w/ CFG)
IS=404.0

VAR (w/o Guidance)
IS=154.3

VAR + CCA (w/o G.)
IS=350.4

VAR (w/ CFGv2)
IS=390.8

Method

$$p(\boldsymbol{x}|\boldsymbol{c}) \longrightarrow \boxed{p^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c})} \propto p(\boldsymbol{x}|\boldsymbol{c}) \left[\frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})}\right]^{s}$$

Cannot be learned due to “Lack of data”

Method

$$p^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c}) \propto p(\boldsymbol{x}|\boldsymbol{c}) \left[\frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})}\right]^s$$

Transform into Learnable forms

What we want

$$\frac{1}{s} \log \frac{p^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x}|\boldsymbol{c})} = \boxed{\log \frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})}}$$

What we have

Conditional Residual,
is Learnable

Theorem 3.1 (Noise Contrastive Estimation, proof in Appendix A). *Let r_θ be a parameterized model which takes in an image-condition pair $(\boldsymbol{x}, \boldsymbol{c})$ and outputs a scalar value $r_\theta(\boldsymbol{x}, \boldsymbol{c})$. Consider the loss function:*

$$\mathcal{L}_\theta^{NCE}(\boldsymbol{x}, \boldsymbol{c}) = -\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c})} \log \sigma(r_\theta(\boldsymbol{x}, \boldsymbol{c})) - \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{c})} \log \sigma(-r_\theta(\boldsymbol{x}, \boldsymbol{c})). \tag{8}$$

Given unlimited model expressivity for r_θ, the optimal solution for minimizing $\mathcal{L}_\theta^{NCE}$ satisfies

$$r_\theta^*(\boldsymbol{x}, \boldsymbol{c}) = \log \frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})}. \tag{9}$$

Method

- New Parameterization

$$\log \frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})} \xleftarrow{\text{learn}} r_\theta(\boldsymbol{x},\boldsymbol{c}) := \frac{1}{s}\log\frac{p_\theta^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x}|\boldsymbol{c})}$$

- Alignment Loss

$$\mathcal{L}_\theta^{\text{CCA}} = -\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c})}\log\sigma\Big[\frac{1}{s}\log\frac{p_\theta^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x}|\boldsymbol{c})}\Big] - \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{c})}\log\sigma\Big[-\frac{1}{s}\log\frac{p_\theta^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x}|\boldsymbol{c})}\Big]$$

Method

- New Parameterization

$$\log \frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})} \xleftarrow{\text{learn}} r_\theta(\boldsymbol{x}, \boldsymbol{c}) := \frac{1}{s} \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x}|\boldsymbol{c})}$$

- Alignment Loss

$$\mathcal{L}_\theta^{\text{CCA}} = -\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c})} \log \sigma \Big[\frac{1}{s} \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x}|\boldsymbol{c})} \Big] - \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{c})} \log \sigma \Big[-\frac{1}{s} \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x}|\boldsymbol{c})} \Big]$$

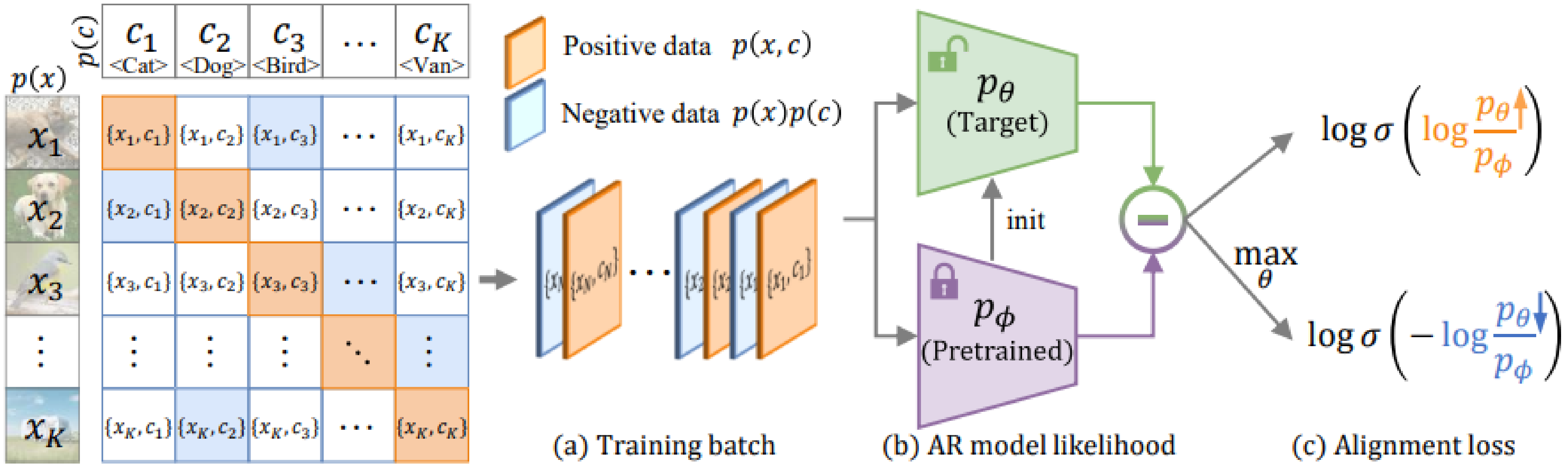
Image & Condition

(Pretraining data)

Image & Random Condition

(Shuffled Data)

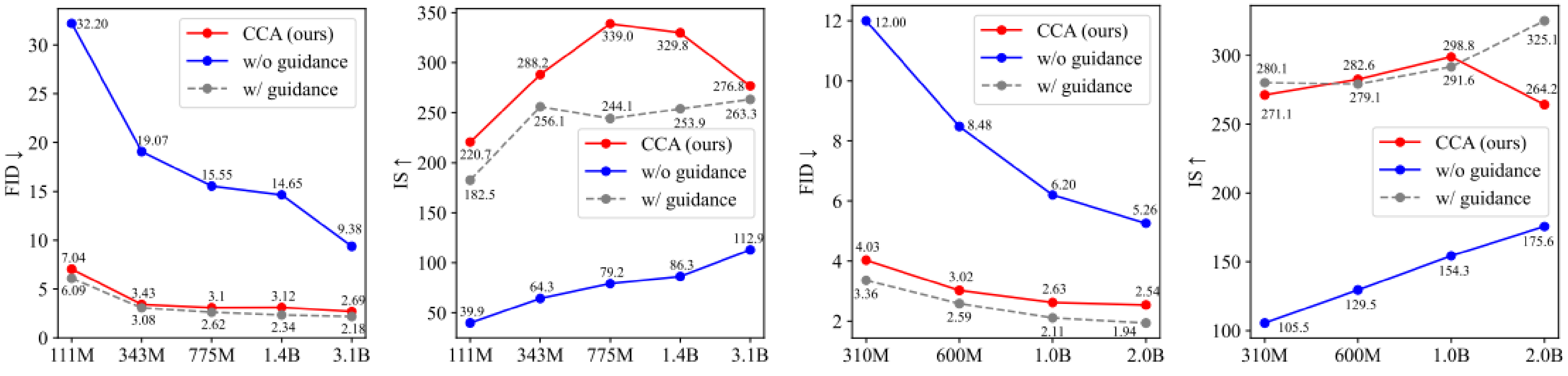
Condition Contrastive Alignment



$$\mathcal{L}_\theta^{\text{CCA}}(\boldsymbol{x}_k, \boldsymbol{c}_k, \boldsymbol{c}_k^{\text{neg}}) = -\underbrace{\log \sigma \left[\beta \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}_k | \boldsymbol{c}_k)}{p_\phi(\boldsymbol{x}_k | \boldsymbol{c}_k)} \right]}_{\text{relative likelihood for positive conditions} \uparrow} \underbrace{-\lambda \log \sigma \left[-\beta \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}_k | \boldsymbol{c}_k^{\text{neg}})}{p_\phi(\boldsymbol{x}_k | \boldsymbol{c}_k^{\text{neg}})} \right]}_{\text{relative likelihood for negative conditions} \downarrow}$$

Result

- How good is CCA?



(a) LlamaGen (b) VAR

Result

- How good is CCA?

	Model	w/o Guidance				w/ Guidance	
		FID↓	IS↑	Precision↑	Recall↑	FID↓	IS↑
Diffusion	ADM (Dhariwal & Nichol, 2021)	7.49	127.5	0.72	0.63	3.94	215.8
	LDM-4 (Rombach et al., 2022)	10.56	103.5	0.71	0.62	3.60	247.7
	U-ViT-H/2 (Bao et al., 2023)	–	–	–	–	2.29	263.9
	DiT-XL/2 (Peebles & Xie, 2023)	9.62	121.5	0.67	**0.67**	2.27	278.2
	MDTv2-XL/2 (Gao et al., 2023)	5.06	155.6	0.72	0.66	1.58	314.7
Mask	MaskGIT (Chang et al., 2022)	6.18	182.1	<u>0.80</u>	0.51	–	–
	MAGVIT-v2 (Yu et al., 2023)	3.65	200.5	–	–	1.78	319.4
	MAGE (Li et al., 2023)	6.93	195.8	–	–	–	–
Autoregressive	VQGAN (Esser et al., 2021)	15.78	74.3	–	–	5.20	280.3
	ViT-VQGAN (Yu et al., 2021)	4.17	175.1	–	–	3.04	227.4
	RQ-Transformer (Lee et al., 2022)	7.55	134.0	–	–	3.80	323.7
	LlamaGen-3B (Sun et al., 2024)	9.38	112.9	0.69	**0.67**	2.18	263.3
	+CCA (Ours)	<u>2.69</u>	**276.8**	<u>0.80</u>	0.59	–	–
	VAR-d30 (Tian et al., 2024)	5.25	175.6	0.75	0.62	1.92	323.1
	+CCA (Ours)	**2.54**	<u>264.2</u>	**0.83**	0.56	–	–

Table 2: Model comparisons on class-conditional ImageNet 256×256 benchmark.

Result

- How similar is CCA to CFG?

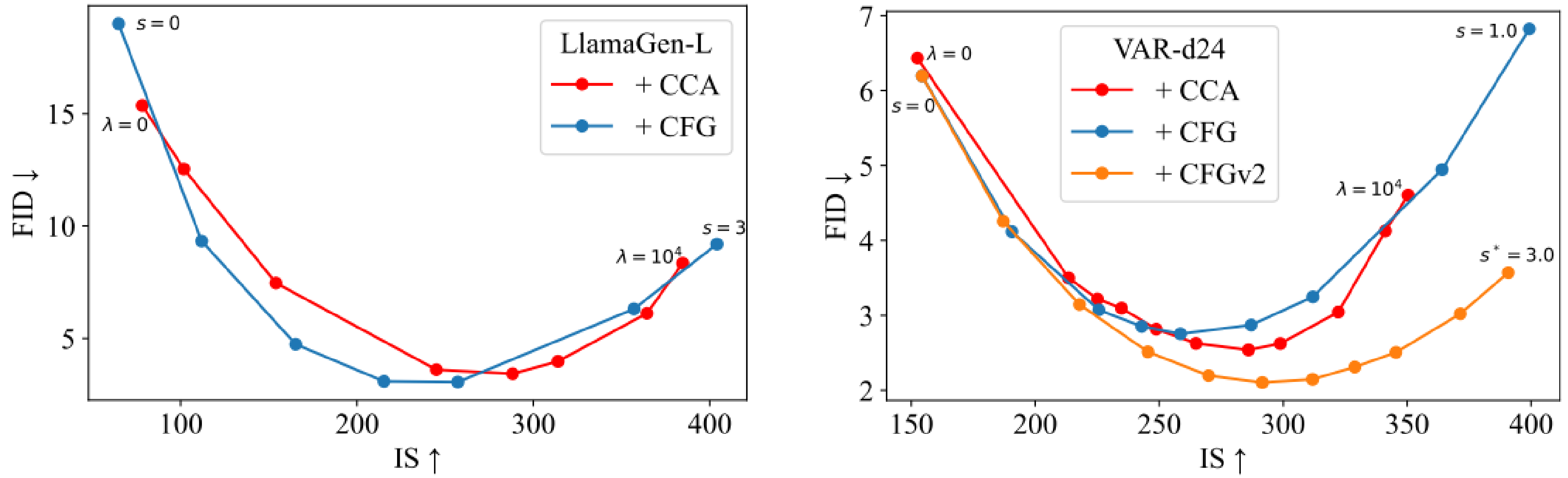


Figure 4: CCA can achieve similar FID-IS trade-offs to CFG by adjusting training parameter λ.

$$\mathcal{L}_\theta^{\text{CCA}}(\boldsymbol{x}_k, \boldsymbol{c}_k, \boldsymbol{c}_k^{\text{neg}}) = -\underbrace{\log \sigma\left[\beta \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}_k|\boldsymbol{c}_k)}{p_\phi(\boldsymbol{x}_k|\boldsymbol{c}_k)}\right]}_{\text{relative likelihood for positive conditions} \uparrow} \underbrace{-\lambda \log \sigma\left[-\beta \log \frac{p_\theta^{\text{sample}}(\boldsymbol{x}_k|\boldsymbol{c}_k^{\text{neg}})}{p_\phi(\boldsymbol{x}_k|\boldsymbol{c}_k^{\text{neg}})}\right]}_{\text{relative likelihood for negative conditions} \downarrow}$$

Result

- Can CCA be combined with CFG?

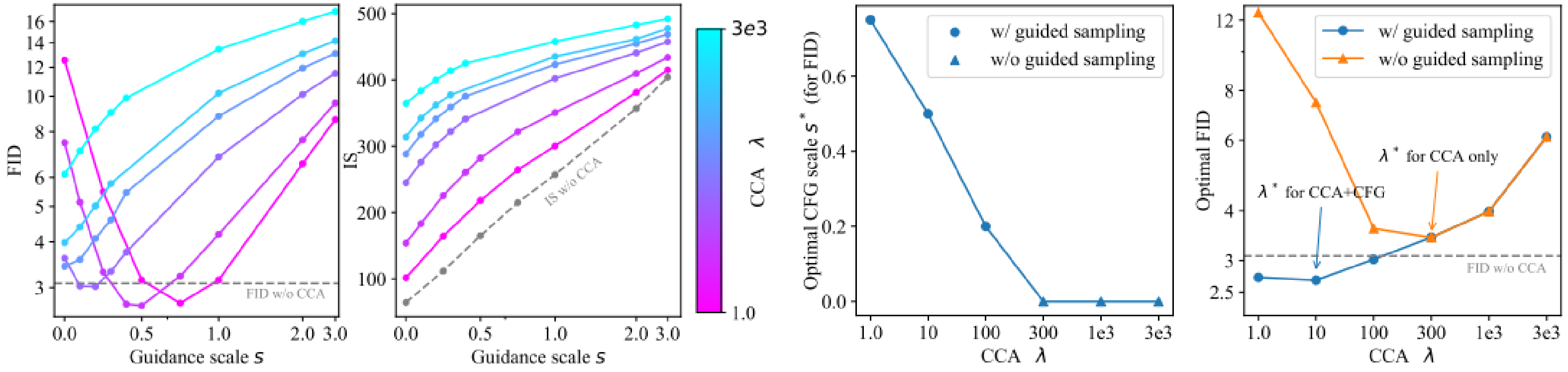


Figure 5: The impact of training parameter λ on the performance of CCA+CFG.

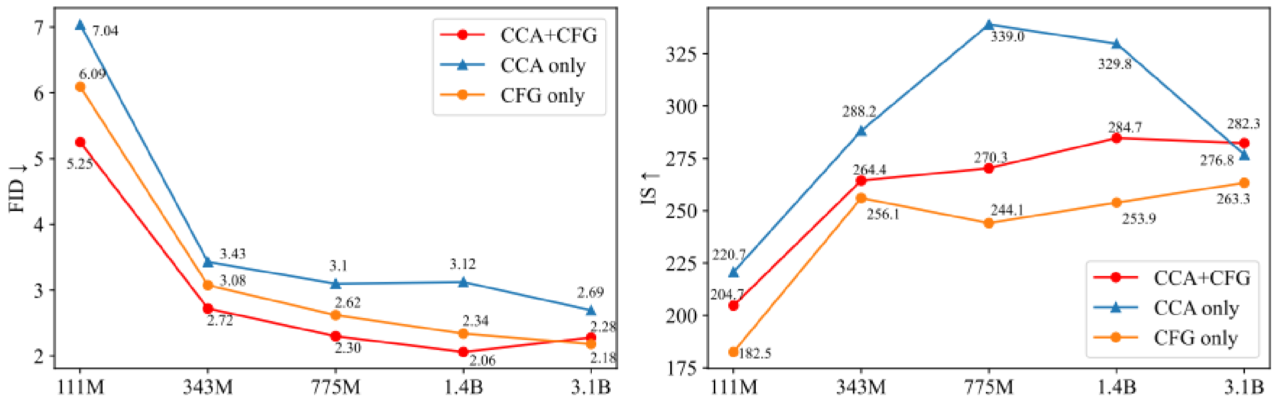


Figure 6: Integration of CCA+CFG yields improved performance over CFG alone.

Summary

- Similar to LLMs, Visual AR can be vastly improved through finetuning.
- Guided Sampling and RL Alignment are inherently connected.

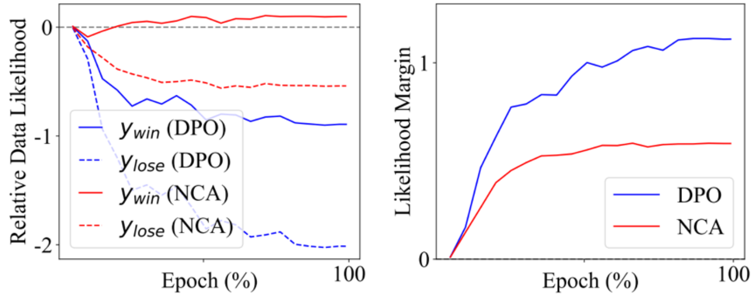
Thank you!

Why not DPO or Classifier Guidance?

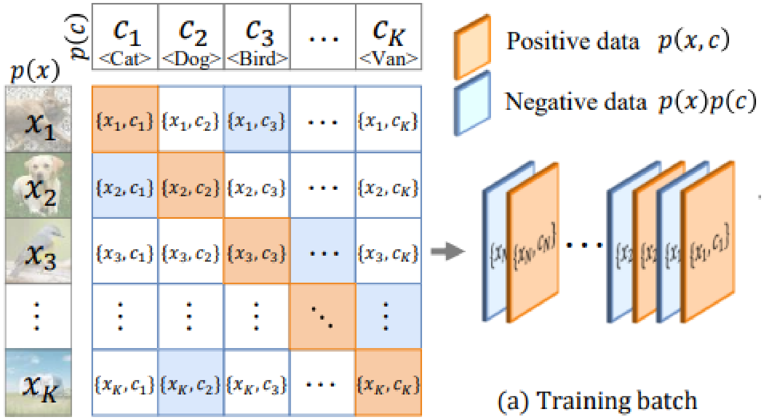
- DPO collapses (likelihood drop issue)

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\{\boldsymbol{s}, \boldsymbol{a}_w \succ \boldsymbol{a}_l\}} \log \sigma(r_\theta^{\text{LM}}(\boldsymbol{s}, \boldsymbol{a}_w) - r_\theta^{\text{LM}}(\boldsymbol{s}, \boldsymbol{a}_l))$$

$$\text{where} \qquad r_\theta^{\text{LM}}(\boldsymbol{s}, \boldsymbol{a}) := \beta \log \frac{\pi_\theta(\boldsymbol{a}|\boldsymbol{s})}{\mu_\phi(\boldsymbol{a}|\boldsymbol{s})}$$



- CG would be too computational expensive
  Almost impossible for text training



Result

- Ablation

Model	FID↓	IS	sFID↓	Precision	Recall
LlamaGen-L	19.00	64.7	8.78	0.61	**0.67**
+DPO	61.69	30.8	44.98	0.36	0.40
+Unlearn	12.22	111.6	7.99	0.66	0.64
+CCA	**3.43**	**288.2**	**7.44**	**0.81**	0.52

Model	FID↓	IS	sFID↓	Precision	Recall
VAR-d24	6.20	154.3	8.50	0.74	**0.62**
+DPO	7.53	232.6	19.10	**0.85**	0.34
+Unlearn	5.55	165.9	8.41	0.75	0.61
+CCA	**2.63**	**298.8**	**7.63**	0.84	0.55

Table 3: Comparision of CCA and LLM alignment algorithms in visual generation.

Guided Sampling

- Diffusion Guidance $$\tilde{\boldsymbol{\epsilon}}_t = (1+w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t)$$
- AR Guidance $$\ell^{\text{sample}} = \ell^c + s(\ell^c - \ell^u)$$
- Sampling Target $$p^{\text{sample}}(\boldsymbol{x}|\boldsymbol{c}) \propto p_\phi(\boldsymbol{x}|\boldsymbol{c}) \left[\frac{p_\phi(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x})}\right]^s$$

Guidance scale increases