STAR
Scale-wise Text-to-image generation via Auto-Regressive representations

Xiaoxiao Ma

2024-09-23

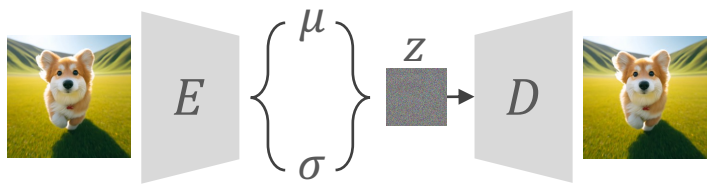
Experiments – generated samples

Figure 1: Generated samples from text prompts. We show 512×512 samples.

Background: Visual Generative Models

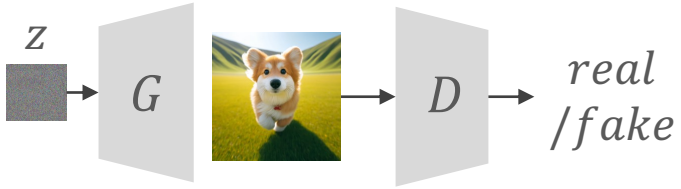
Latent Variable Modeling Learn from **explicitly constrained distribution**

- Variational Auto-Encoders (VAEs)



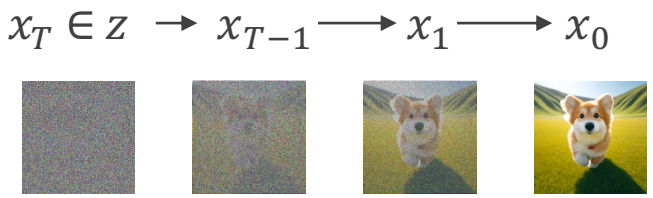
× Blurry results

- Generative Adversial Networks (GANs)



× Instability in training

- Diffusion Models (DMs)

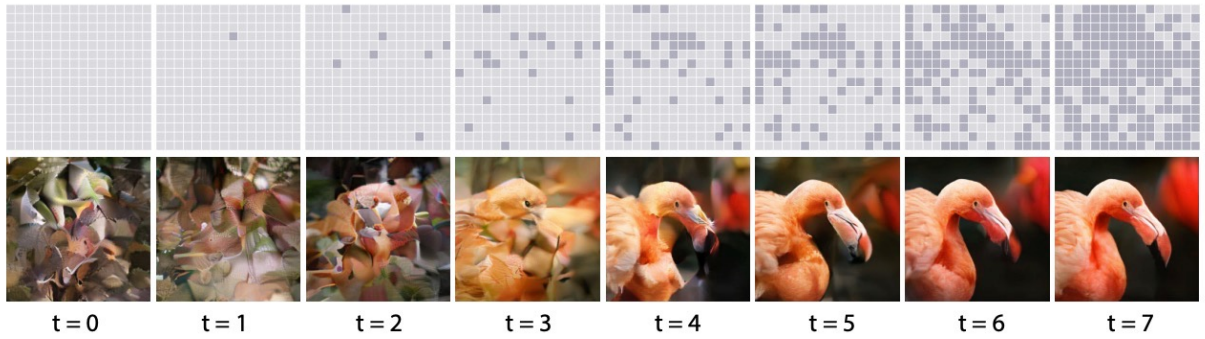


× Time-consuming

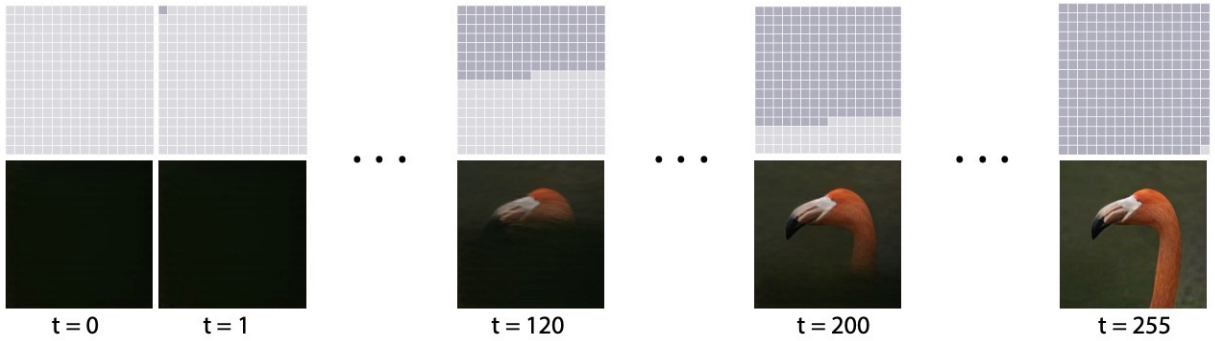
Background: Visual Generative Models

Observed Variable Modeling Learn the **distribution of discrete tokens**

- Mask-based Generation MaskGIT、MUSE

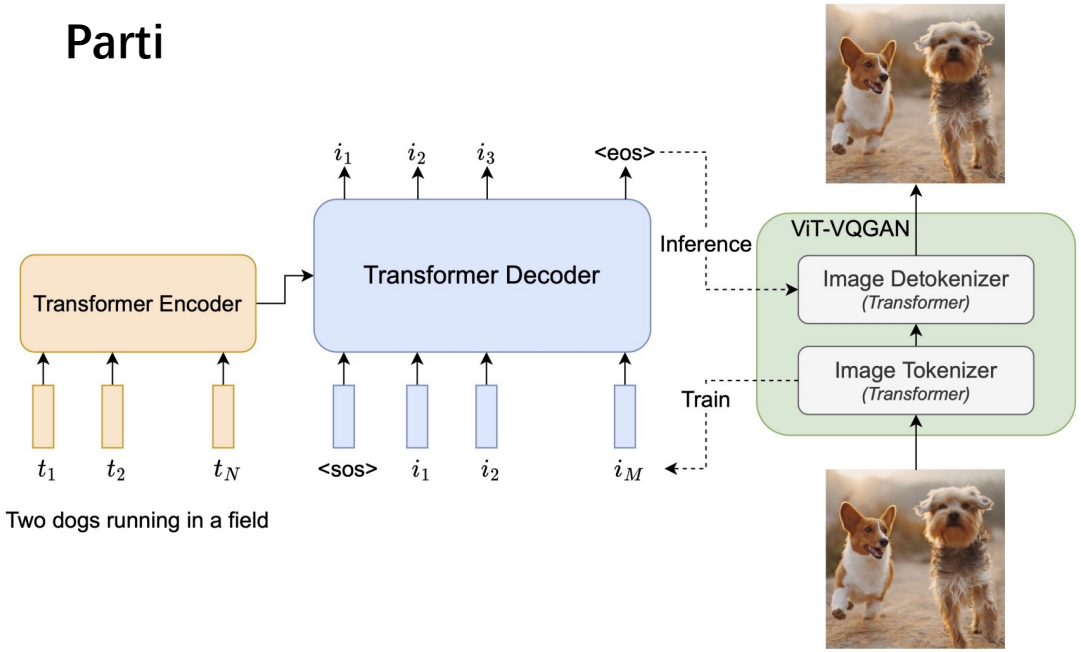


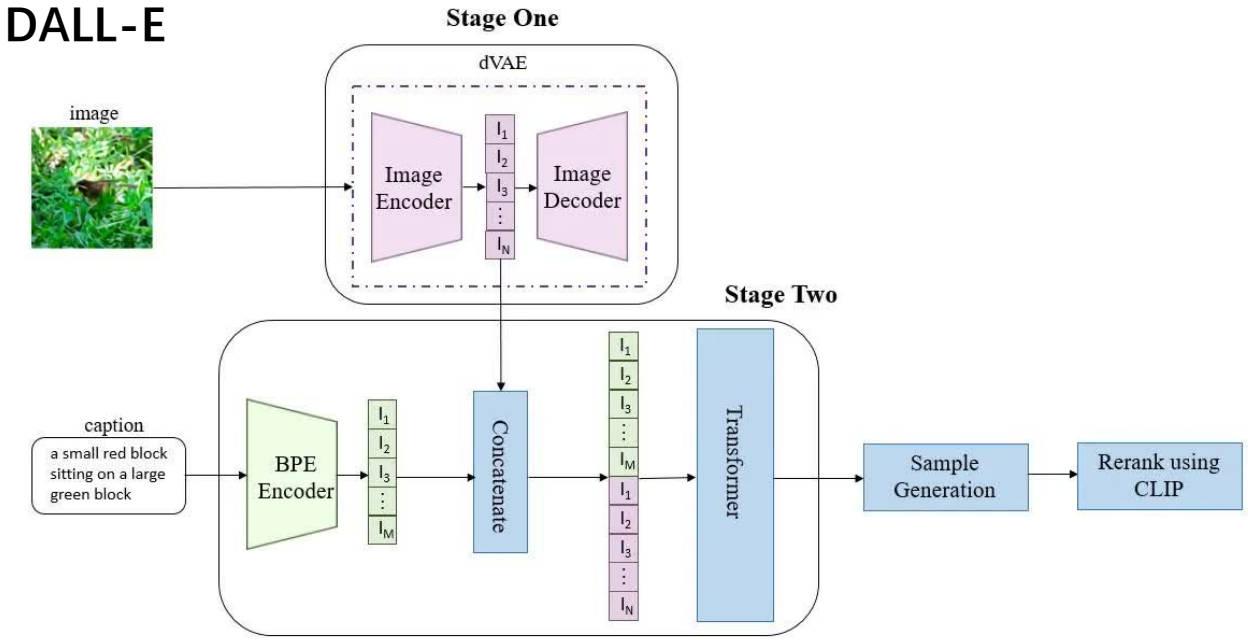
- Auto-Regressive (AR) Generation Taming transformer， DALLE， Parti…



Background: “Next-token prediction” in visual generation

- AR generation has dominated the Text generation field (e.g. GPT)
- An AR model typically includes a image tokenizer, a transformer decoder

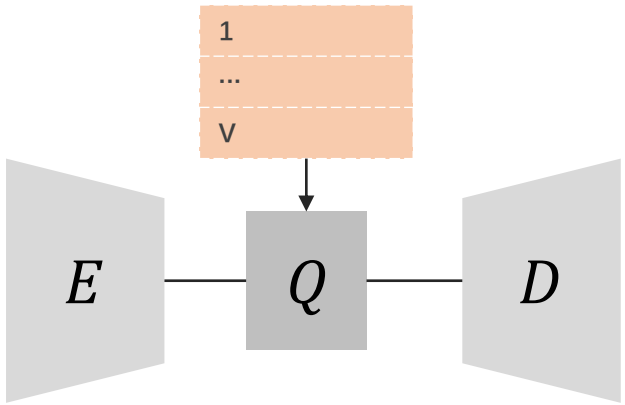




- Probability of each token is predicted based on all previous tokens
- Then the new token is sampled via sampling strategy (e.g. top-p/top-k)

Background: Image Tokenization

- VQ-GAN



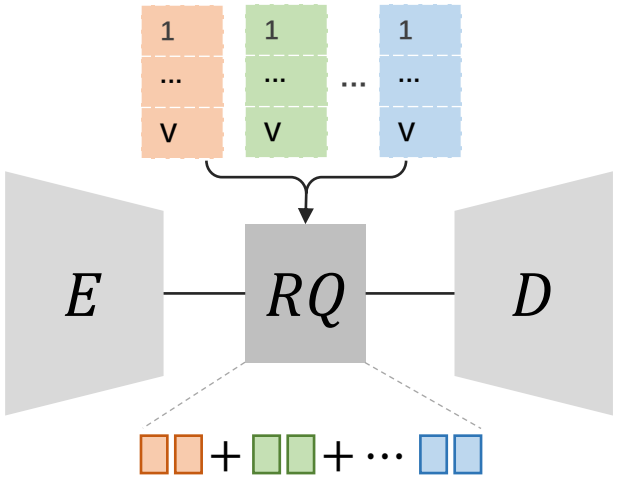
Codebook $[V]$

$$\hat{z} = E(x)$$

$$z_q = Q(\hat{z}) = \left(\arg\min_{z_k \in [V]} \left|\left|\hat{z}_{ij} - z_k\right|\right|\right)$$

$$\hat{x} = D(z_q)$$

- RQ-VAE



Total D codebooks, for each:

$$RQ(z; D) = (k_1, \ldots, k_D) \in [V]^D$$

$$r_0 = z; \;\; r_d = r_{d-1} - Q(r_{d-1})$$

$$\hat{z} = RQ(z; D) = \sum_{d=1}^{D} Q(r_d)$$

RQ-VAE: Autoregressive Image Generation using Residual Quantization

Background: Auto-regressive image generation

Drawbacks

× Image tokens requires **bi-directional** rather than casual correlation

× **Structural degradation** when simply flatten images to 1d sequences

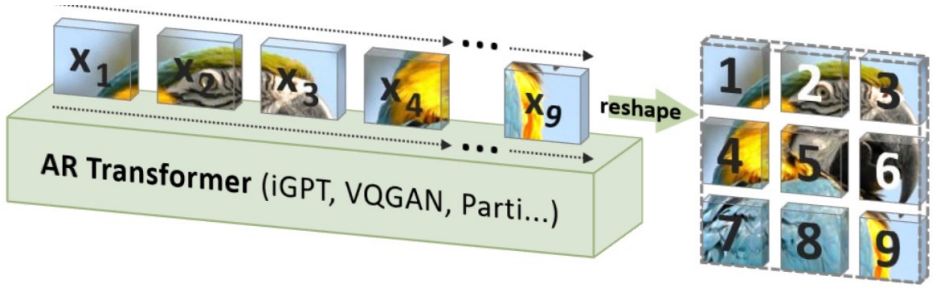
× **Time-consuming** in inference

VAR: from “next-token prediction” to “next-scale prediction”

Quantize latent feature map f into T discrete tokens $(x_1, x_2, \dots x_T)$

$$x_t \in [V]$$

$$p(x_1, x_2, \dots x_k) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \dots, x_{t-1})$$

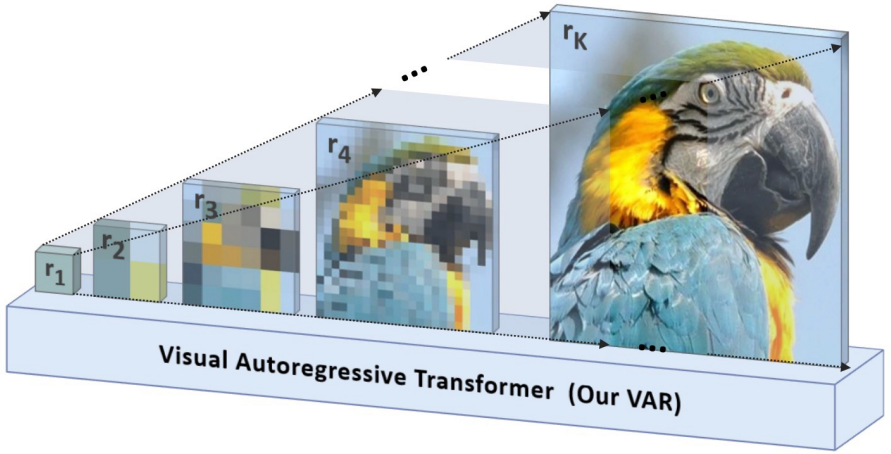


× Time-consuming, Structural degradation···

Quantize latent feature map f into K multi-scale token maps $(r_1, r_2, \dots r_K)$

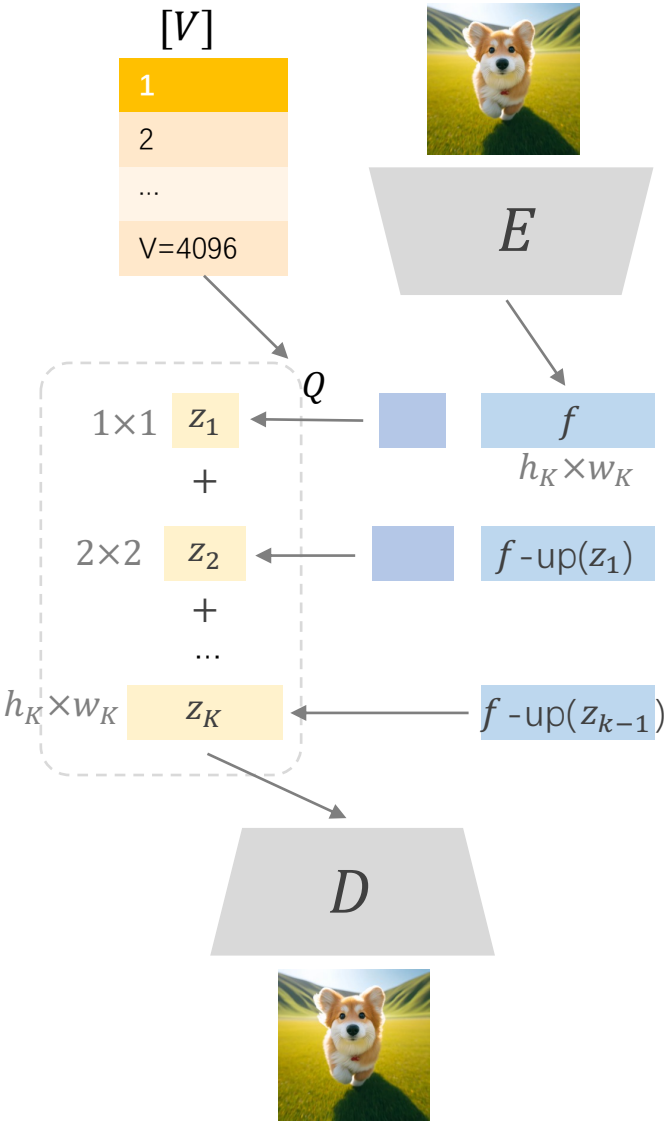
$$r_k \in [V]^{h_k \times w_k}$$

$$p(r_1, r_2, \dots r_k) = \prod_{k=1}^{K} p(r_k \mid r_1, r_2, \dots, r_{k-1})$$



✓ Efficient sampling, Bi-directional correlation, Better scalability···

VAR: Multiscale Image quantizer

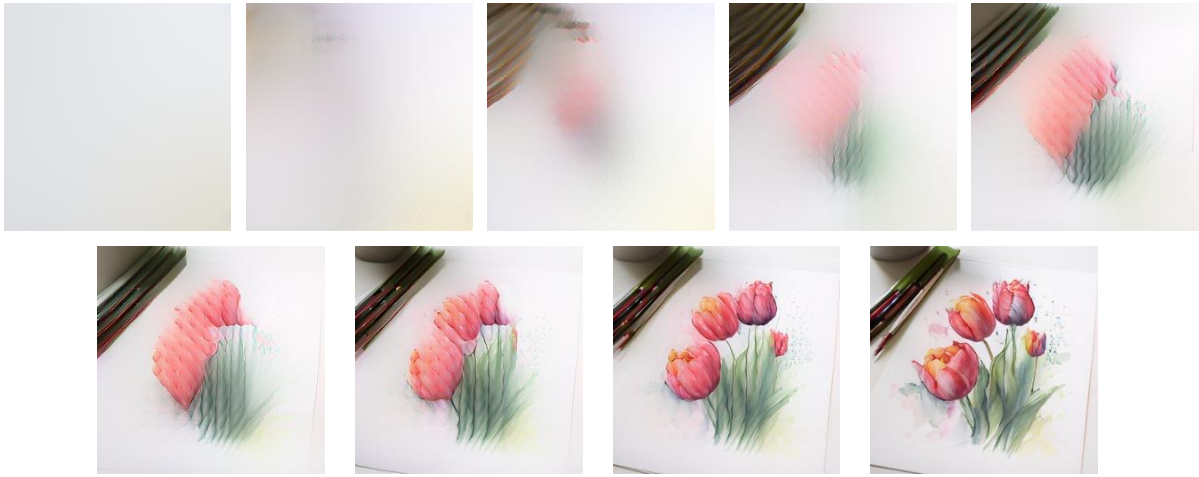


Algorithm 1: Multi-scale VQVAE Encoding

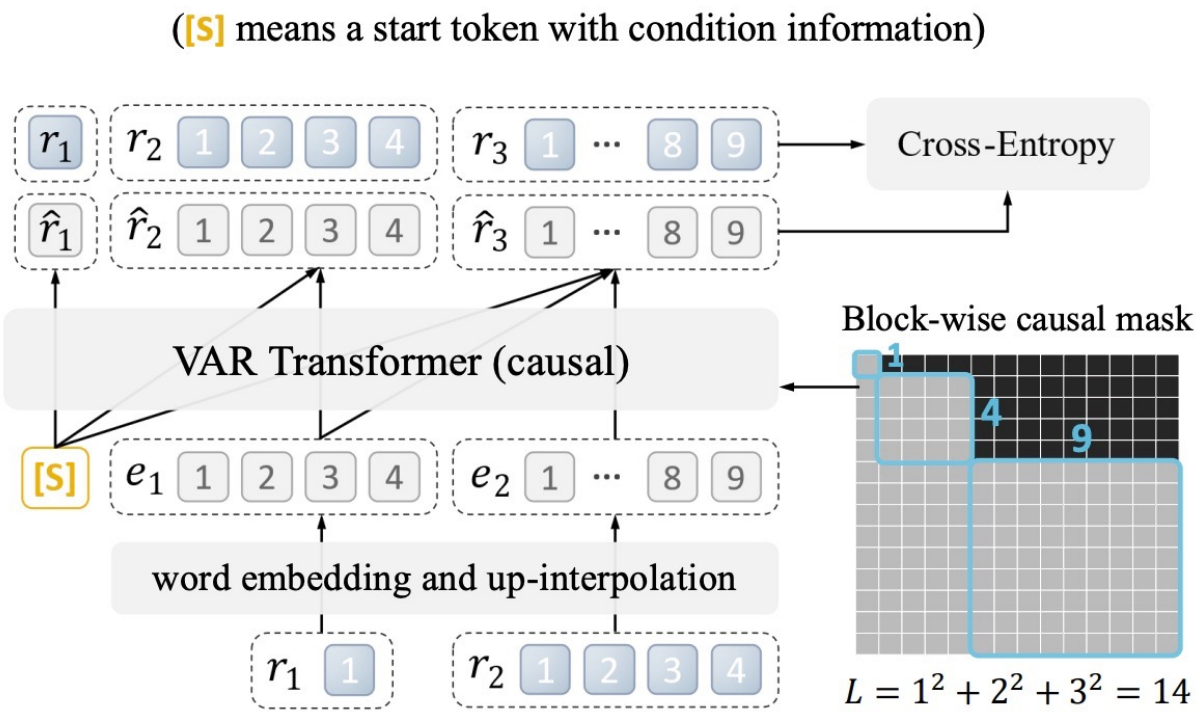
Inputs: raw image im;
Hyperparameters: steps K, resolutions $(h_k, w_k)_{k=1}^{K}$;
$f = \mathcal{E}(im)$, $R = []$;
for $k = 1, \cdots, K$ **do**
  $r_k = \mathcal{Q}(\text{interpolate}(f, h_k, w_k))$;
  $R = \text{queue\_push}(R, r_k)$;
  $z_k = \text{lookup}(Z, r_k)$;
  $z_k = \text{interpolate}(z_k, h_K, w_K)$;
  $f = f - \phi_k(z_k)$;
Return: multi-scale tokens R;

Algorithm 2: Multi-scale VQVAE Reconstruction

Inputs: multi-scale token maps R;
Hyperparameters: steps K, resolutions $(h_k, w_k)_{k=1}^{K}$;
$\hat{f} = 0$;
for $k = 1, \cdots, K$ **do**
  $r_k = \text{queue\_pop}(R)$;
  $z_k = \text{lookup}(Z, r_k)$;
  $z_k = \text{interpolate}(z_k, h_K, w_K)$;
  $\hat{f} = \hat{f} + \phi_k(z_k)$;
$\hat{im} = \mathcal{D}(\hat{f})$;
Return: reconstructed image $\hat{im}$;



VAR: Auto-regressive sampling



Generate tokens in same scale at a time

Step 1:

Step 2:

Step 3:

Total K forward passes, with top-p, top-k sampling

$$p(r_1, r_2, \ldots r_k) = \prod_{k=1}^{K} p(r_k \mid r_1, r_2, \ldots, r_{k-1})$$

Limitations:

Rely heavily on the condition token [s]

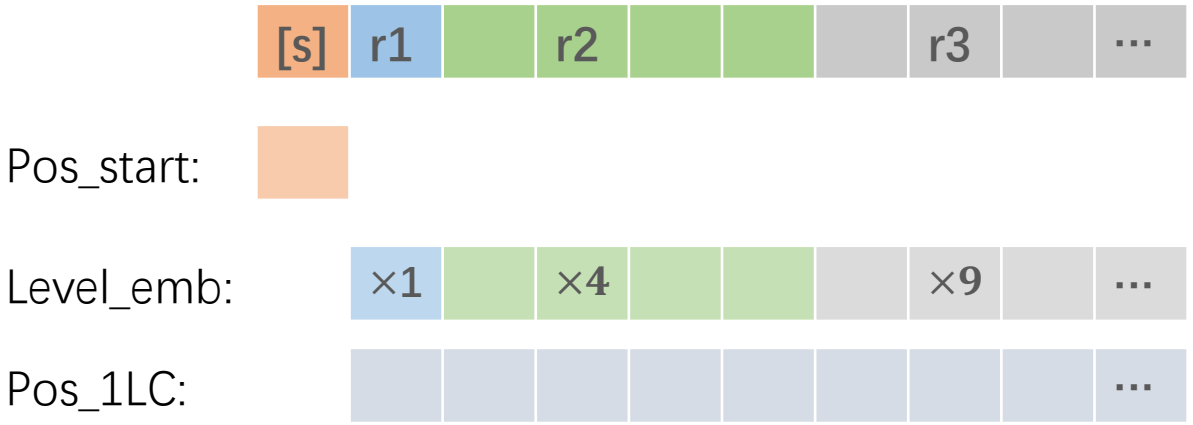
Large attention map (hard to train, especially under high-resolution)

Possibly unstable for sampling, hard to generate complex scenes…

VAR: Positional encodings:

Positional encoding in VAR includes three parts:

- Pos_start: for [s], $length = 1$
- Level_emb: for each scale, $length = K$
- Pos_1LC: for each tokens at each scale, $length = \sum_{k=1}^{K} h_k \times w_k$



Limitations: Unefficiency, duplicated, hard to adapt to new resolutions…

VAR: Intrinsic Issues

Restricted Supervision

Condition limited in one token [s], thus restricted to fixed categories **(close set)**

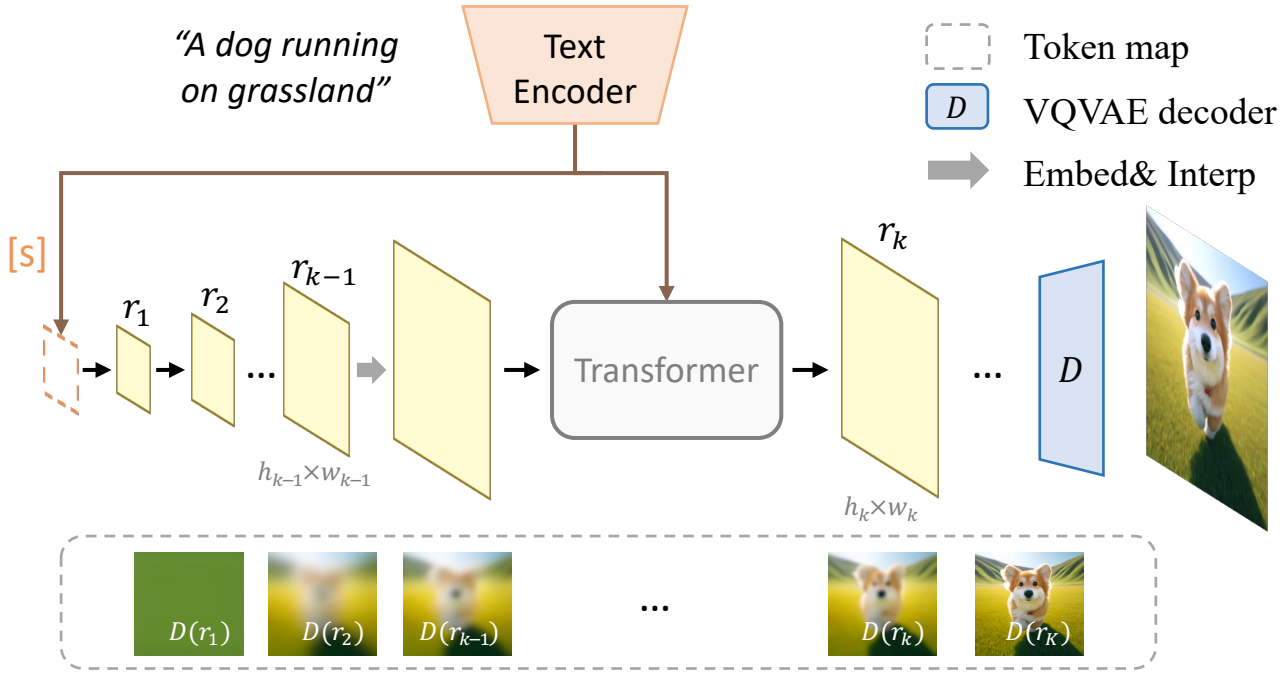
How to generate under text prompts? **(open set, generalizable)**

Inefficiency in Spatial Modeling

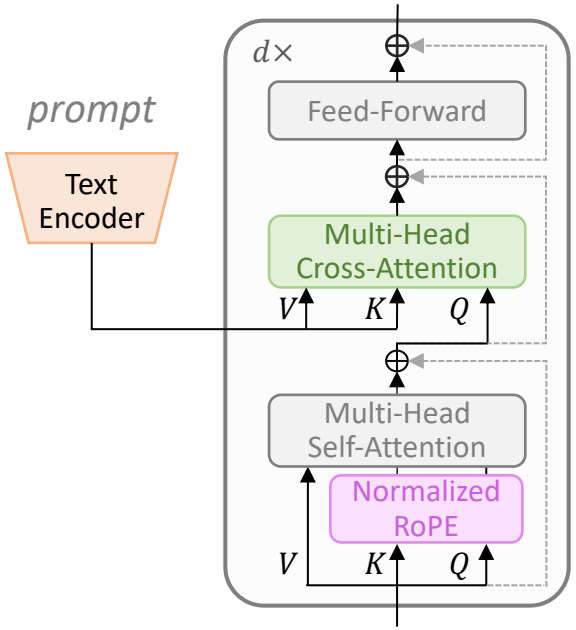
learnable absolute positional encodings (APEs) for each token at each scale

- Extra parameters for training & difficult to learn, **inefficiency**
- Unable to leverage progressive training for **high-resolution generation**

Ours: Scale-wise T2I Gen. via AR Models



(a) Framework of STAR



(b) Transformer architecture

- **Pre-trained text encoder** for handling textual information
- **Cross-Attention** provides detailed textual guidance
- **Normalized RoPE** marks position without extra parameters

STAR: Textual guidance

Pooled text feature for start token [s]

- Provides diverse and generalizable textual descriptions
- Provides global semantic information

Additional cross-attention:

- More detailed textual guidance for each scale (layout, multi-object…)

STAR: Problems in existing positional encodings:

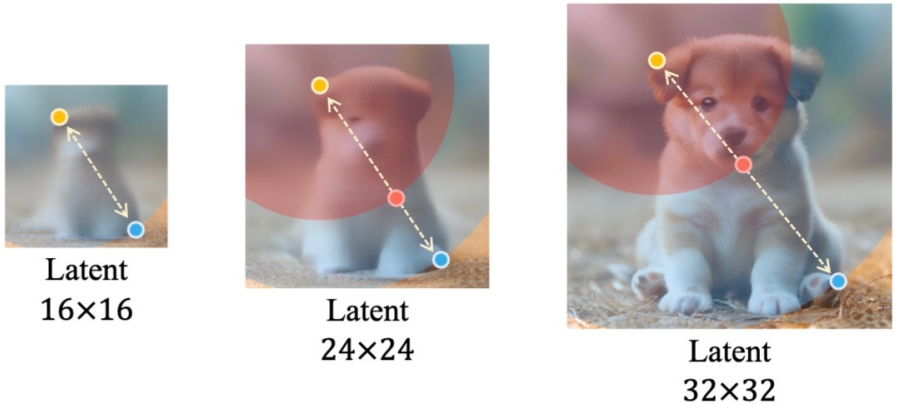
Learnable absolute positional encodings (APEs)

- Only applicable under **fixed image size**
- Require **extra parameters** for training

Sinusoidal positional encodings

- **Confusions** when applied to different scales

(Significantly different encodings in tokens at different scales, while they actually represents the same position in an image)



Ours - Normalized RoPE

For k-th token maps of size $h_k \times w_k$

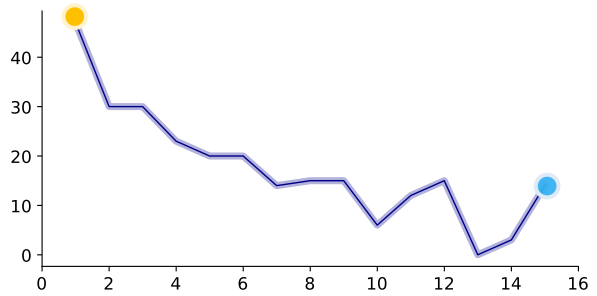
With position $i \in \{1, 2, \ldots, h_k\}, j \in \{1, 2, \ldots, w_k\}$

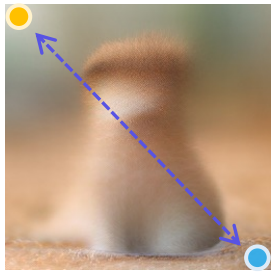
Normalized positional encoding $\mathrm{PE}(i, j)$ is calculated as:

$$\mathrm{PE}(i, j) = \mathrm{RoPE}\left(\frac{i}{h_k} \cdot H\right) \oplus \mathrm{RoPE}\left(\frac{j}{w_k} \cdot W\right)$$

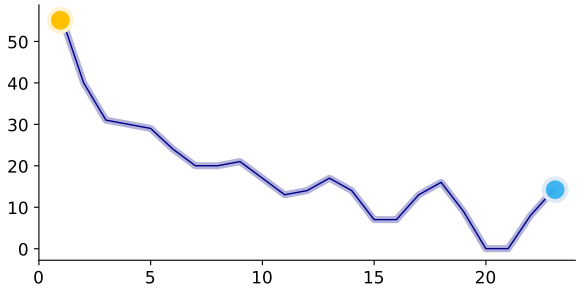
(H, W) represents normalized shape

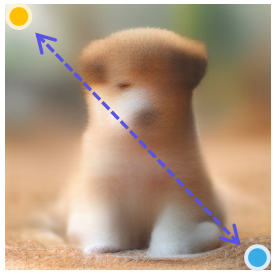
Latent 16×16



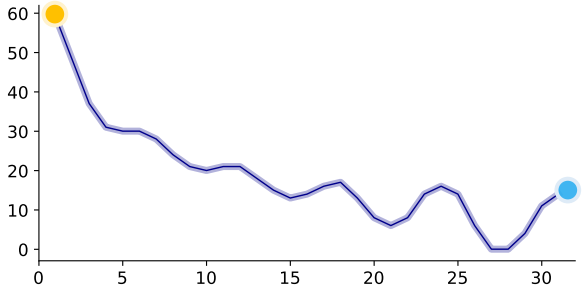


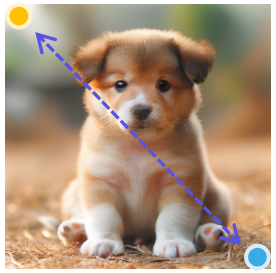
Latent 24×24





Latent 32×32

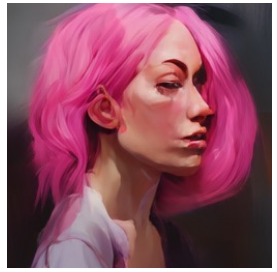




Progressive training

In STAR, only Level_emb is correlated to different scales

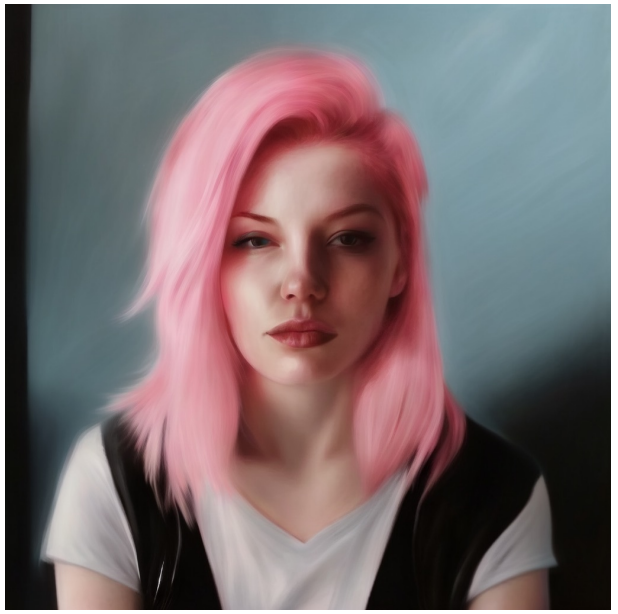
So progressive training is enabled for **high-resolution generation with low training cost**



Trained model under **256x**

Add new scales for **512x** generation

Finetune for several iterations



Repeat process for **1024x** generation

Quantitative Comparisons

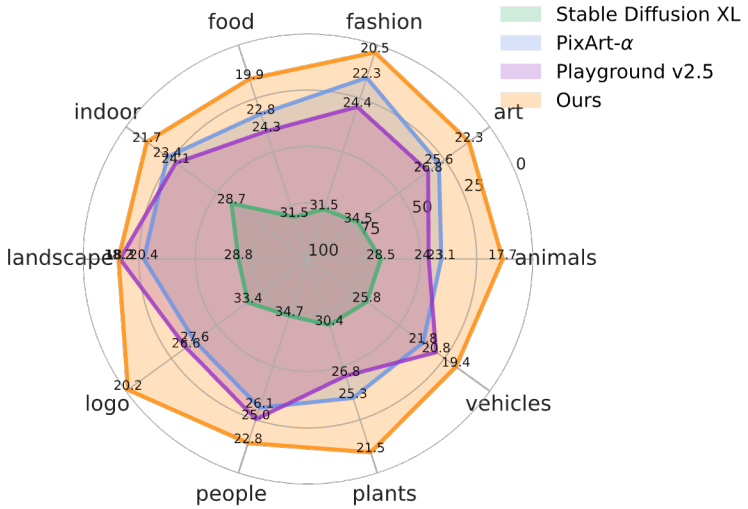
MJHQ-30k benchmark:

Methods	FID ↓	CLIP ↑
Stable Diffusion 2.1	26.96	0.259
Stable Diffusion XL	11.42	**0.291**
Playground v2.5	*6.57*	0.283
PixArt-α	6.64	0.284
Ours	**4.73**	**0.291**

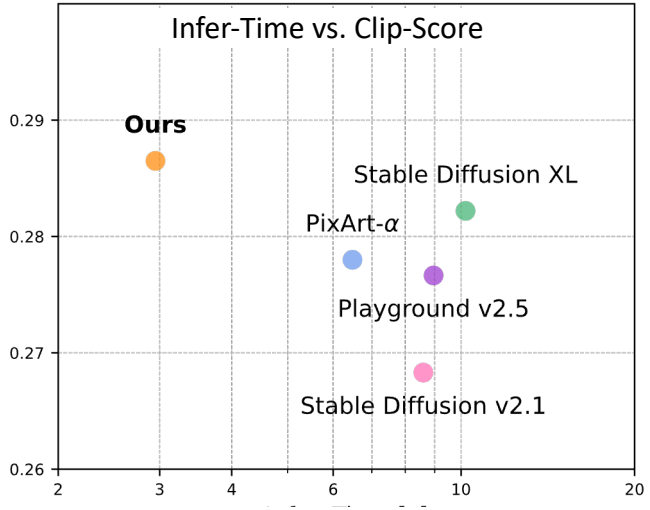
ImageReward v1.0 benchmark:

Methods	IR ↑	CLIP ↑
Stable Diffusion 2.1	0.246	0.268
Stable Diffusion XL	0.416	*0.282*
Playground v2.5	0.693	0.277
PixArt-α	**0.904**	0.278
Ours	*0.866*	**0.287**

MJHQ FID

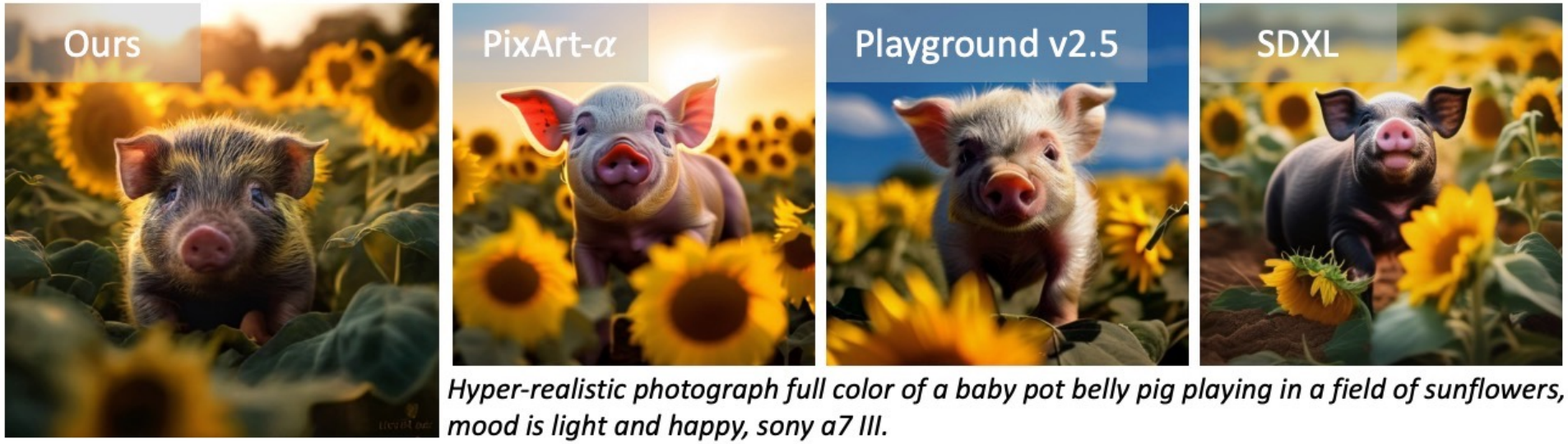


High Efficiency



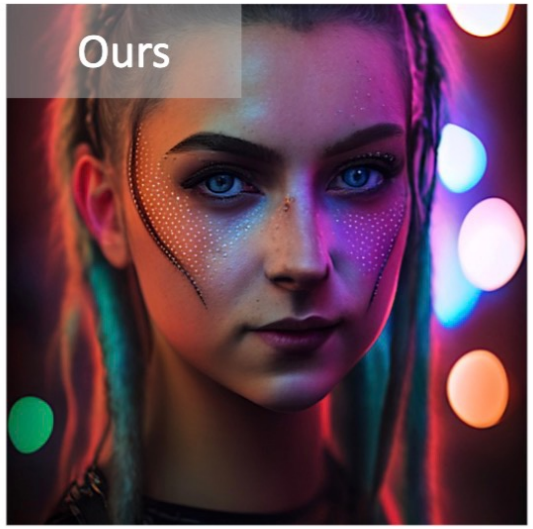
Qualitative Comparisons

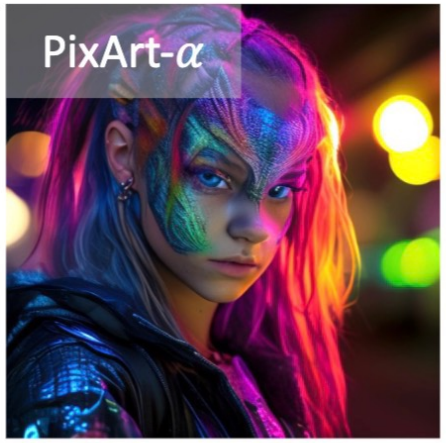
The cutest shiba inu sitting on the ground with fallen leaves, close-up view, high quality.



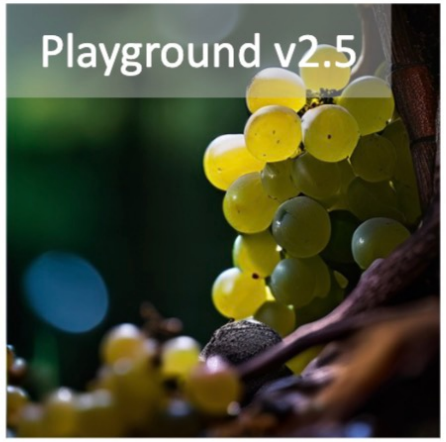
Hyper-realistic photograph full color of a baby pot belly pig playing in a field of sunflowers, mood is light and happy, sony a7 III.

Experiments - qualitative comparisons





Australian rainbow serpent festival during the night time closeup photo of cyberpunk teen girl, photorealistic portrait lens 50mm professional lens.



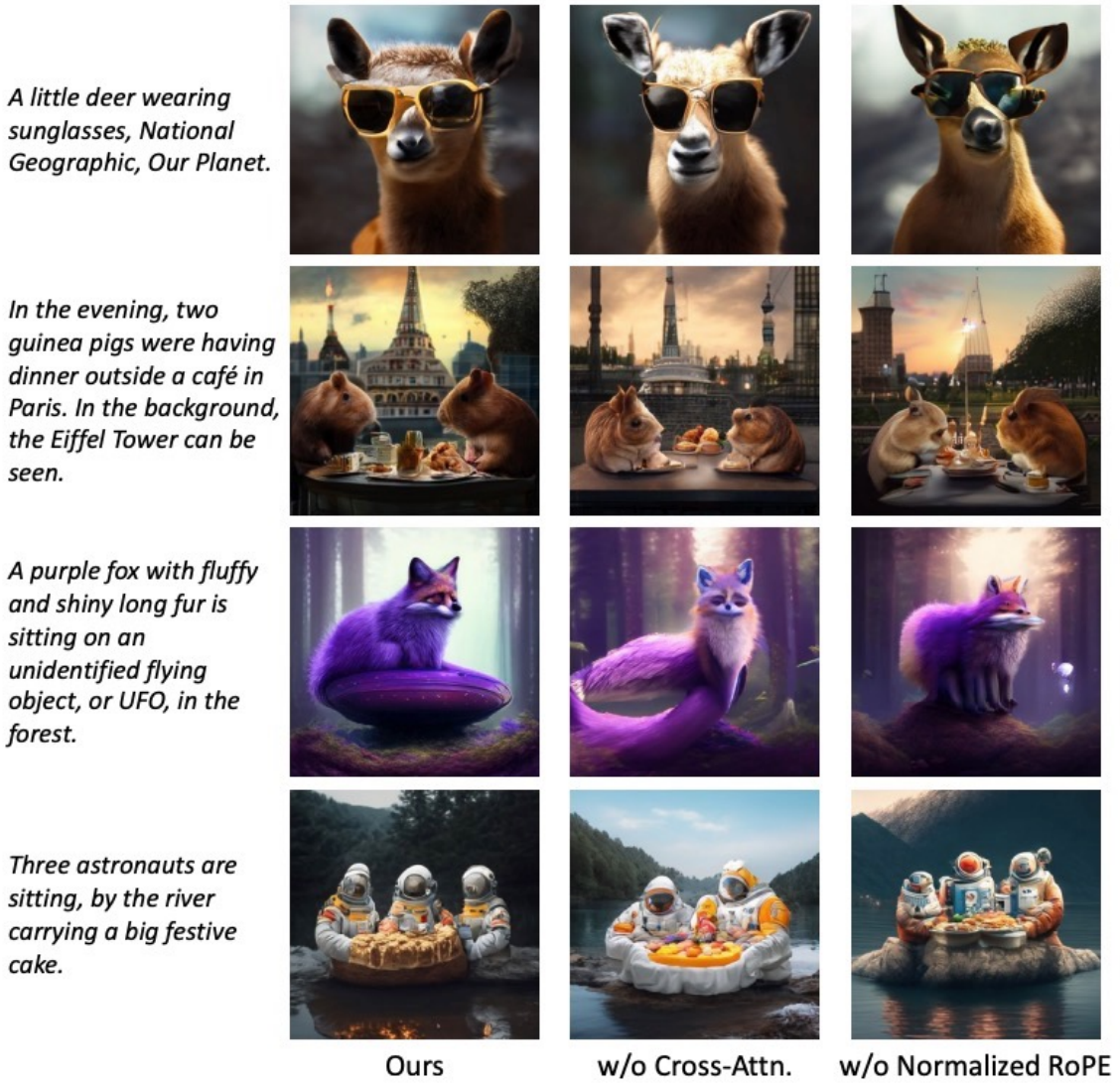
Insanely Realistic, macro, forest floor, professional photography, award winning photo, 60mm photo stacking, grapes in full focus, bokeh in the background, nature background.

Ablation studies

Model size & resolutions:

Depth	#Reso	#Param	CLIP ↑	FID ↓
16	256	274M	0.272	6.88
30	256	1.68B	0.286	5.19
30	512	1.68B	0.291	4.73

Cross-attention & Normalized RoPE:



Samples of generating 1024x images










Discussion - Limitations

Image tokenizer - Unsatisfactory reconstruction of image details

Origin Image

KL VAE (from SD)

Multiscale VQ-VAE

Possibly can be solved via advanced quantization (e.g. LFQ, FSQ, MoVQ…)

Unstable sampling results - Deformation especially under multiple objects

Generate images with arbitrary aspect ratios

Related works - Other VAR related papers

ControlVAR: Exploring Controllable Visual Autoregressive Modeling

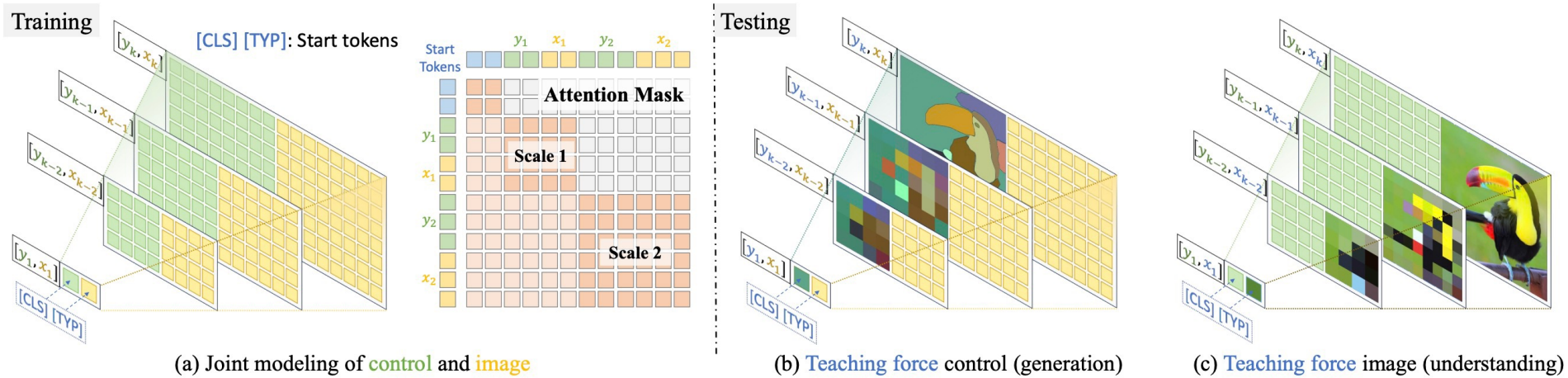


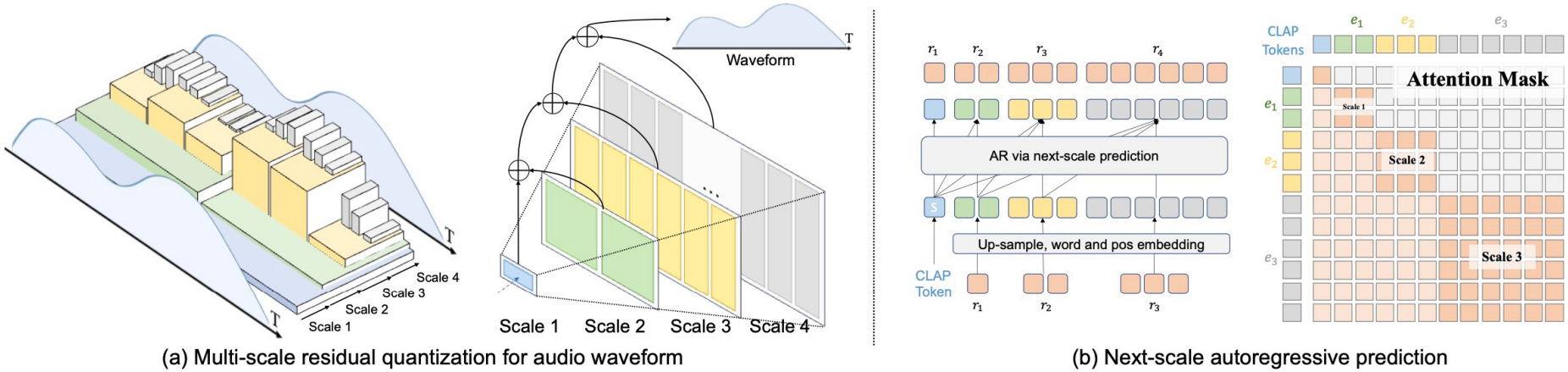
Figure 4: Illustration of ControlVAR. We jointly model the control and image during training with start tokens [CLS] and [TYP] to specify the semantics and control type. We conduct conditional generation by teacher forcing the AR prediction during testing.

Arxiv: https://arxiv.org/abs/2406.09750

Repo: N/A

Related works - Other VAR related papers

Efficient Autoregressive Audio Modeling via Next-Scale Prediction (AAR)



(a) Multi-scale residual quantization for audio waveform

(b) Next-scale autoregressive prediction

Figure 2: Our model involves two distinct training phases. **Stage 1**: Scale-level Audio Tokenizer (SAT) to encode an audio sample into a series of K tokens scales, donated as $\mathcal{R} = (r_1, r_2, \ldots, r_K)$. Each scale encodes information in different frequencies of the audio waveform. **Stage 2**: Acoustic AutoRegressive (AAR) modeling via next-scale prediction relies on the pre-trained SAT to predict each scale-level token r_i by conditioning on all previously predicted scales $r_{<i}$ and a CLAP token (Wu et al. 2023) as the start token. The CLAP token is derived from ground truth audio. During training, we use the standard cross-entropy loss and the attention mask as figured above to ensure that each r_i can only be attributed by $r_{\leq i}$ and the start token.

Arxiv: https://arxiv.org/pdf/2408.09027

Repo: https://github.com/qiuk2/AAR

Related works - Other VAR related papers

G3PT: Unleash the power of Autoregressive Modeling in 3D Generation via Cross-scale Querying Transformer

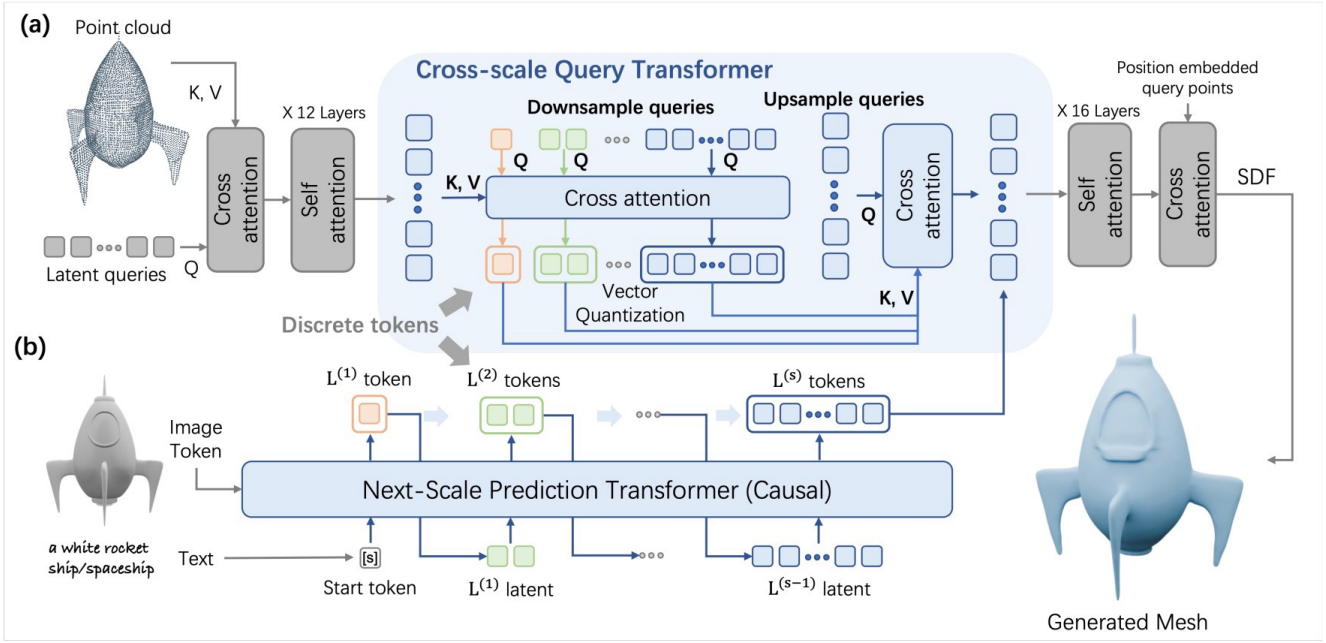


Figure 1: Overall pipeline for processing and generating unordered 3D data. (a) G3PT starts by encoding the input point cloud into discrete scales of token maps, each representing different levels of detail. The proposed Cross-scale Querying Transformer (CQT) utilizes a cross-attention layer with varying numbers of queries to globally connect tokens across different scales, without requiring the tokens to be in a specific order. The final output is the SDF value for each query point. (b) CQT enables 3D generation from coarse to fine scales under various conditions. An autoregressive transformer is trained using next-scale prediction.

Arxiv: https://arxiv.org/pdf/2409.06322 Repo: N/A

Related works – A summary of recent discrete token based Generative Works

The “Next Scale Prediction” scheme can also do Unified Understand & Generation!

Methods	Type	Task	Arxiv
LlamaGen	AR	C2I, T2I	https://arxiv.org/abs/2406.06525
TiTok	Mask	C2I	https://arxiv.org/pdf/2406.07550
MARS	AR	T2I, Multilingual Generation	https://arxiv.org/html/2407.07614v1
Lumina-mGPT	AR	Unified Understand & Gen	https://www.arxiv.org/abs/2408.02657
Show-o	AR+Diff	Unified Understand & Gen	https://arxiv.org/pdf/2408.12528
Transfusion	AR+Diff	Unified Understand & Gen	https://www.arxiv.org/abs/2408.11039
Open-MagViTv2	AR	C2I	https://arxiv.org/html/2409.04410

Thanks!

Q & A