CART: Compositional AutoRegressive Transformer for Image Generation

Siddharth Roheda, Rohit Chowdhury, Aniruddha Bala, Rohan Jaiswal

Samsung Research Institute
Bangalore, India
{sid.roheda, rohit.c, aniruddha.b, r.jaiswal}@samsung.com

Abstract

We propose a novel Auto-Regressive (AR) image generation approach that models images as hierarchical compositions of interpretable visual layers. While AR models have achieved transformative success in language modeling, replicating this success in vision remains challenging due to inherent spatial dependencies in images. Addressing the unique challenges of vision tasks, our method (CART) adds image details iteratively via semantically meaningful decompositions. We demonstrate the flexibility and generality of CART by applying it across three distinct decomposition strategies: (i) Base-Detail Decomposition (Mumford-Shah smoothness), (ii) Intrinsic Decomposition (albedo/shading), and (iii) Specularity Decomposition (diffuse/specular). This "next-detail" strategy outperforms traditional "next-token" and "next-scale" approaches, improving controllability, semantic interpretability, and resolution scalability. Experiments show CART generates visually compelling results while enabling structured image manipulation, opening new directions for controllable generative modeling via physically or perceptually motivated image factorization.

Extended version — https://arxiv.org/abs/2411.10180

Introduction

Recent advancements in Generative AI for image synthesis have garnered significant interest across research and industry. Conventional approaches including Generative Adversarial Networks (GANs) (Goodfellow et al. 2020; Mirza 2014) and Variational Auto Encoders (VAEs) (Kingma 2013; Shao et al. 2020) typically generate entire scenes in a single pass. Recent research has introduced step-wise approaches where each step incorporates a subset of details. Diffusion-based methods (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) initiate with noise and employ denoising models to progressively reveal coherent images. Similarly, Auto-Regressive (AR) models (Van Den Oord, Kalchbrenner, and Kavukcuoglu 2016; Salimans et al. 2017; Gregor et al. 2015; Parmar et al. 2018) tackle generation in a patch-wise manner. Image generation models like VQGAN (Esser et al. 2021) and DALLE (Ramesh et al. 2021) aim to parallel the success of AR models in Large Language Modelling (LLMs) by using visual tokenizers that convert images into 2D token grids enabling next-token prediction.

Despite the success of AR models in Natural Language Processing (NLP), achieving similar vision advancements remains challenging. Recent studies in AR (Tian et al. 2024) highlight that token prediction sequence can significantly impact performance. VAR (Tian et al. 2024) adopts a multi-scale tokenization approach, where token maps at different scales are created within the encoded latent space. A transformer is then trained to predict the next higher-resolution token map, while conditioning on the previously generated token maps. This "next-scale" strategy enables progressive resolution expansion, improving upon raster-scan tokenization. However, while VAR enhances scalability, it refines both global structures and fine details simultaneously at each scale, without explicitly disentangling them. This entanglement of structural and textural features limits fine-grained control over generated image characteristics. Furthermore, such an intertwined representation necessitates retraining or fine-tuning whenever the target generation resolution deviates from that used during VAR training, limiting its flexibility across resolutions.

To address these limitations, we draw inspiration from human perception and visual content creation, which fundamentally follow a compositional approach. For instance, artists typically begin by outlining spatial layouts and global structures, then progressively refine color, textures, and details. Motivated by this process, we propose a novel compositional AR framework that synthesizes images by sequentially predicting constituent visual factors, ranging from structural layouts to appearance refinements. Our approach decomposes training images into physically relevant "base" and "detail" components, encoding them into multi-scale detail token-maps. AR processing initiates with 1×1 tokens, predicting successive token-maps to construct a base component at the target resolution. The model then predicts detail components, incrementally layering them to enhance the base image. Such a "next-detail" generation approach enables the model to synthesize images with significantly higher detail compared to state-of-the-art methods such as (Tian et al. 2024) and (Sun et al. 2024). In addition, it supports training-free high-resolution image generation and facilitates super-resolution of low-quality inputs.

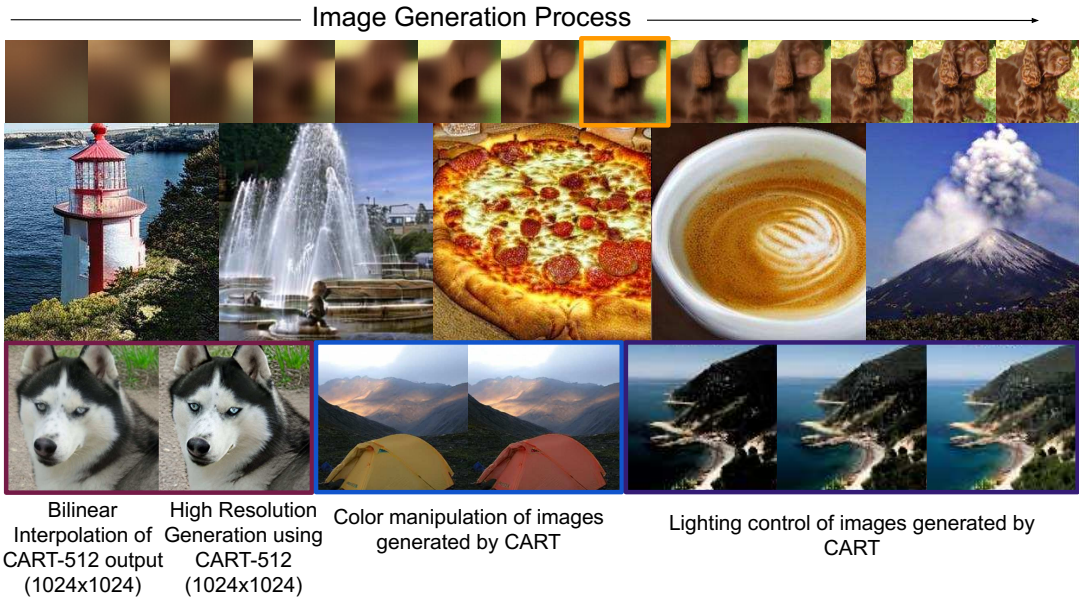


Figure 1: Top row: Image Generation process using CART; orange box indicates generated base component. Middle Row: Generated image samples from CART. Bottom row: Applications of CART: High resolution generation without retraining (Magenta), Recoloring objects in scenes (Blue), Lighting control (Purple)

The **Contributions** of this paper include:

- A novel **iterative image generation** approach aligning with natural image formation order.
- A **hierarchical tokenization strategy** to quantize an image into base and detail layers.
- **High-resolution generation and Super-Resolution without retraining**, demonstrating versatility.
- **Fine-grained control** over image characteristics such as textures, colors, and lighting.

Related Work

Generative Models

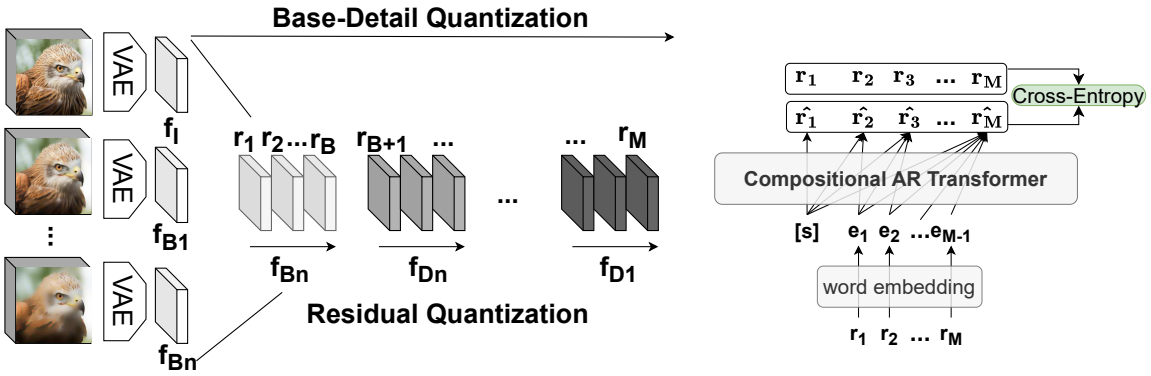
Generative models for image synthesis have advanced rapidly, enabling both unconditional and conditional generation based on priors. VAEs (Kingma 2013; Shao et al. 2020) and GANs (Goodfellow et al. 2020; Mirza 2014) established foundational approaches, with GANs generating high-quality images via adversarial training. Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Ho et al. 2022) introduce sequential denoising processes, gradually refining noise into realistic images. Their remarkable ability to synthesize high quality images with fine-grained visual details have enabled applications in text-to-image generation (Zhang et al. 2023; Zhu et al. 2023), inpainting (Lugmayr et al. 2022; Corneanu, Gadde, and Martinez 2024; Yang, Chen, and Liao 2023), super-resolution (Yue, Wang, and Loy 2024; Li et al. 2022), 3D reconstruction (Anciukevičius et al. 2023; Zhou and Tulsiani 2023), and image editing (Brooks, Holynski, and Efros 2023; Kawar et al. 2023; Bala et al. 2024). However, their many iterative steps add computational overhead, limiting scalability for real-time, high-resolution synthesis.

Auto-Regressive Generative Models

Auto-Regressive (AR) models attempt to predict next tokens in a sequence while conditioned on previous tokens. GPT models (Brown 2020; Radford et al. 2019) using transformers (Vaswani 2017) achieved revolutionary success in language tasks, motivating computer vision applications. Early attempts included DRAW (Gregor et al. 2015) with sequential variational auto-encoding using RNNs, and pixel-level prediction approaches (PixelCNN (Salimans et al. 2017), PixelRNN (Van Den Oord, Kalchbrenner, and Kavukcuoglu 2016), and Image Transformer (Parmar et al. 2018)). However, sequentially predicting billions of pixels proved computationally prohibitive, and Image-GPT (Chen et al. 2020) with 6.8B parameters only achieved image generation at 96×96 resolution. Vector Quantized VAE (VQ-VAE) (Van Den Oord, Vinyals et al. 2017) addressed scalability by compressing images into discrete token sequences. In (Parmar et al. 2018) transformer decoder was utilized to enable AR generation using VQ-VAE tokens. VAR (Tian et al. 2024) demonstrated that token ordering critically impacts AR image generation, and proposed multi-scale tokenization with "next-scale" prediction.

Vector Quantized VAE (VQ-VAE)

In order to perform AR modeling of images via next-token prediction, VQ-VAE is utilized to tokenize the image into discrete tokens. The encoder $\boldsymbol{\mathcal{E}}$, converts images to feature maps $\boldsymbol{f} = \boldsymbol{\mathcal{E}}(\boldsymbol{I}) \in \mathbb{R}^{h \times w \times C}$, followed by quantization to



(a) Encoding and Tokenization (b) Next Detail Prediction

Figure 2: Overview of the CART Approach.

discrete tokens $\boldsymbol{q} = \mathcal{Q}(\boldsymbol{f}) \in [V]^{h \times w}$ using learnable codebook $\mathcal{Z} \in \mathbb{R}^{V \times C}$ with V vectors,

$$q^{(i,j)} = \left(arg \min_{v \in [V]} ||\text{look-up}(\mathcal{Z}, v) - f^{(i,j)}||_2\right) \in [V], \quad (1)$$

where look-up$(\mathcal{Z}, v)$ refers to taking the v^{th} vector in codebook $\mathcal{Z}$. Reconstruction involves codebook lookup $\hat{\boldsymbol{f}} = \text{look-up}(\mathcal{Z}, \boldsymbol{q})$ and decoding $\hat{\boldsymbol{I}} = \mathcal{D}(\hat{\boldsymbol{f}})$. Training a VQ-VAE involves the minimization of a compound loss,

$$||\boldsymbol{I} - \hat{\boldsymbol{I}}||_2 + ||\boldsymbol{f} - \hat{\boldsymbol{f}}||_2 + \lambda_p \mathcal{L}_p(\hat{\boldsymbol{I}}) + \lambda_G \mathcal{L}_G(\hat{\boldsymbol{I}}), \quad (2)$$

where $\mathcal{L}_p$ is perceptual loss (LPIPS (Zhang et al. 2018)), $\mathcal{L}_G$ is discriminative loss (StyleGAN (Karras, Laine, and Aila 2019)), and λ_p and λ_G are the corresponding loss weights.

Mumford-Shah Functional

The Mumford-Shah functional (Mumford and Shah 1989) provides a form of all regularizers aiming at discontinuity-preserving smoothing given a bounded set $\Omega \in \mathbb{R}^d$,

$$\min_{u,K} \int_\Omega |u - f|^2 dx + \alpha \int_{\Omega/K} |\nabla u|^2 dx + \lambda |K|, \quad (3)$$

This approximates vector-valued input image $f : \Omega \to \mathbb{R}^k$ with function $u : \Omega \to \mathbb{R}^k$, which is smooth everywhere except at $(d-1)$-dimensional jump set K. $\lambda > 0$ controls the length of K. A common approach to solve the Mumford-Shah functional is the Ambrosio-Tortorelli approach (1990),

$$\begin{aligned} \min_{u,s} \int_\Omega |u - f|^2 dx + \alpha \int_\Omega (1-s)^2 |\nabla u|^2 dx \\ + \lambda \int_\Omega (\epsilon |\nabla s|^2 + \frac{1}{4\epsilon} s^2) dx, \end{aligned} \quad (4)$$

with a small parameter $\epsilon > 0$ and an edge set indicator $s : \Omega \to \mathbb{R}$. The points $x \in \Omega$ are part of the edge set K if $s(x) \approx 1$ and part of smooth region if $s(x) \approx 0$. The variables u and s are found by alternating minimization.

Proposed Approach

We propose a novel approach for autoregressive image generation where the model initially generates a base image focusing on global structure, and subsequently refines it through iterative detail addition. Our training methodology comprises three steps: **(1) Decomposition**: Each training image is decomposed into n hierarchical factors representing progressive detail layers, **(2) Encoding and Tokenization**: The factors are encoded into a latent space using a VQ-VAE, preserving essential features while reducing dimensionality, **(3) Iterative Prediction**: A Transformer decoder is trained to predict successive detail factors (token-maps), enabling incremental detail addition.

Hierarchical Base-Detail Decomposition

An image can be represented as a linear combination of factor images capturing distinct properties of the image. We decompose an image into a base and a detail factor,

$$\boldsymbol{I} = \boldsymbol{B} + \boldsymbol{D}, \quad (5)$$

where $\boldsymbol{I}, \boldsymbol{B}, \boldsymbol{D} \in \mathbb{R}^{H \times W \times 3}$ denote a training image and its corresponding base and detail factors. The base factor $\boldsymbol{B}$ is obtained by minimizing the Mumford-Shah functional via the Ambrosio-Tortorelli approach, as detailed in Eq. 4. This base factor can be recursively decomposed to yield multiple detail factors,

$$\boldsymbol{I} = \boldsymbol{B}_n + \boldsymbol{D}_n + \boldsymbol{D}_{n-1} + ... + \boldsymbol{D}_1, \quad (6)$$

where, $\boldsymbol{B}_{k-1} = \boldsymbol{B}_k + \boldsymbol{D}_k$, $\forall k \in \{1, ..., n\}$. Equation 6 defines the n^{th} order decomposition of $\boldsymbol{I}$. In this decomposition, the base factor $\boldsymbol{B}_n$ captures the image's overall structure, composition, and global features, while the detail factors $\{\boldsymbol{D}_k\}_{k=1}^n$ represent local features that contribute to the finer details of the image. Figure 4 shows the hierarchical base-detail decomposition process.

We adopt edge-aware smoothing over frequency-based decomposition methods to preserve structural integrity in base images. While frequency-domain approaches such as Discrete Cosine Transform (DCT) (Nash and et al. 2021) and Wavelet Transforms (Yu et al. 2021b) provide computational efficiency, they exhibit fundamental limitations

Figure 3: Generated Samples from CART-256

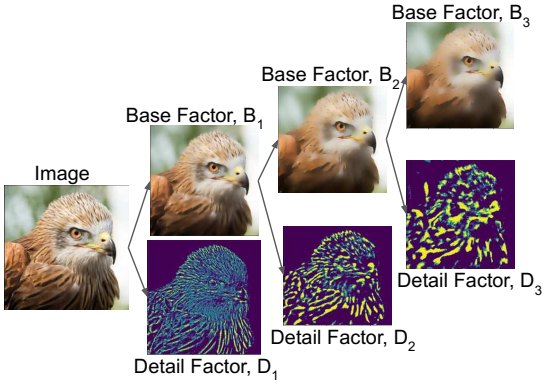


Figure 4: Hierarchical Base-Detail Decomposition

for our compositional framework. DCT-based decomposition applies uniform smoothing across both global structures and local features, failing to distinguish between semantically important edges and fine-grained textures. Additionally, the inverse DCT/Wavelet transformation introduces ringing artifacts that compromise image quality in the reconstructed base component. In contrast, Mumford-Shah smoothing provides selective regularization that preserves global edges and large-scale structural elements while effectively smoothing textural and local features. This edge-preserving property enables successful disentanglement of structural information (captured in base factors) from fine-grained details (captured in detail factors), which is critical for our iterative refinement approach (Further discussion in supplement). Our framework maintains flexibility by supporting various image decomposition techniques within the general formulation of Equation 6. This modularity allows for domain-specific decomposition strategies while preserving the core auto-regressive generation mechanism.

Encoding and Tokenization

In our approach, each image is represented by token maps $\{r_1, r_2, ..., r_{\mathcal{M}}\}$ within the latent space of a Vector Quantized Variational AutoEncoder (VQ-VAE), rather than single tokens. This token-map representation preserves the spatial coherence of the feature map and reinforces the spatial structure inherent in the image. Departing from the multi-scale approach in VAR (Tian et al. 2024), we propose a tokenization scheme such that these token maps represent the base and detail factors. Specifically, the image representation is comprised of $\mathcal{B}$ base token maps, $(r_1, ..., r_{\mathcal{B}})$, where $\mathcal{B} < \mathcal{M}$ and $(\mathcal{M} - \mathcal{B})$ detail token maps, $(r_{\mathcal{B}+1}, ..., r_{\mathcal{M}})$.

Following the Base-Detail Decomposition, we encode the original image I along with the Base Factors $\{\boldsymbol{B}_k\}_{k=1}^{n}$ using a VAE,

$$\boldsymbol{f}_{\boldsymbol{B}_k} = \boldsymbol{\mathcal{E}}(\boldsymbol{B}_k), \tag{7}$$

where $\boldsymbol{f}_{\boldsymbol{B}_k} \in \mathbb{R}^{h \times w \times c}\ \forall k \in \{1, ..., n\}$. The token maps representing the base factor $\boldsymbol{B}_n$, $\{r_1, ..., r_{\mathcal{B}}\}$ are created by performing residual quantization (Lee et al. 2022) on the encoded feature map $\boldsymbol{f}_{\boldsymbol{B}_n}$ with a quantization depth of $\mathcal{B}$. The encoded representation of the k^{th} detail factor is then determined as follows,

$$\boldsymbol{f}_{\boldsymbol{D}_k} = \boldsymbol{f}_{\boldsymbol{B}_{k-1}} - \boldsymbol{f}_{\boldsymbol{B}_k}, \tag{8}$$

where, $\boldsymbol{f}_{\boldsymbol{B}_k}$ is the encoded representation of the k^{th} base factor and $\boldsymbol{f}_{\boldsymbol{B}_0} = \boldsymbol{f}_{\boldsymbol{I}}$. Each detail factor is quantized with quantization depth $\frac{(\mathcal{M}-\mathcal{B})}{n}$, yielding the remaining tokens, as illustrated in Figure 2(a). The full algorithm for extracting token maps from a given image is presented in Algorithm 1.

Iterative Detail Learning

We employ an auto-regressive approach to predict each successive "next-detail" token map. Given the set of tokens $\{r_1, r_2, ..., r_{\mathcal{M}}\}$, the autoregressive likelihood is defined as,

$$P(r_1, ..., r_{\mathcal{M}}) = \prod_{m=1}^{\mathcal{M}} P(r_m | r_1, ..., r_{m-1}). \tag{9}$$

where each autoregressive unit, $r_m \in [V]^{h_m \times w_m}$ is a token map containing $h_m \times w_m$ tokens.

Figure 5: Comparison of samples generated by VAR-256 (top) and CART-256 (bottom).

Algorithm 1: Base-Detail VQ-VAE Encoding

Input:
- Raw image, I
- Target Image dimensions, $h_{\mathcal{M}}, w_{\mathcal{M}}$
- Base Image, B_n

Hyperparameters:
- Total number of tokens, $\mathcal{M}$
- number of base tokens, $\mathcal{B}$
- number of detail factors, n

begin
- $f_I \leftarrow \varepsilon(I)$
- $f_B \leftarrow \varepsilon(B)$
- $f_{D_i} \leftarrow f_{B_{i-1}} - f_{B_i}\ \forall i \in \{1, ..., n\}$
- $t \leftarrow -1$
- **for** *k=1:*$\mathcal{M}$ **do**
  - **if** $k \leq \mathcal{B}$ **then**
    - $r_k \leftarrow \mathcal{Q}(\text{interpolate}(f_B, h_k, w_k))$
    - $R \leftarrow queue_{push}(R, r_k)$
    - $z_k \leftarrow \text{LookUp}(r_k, \mathcal{Z})$
    - $z_k \leftarrow \text{interpolate}(z_k, h_{\mathcal{M}}, w_{\mathcal{M}})$
    - $f_B \leftarrow f_B - \phi_k(z_k)$
  - **end**
  - **else**
    - **if** *mod(k,* $\frac{\mathcal{M}-\mathcal{B}}{n}$*)* $= 0$ **then**
      - $t \leftarrow t + 1$
    - **end**
    - $r_k \leftarrow \mathcal{Q}(f_{D_t})$
    - $R \leftarrow queue_{push}(R, r_k)$
    - $z_k \leftarrow \text{LookUp}(r_k, \mathcal{Z})$
    - $f_{D_t} \leftarrow f_{D_t} - \phi_k(z_k)$
  - **end**
- **end**
- **return** base-detail tokens R.

end

For the model architecture, we utilize a standard decoder-only Transformer architecture similar to that in GPT-2 (Radford et al. 2019), VQ-GAN (Esser et al. 2021), and VAR (Tian et al. 2024). At each auto-regressive step, the Transformer decoder predicts the distribution over all $h_m \times w_m$ tokens in parallel as depicted in Figure 2(b). To enforce causality, we apply a causal attention mask, ensuring that each token map r_m only attends to its preceding tokens $r_{\leq m}$.

Algorithm 2: Base-Detail VQ-VAE Reconstruction

Input: Base-Detail Tokens, R.
Hyperparameters:
- Total number of tokens to represent the image, $\mathcal{M}$
- number of base tokens, $\mathcal{B}$
- number of detail factors, n

begin
- $\hat{f} \leftarrow 0$
- **for** *k=1:*$\mathcal{M}$ **do**
  - **if** $k \leq \mathcal{B}$ **then**
    - $r_k \leftarrow queue_{\text{pop}}(R)\ z_k \leftarrow lookup(Z, r_k)$
    - $z_k \leftarrow \text{interpolate}(z_k, h_k, w_k)$
    - $\hat{f} \leftarrow \hat{f} + \phi_k(z_k)$
  - **end**
  - **else**
    - $r_k \leftarrow queue_{\text{pop}}(R)\ z_k \leftarrow lookup(Z, r_k)$
    - $\hat{f} \leftarrow \hat{f} + \phi_k(z_k)$
  - **end**
- **end**
- $\hat{I} \leftarrow \mathcal{D}(\hat{f})$
- **return** reconstructed image $\hat{I}$

end

Experiments

Implementation Details

For detail decomposition of training images, Mumford-Shah smoothing (Equation 4) with $\alpha = 1$, $\lambda = 0.01$ is used. Each training image is decomposed iteratively to obtain a 3^{rd} order decomposition, $\boldsymbol{I} = \boldsymbol{B}_3 + \boldsymbol{D}_3 + \boldsymbol{D}_2 + \boldsymbol{D}_1$. Note that

Figure 6: High Resolution image generation using patchwise detail prediction using CART-d30-256 and CART-d30-512. Zoom-in recommended to observe finer details.

the computational overhead due to Mumford-Shah decomposition is a one-time cost, as the decomposition is only utilized during the training process and not during inference. A Vanilla VQ-VAE (Van Den Oord, Vinyals et al. 2017) is used along with $\mathcal{M}$ extra convolutions to realize the Base-Detail quantization scheme as depicted in Figure 2(a) and Algorithm 1. To mitigate information loss when upscaling z_k to the target resolution $h_{\mathcal{M}} \times w_{\mathcal{M}}$, we introduce an additional set of $\mathcal{M}$ convolutional layers, denoted as $\{\phi_k\}_{k=1}^{\mathcal{M}}$ which enhance feature refinement and preserve structural details. The base and detail factors share the same code book with $V = 4096$. As in (Tian et al. 2024; Esser et al. 2021), the tokenizer is trained on OpenImages (Kuznetsova et al. 2020) with Compound loss (Equation 2) and spatial downsample of $16\times$.

The tokenized base-detail factors are then utilized to train a Transformer Decoder architecture which learns to predict the "next-detail" token. A standard decoder-only transformer architecture is used similar to GPT-2 (Radford et al. 2019) and VQGAN (Esser et al. 2021). We use a total of 14 steps to generate an image, including 8 steps to generate the base factor and 6 steps to generate the detail factors. During inference, the Transformer predicts the codes and the VQ-VAE decoder decodes the generated image. The decoding process is summarized in Algorithm 2. The depth of the transformer is varied from 16 to 30 to obtain models with varying complexity. The model is trained with initial learning rate of $1e^{-4}$. For training, we use 16 A100 GPUs with a global batch size of 768 for CART-d30-256 and batch size of 384 for CART-d30-512. All visual results are generated with seed 42 and quantitative results are averaged over 10 randomly selected seeds.

Empirical Results

The proposed CART model was evaluated on ImageNet (Deng et al. 2009) at 256×256 (CART-256) and 512×512 (CART-512) resolutions for benchmarking against SOTA image generation methods. Comparative results in Tables 1 and 2 show that CART outperforms SOTA AR and Diffusion models and achieves FID lower than ImageNet validation set while maintaining comparable complexity and generation steps. CART benefits from base-detail decomposition that disentangles global structures from local details, simplifying the learning process through more natural token ordering. Figure 3 depicts generated images using our method, while Figure 5 compares VAR (Tian et al. 2024) and CART outputs. CART produces images with enhanced details and structure compared to VAR's "next-scale" prediction scheme. CART surpasses both Diffusion Transformer (Peebles and Xie 2023; Zhang 2024; Hatamizadeh et al. 2024) and SOTA VAR (Tian et al. 2024) in autoregressive image generation. Comprehensive results are provided in the extended version.

Other Applications

Generalizing to Higher Resolutions A key advantage of employing base-detail decomposition is the explicit disentanglement of global and local image features, facilitating high-resolution image synthesis and image super-resolution even when trained on lower-resolution inputs. Empirically, we observe that the base factor encapsulates global attributes, including class-conditional structure and overall color composition, while the detail factor captures local features such as textures and fine-grained details (see Figure 10). This decomposition allows the base factor to be upscaled without loss of essential global information, while the detail factor is generated in a patchwise manner. Since the detail factor inherently lacks dependencies on global structures, patchwise synthesis does not introduce any discontinuities. Figure 6 compares bilinear upscaling and VAR (Tian et al. 2024) with our method, demonstrating that reusing lower-resolution base images and introducing patchwise details at target resolution effectively preserves content while enhancing fine details. Table 2 presents performance comparisons against state-of-the-art methods. "CART-256-NOV" refers to non-overlapping patchwise detail generation at 512×512, while "CART-256-OV" employs 50% overlapping patches for improved continuity. "CART-512" corre-

Type	Model	FID ↓	IS ↑	Params	Steps
GAN	BigGAN (Brock 2018)	6.95	224.5	112M	1
GAN	GigaGAN (Kang et al. 2023)	3.45	225.5	569M	1
GAN	StyleGAN-XL (Sauer et al. 2022)	2.30	265.1	166M	1
Diffusion	ADM (Dhariwal and Nichol 2021)	10.94	101.0	554M	250
Diffusion	CDM (Ho et al. 2022)	4.88	158.7	-	8100
Diffusion	LDM-4-G (Rombach et al. 2022)	3.60	247.7	400M	250
Diffusion	DiT-XL/2 (Peebles and Xie 2023)	2.27	278.2	675M	250
Diffusion	L-DiT-3B (Zhang 2024)	2.10	304.4	3.0B	250
Diffusion	L-DiT-7B (Zhang 2024)	2.28	316.2	7.0B	250
Diffusion	DiffiT (Hatamizadeh et al. 2024)	1.73	276.5	561M	250
Mask	MaskGIT (Chang et al. 2022)	6.18	182.1	227M	8
Mask	RCG (Li, Katabi, and He 2023)	3.49	215.5	502M	20
AR	VQVAE-2 (Razavi et al. 2019)	31.11	-	13.5B	5120
AR	DCTransformer (Nash and et al. 2021)	36.51	-	738M	-
AR	VQGAN-re (Esser et al. 2021)	5.20	280.3	1.4B	256
AR	ViTVQ-re (Yu et al. 2021a)	3.04	227.4	1.7B	1024
AR	RQTran-re (Lee et al. 2022)	3.80	323.7	3.8B	68
AR	LlamaGen (Sun et al. 2024)	2.18	263.3	3.1B	576
AR	SpectralAR-d24 (Huang et al. 2025)	2.13	307.7	1.0B	64
VAR	VAR-d16 (Tian et al. 2024)	3.30	274.4	310M	10
VAR	VAR-d30 (Tian et al. 2024)	1.92	323.1	2B	10
VAR	VAR-d30-re (Tian et al. 2024)	1.73	350.2	2B	10
VAR	VAR-d30-re (Tian et al. 2024)	1.70	352.8	2B	14
CART	CART-d16	2.89	293.0	310M	14
CART	CART-d24	1.90	328.1	1.0B	14
CART	CART-d24-re	1.77	345.7	1.0B	14
CART	CART-d30	1.65	366.8	2.0B	14
CART	CART-d30-re	1.61	377.5	2.0B	10
CART	CART-d30-re	**1.57**	**381.9**	2.0B	14
	(val. data)	1.78	236.9		

Table 1: Quantitative results on ImageNet 256×256. Suffix '-re' refers to models that use rejection sampling

Figure 7: Comparison of CART for super-resolution with ResShift. Zoom-in recommended to observe finer details.

(a) Recoloring of generated images by manipulating albedo (b) Recoloring of "bell-pepper" and "tent" classes

Figure 8: Recoloring of generated images when using intrinsic decomposition to tokenize the image.

Type	Model	FID ↓	IS ↑
GAN	BigGAN (2018)	8.43	177.9
Diff.	ADM (2021)	23.24	101.0
Diff.	DiT-XL/2 (2023)	3.04	240.8
Mask	MaskGiT (2022)	7.32	156.0
AR	VQGAN (2021)	26.52	66.8
VAR	VAR-d36-s (2024)	2.63	303.2
CART	CART-NOV-256-d30	2.85	297.1
CART	CART-OV-256-d30	2.54	305.7
CART	CART-512-d30	**2.40**	**315.5**

Table 2: Quantitative results on ImageNet 512×512.

Model	PSNR ↑	SSIM ↑	CLIPIQA ↑
Real-ESRGAN (2021)	24.04	0.665	0.523
ResShift-15 (2024)	**24.90**	**0.673**	0.603
Sin-SR (2024)	24.56	0.657	0.611
CART-256-d30	24.16	0.633	0.594
CART-512-d30	24.65	0.660	**0.672**

Table 3: Comparison of CART models with specialized Super-Resolution models. Metrics are reported for SR from 128×128 to 512×512 resolution on ImageNet Test Set.

sponds to full training on 512×512 images. Notably, CART (trained at 256×256) outperforms VAR models trained from scratch at 512×512. Further details and results are provided in the extended version.

Super-Resolution (SR) Given a Low-Resolution (LR) image, we encode it using Base-Detail VQVAE token maps $\{r_1, ..., r_k\}$, where r_k has resolution $h/p \times w/p$ (h, w are LR image dimensions, p is the down-sampling factor of VQ-VAE). $\{r_1, ..., r_k\}$ are appended as past tokens to CART's prediction sequence, which then predicts subsequent token maps $\{r_{k+1}, ..., r_{\mathcal{M}}\}$ in unconditional generation setting. For target resolutions exceeding training resolution, we apply the high-resolution generation strategy described above. Table 3 compares our SR results with SOTA generative methods specifically trained or fine-tuned for super-resolution tasks and Figure 7 provides visual comparison. Although CART yields lower PSNR than ResShift (2024), it surpasses all competing methods in CLIP-IQA score (Wang, Chan, and Loy 2023), indicating superior perceptual image quality as assessed by human visual preference. Comprehensive results are provided in the extended version.

Recoloring Generated Images Intrinsic image decomposition (Careaga and Aksoy 2023) provides a principled approach to disentangle images into reflectance (albedo) and illumination (shading) components, enabling semantically meaningful manipulations. CART adopts this decomposition during training to encourage controllable generation. The observed image $\boldsymbol{I}$ is modeled as the composition of albedo and shading map, $\boldsymbol{I} = \boldsymbol{A} \star \boldsymbol{S}$. Where $\boldsymbol{A} \in \mathbb{R}^{H \times W \times 3}$ encodes illumination-invariant properties (object color and structure) and $\boldsymbol{S} \in \mathbb{R}^{H \times W \times 1}$ captures illumination-dependent effects. To facilitate learning and component-wise manipulation, we convert the multiplicative decomposition to additive form via logarithmic transformation, $\log \boldsymbol{I} = \log \boldsymbol{A} + \log \boldsymbol{S}$. CART leverages this formulation by learning to predict the log-image and reconstructing via exponentiation, $\boldsymbol{I} = \exp(\log \boldsymbol{A} + \log \boldsymbol{S})$. Images are tokenized into 14 steps comprising 7 albedo and 7 shading token maps. This layered approach enables explicit learning of color and lighting factors. The decomposition and separate supervision apply only during training to induce generative factor separation. At inference, CART directly generates compositional outputs without explicit decomposition. By structurally separating these factors during training, CART supports controllable color and illumination in generated images while compositional constraints ensure globally coherent synthesis. Figure 8(a) depicts the process of image generation and color manipulation using this decomposition. Figure 8(b) depicts more instances of recoloring.

Lighting Control of Generated Images Replacing base-detail decomposition in Equation 5 with Specularity decomposition (Saini and Narayanan 2024) enables explicit lighting control in generated images. Following the dichromatic reflection model (Tominaga 1994), images consist of diffuse $(\boldsymbol{A})$ and specular $(\boldsymbol{E})$ components: $\boldsymbol{I} = \boldsymbol{A} + \boldsymbol{E}$. We employ 4^{th}-order decomposition for four illumination control levels: $\boldsymbol{I} = \boldsymbol{A}_4 + \boldsymbol{E}_4 + \boldsymbol{E}_3 + \boldsymbol{E}_2 + \boldsymbol{E}_1$. Generation uses 16

Figure 9: Lighting control via specularity decomposition: first column shows the diffuse base; remaining columns add specular terms to vary global illumination.

autoregressive steps: 8 for base factor generation and 8 for controlled lighting refinement. Figure 9 demonstrates synthesized images with varying illumination while maintaining structural consistency for classes "cliff" and "volcano". Comprehensive results are provided in the extended version.

Ablation Study

Table 4 evaluates the impact of various CART model components. While employing multi-scale tokenization for the base factor yields only marginal gains in FID, this approach significantly reduces memory usage and accelerates generation, offering practical advantages for larger models. Table 5 compares CART performance across different decomposition orders. Decomposition order 0 is equivalent to VAR. Best performance occurs with 3^{rd} order Base-Detail decomposition. Beyond 3^{rd} order, the base image becomes oversmoothed and loses essential global structural details, leading to sub-optimal learning.

Figure 10 compares intermediate generations and self-attention maps for CART and VAR, the latter operating with multi-scale tokenization. VAR jointly refines global structure and local texture at each step, yielding entangled representations that hinder factor-wise control and scaling across resolutions. In contrast, CART first synthesizes a piecewise-smooth base capturing global structure, then incrementally adds detail factors, leading to an explicit hierarchy from structure to texture. This separation improves high-resolution synthesis via base upscaling with patch-wise detail prediction and enhances adaptability to target resolutions unseen during training. The tokenization order aligns with human perceptual organization, prioritizing coarse structures before fine details, and is reflected in progressively localized attention patterns in later steps.

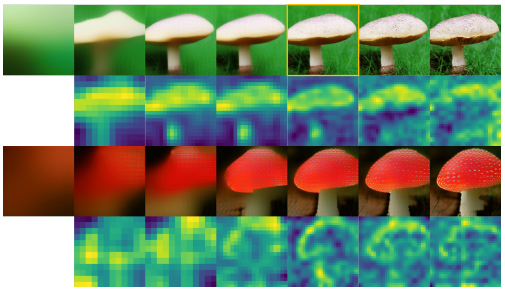


Figure 10: Top row: Intermediate CART results (base image in yellow). 2^{nd} row: CART self-attention maps. 3^{rd}: Intermediate VAR results. Bottom row: VAR self-attention maps.

Model	CFG	MS Tokens	BD Tokens	FID
AR (2021)	✗	✗	✗	18.65
VAR-d16 (2024)	✓	✓	✗	3.30
CART-d16	✓	✗	✓	2.90
CART-d16	✓	✓	✓	2.89
VAR-d30 (2024)	✓	✓	✗	1.70
CART-d30	✓	✓	✓	1.57

Table 4: Ablation Study of CART

Decomposition Order	FID
0 (Special case of VAR)	1.70
1	1.65
2	1.62
3	1.57
4	1.60

Table 5: Impact of decomposition order on CART model.

Figure 11 shows that the Base–Detail VQ-VAE achieves noticeably lower reconstruction error than Vanilla and MS-VQ-VAE. Unlike the baselines, whose MSE grows with quantization depth, the base–detail scheme monotonically reduces error, indicating more effective residual allocation and higher fidelity.

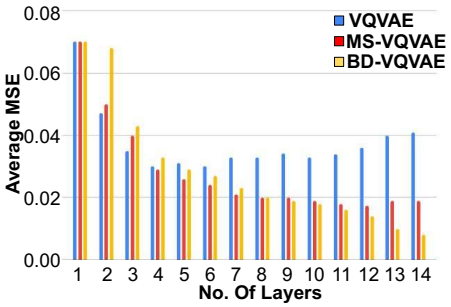


Figure 11: Reconstruction MSE of Vanilla (blue), Multiscale (red), and Base-Detail (yellow) VQ-VAE.

Conclusion

In this paper, we presented a novel auto-regressive framework with next-detail prediction and structured base-detail decomposition, enabling efficient, high-resolution image synthesis through iterative refinement. Our tokenization strategy of separately quantizing base and detail layers, preserves spatial integrity and enhances AR efficiency. Experiments show SOTA image generation and trainig-free extension to editing applications, surpassing limitations of next-token and next-scale approaches for accuracy and efficiency.

References

Ambrosio, L.; and Tortorelli, V. M. 1990. Approximation of functional depending on jumps by elliptic functional via t-convergence. *Communications on Pure and Applied Mathematics*, 43(8): 999–1036.

Anciukevičius, T.; Xu, Z.; Fisher, M.; Henderson, P.; Bilen, H.; Mitra, N. J.; and Guerrero, P. 2023. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12608–12618.

Bala, A.; Jaiswal, R.; Rashid, L.; and Roheda, S. 2024. GalaxyEdit: Large-Scale Image Editing Dataset with Enhanced Diffusion Adapter. *arXiv preprint arXiv:2411.13794*.

Brock, A. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Careaga, C.; and Aksoy, Y. 2023. Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics*, 43(1): 1–24.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11315–11325.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020. Generative pretraining from pixels. In *International conference on machine learning*, 1691–1703. PMLR.

Corneanu, C.; Gadde, R.; and Martinez, A. M. 2024. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4334–4343.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Esser, P.; Rombach, R.; Ommer, B.; and et al. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D.; and Wierstra, D. 2015. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, 1462–1471. PMLR.

Hatamizadeh, A.; Song, J.; Liu, G.; Kautz, J.; and Vahdat, A. 2024. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, 37–55. Springer.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47): 1–33.

Huang, Y.; Chen, W.; Zheng, W.; Duan, Y.; Zhou, J.; and Lu, J. 2025. SpectralAR: Spectral Autoregressive Visual Generation. *arXiv preprint arXiv:2506.10962*.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10124–10134.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kingma, D. P. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7): 1956–1981.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11523–11532.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Li, T.; Katabi, D.; and He, K. 2023. Self-conditioned image generation via generating representations. *arXiv preprint arXiv:2312.03701*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Mirza, M. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mumford, D. B.; and Shah, J. 1989. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*.

Nash, C.; and et al. 2021. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*.

Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; and Tran, D. 2018. Image transformer. In *International conference on machine learning*, 4055–4064. PMLR.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Razavi, A.; Van den Oord, A.; Vinyals, O.; and et al. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saini, S.; and Narayanan, P. 2024. Specularity factorization for low-light enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1–12.

Salimans, T.; Karpathy, A.; Chen, X.; and Kingma, D. P. 2017. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*.

Sauer, A.; Schwarz, K.; Geiger, A.; and et al. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Shao, H.; Yao, S.; Sun, D.; Zhang, A.; Liu, S.; Liu, D.; Wang, J.; and Abdelzaher, T. 2020. Controlvae: Controllable variational autoencoder. In *International conference on machine learning*, 8655–8664. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; and Yuan, Z. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37: 84839–84865.

Tominaga, S. 1994. Dichromatic reflection models for a variety of materials. *Color Research & Application*, 19(4): 277–285.

Van Den Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. In *International conference on machine learning*, 1747–1756. PMLR.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2555–2563.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1905–1914.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2024. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25796–25805.

Yang, S.; Chen, X.; and Liao, J. 2023. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3190–3199.

Yu, J.; Li, X.; Koh, J. Y.; Zhang, H.; Pang, R.; Qin, J.; Ku, A.; Xu, Y.; Baldridge, J.; and Wu, Y. 2021a. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

Yu, Y.; Zhan, F.; Lu, S.; Pan, J.; Ma, F.; Xie, X.; and Miao, C. 2021b. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14114–14123.

Yue, Z.; Wang, J.; and Loy, C. C. 2024. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

Zhang, R. 2024. Alpha-VLLM. Large-dit-imagenet. https://github.com/Alpha-VLLM/LLaMA2-Accessory/tree/main/Large-DiT-ImageNet. Accessed: 2025-07-26.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhou, Z.; and Tulsiani, S. 2023. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12588–12597.

Zhu, Y.; Li, Z.; Wang, T.; He, M.; and Yao, C. 2023. Conditional text image generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14235–14245.