VAR-CLIP: Text-to-Image Generator with Visual Auto-Regressive Modeling

Qian Zhang[1]   Xiangzi Dai[2]   Ninghua Yang[2]
Xiang An[2]   Ziyong Feng[2]   Xingyu Ren[3]

[1] Institute of Applied Physics and Computational Mathematics
[2] DeepGlint [3] Shanghai Jiao Tong University

zhangqian18@iapcm.ac.cn, xiangzidai@deepglint.com

Figure 1. Exemplar images generated from text prompt by VAR-CLIP. We show 256×256 samples.

Abstract

VAR is a new generation paradigm that employs 'next-scale prediction' as opposed to 'next-token prediction'. This innovative transformation enables auto-regressive (AR) transformers to rapidly learn visual distributions and achieve robust generalization. However, the original VAR model is constrained to class-conditioned synthesis, relying solely on textual captions for guidance. In this paper, we introduce VAR-CLIP, a novel text-to-image model that integrates Visual Auto-Regressive techniques with the capabilities of CLIP. The VAR-CLIP framework encodes captions into text embeddings, which are then utilized as textual conditions for image generation. To facilitate training on extensive datasets, such as ImageNet, we have constructed a substantial image-text dataset leveraging BLIP2. Furthermore, we delve into the significance of word posi-
tioning within CLIP for the purpose of caption guidance. Extensive experiments confirm VAR-CLIP's proficiency in generating fantasy images with high fidelity, textual congruence, and aesthetic excellence. Our project page are https://github.com/daixiangzi/VAR-CLIP

1. Introduction

Text-to-image task (T2I), which aims to generate natural and realistic images while understanding textual captions, has been an engaging challenge in the computer vision community. T2I trains on large-scale data to identify data distribution and latent space within it. It then uses text embeddings as a condition to sample the latent distribution of images, achieving satisfying image generation. Text control [30, 43, 55, 57] and image generation [16, 23, 36] play

crucial roles in T2I tasks.

Existing text-to-image methods primarily fall into three categories: Generative Adversarial Network (GAN) [21, 22], Diffusion Model (DM) [7, 41], and Auto-Regressive model (AR) [19, 47]. GANs incorporate a discriminator to regulate image generation by measuring the disparity between real and generated images [14]. While GANs excel at simplifying models, speeding up inference, and enhancing image quality, they struggle with issues like model collapse and limited diversity. DMs gradually eliminate noise from Gaussian noise to produce diverse images [16, 44]. The attention module in DMs helps in aggregating essential visual concepts as outlined in the text [12, 34, 41, 51, 58]. DMs represent a great implementation with superior evaluation metrics due to their generation capabilities [15, 18, 30, 33, 42]. However, the diffusion process incurs high computational costs and inference time owing to the iterative diffusion steps.

Auto-Regressive models (ARs) generate images by predicting the next token from a discrete prefix [1, 6, 38]. The alignment between ARs and large language models (LLMs) offers a unique advantage for cross-modality fusion in textual and graphical generation. Traditional ARs like Pixel-RNN [49], Pixel-CNN [48], and GPT-2-like transformers [37] have made strides in image quality and inference time. Recently, models like LlamaGen [45], VAR [46] and MAR [27] have demonstrated significant prowess in image generation, surpassing Diffusion Models (DMs). VAR introduces a novel auto-regressive model paradigm, shifting from next-token prediction to next-scale prediction, thereby enhancing computational performance and image quality. However, VAR is primarily used for class-conditional image generation, raising the question of its suitability for the T2I task.

To address this question, we introduce VAR-CLIP, a text-to-image framework that leverages CLIP to guide VAR in generating images containing textual information. VAR-CLIP adopts a two-stage training approach: first training a multi-scale VQVAE/VQGAN and then using the CLIP text encoder to extract representations of text captions as conditional tokens for image generation. We also explore how the word position in CLIP influences embeddings, noting that the first 20 tokens carry more weight than the others. Overall, our main contributions are summarized as follows:

- We propose a framework VAR-CLIP for high-quality text-to-image generation with minimal inference time, which use CLIP to obtain text embedding as the condition of VAR to generate images.

- We have created a text-image pair dataset for ImageNet using BLIP-2, enabling ImageNet to support the T2I task.

- We investigate the importance of word position in CLIP. CLIP supports a maximum of 77 tokens, but they have the importance of imbalance for each tokens, the first 20 tokens without start of token and end of token contribute more to the caption.

2. Related Work

Text-to-Image Generation

Text-to-image (T2I) generation techniques can be differentiated based on their probability distribution acquisition methods, falling primarily into three categories: Generative Adversarial Networks (GANs) [3, 22], diffusion models [11], and auto-regressive models [28, 29]. GANs initiate the image generation process from stochastic noise, leveraging a discriminator to progressively mold the noise into coherent imagery. Despite their potential, GANs are frequently hampered by training instability. GigaGAN [20] has exemplified the efficacy of GANs in T2I synthesis, delivering swift inference times, high-fidelity images, and a spectrum of latent space editing capabilities.

Diffusion models have recently gained prominence in the field of image generation, particularly for text-to-image (T2I) synthesis. Unlike GANs, these models introduce noise incrementally to create a Gaussian distribution, from which high-quality and diverse images are generated through a denoising process. Variants such as SD3.0 [9], SORA [4] and DALL-E 3 [2] have demonstrated excellence in T2I tasks. However, challenges remain in improving performance, primarily due to the computational demands of multi-stage denoising and the complexity of integrating with language models.

Auto-regressive methods generate images by predicting the next token from a codebook, utilize architectures of Large Language Models (LLMs) like GPT [13], BERT [6], and LLaMA [47, 47]. The Vector Quantized Variational Autoencoder (VQVAE) [50] enables unsupervised, efficient, and interpretable image representation through discrete tokens. VQGAN [10] improves upon this with transformer integration, enhancing high-resolution image quality beyond PixelCNN's capabilities. VQVAE-2 [40] expands on this for large-scale synthesis, and Yu et al. [53] refine efficiency and accuracy by incorporating Vision Transformers (ViTs) [8]. Masked generation techniques [5, 26] have notably increased auto-regressive decoding speeds on ImageNet by 48 times.

Building on Li's observation that discrete-value spaces are unnecessary for auto-regressive models, Masked Auto-regressive Models (MARs) [27] have been developed, capitalizing on the swiftness of sequence modeling. However, MARs' 1D sequence encoding may neglect 2D image spatiality. Tian's Visual Autoregressive Modeling (VAR) [46]
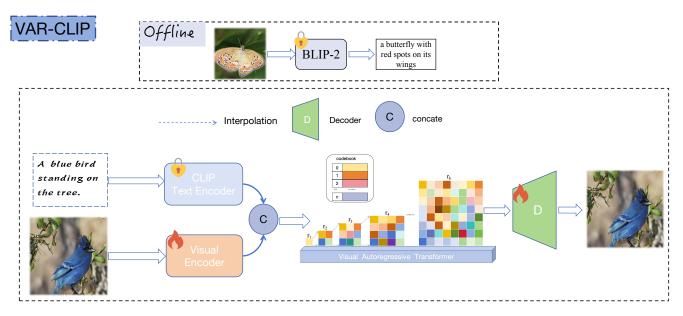
Figure 2. An illustration of VAR-CLIP. For a given text prompt and image, VAR-CLIP generates text embeddings from a pre-trained CLIP model and visual embeddings from a VAR encoder. The text embedding serves as a condition to guide the generation of multi-scale tokens and the final image. The Visual Autoregressive Transformer (VAR) generates these multi-scale tokens through next-scale prediction. During training, we utilize BLIP-2 to obtain text captions.

addresses this with a "next-scale prediction" approach, surpassing diffusion transformers in image generation quality. The integration of VAR in text-conditioned synthesis presents an opportunity for future research.

3. Approach

In this section, we introduce the details about our text-to-image generation framework and the two-stage training strategy.

In Fig. 2, our framework consists of three components: a pre-trained text encoder such as the Contrastive Language–Image model connecting text and images (CLIP), a multi-scale image tokenizer(multi-scale VQVAE) and a conditional visual autoregressive transformer (VAR). During training, a text is initially encoded as embedding $\mathbf{e}_c$ by the text encoder and as the condition start tokens $\mathbf{c}$, Subsequently, we generate multi-scale image tokens through a next-scale prediction strategy using VAR, and obtain residual design on $\hat{f}$ through multi-scale, Finally, we reconstruct the same image using the decoder with $\hat{f}$. Specifically, we obtain image captions using BLIP-2.

3.1. Pre-trained Text Encoder

Pre-trained text encoder map the text inputs into embedding space. Contrastive Language–Image Pre-training (CLIP) [35] learns visual concepts from natural language supervision, establishing a link between text and images. For any input (a text T, denoted as x), CLIP converts it into

a latent embedding $\mathbf{e}_c$ using the text encoder:

$$\mathbf{e}_c = f_{\text{CLIP}}(x), \text{where } x \in \{T\}. \tag{1}$$

The pre-trained CLIP model is trained on 400 million images and a wide variety of natural language text. By maximize the cosine similarity of the text and image embedding, CLIP learns a multi-modal embedding space.

In text-to-image generation, limited text-image training data makes it challenging to utilize large image datasets like ImageNet. CLIP helps bridge text and image by assuming they share embeddings in a latent space. In this context, the ViT-L/14 variant of CLIP [35] is employed for both training and inference.

3.2. Multi-Scale Image Tokenizer

VQVAE [31] transforms image into discrete image tokens to generate high-quality result. The next token can be predicted by its prefix from the transformer, contributing to the generation process. Different from VQVAE, multi-scale VQVAE employs a multi-scale quantization autoencoder to encode an image into K multi-scale discrete token maps for efficient and effective generation. This multi-scale approach enhances both quality and speed. The autoencoder then assists an autoregressive transformer in predicting the next-scale prediction, further improving the generation process.

The architecture used is similar to VQVAE but differs in having a multi-scale quantization layer instead of a sequence of tokens. During encoding, an image is trans-

formed into K token maps $R = (r_1, r_2, r_3, ..., r_K)$, r_k ranging from small to large, r_1 is the start tokens with 1×1 token map, and size of r_K is the $h_K \times w_K$, where $h_K \times w_K$ is the origin image size. The next-scale token depends on its prefix.

$$p(r_1, r_2, ..., r_k) = \prod_{k=1}^{K} p(r_k|r_1, r_2, ..., r_{k-1}) \quad (2)$$

All token maps share the same codebook $\mathcal{Z} \in \mathbb{R}^{V,C}$ through optimizing encoder (E) and decoder (D) parameters. Here, V represents the vocabulary, and C is the vocabulary channel.

For an image I, start by converting it into the embedding feature map $f = E(I) \in \mathbb{R}^{dim_z \times h \times w}$, and then, the image feature map f is convert to multi-scale discrete tokens $r = (r_1, r_2, ..., r_K)$, where $r \in [V]^{h \times w}$. The feature $f^{(i, j)}$ is mapped to the code index $r_q^{(i,j)}$ of the codebook $\mathcal{Z}$ based on its nearest code in terms of Euclidean distance:

$$r_q^{(i,j)} = \underset{v \in [V]}{\arg\min} \|lookup(\mathcal{Z}, v) - f^{i,j}\|_2. \quad (3)$$

where $lookup(\mathcal{Z}, v)$ means the v-th vector in the codebook $\mathcal{Z}$.

During the training of the multi-scale autoencoder, each $\mathcal{Z}$ is used to look up $r_q^{(i,j)}$ in order to acquire $z^{(i,j)}$, for an approximation of the original image I, Subsequently, a reconstructed image $\hat{I}$ is generated by the decoder (D). The multi-scale VQVAE process can be described as follows:

$$z_i = lookup(\mathcal{Z}, r_i), \quad (4)$$

$$z_i = interpolate(z_i, h_K, w_K), \quad (5)$$

$$\hat{f} = \sum_{r=1}^{K} \phi(z_r), \quad (6)$$

$$\hat{I} = D(\hat{f}) \quad (7)$$

In this context, r_i represents the tokens at the i-th scale. During reconstruction, the decoding process incorporates a residual design on $\hat{f}$ and utilizes it as input to the decoder for image reconstruction.

After training model, the multi-scale feature maps and codebooks serve as a multi-scale image tokenizer, with the decoder reconstructing an image from its multi-scale tokens. The multi-scale VQVAE model optimized using the loss function $\mathcal{L}$:

$$\mathcal{L} = \|I - \hat{I}\|_2 + \|\hat{z} - z_q\|_2 + +\lambda_p \mathcal{L}_p(\hat{I}) \quad (8)$$

where $\mathcal{L}_p(\cdot)$ represents the perceptual loss, such as LPIPS [56], and λ_p denotes the weight assigned to this loss.

3.3. Conditional visual autoregressive transformer

Given a sequence of discrete tokens $x = (x_1, x_2, ..., x_n)$, where each token $x_n \in [V]$ and V represents the integer vocabulary, the autoregressive transformer model predicts the highest probability of the next token as V. This implies that the next token x_n is dependent on the prefix $(x_1, x_2, ..., x_{n-1})$. While this assumption views image tokens as sentences in a left-to-right order, the reality is more intricate, and image tokens exhibit spatial relationships. Each token $x^{i,j}$ is associated with its 4 neighbors: $x^{i-1,j}$, $x^{i+1,j}$, $x^{i,j-1}$ and $x^{i,j+1}$.

The conditional autoregressive transformer is designed to predict image tokens based on its CLIP embedding. In the conditional visual autoregressive transformer (VAR) [46], it forecasts "next-scale tokens" instead of "next-token". By quantizing a feature map $f \in mathbb^{h \times w \times C}$ into K multi-scale token maps $(r_1, r_2, ..., r_K)$, each at increasingly higher resolutions of $h_k \times w_k$, culminating in r_K matching the original feature map's resolution of $h \times w$. we anticipate the next scale token as:

$$p(\mathbf{r}|\mathbf{c}) = \prod_k p(r_k|r_{<k}, \mathbf{e}_c) \quad (9)$$

where each autoregressive scale token $r_k \in [V]^{h_k \times w_k}$ represents the token map at scale k, with the sequence $r_{<k}$ serving as the prefix. We incorporate the CLIP image encoder $\mathbf{e}_c = f_{\text{CLIP}(x)}$ as a condition to guide the next scale token. During the k-th autoregressive step, all distributions in r_K are interdependent and will be generated in parallel.

3.4. Training Strategy and Image Caption Generation

This task employs the two-stage training strategy:

First Stage We first train a multi-scale VQVAE with the image dataset in a self-supervised manner. As mentioned in Sec. 3.2, we utilize the pre-trained multi VQVAE model from VAR, since they have not made the training code publicly available.

Second Stage The conditional visual autoregressive transformer is trained at second stage. Since we have paired input-output data (embedding→text), the same with autoregressive transformer(AR), our objective is to maximize the likelihood of the corresponding image token.

The maximum-likelihood of the token sequence is enforce with

$$\mathcal{L}_{\text{Transformer}} = \mathbb{E}_{x \sim p(x)}[-\log p(s)] \quad (10)$$

Image Caption Generation Large datasets like ImageNet [24] lack captions for each image. One approach is

training with image embeddings and inference with text embeddings [52]. However, due to the richness of image data, the generation results often lack quality and fail to meet requirements. Another method involves generating captions for training. BLIP-2 [25] bridges the gap between modalities by processing both text and images. Given an image I, BLIP-2 effectively understands images and generates a textual description T. The text caption doesn't follow a template like "a photo of"; instead, it describes the image, for example, "a butterfly with red spots on its wings."

Classifier-Free Guidance Classifier-free guidance (cfg) [17] empowers generative diffusion models to produce samples of exceptionally high fidelity. Instead of relying on the gradient direction of an image classifier for sampling, this approach integrates the score estimates from both a conditional diffusion model and an unconditional model that is trained concurrently. Inspired by techniques like those used in DALL-E 2 [39], which occasionally sets CLIP embeddings to zero (or utilizes a learned embedding) and randomly omits the text caption during training, we have adapted a similar classifier-free guidance strategy. In our training of VAR-CLIP, we introduce randomness by occasionally replacing CLIP embeddings with Gaussian noise. This methodology has been shown to markedly improve visual quality. The inference cfg is represented as follow:

$$e_c = (1 + t)e_c - te_n \tag{11}$$

where e_c is the text conditional embeddings, t is the weight, and e_n is the Gaussian noise embedding.

4. Experiment

This section describes how we evaluate our method and compare with previous approaches. First, we introduce the dataset we used and the implementation details of our approach.

4.1. Dataset and Implementation Details

Dataset We train and evaluate our method exclusively on the ImageNet dataset [24]. ImageNet 256×256 is utilized for evaluating the conditional generation task, with 1.2 million images across 1000 classes. The entire dataset is employed to train VAR-CLIP. Captions are generated by BLIP-2, with evaluations conducted using captions styled similarly to BLIP-2, such as "a train passing through a field with a steam geyser."

Implementation Details We utilize a ViT-L/14 [35] variant of CLIP as a text encoder to map to an embedding space. It supports 77 tokens and has a latent channel width of 768. Our method is based on VAR, trained by Tian [46], with a configuration featuring $d = 16$ and 310 million parameters.

Following a GPT-2 [37] style transformer, we implement adaptive normalization [32]. To incorporate classifier-free guidance [17], we substitute 10% of text embeddings with noise. During training, this model is trained with a learning rate of 10^{-4}, $\beta_1 = 0.95$, $\beta_2 = 0.95$, and a decay rate of 0.05, consistent with the VAR parameters. We trained our network across 48 A100(80G) GPU machines, with a batch size of 768 per machine, over 1000 epochs, which consumed 4.151 days.

4.2. Result

Fig. 3 presents qualitative outcomes, demonstrating VAR-CLIP's capability to generate images from a diverse array of textual prompts, encompassing flora, fauna, and architectural structures, as well as landscapes. VAR-CLIP adeptly renders images that align with the semantic content of the text. For instance, given the caption "a tower in front of a mountain", VAR-CLIP accurately produces an image featuring a tower with a mountain backdrop. Moreover, VAR-CLIP harnesses the text's descriptive power to evoke specific times of day. As depicted in Figure Fig. 3, captions such as "a hot air balloon flying in the morning" and "a sailboat with a sail at sunset" elicit images that not only capture the essence of morning and dusk but also reflect the corresponding lighting conditions—a bright, hopeful sunrise and the soft, warm hues of a sunset. In terms of image quality, VAR-CLIP delivers high-fidelity results, offering detailed visualizations that bring the fantastical elements of the text to life. The model's proficiency in interpreting and visualizing textual nuances is evident in the rich details and clarity of the generated images.

However, our VAR-CLIP can produce failure images with noticeable artifacts in Fig. 4. This limitation is commonly observed by VAR (poor representation of animals' eyes, incomplete chairs), text-to-image tasks (fish may lose body parts when the number exceeds one), and academic datasets (e.g., ImageNet). The issue of poor representation of animals' eyes can be mitigated by using deeper networks. When $d = 30$, this problem can be controlled to generate better results. Research-driven models trained on ImageNet still exhibit significant differences in visual quality compared to commercial models trained on extensive data. Utilizing clean and information-rich datasets may lead to improved outcomes.

We explore the influence of different positions in the caption. CLIP supports 77 tokens, which include a start token and an end token. We assign the same embeddings to investigate each token's contribution to the caption. CLIP has 12 layers, and we display the token scores for each layer. In Fig. 5a, layer 11 indicates that the initial 25 tokens have a high score contributing to text embeddings. Long-CLIP [54] increases the maximum input length of CLIP from 77 to 248 tokens. Fig. 5b shows the token scores

a tower in front of a mountain

a big butterfly in the purple flower

a red mushroom on the grass

a spider with yellow and black stripes on its back

a valley with a river and a mountain

a hot air balloon flying in the morning

a bridge on the river

a bedroom with a bed and a lamp

a sailboat with a sail at sunset

a colorful starfish on the sea

Figure 3. Generate samples based on ten text captions trained on the ImageNet dataset, resembling those generated from BLIP-2.

Figure 4. Failure cases. Our method can produce noticeable artifacts in the image.

(a) Clip text position score
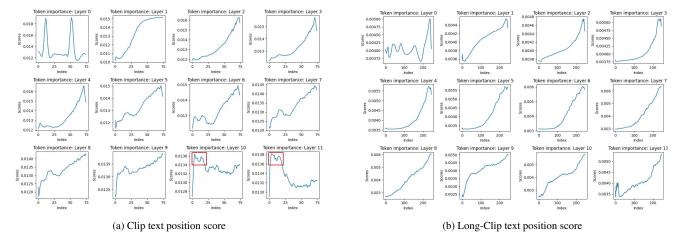
(b) Long-Clip text position score

Figure 5. Clip position score. The different positions in a sentence have varying impacts on the weight of the sentence.

for the 12 layers. The first 75 tokens exhibit similar trends, while the remaining tokens show an increase. Based on this, we suggest that more attention should be paid to the initial 20 tokens in the text caption.

5. Conclusion

In this paper, we introduce VAR-CLIP, an innovative model for text-to-image (T2I) generation. To support this framework, we have created an extensive image-text dataset using BLIP2, enhancing ImageNet's capability to facilitate T2I tasks. Furthermore, we delve into the significance of word positioning in CLIP for image generation, demonstrating VAR-CLIP's ability to produce fantasy images characterized by high fidelity, textual congruence, and aesthetic excellence.

Despite these advancements, VAR-CLIP faces certain

limitations, such as the precision of captioning and the alignment of the text encoder model with the image generation process. To enhance captioning, we propose the adoption of a more sophisticated image description model capable of producing more detailed captions. Addressing text-image alignment in auto-regressive models requires a dual-pronged approach: utilizing advanced language models for text comprehension and conducting focused research akin to that in diffusion models to improve alignment.

Looking ahead, we aim to tackle the limitations identified. Our future work will focus on generating high-quality captions and aligning complex textual and visual elements, including color, spatial arrangement, and associated objects, to further refine the T2I synthesis process.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, pages 11315–11325, 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021.

[11] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[12] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *CVPR*, 2024.

[13] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[19] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[20] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[26] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *CVPR*, pages 2142–2152, 2023.

[27] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

[28] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024.

[29] Xiaoxiao Ma, Mohan Zhou, Tao Liang, Yalong Bai, Tiejun Zhao, Huaian Chen, and Yi Jin. Star: Scale-wise text-to-image generation via auto-regressive representations. *arXiv preprint arXiv:2406.10797*, 2024.

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[31] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019.

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[34] Leigang Qu, Wenjie Wang, Yongqi Li, Hanwang Zhang, Liqiang Nie, and Tat-Seng Chua. Discriminative probing and tuning for text-to-image generation. In *CVPR*, pages 7434–7444, 2024.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[36] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[37] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[40] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[42] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[45] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[46] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

[47] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[48] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[49] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016.

[50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[51] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, pages 6232–6242, 2024.

[52] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*, 2022.

[53] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

[54] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023.

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[57] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023.

[58] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *CVPR*, pages 6818–6828, 2024.