Computational Visual Media
https://doi.org/10.26599/CVM.2025.9450377 Vol. 11, No. 2, April 2025, 405–422

Research Article

Text to image generation with bidirectional Multiway Transformers

Hangbo Bao[1], Li Dong[2], Songhao Piao[1] (✉), and Furu Wei[2]

Abstract In this study, we explore the potential of Multiway Transformers for text-to-image generation to achieve performance improvements through a concise and efficient decoupled model design and the inference efficiency provided by bidirectional encoding. We propose a method for improving the image tokenizer using pretrained Vision Transformers. Next, we employ bidirectional Multiway Transformers to restore the masked visual tokens combined with the unmasked text tokens. On the MS-COCO benchmark, our Multiway Transformers outperform vanilla Transformers, achieving superior FID scores and confirming the efficacy of the modality-specific parameter computation design. Ablation studies reveal that the fusion of visual and text tokens in bidirectional encoding contributes to improved model performance. Additionally, our proposed tokenizer outperforms VQGAN in image reconstruction quality and enhances the text-to-image generation results. By incorporating the additional CC-3M dataset for intermediate finetuning on our model with 688M parameters, we achieve competitive results with a finetuned FID score of 4.98 on MS-COCO.

Keywords text to image generation; VQ-VAE; Transformer; generative models

1 Introduction

In recent years, there have been many attempts and remarkable progress in the text-to-image generation problem, such as DALL-E [1], CogView [2], Parti [3], and Muse [4]. These studies differ from GAN [5–7] and diffusion-based [8, 9] approaches by utilizing image tokenizers to transform input images into discrete visual tokens and then using Transformers [10] to perform text-to-image modeling. The generated visual tokens can be converted back into images using an image tokenizer.

Based on the model encoding and inference methods employed in these approaches, the models can be divided into two distinct types: autoregressive and non-autoregressive. The autoregressive methods [1–3] typically involve a unidirectional decoder that sequentially decodes tokens one-by-one during the inference stage. Non-autoregressive methods [4, 11, 12] employ bidirectional encoding, which allows for the simultaneous prediction of multiple tokens, thereby reducing the number of iterations required to generate an image. Compared to autoregressive encoding methods such as decoder-based [1, 2] and encoder–decoder approaches [3], these non-autoregressive methods [4, 11, 12] which adopt a Transformer encoder for bidirectional encoding, demonstrate a more comprehensive information fusion process. In particular, unidirectional encoding can only attend to previously generated tokens in the multi-head attention mechanism of the Transformer model, while bidirectional encoding enables fusion between any two tokens. Similarly, BERT [13] highlighted the importance of bidirectional encoding in natural language understanding tasks.

These non-autoregressive methods [4, 11, 12] all opt for employing a shared vanilla Transformer [10] encoder to concurrently process visual and text tokens, rather than using a decoupled design such as encoder–decoder-based approaches [3, 14], where dedicated modules for modeling text tokens or visual tokens exist within these models. However, processing visual and textual tokens using the same parameters may make it difficult to optimize the model for

1 Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China. E-mail: H. Bao, hangbobao@gmail.com; S. Piao, piaosh@hit.edu.cn (✉).
2 Microsoft Research Asia, Beijing 100080, China. E-mail: L. Dong, lidong1@microsoft.com; F. Wei, fuwei@microsoft.com.
Manuscript received: 2023-06-14; accepted: 2023-09-01

different modalities. Some studies [15, 16] proposed methods that use distinct parameters to handle different modalities to improve the performance of multimodal understanding tasks [17, 18]. This suggests that for text-to-image models based on encoder structures, there exists the potential to enhance model performance through well-designed decoupled model architectures.

In this study, we aim to harness the performance improvements offered by a concise and efficient decoupled model design and the inference efficiency provided by bidirectional encoding. To achieve this, we explore the potential and performance of Multiway Transformers [16] for text-to-image generation tasks. Our approach can be divided into two parts: image tokenizer and text-to-image modeling. First, we propose a method for improving the image tokenizer: VQGAN Finetuning. Using pretrained Vision Transformers [19] (such as BEiT-3 [15]) for the initialization of the encoder and decoder within the autoencoder during VQGAN training [20], we can enhance the quality of the reconstructed images. Subsequently, we employ the obtained $\text{VQGAN}_{\text{BEiT-3}}$ as image tokenizer to convert 256×256 resolution images into 16×16 visual tokens. Next, we randomly mask the resulting image tokens and combine them with unmasked text tokens to form a token sequence that serves as the input for the bidirectional Multiway Transformers. The model then attempts to restore the masked visual tokens. During the model training process, the parameters were initialized using the BEiT-3 model. Owing to its bidirectional encoding property, the inference stage can utilize non-autoregressive decoding methods to improve the inference efficiency, such as requiring only 24 decoding steps versus 256 steps for autoregressive encoding.

On the MS-COCO benchmark [21], we empirically demonstrate that Multiway Transformers outperform vanilla Transformers, achieving superior FID scores [22]. This confirms the efficacy of the modality-specific parameter computation design within Multiway Transformers, which helps to enhance the performance in text-to-image generation tasks. Furthermore, our ablation studies reveal that the visual tokens to text tokens fusion of self-attention in bidirectional encoding contributes to improved model performance, an outcome that cannot be achieved using methods that rely on text encoders and visual decoders. Furthermore, the proposed $\text{VQGAN}_{\text{BEiT-3}}$ outperforms VQGAN in terms of image reconstruction quality on ImageNet validation set, as evidenced by FID scores of 1.88 compared with 4.98, and it also enhances the generation of text-to-image results. By incorporating the additional CC-3M [23] dataset for intermediate finetuning on our Multiway Transformer model with 688M parameters, we obtain a competitive result, achieving a finetuned FID score of 4.98 on MS-COCO. In addition to empirically demonstrating the effectiveness of our method, we provide a theoretical explanation in Section 2.3 of the proposed approach from the perspective of variational autoencoders [50].

2 Method

As illustrated in Fig. 1, our approach comprises two main components: VQGAN Finetuning for the image tokenizer and Multiway Transformers for text-to-image modeling.

2.1 Image tokenizer

We follow recent studies [1, 20, 26] and use a discrete variational autoencoder (dVAE) [27] as an image tokenizer to convert images from pixel space $x \in \mathbb{R}^{H \times W \times C}$ to discrete tokens $z = [z_i]_{i=1}^{N}$, with each token z_i belonging to a visual vocabulary $\mathcal{V} = \{1, 2, \ldots, |\mathcal{V}|\}$. During the training phase of the text-to-image model, the image tokenizer converts the input images into discrete visual tokens for the input representation and supervision of the masked visual tokens. During the inference phase of the text-to-image model, the image tokenizer converts discrete tokens into pixels. We adopt the same model architecture as ViT-VQGAN [26] by utilizing Vision Transformers (ViTs) [19] for both encoding and decoding within the image tokenizer.

In this study, we transform input images with a resolution of 256×256 pixels into 16×16 visual tokens. The size of the visual vocabulary is set to $|\mathcal{V}| = 8192$.

VQGAN finetuning. In contrast to previous studies [20, 26] that employ random initialization, we adhere to the widely-used pretrain-finetune paradigm [13, 28, 29], which has been successfully applied in various tasks [21, 30–32] across natural language processing and computer vision. During training of the image tokenizer, we initialize the

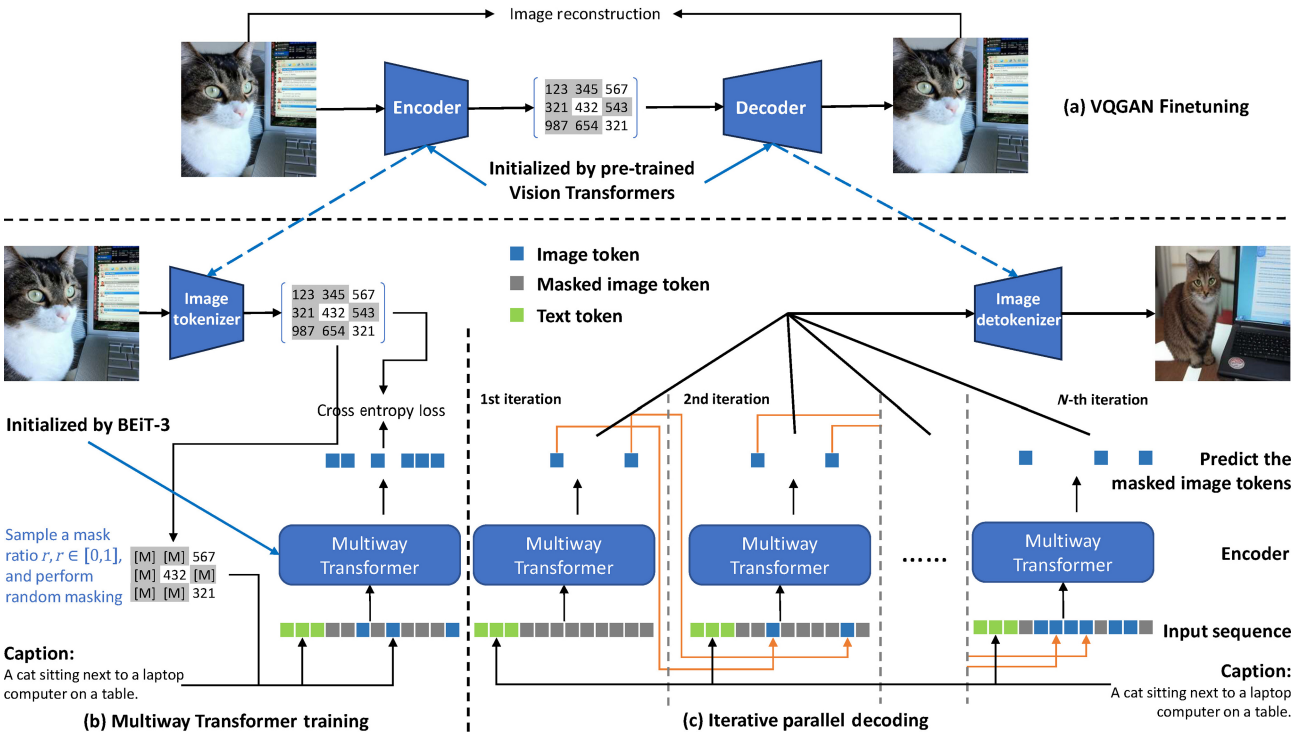

Fig. 1 An overview of our proposed approach entails (a) obtaining the $\text{VQGAN}_{\text{BEiT-3}}$ as the image tokenizer via VQGAN Finetuning, (b) training the bidirectional Multiway Transformer encoder for text-to-image generation tasks, and (c) performing iterative parallel decoding.

encoder using a pretrained Vision Transformer model, such as BEiT-3 [15]. The finetuned image tokenizer obtained through this process is referred to as $\text{VQGAN}_{\text{BEiT-3}}$.

Following previous work [20, 26], during the finetuning process, we employ ℓ_2 loss $\mathcal{L}_{\ell_2}$, perceptual loss $\mathcal{L}_{\text{p}}$ [33, 34], and patch-based adversarial loss $\mathcal{L}_{\text{adv}}$ [35]. We apply the Gumbel-Softmax reparameterization technique [1, 36, 37] to learn the visual vocabulary. To improve the utilization of the visual vocabulary, we also employ a diversity loss $\mathcal{L}_{\text{d}}$ [38]. The final finetuning loss function is given by

$$\mathcal{L} = 1.0\mathcal{L}_{\ell_2} + 1.0\mathcal{L}_{\text{p}} + 0.1\mathcal{L}_{\text{adv}} + 0.01\mathcal{L}_{\text{d}} \quad (1)$$

Decoder finetuning. Inspired by previous work [3, 4], we finetune a larger size pretrained Vision Transformer as the decoder for the image tokenizer once the visual vocabulary has been optimized and fixed. By keeping the encoder and visual vocabulary classifier parameters frozen, we can employ a larger pretrained Vision Transformer model, such as BEiT-3_{L} [15], as a decoder for finetuning. This strategy allows us to achieve enhanced image reconstruction quality, which in turn enables the generation of higher-quality images from textual descriptions.

2.2 Text-to-image modeling

We employ Multiway Transformer as a bidirectional encoder to generate the corresponding visual tokens based on the given textual input. The $\text{VQGAN}_{\text{BEiT-3}}$ serves as an image tokenizer that handles the conversion between visual tokens and pixels.

Input representations. For a given image–text pair (x, y), we process both tokens to serve as the input for the bidirectional encoder. Following the BERT settings [13] for the text, we process the input text using a sentencepiece model [39] to form a sequence composed of the BPE [40] tokens t. Additionally, an `[SOS]` token is added at the beginning to indicate the start of the sequence, and an `[EOS]` token is added at the end to indicate the end of the sequence. The image tokenizer converts the image into a sequence of visual tokens z by following previous study [1]. During the training and inference process, the `[MASK]` token is used to represent the visual tokens that need to be predicted. In the training stage, we mask the image tokens by randomly sampling a mask ratio $r \in [0, 1]$ from the density function $p(r) = \frac{2}{\pi\sqrt{1-r^2}}$ represents a truncated arccos distribution [4]. Here, we represent the masked positions with $m = [m_i]_{i=1}^{N}$ and the masked visual

tokens $\tilde{z}$. For any given position $i \in \{1, \ldots, N\}$, if $m_i = 1$, the visual token at this position is masked and $\tilde{z}_i =$ [MASK]. Conversely, if $m_i = 0$, the token remains unmasked at this position and $\tilde{z}_i = z_i$. Finally, these text tokens $t = [t_i]_{i=1}^{|t|}$ and masked visual tokens $\tilde{z} = [\tilde{z}_i]_{i=1}^{N}$ are concatenated into a single sequence $\boldsymbol{s} = [t; \tilde{z}]$, and their corresponding embedding vectors, combined with position vectors, serve as input $\boldsymbol{H}_0 = [\boldsymbol{H}_0^t; \boldsymbol{H}_0^{\tilde{z}}]$ for the bidirectional encoder.

Backbone network: Multiway Transformers. We utilize Multiway Transformers [16] as the encoder for text-to-image modeling. The encoder contains L Multiway Transformer layers $\boldsymbol{H}^l = \text{Layer}(\boldsymbol{H}^{l-1})$, where $l = 1, 2, \ldots, L$. The output vectors of the last layer, $\boldsymbol{H}^L = [\boldsymbol{H}_L^t; \boldsymbol{H}_L^{\tilde{z}}]$, are regarded as text–image encoding representations. As illustrated in Fig. 2, each layer within the Multiway Transformer encoder receives input tokens composed of token sequences from two distinct modalities: image and text. Throughout the multi-head self-attention and feed-forward network module operations, the model selects appropriate parameters for each token based on its modality. For parameter initialization, we utilize the pretrained weights of the general multimodal pretrained model BEiT-3 [15].

Training. For given visual tokens z, corresponding textual description y and the masked positions $m = [m_i]_{i=1}^{N}$. We represent the masked visual tokens as $\tilde{z}$. After applying Multiway Transformer to bidirectionally encode the input, we obtain the output from the final layer, $\boldsymbol{H}^L = [\boldsymbol{H}_L^t; \boldsymbol{H}_L^{\tilde{z}}]$. For each masked position with $m_i = 1$, we use a softmax classifier to predict the masked visual token, $p(z_i|y, \tilde{z}) = \text{softmax}(\boldsymbol{W}_c h_{L,i}^{\tilde{z}} + \boldsymbol{b}_c)$. The training objective is to minimize the negative log-likelihood:

$$\mathcal{L}_{\text{mask}} = -\mathop{\mathbb{E}}_{(z,y,m)\in\mathcal{D}} \Big[\sum_{\forall i\in[1,N],\ m_i=1} \log p(z_i|y,\tilde{z}) \Big] \tag{2}$$

Since the $\text{VQGAN}_{\text{BEiT-3}}$ we use is trained with the Gumbel-Softmax reparameterization technique and can provide soft labels, we utilize soft labels as training targets during the training process instead of relying on one-hot labels, as in VQGAN [20] and ViT-VQGAN [26].

Iterative parallel inference. After the training is completed, we use the Multiway Transformer to iterative parallel decode [4, 47–49] the corresponding image tokens based on the text input and then use the image tokenizer to map them to pixels. As shown in Fig. 1: (1) Initially, all the image tokens are masked, representing a blank space. (2) We concatenate these image tokens with the input text tokens and feed them into the Multiway Transformer encoder for encoding and prediction. This allows us to obtain confidence levels for each masked position, and because of the bidirectional encoding property of the Multiway Transformer encoder, the prediction of multiple masked tokens is achievable. (3) Based on the predicted confidence scores, a subset of tokens is selected as the input for the next iteration. (4) If the input contains masked image tokens, we continue steps two and three until all image tokens have been successfully predicted. This iterative parallel inference process enables effective and efficient generation of images based on textual input.

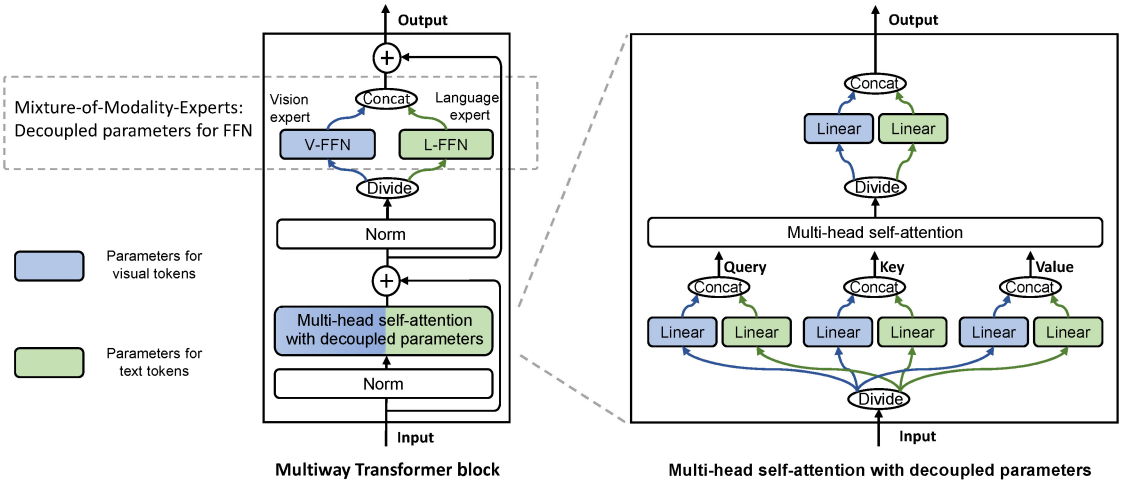

Fig. 2 Multiway Transformer—decoupled architecture.

2.3 From the perspective of variational autoencoder

Our training process maximizes the distribution of the RGB image generated by the given textual description and masked visual tokens. Let x denote the original image, z the corresponding visual tokens, and y the corresponding text description. We use $\tilde{z}$ to denote the masked visual tokens. This training process can be viewed as the variational autoencoder [50] training, and the evidence lower bound (ELB) of the log-likelihood $\log p(x|\tilde{z}, y)$ is

$$\sum_{(x_i,\tilde{z}_i,y_i)\in\mathcal{D}} \log p(x_i|\tilde{z}_i, y_i) \geqslant \sum_{(x_i,\tilde{z}_i,y_i)\in\mathcal{D}} \Big(\mathbb{E}_{z_i\sim q_\phi(\boldsymbol{z}|x_i)} \log p_\psi(x_i|z_i) - D_{\mathrm{KL}}\big(q_\phi(\boldsymbol{z}|x_i), p_\theta(\boldsymbol{z}|\tilde{z}_i, y_i)\big)\Big) \quad (3)$$

where

- $q_\phi(\boldsymbol{z}|x_i)$ denotes the distribution of visual tokens generated by the image tokenizer given the input image[①];
- $p_\psi(x_i|z_i)$ denotes the distribution of the images generated by the image tokenizer given the visual tokens;
- $p_\theta(\boldsymbol{z}|\tilde{z}_i, y_i)$ denotes the distribution of generated visual tokens given the text description and masked visual tokens, modeled by the text-to-image model.

We adopt a two-stage training approach similar to Refs. [4, 29, 41, 47, 51, 52]. In the first stage, we utilize the method described in Section 2.1 to finetune a pretrained visual Transformer as a discrete variational autoencoder [1], serving as the image tokenizer, where the encoder and decoder parameters are denoted by ϕ and ψ, respectively. In the second stage, we first fix the image tokenizer parameters ϕ and ψ, and subsequently employ the method presented in Section 2.2 using Multiway Transformers, learning the parameter θ.

3 Experiments

3.1 Experimental setup

Image tokenizer. Our image tokenizer incorporates both encoder and decoder structures based on Vision Transformers [10, 19], similar to ViT-VQGAN [26]. The encoder is a 12-layer Vision Transformer with 12 attention heads and a hidden layer size of 768. The intermediate size of feed-forward networks is 3072. The visual vocabulary size is 8192 and the quantization layer is an 8192-class linear classifier. The decoder is a 24-layer Transformer model with 16 attention heads and a hidden layer size of 1024. The intermediate size of feed-forward networks is 4096. During the training process, image processing utilized only random cropping for data augmentation. All the input images are processed at a resolution of 256×256. The images are then compressed by the encoder into 16×16 image tokens, and can be restored by the decoder back to 256×256 pixels.

The training of the image tokenizer consists of two stages. In the first stage, the encoder is initialized using the visual parameters from BEiT-3_{B} [15], whereas the decoder uses a Transformer structure identical to that of the encoder. All the parameters are trainable in the first stage. We employ the Adam optimizer [55] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$ for training with a peak learning rate of 2×10^{-4}, cosine learning rate decay of 1×10^{-6}, and training batch size of 256. The weight decay is set to 0.01, and the learning rate warm-up consists of 10,000 steps from zero to the peak learning rate. The relaxation temperature τ decreased linearly from 1.0 to 1/16 [1] during the initial 50,000 steps. During the training process, the dropout [56] and stochastic depth [57] are disabled. In the second stage, the encoder and quantization layer parameters are fixed. The parameters of the decoder obtained in the first stage are discarded, and the visual parameters from BEiT-3_{L} [15] are used to initialize the decoder's parameters. The embedding of the input visual tokens for the decoder is initialized randomly in this stage. A higher learning rate of 1×10^{-3} is used, whereas the other hyperparameters remain the same as those in the first stage.

We train the image tokenizer on both the ImageNet [31] and CC-3M [23] datasets. For the ImageNet training set, we train for 100 epochs (approximately 500k steps) in the first stage and 50 epochs (approximately 250k steps) in the second stage. For the CC-3M training set, we train for 40 epochs (approximately 470k steps) in the first stage and 20 epochs (approximately 235k steps) in the second stage.

Multiway Transformers. We employ base- and large-sized models for text-to-image modeling. Our base

① We assume that the text description y is conditionally independent to this distribution.

model is a 12-layer Multiway Transformer encoder with 12 attention heads and a hidden layer size of 768. The intermediate size of feed-forward networks is 3072. Our large model is based on a 24-layer Multiway Transformer encoder, with each layer containing 16 attention heads and a hidden layer size of 1024. The intermediate size of feed-forward networks is 4096. In each layer of Multiway Transformers self-attention block, there are two sets of parameters for the query, key, value, and output linear layers, as well as layer-norm modules. One set is used for visual token computation and the other set is used for text token computation. In addition, within the feed-forward networks, a vision expert network is employed to process the visual tokens, and a language expert network is used to process the text tokens. Cross-attention is not required in the Multiway Transformers. The input visual vocabulary is set to 8192.

In our experiments, we chose the multimodal pretrained model BEiT-3 [15][②] with the same model size for parameter initialization. Because BEiT-3 is pretrained based on patch embedding, the embedding of visual tokens in our implementation is randomly initialized. During BEiT-3 pretraining, VQKD [58] was used as the image tokenizer, and the latent representation of the CLIP model was restored [59–61], which is fundamentally different from the restored image pixels. Therefore, using the BEiT-3 model for initialization does not imply that supervised text–image generation pretraining is introduced.

For text-to-image generation tasks, the input images were directly resized to a resolution of 256×256. We used the Adam optimizer with a learning rate of 1×10^{-3} and layer-wise learning rate decay [29, 62] of 0.9. During the training process, dropout [56] is disabled, and the stochastic depth [57] is set to 0.1. The training batch size is 256. The learning rate increases linearly from 0 to the peak value within the first 5000 steps and then decays to zero using the cosine schedule. When training large-sized models, techniques such as checkpointing are used to save GPU memory space. For direct training on MS-COCO [21], the number of training epochs is set to 30 and the $\text{VQGAN}_{\text{BEiT-3}}$ trained on ImageNet is used. On the CC-3M dataset, we train for 100 epochs, and the obtained weights are then finetune for another 20 epochs on MS-COCO using the $\text{VQGAN}_{\text{BEiT-3}}$ obtained from training on the CC-3M dataset.

② https://github.com/microsoft/unilm/tree/master/beit3

3.2 Experimental results

We conduct text-to-image experiments on the MS-COCO [21] and CC-3M datasets [23]. We use the Fréchet Inception Distance (FID) [22] as a performance evaluation metric to compare our method with previous studies. We also provide random image samples generated based on MS-COCO prompts without cherry-picking in Fig. 3.

Table 1 presents a comparison between the proposed method and previous approaches for a subset of 30,000 captions sampled from the MS-COCO [21] validation set. Our method, when directly finetuned on MS-COCO without additional data, achieves an FID score of 8.19 with 246M parameters $\text{Multiway Transformer}_{\text{B}}$. When the model parameters increase to 688M, $\text{Multiway Transformer}_{\text{L}}$ attains an FID score of 5.75, demonstrating that our method can further enhance the performance as the model parameters increase. Additionally, after intermediate finetuning [29, 67] on the CC-3M dataset containing three million text–image pairs, $\text{Multiway Transformer}_{\text{L}}$ achieves an FID score of 4.98 on MS-COCO. We utilize the BEiT-3 [15] model for parameter initialization, which employs VQKD [58] as image tokenizer for training to recover the latent in the CLIP model [59, 61]. Consequently, there is no supervised pretraining from text-to-image during the pretraining phase of BEiT-3. The additional labeled data used in our method are minimal compared with other approaches. Despite CoBIT using more than one billion extra data points (more than 300 times ours), we achieve superior performance at the more than 600M model parameter scale, with an FID score of 4.98 (ours) versus 5.06 (CoBIT).

Table 2 presents a comparison between the proposed method and previous approaches for the CC-3M [23] validation set. All the models in the table are trained on the CC-3M dataset. Compared with other methods with more than 600M parameters, our proposed approach achieves the best performance with an FID score of 8.33.

Our method utilizes a bidirectional encoding strategy that allows for non-autoregressive training and inference. Unlike other non-autoregressive methods [4, 11, 12] that handle different modalities using shared para-

Fig. 3 Qualitative comparison with previous studies on MS-COCO prompts. The generated images were taken from Ref. [25]. We do not perform any cherry-picking for our method.

Table 1 Quantitative evaluation of FID score on MS-COCO [21] for 256 × 256 image resolution

Approach	Method type	Model arch	# Params	# Extra labeled data	FID ↓
Zero-short performance					
LDM [8]	Diffusion	—	1.45B	Laion-400M	12.63
GLIDE [25]	Diffusion	—	3.50B	In House 250M	12.24
DALL-E 2 [41]	Diffusion	—	3.50B	In House 250M	10.39
Imagen [42]	Diffusion	—	3.40B	Imagen-460M	7.27
DALL-E [1]	Autoregressive	Shared	12.00B	In House 250M	17.89
Muse [4]	Non-autoregressive	Shared	7.60B	Imagen-460M	7.88
Parti [3]	Autoregressive	Decoupled	20.00B	Laion-400M, ALIGN-1.8B, JFT-4B	7.23
Finetuning performance					
AttnGAN [43]	GAN	—	—	—	35.49
XMC-GAN [24]	GAN	—	—	—	9.33
LAFITE [44]	GAN	—	—	—	8.12
MAGVLT [11]	Non-autoregressive	Shared	447M	CC-3M, CC-12M, SBU, VG	10.74
OFA [45]	Autoregressive	Decoupled	472M	OpenImages, YFCC-100M, ImageNet-21K	10.50
Make-A-Scene [46]	Autoregressive	Shared	4000M	YFCC-100M, CC-12M, CC-3M, Redcaps	7.55
CoBIT_{B} [14]	Autoregressive	Decoupled	626M	ALIGN-1.1B, WebLI-162M	5.06
CoBIT_{L} [14]	Autoregressive	Decoupled	1091M	ALIGN-1.1B, WebLI-162M	4.62
Parti [3]	Autoregressive	Decoupled	20,000M	Laion-400M, ALIGN-1.8B, JFT-4B	3.22
Our finetuning performance					
Multiway $\text{Transformer}_{\text{B}}$	Non-autoregressive	Decoupled	246M	None	8.19
Multiway $\text{Transformer}_{\text{L}}$	Non-autoregressive	Decoupled	688M	None	5.75
Multiway $\text{Transformer}_{\text{L}}$	Non-autoregressive	Decoupled	688M	CC-3M	4.98

Fig. 4 Comparison of images reconstructed by different image tokenizers. The images are sampled from the MS-COCO dataset [21]. The input and recovered image resolutions are 256 × 256, and the latent size for all image tokenizers is 16 × 16.

Table 2 Quantitative evaluation of FID score on CC-3M [23] for 256 ×256 image resolution

Approach	Method type	Model arch	# Params	FID ↓
VQGAN [20]	Autoregressive	Decoupled	600M	28.86
ImageBART [53]	Diffusion+autoregressive	—	2800M	22.61
LDM-4 [8]	Diffusion	—	645M	17.01
RQ-Transformer [54]	Autoregressive	Shared	654M	12.33
Draft-and-revise [12]	Non-autoregressive	Shared	654M	9.65
Muse [4]	Non-autoregressive	Shared	4700M	6.8
Multiway Transformer (ours)	Non-autoregressive	Decoupled	688M	8.33

Table 3 Fréchet Inception Distance (FID) score between the reconstructed images and the original images

Image tokenizer	Initial model encoder–decoder	Latent size	Dataset	Throughput (images/s)	FID on validation	
					ImageNet	COCO
VQGAN-1024 [20]	Random/random	16 × 16	ImageNet	30.0	7.94	9.26
VQGAN-16384 [20]	Random/random	16 × 16	ImageNet	30.0	4.98	5.91
DALL-E dVAE [1]	Random/random	32 × 32	Web data	48.0	32.00	43.41
VQGAN$_{\text{BEiT-3}}$	BEiT-3$_{\text{B}}$/BEiT-3$_{\text{L}}$	16 × 16	ImageNet	**84.5**	**1.88**	**3.07**
VQGAN$_{\text{BEiT-3}}$	BEiT-3$_{\text{B}}$/BEiT-3$_{\text{L}}$	16 × 16	CC-3M	**84.5**	2.73	3.15

meters, Multiway Transformers adopts a decoupled design. We further explore the impact of the decoupled design in Section 3.3.

3.3 Ablation studies

Encoder of image tokenizer. During the VQGAN Finetuning, we investigate the impact of the initialization weight for the encoder. We select multiple weights for initialization: (1) random initialization; (2) contrastive learning-based models, such as MoCo-V3 [65]; (3) self-distillation based models, such as DINO [63]; (4) masked image modeling based models, such as BEiT [29], MAE [64], and BEiT v2 [58]; (5) multimodal pretraining models, such as BEiT-3 [15]; (6) supervised models, such as DeiT [66] trained on ImageNet; (7) image–text contrastive learning models, such as CLIP [61]. All the chosen models use a ViT-base as backbone network with patch size of 16×16. These models are publicly accessible, and we utilize their weights to initialize the encoders for the image tokenizer. Subsequently, we perform VQGAN Finetuning. The image tokenizer is trained on the ImageNet [31] training dataset for 20 epochs in the ablation study unless stated otherwise. We evaluate the quality of recovered images by comparing them to the validation set of MS-COCO [21] and ImageNet [31] datasets, using Frèchet Inception Distance (FID) [22] and Inception Score (IS) [68] as evaluation metrics through auto-encoding.

Table 4 lists the results obtained by initializing the image tokenizer encoder using these models. We discover that random initialization performs the poorest, indicating that incorporating these pretrained Vision Transformers can assist the VQGAN Finetuning in acquiring an image tokenizer capable of superior image reconstruction. We observe that BEiT v2 and BEiT-3 exhibited the best performances. Both BEiT v2 and BEiT-3 were trained using VQKD [58] as the training target for masked image modeling. Considering that the BEiT-3 pretraining dataset was larger than BEiT v2 and that their performances were similar, we chose BEiT-3 for encoder initialization in subsequent experiments. By increasing the number of training epochs to 100, we could further achieve enhanced FID and IS scores for image reconstruction, which will be employed in subsequent experiments.

Unless otherwise specified, we train the image tokenizer on the ImageNet dataset with 20 epochs for ablation. We evaluate the quality using FID and IS between the recovered images and the validation set of the MS-COCO and ImageNet datasets as evaluation metrics through auto-encoding.

Decoder of image tokenizer. After completing the first stage of the image tokenizer, we fix the encoder and quantization layer, re-initialize the decoder parameters, and proceed to the second stage of training for the decoder of the image tokenizer. Similarly, we conduct the second stage of training on the ImageNet dataset for 20 epochs unless otherwise indicated. The evaluation method and metrics are consistent with those employed in the first stage. As illustrated in Table 5, we investigate the impact

Table 4 Comparison of various Vision Transformers leveraged as encoders for the image tokenizer during the first stage of VQGAN Finetuning

Initial model	MS COCO		ImageNet	
	FID ↓	IS ↑	FID ↓	IS ↑
Random	4.89	27.9	4.10	144.3
BEiT [29]	4.04	29.3	3.09	162.8
DINO [63]	4.27	28.9	3.34	158.1
MAE [64]	4.02	29.1	3.06	164.1
MoCo v3 [65]	4.46	28.4	3.54	152.7
DeiT [66]	4.17	28.9	3.28	158.7
CLIP [61]	4.03	29.3	3.08	163.7
BEiT v2 [58]	3.70	29.7	2.76	168.2
BEiT-3 [15]	3.73	29.6	2.85	168.5
BEiT-3* [15]	**3.17**	**30.4**	**2.24**	**175.5**

* denotes a training duration of 100 epochs.

Table 5 Comparison of various Vision Transformers leveraged as decoders for the image tokenizer during the second stage of VQGAN Finetuning

Initial model	MS COCO		ImageNet	
	FID ↓	IS ↑	FID ↓	IS ↑
Base-sized decoder				
Random	3.74	29.4	2.95	166.9
BEiT [29]	3.52	29.9	2.65	170.2
DINO [63]	3.64	29.7	2.97	167.4
MAE [64]	3.53	29.8	2.68	169.7
MoCo v3 [65]	3.58	29.7	2.74	168.7
DeiT [66]	3.61	29.8	2.78	168.4
CLIP [61]	3.60	29.8	2.71	170.1
BEiT v2 [58]	3.42	30.3	2.53	172.3
BEiT-3 [15]	3.46	30.3	2.53	172.4
BEiT-3* [15]	**3.16**	**31.2**	**2.09**	**180.3**
Large-sized decoder				
BEiT-3* [15]	**3.07**	**31.9**	**1.88**	**186.3**

* denotes a training duration of 50 epochs.

of different initialization weights on the decoder during the second stage, with specific initialization models being consistent with those chosen in the first stage. Random initialization yields the poorest results, analogous to the encoder experiment in the first stage. We discover that utilizing pretrained Vision Transformers for parameter initialization for the decoder of the image tokenizer can also enhance the image recovery performance. We observe that utilizing BEiT v2 [58] and BEiT-3 [15] for decoder initialization yielded the most optimal results. We select BEiT-3 for decoder initialization in subsequent experiments. When the number of training epochs was further extended to 50, both FID and IS scores improved. Finally, we attempt to increase the decoder parameters [3, 4] and employ the large version of BEiT-3 for parameter initialization. Increasing the decoder parameters further enhanced the FID and IS scores.

In this study, we employ the $\text{VQGAN}_{\text{BEiT-3}}$, which uses BEiT-3_{B} for first-stage encoder initialization and BEiT-3_{L} for second-stage decoder initialization, as the image tokenizer.

Comparison of image tokenizers. In Table 3, we compare our $\text{VQGAN}_{\text{BEiT-3}}$trained on ImageNet [31] and CC-3M [23] with those proposed in previous studies such as VQGAN [20] and dVAE [1]. The evaluation metric is the FID score calculated between the recovered and original images on the validation sets of the ImageNet [31] and MS-COCO [21] datasets. We observe that our $\text{VQGAN}_{\text{BEiT-3}}$ trained on the ImageNet [31] dataset achieves superior FID scores compared to the VQGAN [20] and dVAE [27] using the random initialization, with a score of 1.88 (ours) vs. 4.98 (VQGAN) on ImageNet. Furthermore, in Fig. 3, we present examples of image reconstruction using different image tokenizers. In these instances, all image tokenizers have a latent size of 16×16, ensuring a fair comparison.

Furthermore, we compare the $\text{VQGAN}_{\text{BEiT-3}}$ obtained in this study with previous methods, such as VQGAN [20] and ViT-VQGAN [26], in terms of their impact on text-to-image generation performance when used as image tokenizers. We use BEiT-3_{L} to initialize Multiway Transformers and subsequently finetune it on the MS-COCO [21] training set for the ablation experiments with a model parameter size of 688M. In the VQGAN experiment, we use a VQGAN checkpoint trained on ImageNet. For the ViT-VQGAN experiment, we employ randomly initialized Vision Transformers as encoders and decoders for the image tokenizer based on our implementation, which we refer to as ViT-VQGAN (our imp).

Table 6 demonstrates that the choice of image tokenizer can affect the text-to-image performance. ViT-VQGAN (our imp) and $\text{VQGAN}_{\text{BEiT-3}}$, which leverage Vision Transformers as the backbone network for encoders and decoders, surpass VQGAN, which utilizes a CNN-based backbone network, with respect to FID scores. The ViT-VQGAN (our imp) trained with BEiT-3 for initialization achieves a superior FID score of 6.07 in comparison to the randomly initialized ViT-VQGAN (our imp), despite having identical model frameworks, number of the parameters, and inference speeds. This indicates that in our proposed VQGAN Finetuning method, utilizing the pretrained BEiT-3_{B} [15] as the initial image tokenizer can enhance the text-to-image model's FID score on the MS-COCO dataset. Furthermore, by keeping the encoder and visual vocabulary fixed, employing the pretrained BEiT-3_{B}as the weight initialization for the decoder can further improve the FID score to 5.88. Moreover, employing a larger pretrained model BEiT-3_{L} [15] contributes to an improved FID score of 5.75.

After establishing the mapping between the continuous pixel space and discrete visual token space using the $\text{VQGAN}_{\text{BEiT-3}}$, powerful pretrained models can be employed to refine the decoder for enhanced recovery capability from discrete visual tokens to pixels. As the encoder and visual vocabulary remain fixed, we can directly upgrade the trained Multiway Transformers model, thereby saving resources that would be consumed in retraining.

Table 6 Comparison of different image tokenizers in text-to-image generation tasks. The three configurations of $\text{VQGAN}_{\text{BEiT-3}}$ employ identical encoder, codebook, and text-to-image model weights. The only difference stems from the decoder of the image tokenizer, which varies among those three configurations

Image tokenizer	Initial model encoder–decoder	MS-COCO FID ↓
VQGAN	Random/random	15.58
ViT-VQGAN (our imp)	Random/random	7.65
$\text{VQGAN}_{\text{BEiT-3}}$ (ours)	BEiT-3_{B}/random	6.07
	BEiT-3_{B}/BEiT-3_{B}	5.88
	BEiT-3_{B}/BEiT-3_{L}	**5.75**

Text-to-image modeling. We conduct ablation studies to analyze the contribution of each component in our model. We experiment with the Multiway Transformer$_{\mathrm{L}}$ model on MS-COCO [21] using the FID scores of 30,000 captions sampled from the MS-COCO [21] validation set as the evaluation metric.

The results of the ablation experiments are presented in Table 7. (1) First, we investigate the impact of the pretrained weights for the text. We find that randomly initializing the text part parameters during model initialization causes some performance degradation (+0.15 on FID), indicating that the pretrained text weights help improve the model performance. (2) Moreover, when we freeze the text part parameters during training, we also observe some degradation (+0.19 FID), suggesting that learnable parameters can enhance performance during the training process. (3) Next, we randomly initialize the visual part parameters and find that, compared to the text part, removing the pretrained visual weights has a more significant impact (+2.51 FID). (4) When we randomly initialize all the parameters in the Multiway Transformer model, we observe a more severe performance degradation of +4.15 FID. This indicates that the initialization from BEiT-3 [15] improves the model's performance. (5) When we replace the soft label provided by VQGAN$_{\mathrm{BEiT\text{-}3}}$ with a one-hot label in text-to-image training, the model performance suffers from a +0.9 FID. This indicates that the soft label provided by the VQGAN$_{\mathrm{BEiT\text{-}3}}$ helps the model understand the relationships between different tokens in the visual vocabulary. (6) In the encoding process of Multiway Transformers, we mask attention from image to text in the self-attention mask to investigate the importance of bidirectional encoding in Multiway Transformers. Removing this interaction leads to a performance drop of 0.41 FID, demonstrating the value of incorporating image-to-text information exchange into the model. In contrast to the approach proposed in this paper, in the encoder–decoder structure methods [3, 4], text tokens within the encoder can only interact with other text tokens inside the encoder, without any fusion with image tokens. Traditional encoder-to-decoder models do not perform image-to-text fusion, which is an advantage of Multiway Transformers. (7) Finally, we replace the Multiway Transformer model architecture with a vanilla Transformer [10] and observe a performance decline of +0.72 FID, indicating that the decoupled model design improves text-to-image generation performance.

Table 7 Ablation studies for Multiway Transformers on MS-COCO text to image generation

	Ablation setting	FID ↓	Δ
	Multiway Transformer	5.75	
(1)	w/o pretrained text weight	5.90	+ 0.15
(2)	w/ freeze text encoder	5.94	+ 0.19
(3)	w/o pretrained vision weight	8.26	+ 2.51
(4)	w/o pretrained weight	9.90	+ 4.15
(5)	w/o softlabel	6.65	+ 0.90
(6)	w/o image-to-text fusion	6.16	+ 0.41
(7)	Vanilla Transformer	6.47	+ 0.72

4 Related work

4.1 Text-to-image generation

Early text-to-image generation studies primarily focused on GAN-based approaches [5, 7, 24, 43, 44]. With the emergence of large-scale, high-quality image-text datasets [1, 42, 69, 70], and the adoption of paradigms [13, 28, 71] from natural language processing, these methods utilize image tokenizers [20, 26, 27, 54, 72] to convert images into visual tokens, and then employ Transformer models to perform text token to visual token modeling. After the model generates the corresponding visual tokens, the image tokenizer converts them back into pixels.

DALL-E [1] first trained a GPT [28, 73] to generate visual tokens using text input as prompts, while other studies, such as Parti [3], adopted a sequence-to-sequence modeling [71] approach using a text encoder to encode input text and an image decoder to generate corresponding visual tokens. Works like DALL-E [1], CogView [2], and RQ-Transformer [54] used shared parameters to encode both visual and text tokens, while others, such as Parti [3] and CoBIT [14], employed decoupled design with separate text and visual modules for different modalities. These methods use autoregressive encoding for training and inference, and generate one token per iteration during the inference stage. By leveraging bidirectional encoding properties [13, 47] and non-autoregressive techniques [48, 49]. Muse [4] can predict multiple visual tokens

in a single iteration, thereby reducing the number of iterations and accelerating the inference speed. By contrast, these methods typically adopt shared-parameter designs. Our study uses a decoupled design through Multiway Transformers [16], which allows for bidirectional encoding and non-autoregressive modeling while still handling different modalities with the corresponding model parameters.

In addition, recent research on diffusion-based methods [8, 9, 41, 42] has shown significant potential.

4.2 Model architecture for multimodal modeling

Recently, several Transformer-based methods have made continuous progress in multimodal domain. Early approaches employed a vanilla Transformer model [10, 19] for both vision and language modeling, sharing the parameters between visual and textual encoding.

For multimodal understanding tasks [17, 18], early methods utilized a shared vanilla Transformer [19] to model visual features [74, 75] and text tokens. Except for the visual feature extraction module [74, 75], the parameters within the Transformer backbone network are shared using the same parameters to process both the visual features and text tokens. ViLT [76] further simplified this approach using patch projection [19] to replace the previous feature extractor [74, 75], resulting in a competitive performance. In contrast to the previous shared-parameter approach for text and visual modeling, some studies [45, 77–79] independently designed decoupled text encoders and visual encoders. Consequently, the text and visual modeling parts are encoded separately without cross-modal deep fusion, such as text-to-image and image-to-text fusion. Unlike the approach of using multiple encoders, VLMo [16] was built upon the ViLT [76] model by introducing Mixture-of-Modality-Experts. This network replaces the single feedforward network in each layer of the vanilla Transformer with multiple feedforward networks to handle different modalities, ultimately improving the performance of the model on downstream tasks.

In the text-to-generation domain, both shared [1, 20, 46, 54] and decoupled design approaches [3, 14] employ autoregressive encoding. However, non-autoregressive methods [4, 11, 12] with higher inference efficiency typically adopt shared-parameter designs. The method proposed in this study, using Multiway Transformers [15, 16] decoupled design, offers better performance than shared parameter approaches while also enabling bidirectional encoding for non-autoregressive training and inference.

4.3 Image tokenizer

Image tokenizers convert images into discrete latent variables using learned deep neural networks, such as Variational Auto-Encoders [72] and discrete Variational Auto-Encoders (dVAEs) [1, 27, 51, 52]. VQGAN [20] applies adversarial loss [35] and perceptual loss [33, 34] to synthesize images using CNNs [80]. ViT-VQGAN [26] builds upon VQGAN, improving the architecture and codebook learning with Vision Transformers [10, 19].

To the best of our knowledge, previous methods have primarily relied on random initialization for training and have not explored or utilized the pretraining and finetuning paradigm [13, 28, 29] based on pretrained models [15, 29, 65, 81].

4.4 Pretrained Vision Transformers

Inspired by the success of GPT [28] and BERT [13] in natural language processing, and the emergence of Vision Transformers [19, 82–84], various pretraining models [15, 29, 63–65, 78, 85, 86] based on Vision Transformers have been developed in the computer vision domain. In contrast to supervised pretraining, where pretraining and downstream tasks are similar, the self-supervised training paradigm of pretraining and finetuning has found widespread applications across different modalities and tasks [21, 30–32, 87]. Some studies [88] have shown that pretrained models can provide better initialization.

In the text-to-image generation domain, a significant portion of studies depend on large-scale supervised datasets [1, 3, 8, 42, 70]. To the best of our knowledge, all these approaches utilize random initialization during the training process of image tokenizers [1, 20, 26]. Pretraining and finetuning paradigm within the text-to-image generation field has not been thoroughly explored.

5 Conclusions

In this study, we explored the potential and performance of bidirectional Multiway Transformers for text-to-image generation, building on recent advancements in the field. Our approach consists of two parts: image tokenizer improvement through VQGAN Finetuning, and text-to-image modeling

using bidirectional Multiway Transformers. We demonstrated superior performance in terms of FID scores on the MS-COCO benchmark, demonstrating the benefits of the modality-specific parameter computation design and bidirectional encoding. Additionally, the proposed image tokenizer outperforms VQGAN in terms of image reconstruction quality, thereby contributing to better text-to-image generation results. By incorporating intermediate finetuning on the CC-3M dataset for our Multiway Transformer model with 688M parameters, we obtained competitive results, achieving a finetuned FID score of 4.98 on MS-COCO.

Availability of data and materials

The datasets and pretrained models used in this study are publicly available.

ImageNet: https://www.image-net.org/download.php or https://www.kaggle.com/c/imagenet-object-localization-challenge

MS-COCO: https://cocodataset.org/

CC-3M: https://ai.google.com/research/ConceptualCaptions or https://github.com/rom1504/img2dataset/blob/main/dataset_examples/cc3m.md

BEiT-3: https://github.com/microsoft/unilm/tree/master/beit3

Author contributions

All authors contributed to the conception and design of this study. Material preparation, data collection, and analysis were performed by Hangbo Bao. The first draft of the manuscript was written by Hangbo Bao and all authors commented on previous versions of the manuscript. All the authors have read and approved the final version of the manuscript.

Acknowledgements

This project received GPU computing resources from Microsoft Corporation.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. *arXiv preprint* arXiv:2102.12092, 2021.

[2] Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. CoView: Mastering text-to-image generation via transformers. In: Proceedings of the 35th Conference on Neural Information Processing Systems, 19822–19835, 2021.

[3] Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint* arXiv:2206.10789, 2022.

[4] Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M. H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint* arXiv:2301.00704, 2023.

[5] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Communications of the ACM* Vol. 63, No. 11, 139–144, 2020.

[6] Karras, T.; Laine, S.; Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401–4410, 2019.

[7] Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint* arXiv:1809.11096, 2018.

[8] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684–10695, 2022.

[9] Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint* arXiv:1503.03585, 2015.

[10] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need, In: Proceedings of the 31st Conference on Neural Information Processing Systems, 5998–6008, 2017.

[11] Kim, S.; Jo, D.; Lee, D.; Kim, J. MAGVLT: Masked generative vision-and-language transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 23338–23348, 2023.

[12] Lee, D.; Kim, C.; Kim, S.; Cho, M.; Han, W. S. Draft-and-revise: Effective image generation with contextual RQ-transformer. *arXiv preprint* arXiv:2206.04452, 2022.

[13] Devlin, J.; Chang, M-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186, 2019.

[14] You, H.; Guo, M.; Wang, Z.; Chang, K. W.; Baldridge, J.; Yu, J. CoBIT: A contrastive bi-directional image-text generation model. *arXiv preprint* arXiv:2303.13455, 2023.

[15] Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 19175–19186, 2023.

[16] Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Wei, F. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint* arXiv:2111.02358, 2021.

[17] Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6325–6334, 2017.

[18] Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 6418–6428, 2019.

[19] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929, 2020.

[20] Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12873–12883, 2021.

[21] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. L. Microsoft COCO: Common objects in context. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8693.* Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 740–755, 2014.

[22] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S.; Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6629–6640, 2017.

[23] Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2556–2565, 2018.

[24] Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; Yang, Y. Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 833–842, 2021.

[25] Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint* arXiv:2112.10741, 2021.

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

[27] Rolfe, J. T. Discrete variational autoencoders. *arXiv preprint* arXiv:1609.02200, 2017.

[28] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. 2018. Available at https://openai.com/index/language-unsupervised/

[29] Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT pre-training of image transformers. *arXiv preprint* arXiv:2106.08254, 2021.

[30] Rajpurkar, P.; Zhang J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2383–2392, 2016.

[31] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* Vol. 115, No. 3, 211–252, 2015.

[32] Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; Torralba, A. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision* Vol. 127, No. 3, 302–321, 2019.

[33] Johnson, J.; Alahi, A.; Li, F.-F. Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9906*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 694–711, 2016.

[34] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595, 2018.

[35] Isola, P.; Zhu, J. Y.; Zhou, T.; Efros, A. A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1125–1134, 2017.

[36] Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with Gumbel-Softmax, *arXiv prirpint* arXiv:1611.01144, 2016.

[37] Maddison, C. J.; Mnih, A.; Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint* arXiv:1611.00712, 2016.

[38] Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint* arXiv:2006.11477, 2020.

[39] Kudo, T.; Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66–71, 2018.

[40] Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725, 2016.

[41] Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint* arXiv:2204.06125, 2022.

[42] Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Seyed Ghasemipour, K.; Ayan, B. K.; Sara Mahdavi, S.; Gontijo Lopes, R.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint* arXiv:2205.11487, 2022.

[43] Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1316–1324, 2018.

[44] Zhou, Y.; Zhang, R.; Chen, C.; Li, C.; Tensmeyer, C.; Yu, T.; Gu, J.; Xu, J.; Sun, T. LAFITE: Towards language-free training for text-to-image generation. *arXiv preprint* arXiv:2111.13792, 2021.

[45] Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint* arXiv:2202.03052, 2022.

[46] Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; Taigman, Y. Make-A-scene: Scene-based text-to-image generation with human priors. In: *Computer Vision – ECCV 2022. Lecture Notes in Computer Science, Vol. 13675*. Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; Hassner, T. Eds. Springer Cham, 89–106, 2022.

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

[48] Gu, J.; Bradbury, J.; Xiong, C.; Li, V. O. K.; Socher, R. Non-autoregressive neural machine translation. *arXiv preprint* arXiv:1711.02281, 2017.

[49] Ghazvininejad, M.; Levy, O.; Liu, Y.; Zettlemoyer, L. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint* arXiv:1904.09324, 2019.

[50] Kingma, D. P.; Welling, M. Auto-encoding variational Bayes. *arXiv preprint* arXiv:1312.6114, 2013.

[51] van den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural discrete representation learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6309–6318, 2017.

[52] Razavi, A.; van den Oord, A.; Vinyals, O. Generating diverse high-fidelity images with VQ-VAE-2. In: Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019.

[53] Esser, P.; Rombach, R.; Blattmann, A.; Ommer, B. ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *arXiv preprint* arXiv:2108.08827, 2021.

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

[55] Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980, 2015.

[56] Murphy, K.; Schölkopf, B.; Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* Vol. 15, No. 1, 1929–1958, 2014.

[57] Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; Weinberger, K. Q. Deep networks with stochastic depth. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9908.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 646–661, 2016.

[58] Peng, Z.; Dong, L.; Bao, H.; Ye, Q.; Wei, F. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint* arXiv:2208.06366, 2022.

[59] Wei, L.; Xie, L.; Zhou, W.; Li, H.; Tian, Q. MVP: Multimodality-guided visual pre-training. In: *Computer Vision – ECCV 2022. Lecture Notes in Computer Science, Vol. 13690.* Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; Hassner, T. Eds. Springer Cham, 337–353, 2022.

[60] Wei, C.; Fan, H.; Xie, S.; Wu, C. Y.; Yuille, A.; Feichtenhofer, C. Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14668–14678, 2022.

[61] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 8748–8763, 2021.

[62] Clark, K.; Luong, M-T.; Le, Q. V.; Manning, C. D. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint* arXiv:2003.10555, 2020.

[63] Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. *arXiv preprint* arXiv:2104.14294, 2021.

[64] He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15979–15988, 2022.

[65] Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9620–9629, 2021.

[66] Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv preprint* arXiv:2012.12877, 2020.

[67] Pruksachatkun, Y.; Phang, J.; Liu H.; Htut, P. M.; Zhang X.; Pang, R. Y.; Vania, C.; Kann, K.; Bowman, S. R. Intermediate-task transfer learning with pretrained language models: When and why does it work? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5231–5247, 2020.

[68] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; Chen, X. Improved techniques for training GANs. In: Proceedings of the 30th Conference on Neural Information Processing Systems, 2016.

[69] Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning, 4904–4916, 2021.

[70] Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint* arXiv:2111.02114, 2021.

[71] Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to sequence learning with neural networks. In: Proceedings of the 28th International Conference on Neural Information Processing System, 3104–3112, 2014.

[72] Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint* arXiv:1312.6114, 2013.

[73] Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; Sutskever, I. Generative pretraining from pixels. In: Proceedings of the 37th International Conference on Machine Learning, 1691–1703, 2020.

[74] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 6, 1137–1149, 2017.

[75] Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; Fu, J. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint* arXiv:2004.00849, 2020.

[76] Kim, W.; Son, B.; Kim, I. ViLT: Vision-and-language transformer without convolution or region supervision. In: Proceedings of the 38th International Conference on Machine Learning, 5583–5594, 2021.

[77] Li, J.; Selvaraju, R. R.; Gotmare, A. D.; Joty, S.; Xiong, C.; Hoi, S. Align before fuse: Vision and language

representation learning with momentum distillation. *arXiv preprint* arXiv:2107.07651, 2021.

[78] Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive captioners are image-text foundation models. *arXiv preprint* arXiv:2205.01916, 2022.

[79] Zeng, Y.; Zhang, X.; Li, H. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint* arXiv:2111.08276, 2021.

[80] LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. In: *The hand book of brain theory and neural networks.* Arbib, M. A. Eds. MIT Press, 255–258, 2022.

[81] Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint* arXiv:2002.05709, 2020.

[82] Wang, W.; Xie, E.; Li, X.; Fan, D. P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media* Vol. 8, No. 3, 415–424, 2022.

[83] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B.; Zhang, Q.; Yang, Y.; et al. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint* arXiv:2103.14030, 2021.

[84] Guo, M. H.; Xu T. X.; Liu, J. J.; Liu, Z. N.; Jiang, P. T.; Mu, T. J.; Zhang, S. H.; Martin, R. R.; Cheng. M. M.; Hu, S. M. Attention mechanisms in computer vision: A survey. *Computational Visual Media* Vol. 8, No. 3, 331–368, 2022.

[85] Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. SimMIM: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9653–9663, 2022.

[86] Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; Kong, T. iBOT: Image BERT pre-training with online tokenizer. *arXiv preprint* arXiv:2111.07832, 2021.

[87] Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 353–355, 2018.

[88] Kong, X.; Zhang, X. Understanding masked image modeling via learning occlusion invariant feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6241–6251, 2023.

Hangbo Bao is currently a Ph.D. student with the Department of Computer Science, Harbin Institute of Technology, where he also obtained his Bachelor of Science degree in 2017. His main research interests include multimodal learning, representation learning, natural language processing, and computer vision.

Li Dong is a researcher at Microsoft Research. Previously, he received his Ph.D. degree in School of Informatics at University of Edinburgh. He has been focusing on large-scale self-supervised learning across tasks, languages, and modalities. Li's research has been recognized through the AAAI/ACM SIGAI Doctoral Dissertation Award Runner Up, the ACL-2018 Best Paper Honourable Mention, the AAAI-2021 Best Paper Runner Up, and fellowship from Microsoft. He also served as an area chair for NeurIPS, ACL, EMNLP, and NAACL multiple times.

Songhao Piao received his Ph.D. degree from the Harbin Institute of Technology (HIT), in 2004, where he is currently a professor and a doctor supervisor with the School of Computer Science and Technology. From 2006 to 2009, he held a postdoctoral position in national key technology in robot technology and system at HIT. His research interests mainly include robot intelligence control, pattern recognition, motion planning, and robot vision.

Furu Wei is a partner research manager at Microsoft Research, where he leads and oversees research on Foundation Models (across tasks, languages and modalities) and AGI, NLP, MT, Speech and Multimodal AI. More recently, he has also been driving the mission-focused research on AGI, focusing on fundamental research of the Foundation of AGI. Furu received his B.S. and Ph.D. degrees in computer science from Wuhan University in 2004 and 2009, respectively. He was a staff researcher at IBM Research - China (IBM CRL) from Jul. 2009 to Nov. 2010, and a research assistant at Department of Computing, The Hong Kong Polytechnic University from Jan. 2007 to Jun. 2009.

To submit a manuscript, please go to `https://jcvm.org`.