UFC-BERT: Unifying Multi-Modal Controls for Conditional Image Synthesis

Anonymous Author(s)
Affiliation
Address
email

Abstract

Conditional image synthesis aims to create an image according to some multi-modal guidance in the forms of textual descriptions, reference images, and image blocks to preserve, as well as their combinations. In this paper, instead of investigating these control signals separately, we propose a new two-stage architecture, UFC-BERT, to unify any number of multi-modal controls. In UFC-BERT, both the diverse control signals and the synthesized image are uniformly represented as a sequence of discrete tokens to be processed by Transformer. Different from existing two-stage autoregressive approaches such as DALL-E and VQGAN, UFC-BERT adopts non-autoregressive generation (NAR) at the second stage to enhance the holistic consistency of the synthesized image, to support preserving specified image blocks, and to improve the synthesis speed. Further, we design a progressive algorithm that iteratively improves the non-autoregressively generated image, with the help of two estimators developed for evaluating the compliance with the controls and evaluating the fidelity of the synthesized image, respectively. Extensive experiments on a newly collected large-scale clothing dataset M2C-Fashion and a facial dataset Multi-Modal CelebA-HQ verify that UFC-BERT can synthesize high-fidelity images that comply with flexible multi-modal controls.

1 Introduction

Conditional image synthesis aims to create an image according to the given control signals. With the increasing demand for flexible conditional image synthesis, various kinds of control signals have been introduced into this field, which can be divided into three main modalities: (i) *textual controls (TC)*, including the class labels [1] and natural language descriptions [62, 54]; (ii) *visual controls (VC)*, such as a spatially-aligned sketch map for reference [17, 60] or another image for style transfer [15, 27]; (iii) *preservation controls (PC)*, which require the synthesized image to preserve some given image blocks, e.g., image outpainting and inpainting [63, 67].

However, control signals of various modalities possess different characteristics. Existing works [62, 26, 61] hence typically design separate methods customized for each control modality. Moreover, most of these approaches only utilize one type of control signal and cannot simultaneously combine multiple types of controls in a concise and versatile model. This begs the question: can we integrate any number of multi-modal control signals into a unified framework for flexible conditional image synthesis? There are two inevitable challenges in this setting: (i) how to unify the multi-modal controls and represent them in a unified form, especial when employing multiple control signals from different modalities concurrently; (ii) how to guarantee the fulfillment of the multi-modal controls while ensuring the fidelity of the synthesized image.

Recently, two-stage image synthesis [42, 48, 3, 13, 47] has made great progress. The first stage learns a convolutional autoencoder with quantized latent representations for converting an image into a

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

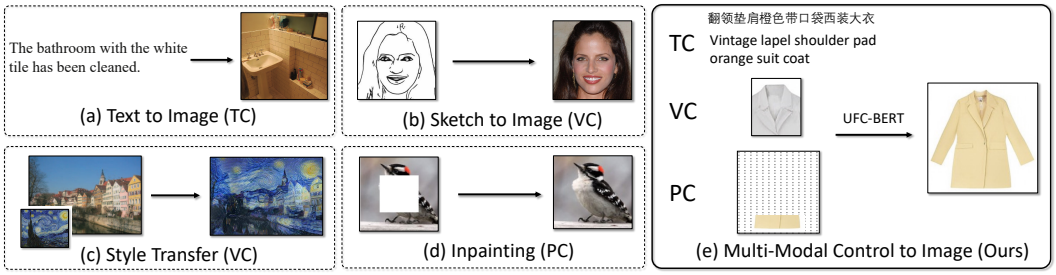

Figure 1: The three main modalities of control signals for conditional image synthesis: Textual Controls (TC), Visual Controls (VC), and Preservation Controls (PC).

sequence of discrete tokens, e.g., for compressing a 256×256 image into a sequence of 32×32 tokens where each token correlates mainly with an 8×8 block of the image. Converting a sequence of tokens back into an image is also supported. The second stage then typically adopts an autoregressive model, e.g., PixelCNN [41] or a unidirectional Transformer decoder [55], to capture the distribution over sequences of tokens. Particularly, the Transformer-based methods [3, 13, 47] exploit the global expressivity of Transformer to capture long-range relationships between local constituents.

In this paper, we make two key observations about the two-stage framework. First, the two-stage framework has the advantage that it can potentially unify the multi-modal control signals and the generated image into a single sequence of discrete tokens. However, existing works [42, 48, 13, 47] largely neglect this advantage of the two-stage framework over the traditional one-stage approaches such as those based mainly on the generative adversarial networks (GAN) [18]. Second, the *autoregressive (AR)* approach to sequence generation, adopted by the existing two-stage methods such as DALL-E [47] and VQGAN [13], brings undesirable shortcomings: (i) the token-by-token synthesis procedure leads to slow generation speed, especially for the heavyweight Transformer [3, 13, 47]; (ii) each generated token can only catch sight of the previously generated tokens and cannot incorporate bidirectional contexts, which may affect the holistic consistency of image synthesis; (iii) the fixed left-to-right order of autoregressive decoding cannot respond to the preservation control signals unless the image blocks to be preserved are at the beginning of the sequence. Notably, different from AR generation, *non-autoregressive (NAR)* sequence generation with bidirectional Transformer, i.e., BERT [9], can naturally avoid the three shortcomings.

Based on the aforementioned observations, we propose UFC-BERT, a novel BERT-based two-stage framework to **U**ni**F**y any number of multi-modal **C**ontrols for conditional image synthesis. Concretely, the textual, visual, and preservation control signals, as well as the generated image, are uniformly represented as a sequence of discrete tokens, as shown in Figure 2. The textual control consists of word tokens for class labels or natural language descriptions. The visual control(s) and the generated image are both represented as discrete tokens due to the first stage, where each token corresponds to a block within the reference image(s) or the generated image. Zero, one, or more reference images are supported. To preserve a given image block within the generated image, we encode the given image block into discrete tokens and fix corresponding parts of the generated sequence to the tokens.

We train UFC-BERT via the masked sequence modeling task, which predicts a masked subset of the target image's tokens conditioned on both the multi-modal control signals and the generation target's unmasked tokens. During inference, we adopt Mask-Predict, a NAR generation algorithm [16, 21, 7], which predicts all target tokens at the first iteration and then iteratively re-mask and re-predict a subset of tokens with low confidence scores. To further improve upon the NAR generation algorithm, we exploit the discriminative capability of the BERT architecture [11, 68] and add two estimators (see Figure 2), where one estimator estimates the relevance between the generated image and the control signals, and the other one estimates the image's fidelity. The two estimators help improve the quality of the synthesized image, because at each iteration we can generate multiple samples and keep only the highly-scored one before starting the next iteration. The two estimators also help save the number of iterations needed, since the algorithm can dynamically terminate if running for more iterations no longer improves the scores.

The extensive experiments on M2C-Fashion, a newly collected clothing dataset with tens of millions of image-text pairs, as well as on Multi-Modal CelebA-HQ [28, 61], a public facial dataset, demonstrate UFC-BERT can synthesize high-quality images that comply with various multi-modal controls.

2 Related Works

We have discussed the connection between our work and **Two-Stage Image Synthesis** in the Introduction. In this section, we further discuss related works from other fields.

Conditional Image Synthesis. A variety of control signals have been introduced into conditional image synthesis. The class-conditional generation task [1, 39] adopts class labels as control signals. The text-to-image synthesis task [64, 65, 30, 62, 70, 54] further employs natural language descriptions as controls. The image-to-image translation task generates photo-realism images from visual controls, such as a sketch map [17, 60], semantic label map [26, 4, 58], human pose [37] or another image for style transfer [15, 27]. Moreover, image outpainting and inpainting [25, 63] can be regarded as image synthesis conditioned on preservation control signals, where some image blocks of the desired image are already specified and need to be preserved in the generated image. However, these works only utilize one kind of control signal and design their methods customized for each kind of control. Text-guided image manipulation [10, 40, 67, 31, 61] semantically edits an image, where the text description and the original image serve as control signals. But they still fail to unify multiple modalities in a universal form and cannot easily extend to more control modalities. To promote versatility and extensibility, we propose UFC-BERT to unify any number of multi-modal controls.

Visual-Language Transformer. With great progress in language tasks [55, 44, 45, 2], the transformer architecture is being rapidly transferred to other fields such as vision [3, 12] and audio [6]. Recently, pretraining visual-language transformer [43, 24, 68, 7, 35, 53] (e.g. multi-modal BERT) has achieved significant improvements on a variety of downstream tasks, e.g. visual question answering, image captioning [68], and text-to-image generation [7]. Among them, the single-stream architecture [52, 32, 46, 5, 43, 24] uses a single transformer to jointly model a pair of text and image, while the two-stream architecture [35, 36, 53] applies two transformers to separately learn the representations of the text and the image, respectively. Our UFC-BERT is also a variant of the single-stream visual-language transformer, but focuses on flexible multi-modal image synthesis instead of multi-modal pretraining.

Non-Autoregressive Sequence Generation. Though it is natural to autoregressively predict tokens from left to right when generating a sequence, autoregressive decoding suffers from the slow speed and sequential error accumulation. Thus, the non-autoregressive generation (NAR) paradigm is proposed to avoid these drawbacks in neural machine translation [19, 20, 29, 16], image captioning [14, 22, 68], and speech synthesis [50, 49]. These approaches often employ the bidirectional Transformer (i.e. BERT) as it is not trained with a specific generation order. Our progressive NAR generation algorithm improves upon the Mask-Predict non-autoregressive algorithm [57, 16, 38, 33], by introducing the relevance estimator and the fidelity estimator to facilitate sample selection and dynamic termination.

3 UFC-BERT For Multi-Modal Image Synthesis

3.1 Background: Two-Stage Image Synthesis

In this section, we review the two-stage architecture [42, 48, 13, 47] for image synthesis.

At the first stage, a codebook $\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^{K}$ for vector quantization is learned, where $\mathbf{z}_k \in \mathbb{R}^{n_z}$ is the k-th code-word in the codebook and K is the number of code-words. An image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ can be transformed into (or from) a collection of code-words $\mathbf{Z} \in \mathbb{R}^{h \times w \times n_z}$. Concretely, a convolutional encoder E first encodes the original image $\mathbf{X}$ as $\hat{\mathbf{Z}} = E(\mathbf{X}) \in \mathbb{R}^{h \times w \times n_z}$. Then an element-wise quantization step $\mathbf{q}(\cdot)$ is applied to each element $\hat{\mathbf{Z}}_{ij}$ to obtain the element's closest code-word $\mathbf{z}_k$, i.e., $\mathbf{q}(\hat{\mathbf{Z}}_{ij}) = \arg\min_{\mathbf{z}_k \in \mathcal{Z}} \|\hat{\mathbf{Z}}_{ij} - \mathbf{z}_k\|$. For reconstruction, a convolutional decoder D is also learned for recovering image $\hat{\mathbf{X}} \in \mathbb{R}^{H \times W \times 3}$ from $\mathbf{Z}$ such that $\hat{\mathbf{X}}$ is close to $\mathbf{X}$. The first stage can be denoted by

$$\mathbf{Z} = \mathbf{q}(E(\mathbf{X})),\ \hat{\mathbf{X}} = D(\mathbf{Z}). \tag{1}$$

Due to the convolutional layers, each of the $h \times w$ elements of $\hat{\mathbf{Z}}$ mainly correlates with an $\frac{H}{h} \times \frac{W}{w}$ block of the image, though its receptive field may be larger if multiple convolutions are stacked.

At the second stage, image $\mathbf{X}$'s quantized representation $\mathbf{Z}$ can be rewritten as a sequence of codes $\mathbf{I} \in \{0, \ldots, |\mathcal{Z}|-1\}^{N_I}$, composed of N_I $(= h \times w)$ indices from the codebook $\mathcal{Z}$. Thus, image synthesis can be formulated as autoregressive sequence generation, i.e. predicting the distribution $\Pr(I_i | \mathbf{I}_{<i}, \mathbf{C})$ of the next token I_i conditioned on the preceding tokens $\mathbf{I}_{<i}$ and the control signals $\mathbf{C}$. The

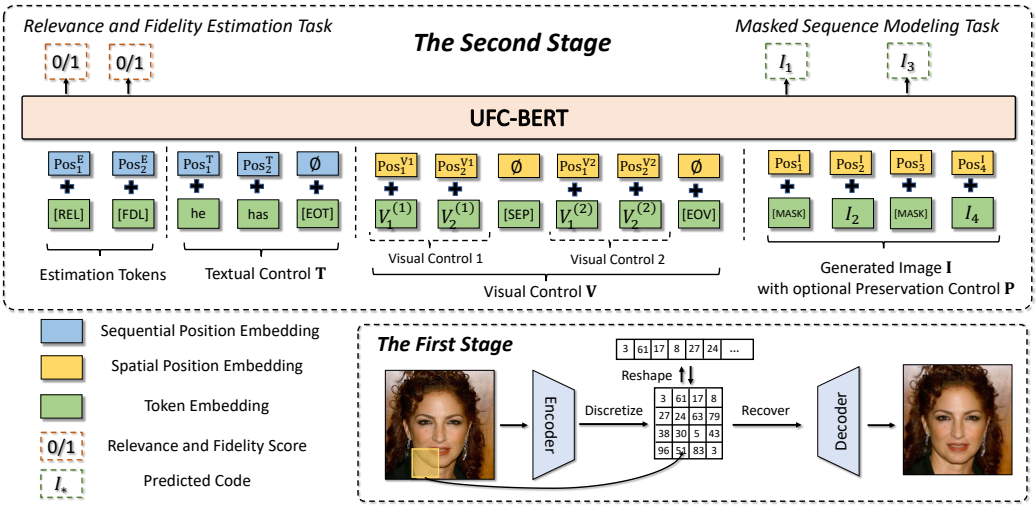

Figure 2: The framework of UFC-BERT, where the textual control (TC), vsiual control (VC), and preservation control (PC), as well as the image to generate, collectively form a sequence of tokens.

distribution is typically modeled using a unidirectional Transformer. The likelihood is then $\Pr(\mathbf{I}|\mathbf{C}) = \prod_i \Pr(I_i|\mathbf{I}_{<i}, \mathbf{C})$. Parameters are learned by minimizing $\mathcal{L}_{\text{AR}} = \mathbb{E}_{\mathbf{I}\sim data}\left[-\log \Pr(\mathbf{I}|\mathbf{C})\right]$.

We focus on improving the second stage. Specifically, the autoregressive paradigm adopted by the existing two-stage works [42, 48, 13, 47] suffers from slow generation speed, fails to capture bidirectional contexts, and cannot fully support preservation control signals. We thus propose UFC-BERT, a novel NAR approach for stage two, to unify any number of multi-modal controls and tackle the shortcomings of AR. As for stage one, we directly follow VQGAN's design [47], which improves upon VQVAE [42] by incorporating a perceptual loss [27] and patch-based adversarial training [26].

3.2 Problem Formulation

Conditional image synthesis aims to generate an image that satisfies a set of control signals $\mathbf{C}$. We consider three major modalities of control signals. A *Textual Control (TC)* consists of a sequence of words $\mathbf{T} \in \{0, \ldots, |\mathcal{W}| - 1\}^{N_T}$, where $\mathcal{W}$ is the vocabulary and N_T is the number of words in the text. In the two-stage framework, an image can be converted into a sequence of code-words (i.e. tokens) based on stage one's encoder E and codebook $\mathcal{Z}$. Thus, a *Visual Control (VC)* is denoted by a sequence $\mathbf{V} \in \{0, \ldots, |\mathcal{Z}| - 1\}^{N_V}$ consisting of code-words from the codebook $\mathcal{Z}$, where N_V is the sequence length. Similarly, the target (i.e., the image to generate) is a sequence of code-words $\mathbf{I} \in \{0, \ldots, |\mathcal{Z}| - 1\}^{N_I}$. We support zero, one, or multiple visual controls for flexibility. As for the *Preservation Control (PC)*, it is a sequence of binary masks $\mathbf{P} \in \{0, 1\}^{N_I}$ with the same length as $\mathbf{I}$, where 1 means that the token is known (i.e., I_i is ground-truth if $P_i = 1$) while 0 means the token needs to be predicted. We aim to design a model at the second stage to synthesize the target image's sequence $\mathbf{I}$ conditioned on $\mathbf{C}$, i.e., a combination of any number of control signals from $\{\mathbf{T}, \mathbf{V}, \mathbf{P}\}$.

3.3 Model Inputs

As shown in Figure 2, our UFC-BERT modifies the original BERT model [9] to accommodate any number of multi-modal controls. Similar to BERT, the backbone is a multi-layer bidirectional Transformer encoder, enabling the dependency modeling between all input elements. The input sequence of UFC-BERT always starts with two special tokens [REL] and [FDL] for relevance estimation and fidelity estimation, then goes on with the word sequence $\mathbf{T}$ of textual controls, code sequence $\mathbf{V}$ of visual controls, and ends with the code sequence $\mathbf{I}$ of the target image to generate. Two special separation tokens [EOT] and [EOV] are appended to the end of the textual and the visual control sequences, respectively. If there are multiple visual controls, another special token [SEP] is inserted to separate them. The sequence $\mathbf{I}$ of the target image to generate may be partially or fully masked by a special token [MASK]. When the preservation control $\mathbf{P}$ is present and $P_i = 1$, token I_i in $\mathbf{I}$ is always set to the code-word corresponding to the given image block to be preserved.

Each input token's representation is the sum of the position and token embeddings:

Position Embedding. Our UFC-BERT learns independent sets of position embeddings for the different kinds of the inputs to achieve better distinguishment between the various modalities. The position embeddings for the word sequence are the same as BERT, i.e., we use sequential position embeddings. For a visual control or the target image, the position embedding of each token is decided according to where this token lies on the $h \times w$ grid, i.e., we use spatial position embeddings.

Token Embedding. For textual controls, we use Byte-Pair Encoding [51] to segment each word into sub-words and then learn sub-word embeddings. Each special token, e.g., [REL] or [MASK], is assigned a dedicated embedding. For visual controls and the target image, we learn an embedding for each code-word. We do not directly use the embeddings from stage one's codebook due to the decoupling of the two stages.

3.4 Training: Masked Sequence Modeling with Relevance and Fidelity Estimation

As shown in Figure 2, we train UFC-BERT via masked sequence modeling, i.e., predicting the masked tokens in the target image conditioned on the controls. A relevance estimator and a fidelity estimator are also trained in the process, and will be key to our progressive NAR generation algorithm.

Task 1: Masked Sequence Modeling. This task is similar to Masked Language Modeling (MLM) in BERT, but incorporates multi-modal control signals when predicting the masked tokens. To construct training samples, we mask parts of the target image $\mathbf{I}$ to predict using four strategies: (1) randomly decide the number of tokens to mask, and then randomly mask the desired number of tokens; (2) mask all tokens; (3) mask the tokens within some boxed areas of the image, where the number of boxes and the box sizes are randomly decided; (4) mask the tokens outside some random boxed areas of the image. We use the four strategies with probability 0.70, 0.10, 0.10, and 0.10, respectively. To construct multi-modal control signal $\mathbf{C}$ for each training sample, there are four different combinations: *<TC, VC>, <TC>, <VC>, <empty>*, where *<TC, VC>* means the textual and visual controls are simultaneously employed, *<TC>* or *<VC>* means only a textual or visual signal is used, and *<empty>* means no textual or visual control is present. Note that the preservation control is already included in the masked sequence modeling task. Since our dataset does not contain ground-truth pairs of visual controls and target images, we crop one or multiple regions of a target image to construct VC for the target image. Because image synthesis from solely textual controls is more challenging than from other signals, we use the four combinations with probability 0.20, 0.55, 0.20, 0.05, respectively, where textual controls get more attention. We feed UFC-BERT's outputs at each position of $\mathbf{I}$ into a softmax classifier over the codebook $\mathcal{Z}$, which produces a probability score $Y_i = \Pr(I_i|\mathbf{I}_U, \mathbf{C})$ for each position $i \in M$, where M is the set of masked positions and U is the unmasked set. Finally, the masked sequence modeling task minimizes the softmax cross-entropy loss $\mathcal{L}_{\text{MSM}} = \mathbb{E}_{\mathbf{I}_M, \mathbf{I}_U}\left[-\log \Pr(\mathbf{I}_M|\mathbf{I}_U, \mathbf{C})\right]$, where $\Pr(\mathbf{I}_M|\mathbf{I}_U, \mathbf{C}) = \prod_{i \in M} Y_i$.

Task 2: Relevance Estimation. This task is to learn a binary classifier that judges whether the generated image is relevant or irrelevant to the given multi-modal control $\mathbf{C}$. Briefly, we add a linear layer on the output corresponding to the special token [REL]. The linear layer outputs a scalar representing the logit, and a binary cross-entropy loss $\mathcal{L}_{\text{REL}}$ is added. During training, the training samples from Task 1 serve as the positive instances (i.e. relevant pairs). We construct negative instances (i.e. irrelevant pairs) by swapping the control signals of two training samples.

Task 3: Fidelity Estimation. This task aims to distinguish whether the generated image is realistic from the view of human visual cognition. Similar to relevance estimation, we feed the output corresponding to [FDL] into a linear layer for binary classification and add another binary cross-entropy loss $\mathcal{L}_{\text{FDL}}$. Since the low-fidelity images (i.e. negative instances) do not exist in the dataset, we run UFC-BERT from previous epochs to synthesize images based solely on textual control signals, and use the synthesized images as negative instances.

We combine the three tasks' losses to train UFC-BERT, i.e.,

$$\mathcal{L}_{\text{UFC-BERT}} = \lambda_1 \mathcal{L}_{\text{MSM}} + \lambda_2 \mathcal{L}_{\text{REL}} + \lambda_3 \mathcal{L}_{\text{FDL}}, \tag{2}$$

where λ_1, λ_2 and λ_3 are set to 1.0, 0.5, and 0.5 to balance the three losses. The masked sequence modeling task ignores the negative instances from the other two tasks, i.e., irrelevant pairs or unrealistic instances. And the fidelity estimation task is added only after a certain number of epochs.

3.5 Inference: Progressive Non-Autoregressive Generation

We design a Progressive Non-Autoregressive Generation (PNAG) algorithm for conditional image synthesis after training, which improves upon Mask-Predict [16, 21, 7]. Mask-Predict predicts all target tokens when given a fully-masked sequence at the first iteration, and then iteratively re-mask and re-predict a subset of tokens with low probability scores for a constant number of iterations. However, Mask-Predict cannot ensure the efficacy of multi-modal controls and the fidelity of the synthesized images, and requires determining the number of iterations. Our PNAG tackles its drawbacks via sample selection and dynamic termination, based on the relevance and fidelity estimators.

Each iteration of our PNAG algorithm consists of a *Mask* step and then a *Predict* step. Let $\mathbf{I}^{(t,in)} = (I_1^{(t,in)}, \ldots, I_{N_I}^{(t,in)})$ and $\mathbf{I}^{(t,out)} = (I_1^{(t,out)}, \ldots, I_{N_I}^{(t,out)})$ be the state of the target image's sequence before and after the t-th iteration, respectively. The tokens in $\mathbf{I}^{(0,out)}$ for $t = 0$ is all set to [MASK] except for the positions that are controlled by the preservation signals, i.e., except for $I_i^{(0,out)}$ that has $P_i = 1$. If a preservation control is present, i.e. $P_i = 1$, we always set $I_i^{(t,in)}$ and $I_i^{(t,out)}$ for all t to be the code-word that corresponds to the provided image block to be preserved.

Mask Step. At the beginning of iteration t ($t \geq 1$), we construct the input sequence $\mathbf{I}^{(t,in)}$ by (re-)masking a subset of tokens in the generated sequence $\mathbf{I}^{(t-1,out)}$ from the last iteration. Similar to beam search, we construct B parallel input sequences $\{\mathbf{I}_1^{(t,in)}, \ldots, \mathbf{I}_B^{(t,in)}\}$ at each iteration. Specifically, to produce each $\mathbf{I}_b^{(t,in)}$, we mask n tokens of $\mathbf{I}^{(t-1,out)}$ by sampling from a multinomial distribution proportional to the probability scores $\mathbf{Y}^{(t-1)} = \{Y_i^{(t-1)}\}_{i=1}^{N_I}$ (see Equation 3). Here $n = N_I \cdot (\alpha + \frac{T-t}{T-1} \cdot \beta)$, where α, β are two hyper-parameters and T is the maximum possible number of iterations, such that the number of tokens to re-mask gradually decreases after every iteration.

Predict Step. Given the control $\mathbf{C}$ and an input sequence $\mathbf{I}_b^{(t,in)}$, UFC-BERT estimates a distribution $\Pr(\hat{I}_i | \mathbf{I}_b^{(t,in)}, \mathbf{C})$ for each masked position i. UFC-BERT also estimates the relevance score S_b^R and fidelity score S_b^F regarding the image that it is about to synthesize, and summarizes the scores into a comprehensive score $S_b^{(t)} = \sigma S_b^R + (1-\sigma) S_b^F$, where σ is a coefficient for adjusting the importance of the two. We perform *sample selection* based on $S_b^{(t)}$, i.e., we select the b-th sequence $\mathbf{I}_b^{(t,in)}$ with the highest $S_b^{(t)}$, discard the others, and then generate $\mathbf{I}^{(t,out)}$ based on the selected $\mathbf{I}_b^{(t,in)}$ as follows:

$$I_i^{(t,out)} \sim \Pr(\hat{I}_i | \mathbf{I}_b^{(t,in)}, \mathbf{C}), \qquad Y_i^{(t)} \leftarrow \Pr(\hat{I}_i = I_i^{(t,out)} | \mathbf{I}_b^{(t,in)}, \mathbf{C}), \tag{3}$$

where each token $I_i^{(t,out)}$ is sampled from the multinomial distribution $\Pr(\hat{I}_i | \mathbf{I}_b^{(t,in)}, \mathbf{C})$ and the corresponding probability is assigned to $Y_i^{(t)}$. Note that we predict tokens for *all* masked positions regardless of the predictions' confidence. We also implement *dynamic termination* based on $S_b^{(t)}$. Specifically, if the current iteration's score $S_b^{(t)}$ is higher than S_{max} (initialized as zero), we set S_{max} to $S_b^{(t)}$ and record the current iteration as t_{max}. If S_{max} does not increase after three consecutive iterations, we select $\mathbf{I}^{(t_{max},out)}$ as the final result and terminate our generation algorithm.

4 Experiments

4.1 Datasets and Hyperparameters

In experiments, we focus on two practical fields of image synthesis: fashionable clothing and human faces. We collect a very large-scale clothing dataset M2C-Fashion with Chinese text descriptions, which contains tens of millions of image-text pairs, much larger than the commonly used text-to-image datasets COCO [34] and CUB [56]. Details of the dataset are provided in the supplementary material. We additionally use another high-resolution facial dataset Multi-Modal CelebA-HQ [28, 61].

Following the model setting of VQGAN [13], we use the 256×256 image size on the two datasets and transform each image to a discrete sequence of 16×16 codes, where the codebook size $|\mathcal{Z}|$ is set to 1024. For the BERT model, we set the number of layers, hidden size, and the number of attention heads to 24, 1024, and 16, respectively. Our UFC-BERT has 307M parameters, same as the Transformer used by VQGAN. As for hyper-parameters of PNAG, we set the parallel decoding

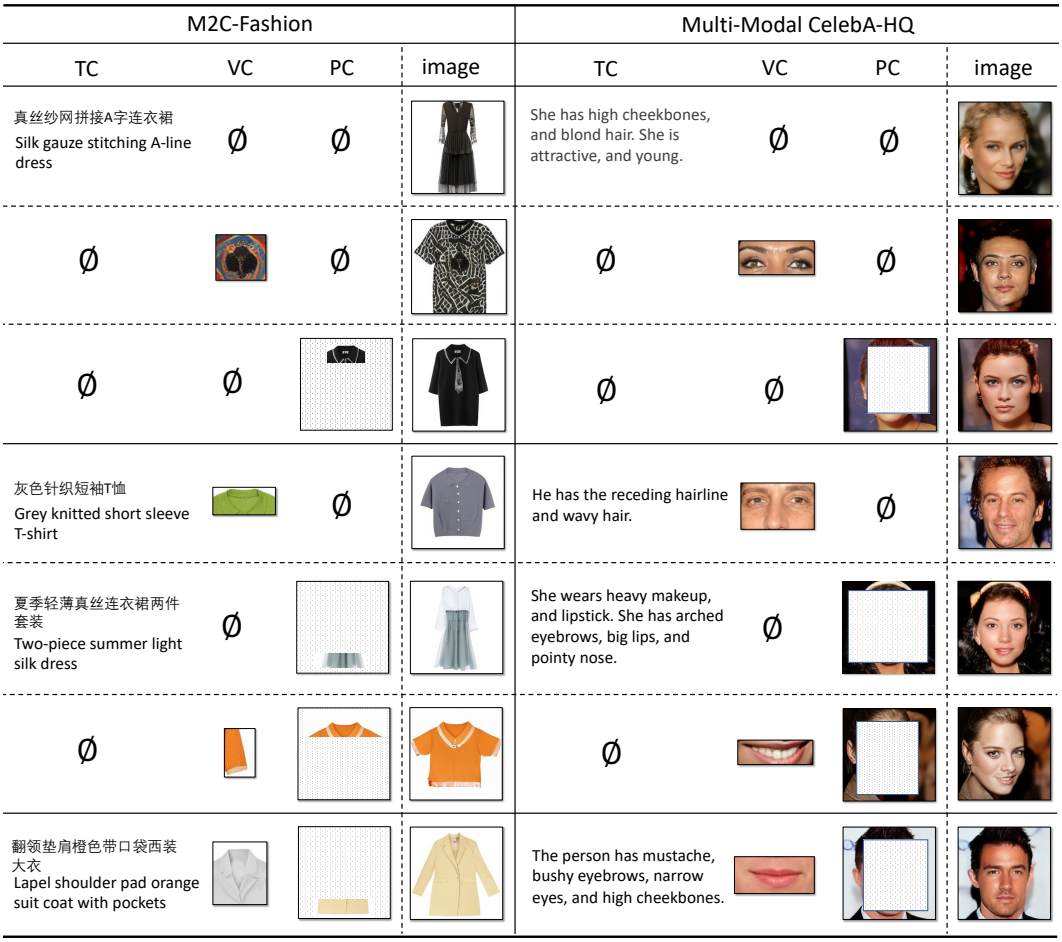

Figure 3: Images generated by our UFC-BERT under various combinations of textual controls (TC), visual controls (VC), and preservation controls (PC). Please see the supplemental material for more showcases, where we also include a study on the diversity of the images generated by UFC-BERT and analyze how the multiple control signals interfere with each other.

number B to 5 and the balance coefficient σ to 0.5. We set the initial mask ratio α, the minimum mask ratio β, and the maximum iteration number T to 0.8, 0.2, and 10, respectively.

4.2 Flexibility of Multi-Modal Controls for Conditional Image Synthesis

In this section, we qualitatively verify the synthesis ability of UFC-BERT with three modalities of control signals, i.e., textual, visual, and preservation controls. The textual controls are the texts paired with the images, which are already provided by the two datasets, while the visual controls are code sequences of cropped regions, e.g. regions that represent logos or texture of clothes.

In Figure 3, we synthesize images conditioned on combinations of the three types of control signals. The results demonstrate UFC-BERT can unify any number of multi-modal controls to synthesize high-quality images. Further, UFC-BERT supports one or multiple visual controls for more flexible synthesis, as shown in Figure 4 where we generate images given 2$\sim$3 visual controls. We observe that UFC-BERT can reasonably fuse multiple visual elements and produce a harmonious image.

4.3 Quantitative Comparison to Existing Methods for Text-to-Image Synthesis

In this section, we investigate how our UFC-BERT quantitatively compares to existing models. Considering most existing methods only utilize one control signal, we select the most common and challenging task *text-to-image synthesis* to compare the synthesis ability.

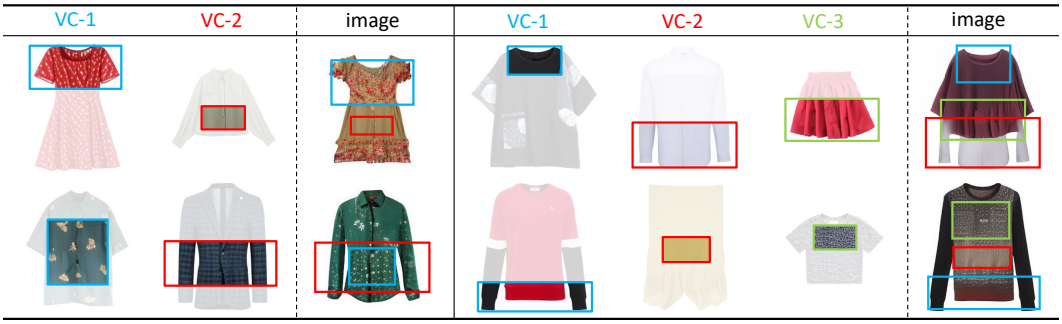

Figure 4: Image synthesis with multiple visual controls, where we crop regions from $2 \sim 3$ images to serve as the visual controls. UFC-BERT synthesizes images that naturally fuse the visual elements.

Table 1: Comparisons with GAN baselines for text-to-image synthesis on Multi-Modal CelebA-HQ.

Method	AttnGAN [62]	ControlGAN [30]	DF-GAN [54]	DM-GAN [69]	TediGAN [61]	UFC-BERT (our)
FID ↓	125.98	116.32	137.60	131.05	106.37	**66.72**
LPIPS ↓	0.512	0.522	0.581	0.544	0.456	**0.448**

Table 2: Comparisons with the autoregressive two-stage method VQGAN for text-to-image synthesis. ↓ means the lower the better, while ↑ means the opposite. We evaluate speed on the same V100 GPU.

Datasets	Methods	Automatic Metrics				Human Pairwise Study		Inference Speed
		FID↓	LPIPS ↓	PSNR↑	SSIM↑	Relevance	Fidelity	
M2C-Fashion	VQGAN (AR)	12.48	0.483	10.80	0.56	38.6%	44.2%	8.73 sec/sample
	UFC-BERT (NAR)	**11.53**	**0.461**	**13.14**	**0.58**	**61.4%**	**55.8%**	**0.81** sec/sample
Multi-Modal CelebA-HQ	VQGAN (AR)	**52.63**	0.503	8.98	0.28	42.7%	46.9%	8.66 sec/sample
	UFC-BERT (NAR)	66.72	**0.448**	**9.56**	**0.29**	**57.3%**	**53.1%**	**0.79** sec/sample

First, we compare our UFC-BERT with GAN-based text-to-image models AttnGAN [62], ControlGAN [30], DF-GAN [54], DM-GAN [69] and TediGAN [61] on the Multi-Modal CelebA-HQ dataset. For evalution, we adopt two automatic metrics FID [23] and LPIPS [66]. We report the results on Table 1 and our UFC-BERT achieves the best performance on the two metrics, even outperforming the TediGAN that uses slow and complex instance-level optimization. This demonstrates the two-stage architecture and non-autoregressive generation of UFC-BERT are suitable for text-to-image synthesis.

Besides, we compare our UFC-BERT with the autoregressive two-stage method VQGAN from three aspects: (i) the automatic metrics FID for image quality, as well as LPIPS, PSNR [59] and SSIM [59] for the similarity between the generated image and the ground truth; (ii) the Relevance and Fidelity metrics are evaluated through a user study, where the users are asked to judge which model's output is more relevant to the textual descriptions, and more photo-realistic; (iii) the synthesis speed of the two approaches. Note that the autoregressive inference implementation of VQGAN has been optimized by caching the preceding computation as in Transformer-XL [8], and UFC-BERT and VQGAN have the same parameter number (307M) for fair com-

Figure 5: Typical examples of UFC-BERT and VQGAN for text-to-image synthesis, including a counterfactual case.

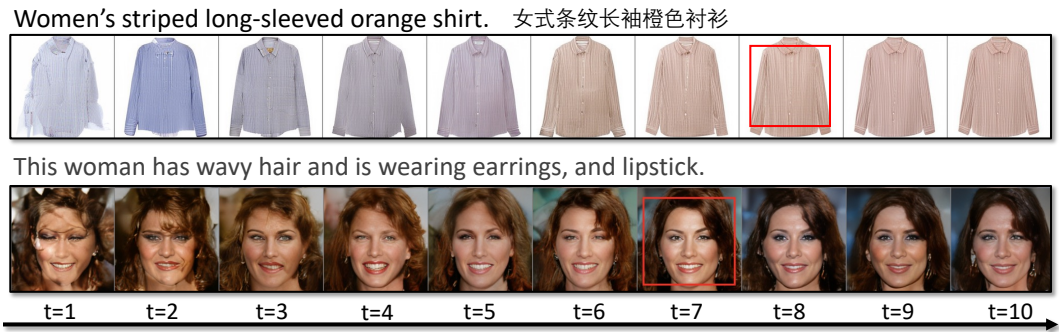

Figure 6: The iterative inference process of our PNAG algorithm. The red bounding box means the image has the highest comprehensive score and is selected as the final output result.

Table 3: Ablation studies of our PNAG inference algorithm. PNAG(w/o. REF) and PNAG(w/o. FDL) set B to the default value 5. MNAG is the original Mask-Predict algorithm [16].

Dataset	Metrics	MNAG [16]	PNAG(w/o. REF)	PNAG(w/o. FDL)	PNAG(B=1)	PNAG(B=5)	PNAG(B=10)
M2C-Fashion	FID ↓	14.77	12.17	13.14	12.72	11.53	**11.14**
	LPIPS ↓	0.488	0.477	0.469	0.479	0.461	**0.456**
Multi-Modal CelebA-HQ	FID ↓	72.04	68.90	70.32	69.49	66.72	**65.30**
	LPIPS ↓	0.514	0.469	0.463	0.475	0.448	**0.445**

parison. For the user study, the two models receive the same textual signals, and each generates 50 images. We collect the pairwise comparison results from five volunteers.

As shown in Table 2, our UFC-BERT achieves better performance for almost all criteria with about 11× speedup. This suggests our non-autoregressive UFC-BERT with progressive NAR generation algorithm can synthesize high-fidelity images relevant to textual descriptions. As for the FID metric, UFC-BERT outperforms VQGAN on M2C-Fashion, but has worse performance on Multi-Modal CelebA-HQ, it may be due to the fact that the autoregressive VQGAN can more easily memorize the pattern of a small dataset (only 30,000 facial images). In Figure 5, we further show typical generated examples to intuitively display the difference between the two approaches, including a case of counterfactual generation. We find that UFC-BERT can synthesize high-quality images, even for the counterfactual case.

4.4 The Effectiveness of Our Progressive NAR Generation Algorithm

In this section, we first visualize in Figure 6 the iterative process of our PNAG inference method based on the relevance and fidelity estimators. The images with red bounding boxes are the final outputs that match the textual control signals. We can find that the fidelity and relevance of the images increase after a few iterations, verifying our PNAG algorithm can guide the inference process towards a better direction and synthesize more realistic images that match the control signals.

We then conduct ablation studies of PNAG. As shown in Table 3, we develop three ablated inference methods PNAG(w/o. REF), PNAG(w/o. FDL) and MNAG, where PNAG(w/o. REF) and PNAG(w/o. FDL) discard the relevance estimator and the fidelity estimator, respectively, and MNAG is the original Mask-Predict method [16] without any estimator. The results demonstrate that the two estimators effectively utilize the discriminative capability of UFC-BERT and do help improve the synthesis quality. Additionally, we vary the crucial hyper-parameter of PNAG B (i.e. the parallel decoding number during inference) from 1 to 10, and the results in Table 3 show that a larger B is beneficial to the synthesis quality.

5 Conclusions

We proposed UFC-BERT to unify any number of multi-modal controls in a universal form for conditional image synthesis. We utilized non-autoregressive generation to improve inference speed, enhance holistic consistency, and support preservation controls. Further, we designed a progressive generation algorithm based on relevance and fidelity estimators to ensure relevance and fidelity.

References

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

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[3] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1520, 2017.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

[6] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[7] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-lxmert: Paint, caption and answer questions with multi-modal transformers. *arXiv preprint arXiv:2009.11278*, 2020.

[8] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference on The North American Chapter of the Association for Computational Linguistics*, 2019.

[10] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.

[11] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*, 2019.

[12] 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.

[13] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[14] Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. Masked non-autoregressive image captioning. *arXiv preprint arXiv:1906.00717*, 2019.

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[16] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

[17] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1171–1180, 2019.

[18] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[19] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

[20] Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. Non-autoregressive neural machine translation with enhanced decoder input. In *Proceedings of the American Association for Artificial Intelligence*, volume 33, pages 3723–3730, 2019.

[21] Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. Incorporating bert into parallel sequence decoding with adapters. In *Advances in Neural Information Processing Systems*, 2020.

[22] Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. *arXiv preprint arXiv:2005.04690*, 2020.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. 2017.

[24] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[25] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016.

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[28] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[29] Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic non-autoregressive neural sequence modeling by iterative refinement. *arXiv preprint arXiv:1802.06901*, 2018.

[30] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Controllable text-to-image generation. *arXiv preprint arXiv:1909.07083*, 2019.

[31] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

[32] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the American Association for Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[33] Yi Liao, Xin Jiang, and Qun Liu. Probabilistically masked language model capable of autoregressive generation in arbitrary word order. *arXiv preprint arXiv:2004.11579*, 2020.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[35] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[36] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[37] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017.

[38] Elman Mansimov, Alex Wang, Sean Welleck, and Kyunghyun Cho. A generalized framework of sequence generation with application to undirected sequence models. *arXiv preprint arXiv:1905.12790*, 2019.

[39] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.

[40] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *arXiv preprint arXiv:1810.11919*, 2018.

[41] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. *arXiv preprint arXiv:1606.05328*, 2016.

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

[43] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[44] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[46] Wasifur Rahman, Md Kamrul Hasan, Amir Zadeh, Louis-Philippe Morency, and Mohammed Ehsan Hoque. M-bert: Injecting multimodal information in the bert structure. *arXiv preprint arXiv:1908.05787*, 2019.

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[48] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019.

[49] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[50] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*, 2019.

[51] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[52] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[53] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[54] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[56] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[57] Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

[58] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[60] Weihao Xia, Yujiu Yang, and Jing-Hao Xue. Cali-sketch: Stroke calibration and completion for high-quality face image generation from poorly-drawn sketches. *arXiv preprint arXiv:1911.00426*, 2019.

[61] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse image generation and manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[62] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[63] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

[64] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.

[65] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[66] 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.

[67] Zijian Zhang, Zhou Zhao, Zhu Zhang, Baoxing Huai, and Jing Yuan. Text-guided image inpainting. In *Proceedings of the ACM International Conference on Multimedia*, pages 4079–4087, 2020.

[68] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the American Association for Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

[69] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.

[70] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see the supplemental material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The dataset and code will be released upon publication.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Some quantitative results of the baselines are directly from the previous works, which do not contain error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]