Improved Transformer for High-Resolution GANs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Attention-based models, exemplified by the Transformer, can effectively model long range dependency, but suffer from the quadratic complexity of self-attention operation, making them difficult to be adopted for high-resolution image generation based on Generative Adversarial Networks (GANs). In this paper, we introduce two key ingredients to Transformer to address this challenge. First, in low-resolution stages of the generative process, standard global self-attention is replaced with the proposed multi-axis blocked self-attention which allows efficient mixing of local and global attention. Second, in high-resolution stages, we drop self-attention while only keeping multi-layer perceptrons reminiscent of the implicit neural function. To further improve the performance, we introduce an additional self-modulation component based on cross-attention. The resulting model, denoted as HiT, has a linear computational complexity with respect to the image size and thus directly scales to synthesizing high definition images. We show in the experiments that the proposed HiT achieves state-of-the-art FID scores of 31.87 and 2.95 on unconditional ImageNet 128×128 and FFHQ 256×256, respectively, with a reasonable throughput. We believe the proposed HiT is an important milestone for generators in GANs which are completely free of convolutions.

1 Introduction

Attention-based models demonstrate notable learning capabilities for both encoder-based and decoder-based architectures [56, 58] due to their self-attention operations which can capture long-range dependencies in data. Recently, Vision Transformer [7], one of the most powerful attention-based models, has achieved a great success on encoder-based vision tasks, specifically image classification [7, 50], segmentation [29, 54], and vision-language modeling [37]. However, applying the Transformer to image generation based on Generative Adversarial Networks (GANs) is still an open problem.

The main challenge of adopting the Transformer as a decoder/generator lies in two aspects. On one hand, the quadratic scaling problem brought by the self-attention operation becomes even worse when generating pixel-level details for high-resolution images. For example, synthesizing a high definition image with the resolution of 1024×1024 leads to a sequence containing around one million pixels in the final stage, which is unaffordable for the standard self-attention mechanism. On the other hand, generating images from noise inputs poses a higher demand for spatial coherency in structure, color, and texture than discriminative tasks, and hence a more powerful yet efficient self-attention mechanism is desired for decoding feature representations from inputs.

In view of these two key challenges, we propose a novel Transformer-based decoder/generator in GANs for high-resolution image generation, denoted as HiT. HiT employs a hierarchical structure of Transformers and divides the generative process into low-resolution and high-resolution stages, focusing on feature decoding and pixel-level generating, respectively. Specifically, its low-resolution stages follow the design of Nested Transformers [61] but enhanced by the proposed multi-axis blocked self-attention to better capture global information. Assuming that spatial features are well decoded

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

after low-resolution stages, in high-resolution stages, we drop all self-attention operations in order to handle extremely long sequences for high definition image generation. The resulting high-resolution stages of HiT are built by multi-layer perceptrons (MLPs) which have a linear complexity with respect to the sequence length. Note that this design aligns with the recent findings [48, 49] that pure MLPs manage to learn favorable features for images, but it simply reduces to an implicit neural function [32, 33, 34] in the case of generative modeling. To further improve the performance, we present an additional cross-attention module that acts as a form of self-modulation [5]. In summary, this paper makes the following contributions:

- We propose HiT, a Transformer-based generator for high-fidelity image generation. Standard self-attention operations are removed in the high-resolution stages of HiT, reducing them to an implicit neural function. The resulting architecture easily scales to high definition image synthesis (with the resolution of 1024×1024) and has a comparable throughput to StyleGAN2 [21].
- To tame the quadratic complexity and enhance the representation capability of self-attention operation in low-resolution stages, we present a new form of sparse self-attention operation, namely multi-axis blocked self-attention. It captures local and global dependencies within non-overlapping image blocks in parallel by attending to a single axis of the input tensor at a time, each of which uses a half of attention heads. The proposed multi-axis blocked self-attention is efficient, simple to implement, and yields better performance than other popular self-attention operations [29, 52, 61] working on image blocks for generative tasks.
- In addition, we introduce a cross-attention module performing attention between the input and intermediate features. This module re-weights intermediate features of the model as a function of the input, playing a role as self-modulation [5], and provides important global information to high-resolution stages where self-attention operations are absent.
- The proposed HiT obtains competitive FID [11] scores of 31.87 and 2.95 on unconditional ImageNet [40] 128×128 and FFHQ [21] 256×256, respectively, highly reducing the gap between ConvNet-based GANs and Transformer-based ones. We also show that HiT not only works for GANs but also can serve as a general encoder for other models such as VQ-VAE [51]. Moreover, we observe that the proposed HiT can obtain more performance improvement from regularization than its ConvNet-based counterparts. To the best of our knowledge, these are the best reported scores for an image generator that is completely free of convolutions, which is an important milestone towards adopting Transformers for high-resolution generative modeling.

2 Related work

Transformers for image generation. There are two main streams of image generation models built on Transformers [53] in the literature. One stream of them [9, 35] is inspired by auto-regressive models that learn the joint probability of the image pixels. The other stream focuses on designing Transformer-based architecture for generative adversarial networks (GANs) [10]. This work follows the spirit of the second stream. GANs have made great progress in various image generation tasks, such as image-to-image translation [47, 66, 62] and text-to-image synthesis [59, 60], but most of them depend on ConvNet-based backbones. One recent attempt [16] builds a pure Transformer-based GAN by a careful design of attention hyper-parameters as well as upsampling layers. However, such a model is only validated on small scale datasets (e.g., CIFAR-10 [24] and STL-10 [6] consisting of 32×32 images) and does not scale to complex real-world data. A concurrent work [14] leverages the Transformer as a plugin component to build the bipartite structure to allow long-range interactions during the generative process, but its main backbone is still a ConvNet based on StyleGAN [21]. To our best knowledge, no existing work has successfully applied a Transformer-based architecture completely free of convolutions for high-resolution image generation in GANs.

Attention models. Great efforts [2, 12, 55] have been made to reduce the quadratic complexity of self-attention operation. When handling vision tasks, some works [29, 52, 61] propose to compute local self-attention in blocked images which takes advantage of the grid structure of images. The most related works to the proposed multi-axis blocked self-attention are [12, 55] where they also compute attention along axes. But our work differs notably in that we compute different attentions within heads on blocked images. Some other works [4, 14, 15] avoid directly applying standard self-attention to the input pixels, and perform attention between the input and a small set of latent

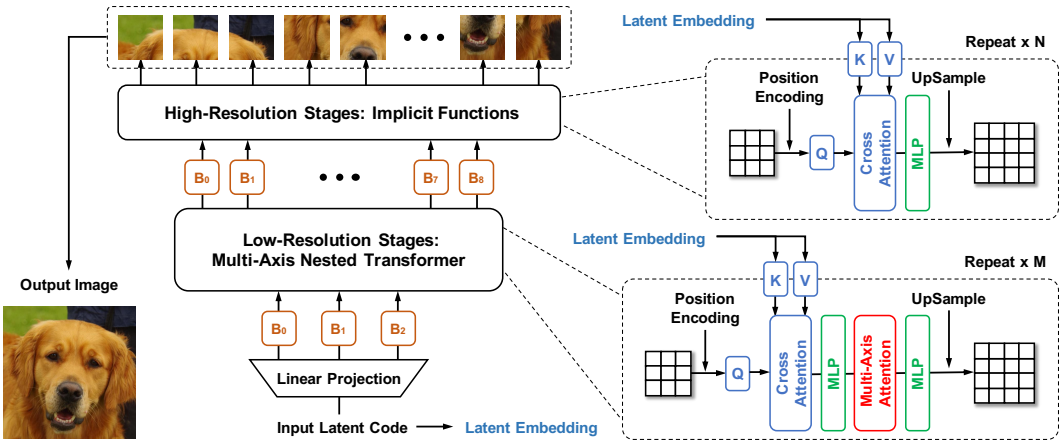

Figure 1: **HiT architecture**. In each stage, input the feature is first organized into blocks (denoted as B_i). HiT's low-resolution stages follow the decoder design of Nested Transformer [61] which is then enhanced by the proposed multi-axis blocked self-attention. We drop self-attention operations in the high-resolution stages, resulting in implicit functions. The model is further boosted by cross-attention modules which allow intermediate features to be modulated directly by the input latent code.

units. Our cross-attention module differs from them in that we apply cross-attention to generative modeling designed for pure Transformer-based architectures.

Implicit neural representations. The largest popularity of implicit neural representations/functions (INRs) is studied in 3D deep learning to represent a 3D shape in a cheap and continuous way [32, 33, 34]. Recent studies [1, 8, 26, 45] explore the idea of using INRs for image generation, where they learn a hyper MLP network to predict an RGB pixel value given its coordinates on the image grid. Among them, [1, 45] are closely related to our high-resolution stages of the generative process. One remarkable difference is that our model is driven by the cross-attention module and features generated in previous stages instead of the hyper-network presented in [1, 45].

3 Approach

3.1 Main Architecture

In the case of unconditional image generation, HiT takes a latent code $z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ as input and generates an image of the target resolution through a hierarchical structure. The latent code is first projected into an initial feature with the spatial dimension of $H_0 \times W_0$ and channel dimension of C_0. During the generation process, we gradually increase the spatial dimension of the feature map while reducing the channel dimension in multiple stages. We divide the generation stages into low-resolution stages and high-resolution stages to balance feature dependency range in decoding and computation efficiency. The overview of the proposed method is illustrated in Figure 1.

In low-resolution stages, we allow spatial mixing of information by efficient attention mechanism. We follow the decoder form of Nested Transformer [61] where in each stage, the input feature is first divided into non-overlapping blocks where each block can be consider as a local patch. After being combined with learnable positional encoding, each block is processed independently via a shared attention module. We enhance the local self-attention of Nested Transformer by the proposed multi-axis blocked self-attention that can produce a richer feature representation by explicitly considering local (within blocks) as well as global (across blocks) relations. We denote the overall architecture of these stages as multi-axis Nested Transformer.

Assuming that spatial dependency is well modeled in the low-resolution stages, the high-resolution stages can focus on synthesizing pixel-level image details purely based on the local features. Thus in high-resolution stage, we remove all self-attention modules and only maintain MLPs which can further reduce computation complexity. The resulting architecture in this stage can be viewed as a implicit neural function conditioned on the given latent feature as well as positional information.

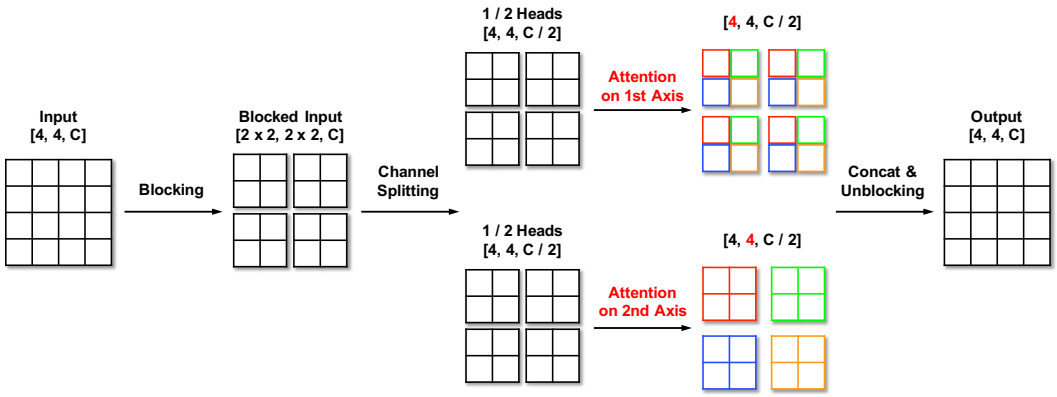

Figure 2: **Multi-axis self-attention architecture.** The different stages of multi-axis self-attention for a $[4, 4, C]$ input with the block size of $b = 2$. The input is first blocked into 2×2 non-overlapping $[2, 2, C]$ patches. Then regional and dilated self-attention operations are computed along two different axes, respectively, each of which uses a half of attention heads. The attention operations run in parallel for each of the tokens and their corresponding attention regions, illustrated with different colors. The spatial dimensions after attention are the same as the input image.

We further enhance HiT by adding a cross-attention module at the beginning of each stage which allows the network to directly condition the intermediate features on the initial input latent code. This kind of self-modulation layer leads to improved generative performance, especially when self-attention modules are absent in high-resolution stages. In the following sections, we provided detailed descriptions for the two main architectural components of HiT: (i) multi-axis blocked self-attention module and (ii) cross-attention module, respectively.

3.2 Multi-Axis Blocked Self-Attention

Different from the blocked self-attention in Nested Transformer [61], the proposed multi-axis blocked self-attention performs attention on more than a single axis. The attentions performed in two axes correspond to two forms of sparse self-attention, namely regional attention and dilated attention. Regional attention follows the spirit of blocked local self-attention [29, 52] where tokens attend to their neighbours within non-overlapped blocks. To remedy the loss of global attention, dilated attention captures long-range dependencies between tokens across blocks: it subsamples attention regions in a manner similar to dilated convolutions with a fixed stride equal to the block size. Figure 2 illustrates an example of these two attentions.

To be specific, given an input of image with size (H, W, C), it is blocked into a tensor X of the shape $(b \times b, \frac{H}{b} \times \frac{W}{b}, C)$ representing (b, b) non-overlapping blocks each with the size of $(\frac{H}{b}, \frac{W}{b})$. Dilated attention mixes information along the first axis of X while keeping information along other axes independent; regional attention works in an analogous manner over the second axis of X. They are straightforward to implement: attention over the i-th axis of X can be implemented by einsum operation which is available in most deep learning libraries.

We mix regional and dilated attention in a single layer by computing them in parallel, each of which uses a half of attention heads. Our method can be easily extended to model more than two axes by performing attention on each of the axis in parallel. Axial attention [12, 55] can be viewed as a special case of our method, where the blocking operation before attention is removed. However, we find blocking is the key to achieve significant performance improvement in the experiment. This is because the blocked structure reveals a better inductive bias for images. Compared with [12, 29, 55] where different attention modules are interleaved in consecutive layers, our approach aggregates local and global information in a single round, which is not only more flexible for architecture design but also shown to yield better performance than interleaving in our experiment.

Balancing attention between axes. In each multi-axis blocked self-attention module, the input feature is blocked in a balanced way such that we have $b \times b \approx \frac{H}{b} \times \frac{W}{b}$. This ensures that regional and dilated attention is computed on an input sequence of a similar length, avoiding half of the

attention heads are attended to a too sparse region. The proposed attention on a balanced blocked image of size $N = H \times W$ has an $\mathcal{O}(\sqrt{N})$ saving in computation over standard self-attention.

3.3 Cross-Attention for Self-Modulation

To further improve the global information flow, we propose to let the intermediate features of the model directly attend to a small tensor projected from the input latent code. This is implemented via a cross-attention operation and can be viewed as a form of self-modulation [5]. The proposed technique has two benefits. First, as shown in [5], self-modulation stabilizes the generator towards favorable conditioning values and also appears to improve mode coverage. Second, when self-attention modules are absent in high-resolution stages, attending to the input latent code provides an alternative way to capture global information when generating pixel-level details.

Formally, let X_l be the first-layer feature representation of the l-th stage. The input latent code z is first projected into a 2D spatial embedding Z with the resolution of $H_Z \times W_Z$ and dimension of C_Z by a linear function. X_l is then treated as the query and Z as the key and value. We compute their cross-attention following the update rule: $X_l' = \text{MHA}(X_l, Z + P_Z)$, where MHA represents the standard mulit-head self-attention, X_l' is the output, and P_Z is the learnable positional encoding having the same shape as Z. Note that Z is shared across all stages. For an input feature with the sequence length of N, the embedding size is a pre-defined hyperparameter which is usually far less than N (i.e., we have $H_Z \times W_Z \ll N$). Therefore, the resulting cross-attention operation has linear complexity $\mathcal{O}(N)$.

4 Experiments

4.1 Experiment Setup

Datasets. We validate the proposed method on three datasets: ImageNet [40], CelebA-HQ [18], and FFHQ [21]. ImageNet (LSVRC2012) [40] contains roughly 1.2 million images with 1000 distinct categories and we down-sample the images to 128×128 and 256×256 by bicubic interpolation. We use random crop for training and center crop for testing. This dataset is challenging for image generation since it contains samples with diverse object categories and textures. We also adopt ImageNet as the main test bed during the ablation study.

CelebA-HQ [18] is a high-quality version of the CelebA dataset [30] containing 30,000 of the facial images at 1024×1024 resolution. To align with [18], we use these images for both training and evaluation. FFHQ [21] includes vastly more variation than CelebA-HQ in terms of age, ethnicity and image background, and also has much better coverage of accessories such as eyeglasses, sunglasses, and hats. This dataset consists of 70,000 high-quality images at 1024×1024 resolution, out of which we use 50,000 images for testing and train models with all images following [21]. We synthesize images on these two datasets with the resolutions of 256×256 and 1024×1024.

Evaluation metrics. We adopt the Inception Score (IS) [42] and the Fréchet Inception Distance (FID) [11] for quantitative evaluation. Both metrics are calculated based on a pre-trained Inception-v3 image classifier [46]. Inception score computes KL-divergence between the real image distribution and the generated image distribution given the pre-trained classifier. Higher inception scores mean better image quality. FID is a more principled and comprehensive metric, and has been shown to be more consistent with human judgments of realism [11, 59]. Lower FID values indicate closer distances between synthetic and real data distributions. In our experiments, 50,000 samples are randomly generated for each model to calculate the inception score and FID on ImageNet and FFHQ, while 30,000 samples are produced for comparison on CelebA-HQ. Note that we follow [42] to split the synthetic images into groups (5000 images per group) and report their averaged inception score.

4.2 Implementation Details

Architecture configuration. In our implementation, HiT starts from an initial feature of size 8×8 projected from the input latent code. We use pixel shuffle [44] for upsampling the output of each stage, as we find using nearest neighbors leads to model failure which aligns with the observation in Nested Transformer [61]. The number of low-resolution stages is fixed to be 4 as a good trade-off between computational speed and generative performance. For generating images larger than the

Table 1: Comparison with the state-of-the-art methods on the ImageNet 128×128 dataset. † is based on a supervised pre-trained ImageNet classifier.

Method	FID ↓	IS ↑
Vanilla GAN [10]	54.17	14.01
PacGAN2 [27]	57.51	13.50
MGAN [13]	58.88	13.22
Logo-GAN-AE [41]	50.90	14.44
Logo-GAN-RC [41]†	38.41	18.86
SS-GAN (sBN) [5]	43.87	-
Self-Conditioned GAN [28]	40.30	15.82
ConvNet-R_1	39.71	18.61
HiT (Ours)	**31.87**	**21.32**

Table 2: Reconstruction FID on the ImageNet 256×256 dataset. We note that VQ-VAE-2 utilizes a hierarchical organization of VQ-VAE and thus has two codebooks $\mathcal{Z}$.

Method	Embedding size and $\lvert\mathcal{Z}\rvert$	FID ↓
VQ-VAE [51]	32, 1024	75.19
DALL-E [38]	32, 8192	34.30
VQ-VAE-2 [39]	64, 512 32, 512	10.00
VQGAN [9]	16, 1024	8.00
VQ-HiT (Ours)	16, 1024	**6.37**

Figure 3: Unconditional image generation results of HiT trained on ImageNet 128×128.

resolution of 128×128, we scale HiT to different model capacities – small, base, and large, denoted as HiT-{S, B, L}. We refer to the supplementary materials for their detailed architectures.

It is worth emphasizing that the flexibility of the proposed multi-axis blocked self-attention makes it possible for us to build smaller models than [29]. This is because they interleave different types of attention operations and thus require the number of attention blocks in a single stage to be even (at least 2). In contrast, our multi-axis blocked self-attention combines different attention outputs within heads which allows us to use an arbitrary number (e.g., 1) of attention blocks in a model.

Training details. In all the experiments, we use a ResNet-based discriminator following the architecture design of [21]. Our model is trained with a standard non-saturating logistic GAN loss with R_1 gradient penalty [31] applied to the discriminator. R_1 penalty penalizes the discriminator for deviating from the Nash-equilibrium by penalizing the gradient on real data alone. The gradient penalty weight is set to 10. Adam [22] is utilized for optimization with $\beta_1 = 0$ and $\beta_2 = 0.99$. The learning rate is 0.0001 for both the generator and discriminator. All the models are trained using TPU for one million iterations on ImageNet and 500,000 iterations on FFHQ and CelebA-HQ. We set the mini-batch size to 256 for the image resolution of 128×128 and 256×256 while to 32 for the resolution of 1024×1204. To keep our setup simple, we do not employ any training tricks for GAN training, such as progressive growing, equalized learning rates, pixel normalization, noise injection, and mini-batch standard deviation that recent literature demonstrate to be cruicial for obtaining state-of-the-art results [18, 21]. When training on FFHQ and CelebA-HQ, we utilize balanced consistency regularization (bCR) [64] for additional regularization where images are augmented by flipping, color, translation, and cutout as in [63]. bCR enforces that for both real and generated images, two sets of augmentations applied to the same input image should yield the same output. bCR is only added to the discriminator with the weight of 10.

Table 3: Comparison with the state-of-the-art methods on CelebA-HQ (**left**) and FFHQ (**right**) with the resolutions of 256×256 and 1024×1024. $\star$ bCR is not applied at the 1024×1024 resolution.

Method	FID ↓ (CelebA-HQ) ×256	×1024
VAEBM [57]	20.38	-
StyleALAE [36]	19.21	-
PG-GAN [18]	8.03	-
COCO-GAN [25]	-	9.49
VQGAN [9]	10.70	-
StyleGAN [20]	-	**5.17**
HiT-B (Ours)	**3.39**	8.83*

Method	FID ↓ (FFHQ) ×256	×1024
U-Net GAN [43]	7.63	-
StyleALAE [36]	-	13.09
VQGAN [9]	11.40	-
INR-GAN [45]	9.57	16.32
CIPS [1]	4.38	10.07
StyleGAN2 [21]	3.83	**4.41**
HiT-B (Ours)	**2.95**	6.37*

Table 4: Comparison with the main competing methods in terms of number of network parameters, throughput, and FID on FFHQ 256×256. The throughput is measured on a single Tesla V100 GPU.

Architecture	Model	#params (million)	Throughput (images / sec)	FID ↓ (FFHQ ×256)
ConvNet	StyleGAN2 [21]	30.03	95.79	3.83
INR	CIPS [1]	45.90	27.27	4.38
	INR-GAN [45]	107.03	266.45	9.57
Transformer	HiT-S	38.01	86.64	3.06
	HiT-B	46.22	52.09	2.95
	HiT-L	97.46	20.67	2.58

4.3 Results on ImageNet

Unconditional generation. We start by evaluating the proposed HiT on the ImageNet 128×128 dataset, targeting the unconditional image generation setting for simplicity. In addition to recently reported state-of-the-art GANs, we implement a ConvNet-based generator following the widely-used architecture from [58] while using the exactly same training setup of the proposed HiT (e.g., losses and R_1 gradient penalty) denoted as ConvNet-R_1 for a fair comparison. The results are shown in Table 1. We can see that HiT outperforms the previous ConvNet-based methods by a notable margin in terms of both IS and FID scores. Note that [5, 28] leverage auxiliary regularization techniques and our method is complementary to them. Examples generated by HiT on ImageNet are shown in Figure 3, from which we observe nature visual details and diversity.

Reconstruction. We are also interested in the reconstruction ability of the proposed HiT and evaluate by employing it as a decoder for the vector quantised variational auto-encoder (VQ-VAE [51]), a state-of-the-art approach for visual representation learning. In addition to the reconstruction loss and adversarial loss, our HiT-based VQ-VAE variant, namely VQ-HiT, is also trained with the perceptual loss [17] following the setup of [9, 65]. Please refer to the supplementary materials for more details on the architecture design and model training. We evaluate the metric of reconstruction FID on ImageNet 256×256 and report the results in Table 2. Our VQ-HiT attains the best performance while providing significantly more compression (i.e., smaller embedding size and fewer number of codes in the codebook $\mathcal{Z}$) than [38, 39, 51].

4.4 Higher Resolution Generation

Baselines. To demonstrate the utility of our approach for high-fidelity images, we compare to state-of-the-art techniques for generating images on CelebA-HQ and FFHQ, focusing on two common resolutions of 256×256 and 1024×1024. The main competing method of the proposed HiT is StyleGAN [20, 21] – a hypernetwork-based ConvNet archieving the best performance on these two datasets. On the FFHQ dataset, apart from our Convnet-based counterparts, we also compare to the most recent INR-based methods including CIPS [1] and INR-GAN [45]. These two models are closely related to HiT as the high-resolution stages of our approach can be viewed as a form of INR.

Figure 4: Synthetic face images by HiT-B trained on CelebA-HQ 1024×1024 and 256×256.

Table 5: Ablation study. We start with the INR-based generator [3, 23] conditioned on the input latent code and gradually improve it with the proposed attention components and their variations. O/M denotes "out-of-memory" error: the model cannot be trained for the batch size of one.

	Model configuration	#params (million)	Throughput (images / sec)	FID ↓	IS ↑
	Latent-code conditioned INR decoder [3, 23]	42.68	110.39	56.33	16.19
+	Cross-attention for self-modulation	61.55	82.67	35.94	19.42
+ one of	All-to-all self-attention [53]	67.60	-	O/M	O/M
+ one of	Axial attention [12, 55]	67.60	74.21	35.15	19.79
+ one of	Blocked local attention [52, 61]	67.60	75.54	33.70	19.96
+ one of	Interleaving blocked regional and dilated attention	67.60	75.54	32.96	20.75
+ one of	Multi-axis blocked self-attention (Ours)	67.60	75.54	32.23	20.96
+	Balancing attention between axes (Full model)	67.60	75.33	**31.87**	**21.32**

Results. We report the results in Table 3. Impressively, the proposed HiT obtains the best FID scores at the resolution of 256×256 and sets the new state of the art on both datasets. Meanwhile, our performance is also competitive at the resolution of 1024×1024 but only slightly shy of StyleGAN. This is due to our finding that conventional regularization techniques such as [64] cannot be directly applied to Transformer-based architectures for synthesizing ultra high-resolution images. We believe this triggers a new research direction and will explore it as our future work. It is also worth mentioning that our method consistently outperforms INR-based models, which suggests the importance of involving the self-attention mechanism into image generation.

Table 4 provides a more detailed comparison to our main competing methods in terms of the number of parameters, throughput, and FID on FFHQ 256×256. We find that the proposed HiT-S has a comparable runtime performance (throughput) with StyleGAN while yielding better generative results. More importantly, FID scores can be further improved when larger variants of HiT are employed. Figure 4 illustrates our synthetic face images on CelebA-HQ.

4.5 Ablation Study

We then evaluate the importance of different components of our model by its ablation on ImageNet 128×128. We start with a baseline architecture without any attention modules which reduces to a pure implicit neural function conditioned on the input latent code [3, 23]. We build upon this baseline by gradually incorporating cross-attention and self-attention modules. We compare the methods that perform attention without blocking (including standard self-attention [53], axial attention [12, 55]) and with blocking (including blocked local attention [52, 61]). We also compare a variant of our method where different types of attention are interleaved other than combined in attention heads.

The results are reported in Table 5. As we can see, the latent code conditional INR baseline cannot fit training data favourably since it lacks an efficient way to capture the latent code during the generative process. Interestingly, incorporating the proposed cross-attention module improves the baseline and achieves the state of the art, thanks to its function of self-modulation. More importantly, we find that blocking is vital for attention: performing different types of attention after blocking can all improve

Table 6: Performance as a function of the number of self-attention stages on ImageNet 128×128. The attention configuration is defined using the protocol $[a, b]$, where a and b refer to the number of stages in the low-resolution and high-resolution stages of the model, respectively.

Attention configuration	$[0, 5]$	$[1, 4]$	$[2, 3]$	$[3, 2]$	$[4, 1]$
#params (million)	61.55	66.01	67.19	67.52	67.60
Throughput (images / sec)	82.67	80.88	80.22	78.06	75.33
FID ↓	35.94	34.16	33.69	32.72	31.87

Table 7: The effectiveness of bCR [64] on both StyleGAN2 and HiT-based variants. † indicates the results of StyleGAN2 are obtained from [19] which uses a lighter-weight configuration of [21].

+ bCR [64]	StyleGAN2 [21]†	HiT-S	HiT-B	HiT-L
✗	5.28	6.07	5.30	5.13
✓	3.91	3.06	2.95	2.58
Δ FID	1.37	3.01	2.35	2.55

the performance by a notable margin while the axial attention cannot. This is due to the fact that the image structure after blocking introduces a better inductive bias for images. At last, we observe that the proposed multi-axis blocked self-attention yields better FID scores than interleaving different attention outputs, and it achieves the best performance after balancing the attention axes.

In Table 6, we investigate another form of our variations, where we incorporate attention in different number of stages across the generator. We can see that the more in the stack the attention is applied (low-resolution stages), the better the model performance will be. This study provides a validation for the effectiveness of self-attention operation in generative modeling. However, for feature resolutions larger than 64×64, we do not observe obvious performance gains brought by self-attention operations but a notable degradation in both training and inference time.

4.6 Effectiveness of Regularization

We further explore the effectiveness of regularization for the proposed Transformer-based generator. Note that different from the recent interest in training GANs in the few-shot scenarios [19, 63], we study the influence of regularization for ConvNet-based and Transformer-based generators in the full-data regime. We use the whole training set of FFHQ 256×256, and compare HiT-based variants with StyleGAN2 [19]. The regularization method is bCR [64]. As shown in Table 7, all variants of HiT achieve much larger margins of improvement than StyleGAN2. This confirms the finding [7, 16, 50] that Transformer-based architectures are much more data-hungry than ConvNets in both classification and generative modeling settings, and thus strong data augmentation and regularization techniques are crucial for training Transformer-based generators.

5 Conclusion

We present HiT, a novel Transformer-based generator for high-resolution image generation based on GANs. To address the quadratic scaling problem of Transformers, we structure the low-resolution stages of HiT following the design of Nested Transformer and enhance it by the proposed multi-axis blocked self-attention. In order to handle extremely long inputs in the high-resolution stages, we drop self-attention operations and reduce the model into implicit functions. We further improve the model performance by introducing a cross-attention module which plays a role of self-modulation. Our experiments demonstrate that HiT achieves highly competitive performance for high-resolution image generation compared with its ConvNet-based counterparts. For future work, we will investigate Transformer-based architectures for discriminators in GANs and efficient regularization techniques for Transformer-based generators to synthesize ultra high-resolution images. We note that the synthetic image generation techniques have the potential to mitigate bias and privacy issues for data collection and annotation. However, such techniques could be misused to produce misleading information, and researchers should explore the techniques responsibly.

References

[1] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. Image generators with conditionally-independent pixel synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[3] Tristan Bepler, Ellen D Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. Explicitly disentangling image content from translation and rotation with spatial-VAE. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15409–15419, 2019.

[4] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021.

[5] Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly. On self modulation for generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 215–223, 2011.

[7] 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[8] Emilien Dupont, Yee Whye Teh, and Arnaud Doucet. Generative models as distributions of functions. *arXiv preprint arXiv:2102.04776*, 2021.

[9] 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 (CVPR)*, 2021.

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6629–6640, 2017.

[12] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[13] Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. MGAN: Training generative adversarial nets with multiple generators. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[14] Drew A Hudson and C Lawrence Zitnick. Generative adversarial transformers. *arXiv preprint arXiv:2103.01209*, 2021.

[15] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

[16] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. TransGAN: Two transformers can make one strong GAN. *arXiv preprint arXiv:2102.07074*, 2021.

[17] 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 (ECCV)*, pages 694–711, 2016.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[23] Marian Kleineberg, Matthias Fey, and Frank Weichert. Adversarial generation of continuous implicit shape representations. In *Eurographics*, 2020.

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[25] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. COCO-GAN: Generation by Parts via Conditional Coordinating. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4512–4521, 2019.

[26] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. InfinityGAN: Towards infinite-resolution image synthesis. *arXiv preprint arXiv:2104.03963*, 2021.

[27] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[28] Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14286–14295, 2020.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[31] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3481–3490, 2018.

[32] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019.

[33] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020.

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019.

[35] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4055–4064, 2018.

[36] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14104–14113, 2020.

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

[38] 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.

[39] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[41] Alexander Sage, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Logo synthesis and manipulation with clustered generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5879–5888, 2018.

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[43] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A U-Net Based Discriminator for Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8207–8216, 2020.

[44] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[45] Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny. Adversarial generation of continuous images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[47] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N Metaxas. CR-GAN: Learning complete representations for multi-view generation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 942–948, 2018.

[48] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. MLP-Mixer: An all-MLP Architecture for Vision. *arXiv preprint arXiv:2105.01601*, 2021.

[49] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. ResMLP: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021.

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[51] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6309–6318, 2017.

[52] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[53] 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 (NeurIPS)*, pages 6000–6010, 2017.

[54] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[55] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 108–126, 2020.

[56] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[57] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[58] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 7354–7363, 2019.

[59] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[60] 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 (ICCV)*, pages 5907–5915, 2017.

[61] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating nested transformers. *arXiv preprint arXiv:2105.12723*, 2021.

[62] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018.

[63] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[64] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for GANs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[65] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN inversion for real image editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 592–608, 2020.

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

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] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (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] This paper does not have theoretical results.
   (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)? [No] Our code contains face synthesis techniques which could be misused to produce misleading information. Thus, we will not release our source code in this submission timeline.
   (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] Reporting variance is less common in the compared benchmarks of related literature. The experiments are very expensive to run. We have checked the results are reproducible by multiple runs and the variance is negligible.
   (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? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We cite the dataset papers and the used datasets are widely-used standard datasets already giving consent to use.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No sensitive data is used by this paper.
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] This work has no research related to human subjects.
   (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]