Adversarial Text to Continuous Image Generation

Kilichbek Haydarov Aashiq Muhamed Xiaoqian Shen Jovana Lazarevic

Ivan Skorokhodov Chamuditha Jayanga Galappaththige Mohamed Elhoseiny

{first_name.last_name}@kaust.edu.sa

King Abdullah University of Science and Technology

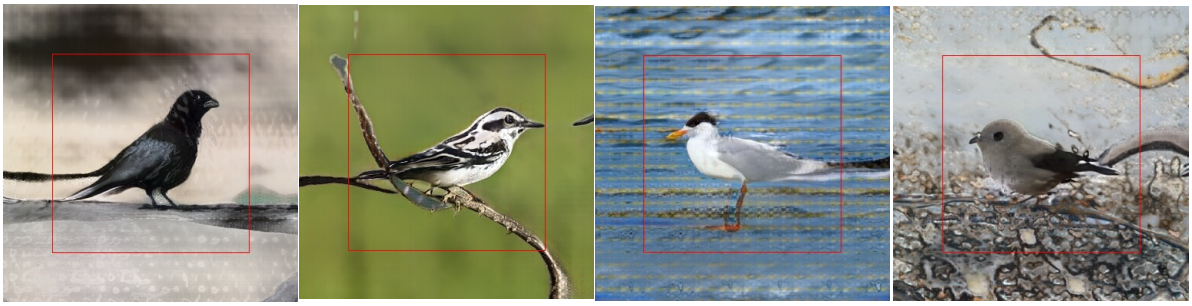

Figure 1. **Text Conditioned Extrapolation outside of Image Boundaries**: The red rectangles indicate the resolution boundaries that our HyperCGAN model was trained. By design, our model can synthesize meaningful pixels at surrounding (x, y) coordinates beyond these boundaries without any explicit training. For example, it can meaningfully extend bird images with more natural details like the tail, background, and the branch of the tree.

Abstract

Existing GAN-based text-to-image models treat images as 2D pixel arrays. In this paper, we approach the text-to-image task from a different perspective, where a 2D image is represented as an implicit neural representation (INR). We show that straightforward conditioning of the unconditional INR-based GAN method on text inputs is not enough to achieve good performance. We propose a word-level attention-based weight modulation operator that controls the generation process of INR-GAN based on hypernetworks. Our experiments on benchmark datasets show that HyperCGAN achieves competitive performance to existing pixel-based methods and retains the properties of continuous generative models. Project page link: https://kilichbek.github.io/webpage/hypercgan

1. Introduction

Humans have the innate ability to connect what they visualize with language or textual descriptions. Text-to-image (T2I) synthesis, an AI task inspired by this ability, aims to generate an image conditioned on text input. Compared to other possible inputs in the conditional generation literature, sentences are an intuitive and flexible way to express visual content that we may want to generate. The main challenge in traditional T2I synthesis lies in learning from the unstructured description and connecting the different statistical properties of vision and language inputs. This field has seen significant progress in recent years in synthesis quality, the size and complexity of datasets used as well as image-text alignment (e.g., [24, 39, 40, 42, 44, 60, 65, 69, 73]).

Despite the significant progress, images in existing T2I approaches are typically represented as a discrete 2D pixel array which is a cropped, quantized version of the true contin-

uous underlying 2D signal. In this paper, we take an alternative view, where we represent images as a continuous signal through an *Implicit Neural Representation* (INR), which provides a natural way to parameterize images using a neural network that predicts the RGB color at an (x, y) image location. Operating directly with INR naturally facilitates several benefits such as extrapolation outside of image boundaries, accelerated inference of low-resolution images, and out-of-the-box superresolution. In addition, INRs do not depend on spatial resolution, allowing for arbitrary-resolution generation while maintaining nearly constant memory requirements. In contrast, discrete-based GANs require both generator and discriminator to scale w.r.t spatial resolution, making training of such models impractical. Figure 2 shows that for discrete-based models, increasing training resolution leads to decreasing effective batch size during training due to GPU memory limits which eventually break. Current diffusion models [19, 55, 57], despite their impressive results, suffer from the same scalability limitations due to dependency on spatial resolution and slower sampling speed compared to GAN-based models. Recent works [21, 47] prove GANs can rival diffusion models when carefully scaled up.

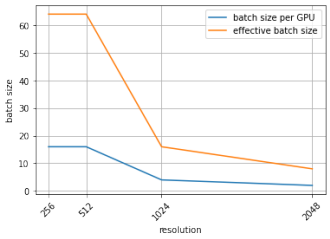

Figure 2. Scalability limitations in discrete decoders: Increasing training resolution decreases batch size/GPU hitting GPU limits. As resolution approaches the value 2048, training becomes invisible due to batch size per GPU approaching 1.

The prevalent T2I models [65, 69, 71] use architecture-specific designs to condition the generator and discriminator on textual information and often introduce additional text-matching losses. These approaches utilize text embeddings $\boldsymbol{c}$ to condition their models by updating a hidden representation $\boldsymbol{h}$. Unlike these approaches, we explore a different paradigm: we use *hypernetworks* [17] to condition the model on textual information $\boldsymbol{c}$ by modulating the model weights. Such a procedure can be viewed as creating a different instance of the model for each conditioning vector $\boldsymbol{c}$ and was recently shown to be more expressive than the embedding-based conditioning approaches [12].

Our proposed HyperCGAN introduces a hypernetwork-based conditioning mechanism for text-to-continuous image (T2CI) generation. It enables unconditional INR-GAN [53] backbone to efficiently generate continuous images conditioned on input text while preserving the desired properties of the continuous signal. A vanilla hypernetwork [6] generates the entire parameter vector $\boldsymbol{\theta}$ from the conditioning signal c, i.e. $\boldsymbol{\theta} = F(\boldsymbol{c})$, where $F(\boldsymbol{c})$ is a *modulating* hypernetwork. However, this quickly becomes infeasible in modern neural networks where $|\boldsymbol{\theta}|$ can easily span millions of parameters. To address this issue, our HyperCGAN instead produces a *tensor-decomposed* modulation $F(\boldsymbol{c}) = \boldsymbol{M}$ of the same size as the weight tensor $\boldsymbol{W}$. This tensor is then used to alter $\boldsymbol{W}$ via an element-wise multiplicative operation $\boldsymbol{W}_c = \boldsymbol{W} \odot \boldsymbol{F}(\boldsymbol{c})$. We develop an attention-based word level modulation (WHAtt) to alter weight tensors $\boldsymbol{W}$ of the INR-based decoder using $F(\boldsymbol{c})$. Figure 1 shows images generated by our HyperCGAN on CUB [64] dataset. By harnessing the power of hypernetwork-based conditioning and leveraging continuous representation via INRs, our HyperCGAN demonstrates its ability to augment bird images with enhanced natural details, e.g., the tail, background, and branches of the tree. This finding poses a promising paradigm for the future progression of generative models, i.e., the natural capability of producing images of arbitrary resolutions while maintaining visual semantic consistency at low training costs. We hope our work paves the way towards efficient conditional image generation at arbitrary resolutions. Our primary contributions are as follows:

- We propose the HyperCGAN framework for synthesizing continuous images from text input. The model is augmented with a novel language-guided mechanism termed *WHAtt*, that modulates weights at the word level.
- We show that our method has a natural ability to meaningfully extrapolate outside the image boundaries, and can outperform most existing discrete methods on CUB, COCO, and ArtEmis datasets, including stacked generators and single generator methods.
- We establish a new affective T2I benchmark based on the ArtEmis dataset [1], which has 455,000 affective utterances collected on more than 80K artworks. ArtEmis contains captions that explain emotions elicited by a visual stimulus, which can lead to more human emotion-aware T2I generative models.

2. Related Work

Text-to-image synthesis: T2I synthesis has been an active area of research since at least [32, 41] proposed a DRAW-based [15] model to generate images from captions. [41] first demonstrated improved fidelity of the generated images from text using GANs [14]. Since then, several works adopted text-conditional GANs approaches for T2I synthesis [24, 60, 65, 68, 69, 71, 73]. With the development of diffusion models [8, 19, 57], autoregressive (AR) transformers [7], and large-scale language encoders [20, 38], T2I synthesis has shown remarkable improvement in zero-shot setting. Both AR-based models (e.g. DALL-E [39] Make-A-Scene [11], CogView [9], Parti [66]) and Diffusion-based models (e.g., GLIDE [35], DALL-E 2 [40], Imagen [44], Stable Diffusion [42]) achieved remarkable results replacing

popular GAN-based architectures, but their iterative sampling process is computationally expensive to synthesize the high-quality images. Although there were attempts to accelerate the sampling process by reducing sampling steps [31, 33, 45, 56], precomputed features [25], or performing the reverse process in low-dimensional latent space instead of pixel space, the reverse process still remains time-consuming and not competitive to GANs inference speed. Recent works show GANs can still be competitive to diffusion and AR methods in zero-shot T2I generation setup by redesigning their architecture for this task [21, 47, 61]. However, these methods are still typically limited to discrete image generation and do not easily support continuous image generation.

Implicit Neural Representation (INR): INRs parametrize any type of signal (e.g. images, audio signals, 3D shapes) as a continuous function that maps the domain of the signal to values at a specified coordinate [13, 34, 51, 52]. For 2D image synthesis, several works have explored ways to enable INRs using generative models [3, 50, 53, 54]. Our goal is to enable INR-based generative models via hypernetwork-based conditioning.

Connection to hypernetworks: Hypernetworks are models that generate parameters for other models. They have been applied to several tasks in architecture search [67], few-shot learning [4], and continual learning [63]. Generative hypernetworks, also called implicit generators [3, 53] were recently shown to rival StyleGAN2 [22] in generation quality. Despite the progress in unconditional INR-based decoders (e.g., [3, 27, 53, 54]), generating high-quality continuous images conditioned on text is less studied compared to discrete image generators. Our hypernetwork-augmented modulation approach facilitates conditioning the continuous image generator on text while preserving the desired INR properties (e.g., superresolution, extrapolation).

Art generation: Synthetically generating realistic artworks with conditional GAN is challenging due to unstructured shapes and their metaphoric nature. Several works have explored learning artistic style representations by conditioning GANs on labels such as artist, genre, style, and emotion [2, 58, 59] or by learning about styles and deviating from style norms [10, 48]. We extend prior work by applying our method to the novel text-to-continuous-image generation task on the challenging ArtEmis [1] dataset, where we leverage verbal explanations as conditioning signals to achieve better human cognition-aware T2I synthesis.

3. Method

Baseline INR-based decoder. We build our approach upon the INR-based generator [53], which consists of two main modules: a hypernetwork $H(\mathbf{z})$ and an MLP model $F_{\theta(z)}(x, y)$. The hypernetwork $H(\mathbf{z})$ samples a noise vector $\mathbf{z} \sim \mathcal{N}(0, I)$ and produces two matrices $\boldsymbol{A}^\ell \in \mathbb{R}^{d^\ell_{out} \times r}$ and $\boldsymbol{B}^\ell \in \mathbb{R}^{r \times d^\ell_{in}}$, and through matrix multiplication obtains modulating matrix $\boldsymbol{W}^\ell_h = \boldsymbol{A}^\ell \times \boldsymbol{B}^\ell$ with rank r for each layer ℓ of $F_{\theta(\mathbf{z})}(x, y)$. The shared parameter matrix $\boldsymbol{W}^\ell_s \in \mathbb{R}^{d^\ell_{out} \times d^\ell_{in}}$ of the MLP model $F_{\theta(\mathbf{z})}(x, y)$ is updated via $\boldsymbol{W}^\ell = \boldsymbol{W}^\ell_s \odot \sigma(\boldsymbol{W}^\ell_h)$, where σ denotes sigmoid function. The MLP model $F_{\theta(\mathbf{z})}(x, y)$ then predicts RGB values at each location (x, y) of a predefined coordinate grid to synthesize an image $\mathbf{x}'$.

Discriminator. We adopt the discriminator proposed in LAFITE [72] because of its simplicity and effectiveness. Given the text features $\mathbf{h}$ from a text encoder, this type of discriminator outputs: $D(\mathbf{x}, \mathbf{h}) = f_d(\mathbf{x}) + \langle \mathbf{h}, f_s(\mathbf{x}) \rangle$, where $f_d(\mathbf{x})$ yields high value when image $\mathbf{x}$ is real, while inner product $\langle \mathbf{h}, f_s(\mathbf{x}) \rangle$ indicates how well the input image $\mathbf{x}$ is semantically aligned with text features $\mathbf{h}$.

Training objectives. We use the standard conditional GAN losses for the generator and discriminator:

$$\begin{aligned} \mathcal{L}_G &= -\sum_{i=1}^{n} \log \sigma(D(\mathbf{x}'_i, \mathbf{h}_i)), \\ \mathcal{L}_D &= -\sum_{i=1}^{n} \log \sigma(D(\mathbf{x}_i, \mathbf{h}_i)) - \sum_{i=1}^{n} \log(1 - \sigma(D(\mathbf{x}'_i, \mathbf{h}_i))), \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ denotes a sigmoid function. Following previous works [21, 69, 72], in order to increase text and image alignment, we use the following contrastive regularizers:

$$\begin{aligned} \mathcal{L}_{ConD} &= -\tau \sum_{i=1}^{n} \log \frac{\exp(\cos(f_s(\mathbf{x}_i), \mathbf{h}_i)/\tau)}{\sum_{j=1}^{n} \exp(\cos(f_s(\mathbf{x}_j), \mathbf{h}_i)/\tau)}, \\ \mathcal{L}_{ConG} &= -\tau \sum_{i=1}^{n} \log \frac{\exp(\cos(f_I(\mathbf{x}'_i), \mathbf{h}_i)/\tau)}{\sum_{j=1}^{n} \exp(\cos(f_I(\mathbf{x}'_j), \mathbf{h}_i)/\tau)}, \end{aligned} \tag{2}$$

where cos denotes cosine similarity between, τ is hyperparameter. $\mathcal{L}_{ConD}$ forces discriminator to output image features $f_s(\mathbf{x}_i)$ that is similar to the corresponding text feature $\mathbf{h}_i$. $\mathcal{L}_{ConG}$ enforces image features from pretrained CLIP image encoder $f_I(\mathbf{x}')$ to be similar to corresponding text features $\mathbf{h}_i$. Our final objective losses for the generator and discriminator are defined as:

$$\begin{aligned} \mathcal{L}'_G &= \mathcal{L}_G + \gamma \mathcal{L}_{ConD} + \lambda \mathcal{L}_{ConG}, \\ \mathcal{L}'_D &= \mathcal{L}_D + \gamma \mathcal{L}_{ConD} \end{aligned} \tag{3}$$

Our initial experimentation reveals that the straightforward conditioning of INR-GAN is insufficient, and training the model with the aforementioned Discriminator and objectives does not yield stable results. Consequently, we introduce a novel approach called word-level modulation, which

enhances the model's learning capability by incorporating word-level hypernetworks in conjunction with our WHAtt attention mechanism. The specifics of this mechanism will be elaborated upon in subsequent sections of this paper.

3.1. Hyper-Conditional GANs (HyperCGANs)

Text Conditioning. In line with previous works [21, 44], our approach also employs conditioning mechanisms for the generator. The choice of text information utilized for conditioning depends on the granularity of language representation, which can be either at the word-level or the sentence-level. To facilitate conditioning, we preprocess the input text by tokenizing it and padding it to a fixed length of $C = 77$. Subsequently, we extract text features from two different sources: (1) features $\mathbf{t}_{proj}$ obtained from the projection layer of the pre-trained CLIP model (which remains fixed during training), (2) contextual features denoted as $\mathbf{t}$ from the penultimate layer of the same CLIP text encoder. Each component in the vector $\mathbf{t}_i$ of $\mathbf{t}$ corresponds to the representation of the i^{th} word in the input sentence. Specifically, we refer to the set of components $\mathbf{t}_{local} = \mathbf{t}_{1:C} \in \mathbb{R}^{C\times 512}$ as capturing local word-level information. Additionally, we use the "end of text" (EOT) component of $\mathbf{t}$ which aggregates global information and is denoted as $\mathbf{t}_{global} \in \mathbb{R}^{512}$.

3.1.1 Conditioning signals for weight modulation.

Sentence-level Conditioning. A direct approach for conditioning the unconditional INR-GAN is to utilize either $\mathbf{t}_{proj}$ or $\mathbf{t}_{global}$, both of which have dimensions d_c, as extracted text embeddings. In this scenario, the Hypernetwork backbone receives the concatenation of the noise vector $\mathbf{z} \sim \mathcal{N}(0, I)$ with the text embedding t, denoted as $[\mathbf{z}, t]$. The value of t can be either $\mathbf{t}_{proj}$ or $\mathbf{t}_{global}$. Subsequently, for each linear layer ℓ within the MLP-decoder $F_{\theta(\mathbf{z})}(x, y)$, separate modulating tensors $M^{\ell}_{\mathbf{z},s}$ are generated through the hypernetwork $H([\mathbf{z}, t])$. These tensors, $\boldsymbol{M}^{\ell}_{\mathbf{z},s}$, are then used to modulate the weights $\boldsymbol{W}^{\ell}_{s}$ of the INR-based decoder at layer ℓ through element-wise multiplication: $\boldsymbol{W}^{\ell} = \boldsymbol{W}^{\ell}_{s} \odot \sigma(\boldsymbol{M}^{\ell}_{\mathbf{z},s})$. However, our preliminary experiments revealed that this form of conditioning resulted in subpar performance and unstable training. We hypothesize that sentence-level information may not provide an adequate level of detail necessary to effectively guide the generation process. Consequently, we propose a novel conditioning mechanism that leverages word features to enhance the synthesis of T2CI models.

Word-level Conditioning. Word embeddings $\mathbf{t}_{local} \in \mathbb{R}^{C\times d_w}$ are represented as a sequence of individual vectors of size d_w for each word in the sentence, where C denotes sequence length of the word embeddings (i.e., the number of tokens). We generate a set of C weight matrices $\{\boldsymbol{W}^{\ell}_{i} \in \mathbb{R}^{d^{\ell}_{out}\times d^{\ell}_{in}}\}^{C}_{i=1}$ for each i-th word in the sequence through a fully connected layer `FC`, then use our a novel Word-level Hyper-Attention mechanism proposed in this work, termed WhAtt, to select more important "word" weights, detailed later in this section.

3.1.2 Extreme Modulating Tensor Factorization (X-factorization).

Producing a full-rank tensor directly $\boldsymbol{W}^{\ell}_{i}$ for each layer ℓ is memory-intensive and infeasible even for modestly sized architectures. For example, if the hidden layer size of our hypernetwork is of size $d_h = 512$ and the weight tensor at layer ℓ is of dimensionality $d_o = c_{\text{out}} \times c_{\text{in}} = 512 \times 512 \approx 0.26$ million, then the output weight matrix in the hypernetwork will be of size $d_o \times d_h \approx 0.134$ billion. To overcome this issue, we propose factorizing the modulating tensor with an *extreme* low-rank tensor decomposition for learning efficiency. The canonical polyadic (CP) decomposition [23] lets us express a rank-R tensor $\mathcal{T} \in \mathbb{R}^{d_1\times\ldots\times d_n}$ as a sum of R rank-1 tensors:

$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{v}^{r}_{1} \otimes \ldots \otimes \mathbf{v}^{r}_{n} \tag{4}$$

where $\otimes$ is the tensor product and $\mathbf{v}^{k}_{r}$ is a vector of length d_k. Thus, we generate separately low-rank factors $\mathbf{v}^{k}_{r}$ and build a modulating tensor out of these low-rank factors. Going back to our previous example, this factorization leads to $d_o = c_{\text{out}} + c_{\text{in}} = 512 + 512 = 1024$. So, the output weight matrix in the hypernetwork will be of size $d_o \times d_h \approx 0.5$ million parameters which leads to $\approx 99.6\%$ decrease in the parameter size. Therefore, each $\boldsymbol{W}^{\ell}_{i} \in \mathbb{R}^{d^{\ell}_{out}\times d^{\ell}_{in}}$ will be the tensor product of two vectors $\mathbf{v}^{\ell}_{1i} \in \mathbb{R}^{d^{\ell}_{out}}$ and $\mathbf{v}^{\ell}_{2i} \in \mathbb{R}^{d^{\ell}_{in}}$: $\boldsymbol{W}^{\ell}_{i} = \mathbf{v}^{\ell}_{1i} \otimes \mathbf{v}^{\ell}_{2i}$.

3.1.3 Word-level Hyper Attention (WHAtt).

In contrast to sentence embedding where words are summarized in one vector, individual word embeddings consist of sequences of individual word encodings, containing fine-grained information that is typically visually grounded to the image. Hence, we focus on how to leverage this information in our model. We introduce a Word-level Hyper Attention mechanism, denoted as WhAtt, that can leverage this word-level as well as information through self-attention. Given a set of weights $\{\boldsymbol{W}^{\ell}_{i} \in \mathbb{R}^{d^{\ell}_{out}\times d^{\ell}_{in}}\}^{C}_{i=1}$ from our hypernetworks, we need to select the most relevant word weight for the current layer ℓ. To do this, we incorporate an attention mechanism. The set of weights can be viewed as a tensor $\boldsymbol{Q}^{\ell} \in \mathbb{R}^{C\times(d^{\ell}_{\text{out}}\times d^{\ell}_{\text{in}})}$, where C denotes sequence lengths. We apply scaled dot product attention mechanism [62] to attend to the relevant word weights to get modulating

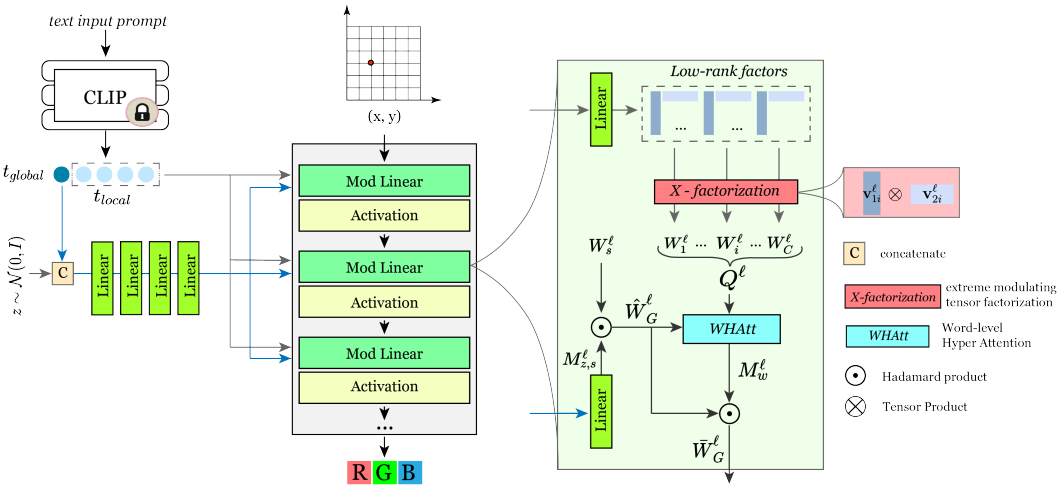

Figure 3. **The architecture of the proposed HyperCGAN:** Linear layers are used as hypernetworks. Overall, given text embeddings and noise vector, hypernetworks generate parameters for modulating weights of INR-based decoder.

weight $\boldsymbol{M}_w^\ell \in \mathbb{R}^{d_{\text{out}}^\ell \times d_{\text{in}}^\ell}$:

$$\boldsymbol{M}_w^\ell = \text{WHAtt}(\boldsymbol{W}^\ell, \boldsymbol{Q}^\ell) = \text{softmax}(\frac{\boldsymbol{W}^\ell (\boldsymbol{Q}^\ell)^T}{\sqrt{d_{\text{out}}^\ell \times d_{\text{in}}^\ell}}) \boldsymbol{Q}^\ell, \tag{5}$$

where $\boldsymbol{W}^\ell$ is the weight matrix at layer ℓ, $\boldsymbol{M}_w^\ell$ is the word-level modulating tensor, $\boldsymbol{W}^\ell$ and $\boldsymbol{M}_w^\ell \in \mathbb{R}^{d_{\text{out}}^\ell \times d_{\text{in}}^\ell}$. Finally, the modulating tensors for the generator for both sentence and word-based modulation are defined by Eq. 6:

$$\begin{aligned} \hat{\boldsymbol{W}}_G^\ell &= \boldsymbol{W}_s^\ell \odot \sigma(\boldsymbol{M}_{z,s}^\ell) \\ \bar{\boldsymbol{W}}_G^\ell &= \hat{\boldsymbol{W}}_G^\ell \odot \sigma(\text{WHAtt}(\hat{\boldsymbol{W}}_G^\ell, \boldsymbol{Q}^\ell)) \end{aligned} \tag{6}$$

where $\bar{\boldsymbol{W}}_G^\ell$ is the modulated weight at layer ℓ for the generator. The first modulation operation can be viewed as obtaining a general context about the image, whereas the attention operation helps to choose the more relevant information. More generally, word-level conditioning benefit for visual-semantic consistency was first demonstrated for discrete decoders in AttnGAN [65]. Our word-level modulation is our proposed mechanism to bring similar properties to text-conditioned continuous image generation. The overall architecture of our model can be seen in Figure 3.

4. Experiments and Results

In this section, we first define the used datasets, metrics, and our baselines following which we compare our model relative to the baselines on the benchmarks, and study the various properties and limitations of our approach.

Datasets. We comprehensively evaluate HyperCGAN on the standard text-to-image benchmarsk MS-COCO [29], CUB [64] datasets, as well as on ArtEmis [1] dataset.

– **COCO** 256^2 contains over 80K images for training and more than 40K images for testing. Each image has 5 associated captions that describe the visual content of the image. We use the splits proposed in [65] to train and test our models.

– **ArtEmis** 256^2 **(introduced T2I benchmark)** contains over 450K emotion attributes and explanations from humans on more than 81K artworks from WikiArt dataset. Each image is associated with at least 5 captions. The unique aspect of the dataset is that utterances are more affective and subjective rather than descriptive. These aspects of the dataset impose additional challenges on T2I generation task. We use the train and test splits provided by the authors and benchmark recent T2I methods on it. Both COCO and ArtEmis are scene-level T2I benchmarks.

– **CUB** 256^2 contains 8,855 training and 2,933 test images of bird species. Each image has 10 corresponding text descriptions. In contrast to COCO and ArtEmis, CUB is an object-level benchmark, yet challenging since this dataset contains fine-grained details about the bird species.

Evaluation Metrics. We evaluate all models in terms of both Image Quality and Text-Image Alignment. Due to the limitations of the Inception score (IS) [46] to capture the diversity and quality of the generation, we report Frechet Inception Distance (FID) [18] score following previous works [60, 69, 70, 72]. Additionally, we compute *R-precision* since image quality scores alone cannot reflect whether the generated image is well conditioned on the given text description. Given a generated image, R-precision measures the retrieval rate for the corresponding caption using a multi-modal network which computes the similarity score between image features and text features. As suggested in [36], we also report the R-precision score where image-text similarity is computed with CLIP [37], dubbed as CLIP-R.

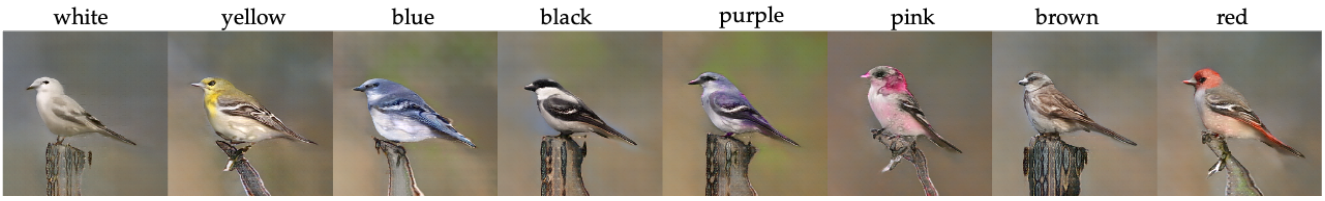

Figure 4. **Exploring Model Sensitivity:** Here, the input noise z is kept fixed while varying color names in the prompt *"a small {color}, bird with white and dark gray wingbars and white breast and long tail"*, aiming to assess the model's sensitivity to word-level modulation.

Out of an input text as a positive and 99 negative captions for the generated image, the CLIP model should give the highest similarity score for the positive caption if the generated image aligns with it.

Configuration	CLIP-R ↑	FID ↓
INR-GAN	-	-
+ $\mathbf{t}_{\text{proj}}$, rank 1	34.81%	78.23
+ $\mathbf{t}_{\text{proj}}$, rank 5	40.13%	69.92
+ $\mathbf{t}_{\text{global}}$, rank 1	45.67%	62.52
+ $\mathbf{t}_{\text{global}}$, rank 5	51.81%	57.25
+ WHAtt	OOM	OOM
+ X-factorization	51.12%	19.13
+ $\mathbf{t}_{\text{global}}$, rank 1	53.78%	18.15
+ $\mathbf{t}_{\text{global}}$, rank 5	51.87%	14.12

Table 1. **T2CI Performance on CUB** 256^2**.** Our hypernetwork-based conditioning makes it possible to use word-level conditioning, which is crucial in achieving good results.

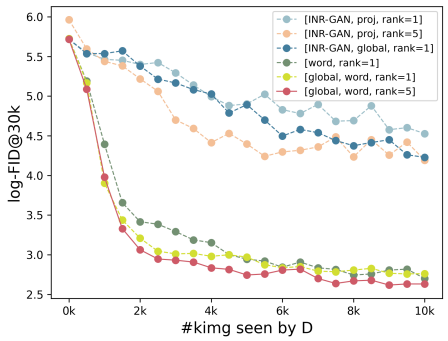

Figure 5. **FID scores (in log scale) on CUB** 256^2**.** Using our word-level conditioning gives a clear advantage. "kimg" denotes the number of images seen by D.

Effectiveness of Word-level modulation. Our study begins by evaluating the effectiveness of our hypernetwork-based word-level attention mechanism through an ablation study on the CUB dataset (Table 1). Since our work is the first attempt at T2CI, we start transforming unconditional INR-GAN to be conditioned on either sentence embeddings t_{proj} or global embeddings t_{global} and adopt it as a baseline. In this transformation, this baseline simply takes the concatenated noise vector and sentence/global embeddings and then generates parameters for the decoder to synthesize an image. For consistency, we employ the discriminator from LAFITE [72] in all our experiments. We observe that simply increasing the rank of the INR does not yield improvements in both FID and CLIP-R results. In contrast, our proposed word-level conditioning mechanism enhances convergence rate (refer to Figure 5) and FID scores. Note that without factorization, the models that use word conditioning fail due to the Out-Of-Memory (OOM) error. We hypothesize that our word-level modulation has significantly better performance due to the improved granularity connecting the generated images to the input text. Moreover, incorporating additional global embeddings and increasing the rank further improves the results. Figure 4 shows that word-level modulation effectively captures the color change for fine-grained generation.

5. Analysis on Continuous Image Synthesis

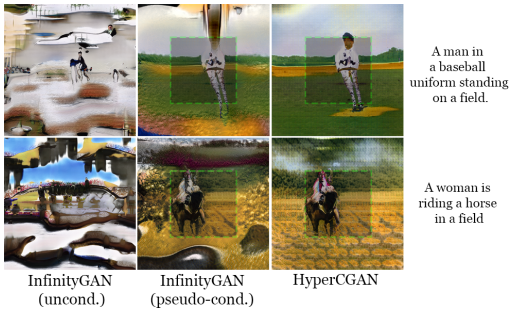

Figure 6. Qualitative results on extrapolation (from 256^2 to 512^2)

In this section, we take a close look at the capability of our models in terms of continuous image synthesis: extrapolation and superresolution.

Extrapolation. In this section, we evaluate the capability of our model in generating images *beyond* the resolution encountered during training. HyperCGAN is trained on a *fixed* coordinate grid within the range $[-1, 1]^2$. During inference, we extend the grid beyond this range (e.g., $[-1.5, 1.5]^2$) to do extrapolation. While there is no clear *text-conditioned* counterpart model for direct comparison, we use InfinityGAN [28], designed for generating images of arbitrary di-

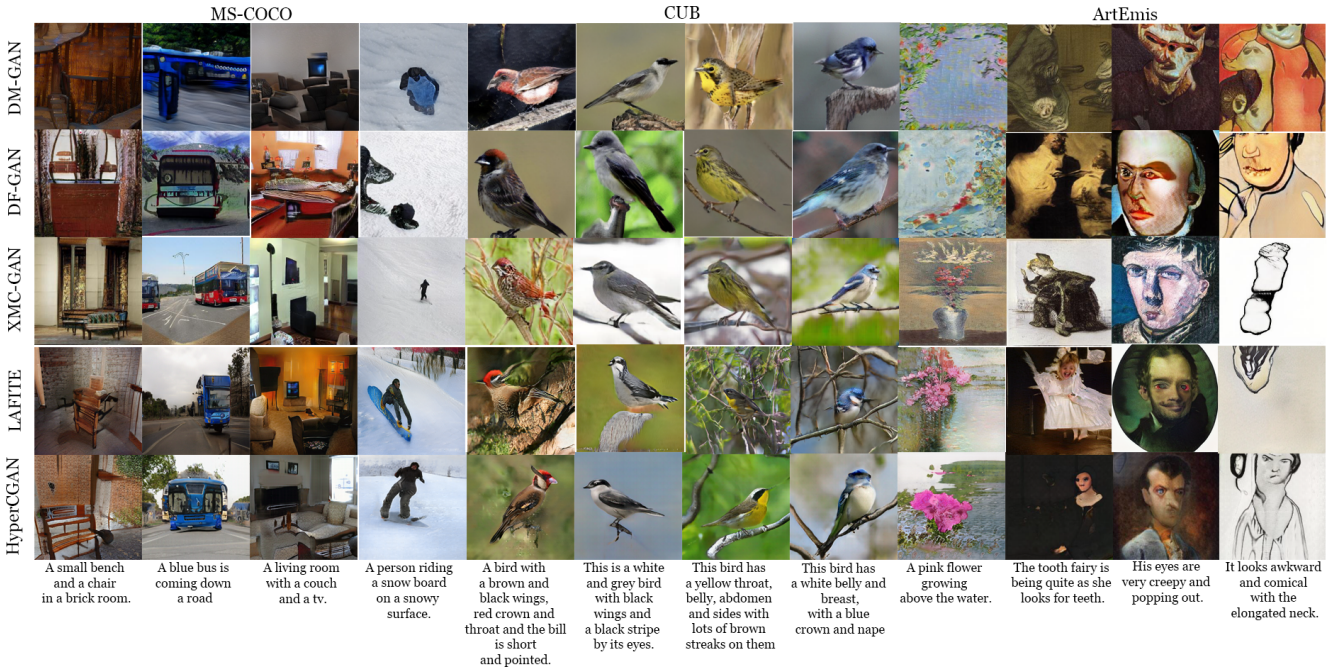

Figure 7. Qualitative results on three datasets: MS-COCO 256^2, CUB 256^2, and ArtEmis 256^2.

mensions, as a reference point. We train InfinityGAN from scratch on the MS-COCO dataset. During the inference, owing to its *unconditional* nature, we conduct two types of image generation with InfinityGAN: 1) Unconditional (uncond.) arbitrary-sized image generation, and 2) Pseudo-conditional (Pseudo-cond.) generation. For the latter, we employ images generated by HyperCGAN as initial images, invert them, and perform extrapolation using InfinityGAN.

The results are quantified in terms of Scale Inverse FID (ScaleInv-FID), as suggested by InfinityGAN, and presented in Table 2. Remarkably, our model demonstrates superior performance compared to InfinityGAN, even though it was not explicitly trained for this task. For a visual representation of the results, please refer to Figure 6. Notably, InfinityGAN tends to blend disparate styles in unconditional generation, resulting in inconsistencies. On the other hand, the pseudo-conditional version of InfinityGAN shows improved results, although enhancements are still needed for extended regions. Our method, even when employed in a zero-shot manner, outperforms InfinityGAN with fewer parameters.

High-resolution sampling. Another useful property of our model is to generate images at any resolution, even though it was trained on lower resolution. High-resolution synthesis can be achieved by sampling *denser* coordinate grids within range $[-1, 1]^2$. We evaluate our model and compare against unconditional AnyResGAN [5] as well as SD-Upsampler [42] on COCO dataset. AnyResGAN was trained from scratch for this comparison. As input to SD-Upsampler, we utilized outputs from our model. We report

Table 2. ScaleInv-FID Results on Extrapolation on COCO: Models trained on 256^2 resolution and evaluated on 2x and 4x extrapolation.

Method	1x	2x	4x	NoP
InfinityGAN [28](uncond)	76.94	103.13	153.64	73M
InfinityGAN [28] (pseudo-cond)	41.71	132.24	120.32	73M
HyperCGAN	29.92	62.01	85.4	57M

Table 3. High-resolution sampling results $256^2 \rightarrow 1024^2$ on COCO. Inference time (Inf time) is computed in GPU.

Method	pFID	Inf time	NoP
SD-Upsampler	21.12	14.8 s	846 M
AnyResGAN	34.68	0.006s	61 M
HyperCGAN (ours)	34.64	0.019s	57 M

patch-FID (pFID) scores in Table 3. The results reveal that our method achieves comparable results to AnyResGAN, a model specifically trained with two-stage patch-based training for high-resolution synthesis (training details in Supplementary). It's important to note that our model was not trained for superresolution but rather generation was done in zero-shot fashion, while having fewer parameters. However, a fair comparison with SD-Upsampler is challenging, as this model is trained on a 10M subset of LAION containing images of resolutions $> 2048^2$ and involves significantly more parameters.

Comparison to the State-of-the-Art. In our final evaluation, we benchmark HyperCGAN against discrete state-of-

Table 4. Comparison to SOTA Discrete T2I models. **Bold**, blue, and cyan indicates 1st, 2nd, and 3rd places. VQ-Diffusion-F* was pre-trained on CC dataset [49] and all other methods are trained from scratch.

Model		COCO 256^2		ArtEmis 256^2		CUB 256^2			
	Year	FID ↓	CLIP-R ↑	FID ↓	CLIP-R ↑	FID ↓	CLIP-R ↑	NoP ↓	Cont
AttnGAN [65]	2018	35.49	29.31%	45.64	7.11%	23.98	31.23%	230M	✗
ControlGAN [24]	2019	34.52	24.96%	42.01	7.38%	22.85	35.71%	250M	✗
DM-GAN [73]	2019	32.64	40.31%	31.4	12.92%	16.09	45.07%	46M	✗
DAE-GAN [43]	2021	28.12	-	-	-	15.19	-	98M	✗
TIME [30]	2021	31.14	-	-	-	14.30	-	120M	✗
DF-GAN [60]	2022	19.32	26.13%	25.4	9.81%	14.81	28.39%	**19M**	✗
SSA-GAN [26]	2022	19.37	30.28%	-	-	15.61	29.60%	109M	✗
XMC-GAN [69]	2021	9.87	48.31%	15.47	36.68%	15.56	30.40%	166M	✗
LAFITE [72]	2022	8.12	95.59%	**12.04**	**88.93%**	10.48	**59.08%**	75M	✗
GALIP [61]	2023	**5.85**	**99.84%**	-	-%	**10.08**	-%	82M	✗
VQ-Diffusion-F [16]*	2022	13.86	60.32%	-	-	10.32	43.13%	370M	✗
HyperCGAN (ours)		13.54	85.12%	15.89	55.23%	14.12	51.87%	57M	✓
Real Images		-	89.43%	-	45.12%	-	26.20%		

the-art approaches [16, 24, 60, 61, 65, 69, 71, 73]. Figure 7 visually compares the qualitative results of our model to these state-of-the-art methods, showcasing comparable generation qualities. Table 4 provides a comprehensive overview, demonstrating that our models outperform many of the comparison methods, including most 2022 ones. Notably, our model achieves competitive results against the diffusion-based model VQ-Diffusion-F in terms of FID on COCO. It's crucial to consider that VQ-Diffusion contains 370M parameters, undergoing training on 7M samples from the Conceptual Captions dataset and fine-tuning on COCO and CUB datasets. When compared to recent advancements such as LAFITE and GALIP, our models exhibit higher FID values. It is important to note that in contrast, our models utilize significantly fewer parameters and offer additional continuous properties like superresolution and extrapolation, which uniquely characterize our method.

6. Limitations and Discussions

Figure 8. **Failure Cases:** blob patches, ignoring some words, not counting objects.

Our studies shed light on the potential to narrow the performance gap between discrete and continuous text-to-image synthesis paradigms by leveraging our innovative conditioning mechanism for INR-based models. This mechanism holds promise for advancing continuous text-to-image generation. While our model captures the semantic meaning of inputs and offers competitive results, there is room for improvement in visual quality to further reduce the gap compared to the recent discrete state-of-the-art. Furthermore, our model occasionally struggles with accurately capturing counting and the correct composition of objects. Figure 8 illustrates instances of failure cases. One contributing factor to the visual limitations might be the fact that our model generates pixels *independently*, lacking spatial local context. Also, common artifacts associated with INR-based GANs, such as wavy or patterned textures, and stains can be observed, especially during superresolution and extrapolation tasks. To address this, incorporating specialized training techniques, akin to those proposed in [5], may help improve extrapolation/superresolution performance.

7. Conclusion

In this paper, we propose HyperCGAN, a novel HyperNet-based conditional continuous GAN. HyperCGAN is a text-to-continuous-image generative model with a single generator that operates with a novel language-guided tensor modulation operator for sentence-level and word-level attention mechanisms. To our knowledge, HyperCGAN is the first approach that facilitates text-to-continuous-image generation for objects and complex scenes, and we show its ability to meaningfully extrapolate images beyond training image dimension while maintaining alignment with the input language description. We showed that HyperCGAN achieves comparable performance compared to most of the existing discrete-based text-to-image synthesis baselines. We hope that our method may encourage future work on hypernetworks on text-to-continuous Image Generation (T2CI).

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5

[2] David Alvarez-Melis and Judith Amores. The emotional gan: Priming adversarial generation of art with emotion. In *2017 NeurIPS Machine Learning for Creativity and Design Workshop*, 2017. 3

[3] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 14278–14287, 2021. 3

[4] Luca Bertinetto, João F. Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *Advances in Neural Information Processing Systems 29*, pages 523–531. Curran Associates, Inc., 2016. 3

[5] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *European Conference on Computer Vision*, 2022. 7, 8

[6] Oscar Chang, Lampros Flokas, and Hod Lipson. Principled weight initialization for hypernetworks. In *International Conference on Learning Representations*, 2020. 2

[7] 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. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[10] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017. 3

[11] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 2

[12] Tomer Galanti and Lior Wolf. On the modularity of hypernetworks. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[13] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019. 3

[14] 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. 2

[15] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471. PMLR, 2015. 2

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 8

[17] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 2

[18] 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. *arXiv preprint arXiv:1706.08500*, 2017. 5

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[20] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013. 2

[21] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4

[22] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[23] Henk AL Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3):105–122, 2000. 4

[24] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Controllable text-to-image generation. 2019. 1, 2, 8

[25] Muyang Li, Ji Lin, Chenlin Meng, Stefano Ermon, Song Han, and Jun-Yan Zhu. Efficient spatially sparse inference for conditional gans and diffusion models. *arXiv preprint arXiv:2211.02048*, 2022. 3

[26] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18187–18196, 2022. 8

[27] 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/CVF International Conference on Computer Vision*, pages 4512–4521, 2019. 3

[28] Chieh Hubert Lin, Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, and Ming-Hsuan Yang. InfinityGAN: Towards infinite-pixel image synthesis. In *International Conference on Learning Representations*, 2022. 6, 7

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[30] Bingchen Liu, Kunpeng Song, Yizhe Zhu, Gerard de Melo, and Ahmed Elgammal. Time: Text and image mutual-translation adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2082–2090, 2021. 8

[31] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 3

[32] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 2

[33] Chenlin Meng, Ruiqi Gao, Diederik P Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. *arXiv preprint arXiv:2210.03142*, 2022. 3

[34] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 3

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[36] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 5

[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. 5

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 1, 2

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

[41] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 2

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 7

[43] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13960–13969, 2021. 8

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2, 4

[45] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3

[46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. 5

[47] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. 2023. 2, 3

[48] Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, and Camille Couprie. Design: Design inspiration from generative networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[49] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 8

[50] Rajhans Singh, Ankita Shukla, and Pavan Turaga. Polynomial implicit neural representations for large diverse datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2051, 2023. 3

[51] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[52] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 3

[53] Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny. Adversarial generation of continuous images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10753–10764, 2021. 2, 3

[54] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. *arXiv preprint arXiv:2104.06954*, 2021. 3

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[57] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[58] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE, 2017. 3

[59] Wei Ren Tan, Chee Seng Chan, Hernan E Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2018. 3

[60] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16515–16525, 2022. 1, 2, 5, 8

[61] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. *arXiv preprint arXiv:2301.12959*, 2023. 3, 8

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 4

[63] Johannes von Oswald, Christian Henning, João Sacramento, and Benjamin F. Grewe. Continual learning with hypernetworks. In *International Conference on Learning Representations*, 2020. 3

[64] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5

[65] 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. 1, 2, 5, 8

[66] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[67] Chris Zhang, Mengye Ren, and Raquel Urtasun. Graph hypernetworks for neural architecture search. In *International Conference on Learning Representations*, 2019. 3

[68] 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. 2

[69] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. 2021. 1, 2, 3, 5, 8

[70] Zhenxing Zhang and Lambert Schomaker. Dtgan: Dual attention generative adversarial networks for text-to-image generation. *arXiv preprint arXiv:2011.02709*, 2020. 5

[71] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17907–17917, 2022. 2, 8

[72] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 3, 5, 6, 8

[73] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 1, 2, 8