Text-to-Image Generation via Implicit Visual Guidance and Hypernetwork

Xin Yuan,[1] **Zhe Lin,**[2] **Jason Kuen,**[2] **Jianming Zhang,**[2] **John Collomosse**[2, 3]

[1]University of Chicago, [2]Adobe Research [3]University of Surrey
yuanx@uchicago.edu, {zlin, kuen, jianmzha}@adobe.com, j.collomosse@surrey.ac.uk

Abstract

We develop an approach for text-to-image generation that embraces additional retrieval images, driven by a combination of implicit visual guidance loss and generative objectives. Unlike most existing text-to-image generation methods which merely take the text as input, our method dynamically feeds cross-modal search results into a unified training stage, hence improving the quality, controllability and diversity of generation results. We propose a novel hypernetwork modulated visual-text encoding scheme to predict the weight update of the encoding layer, enabling effective transfer from visual information (e.g. layout, content) into the corresponding latent domain. Experimental results show that our model guided with additional retrieval visual data outperforms existing GAN-based models. On COCO dataset, we achieve better FID of 9.13 with up to 3.5× fewer generator parameters, compared with the state-of-the-art method.

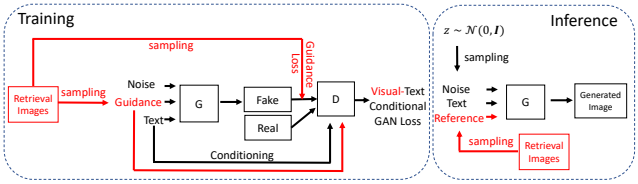
1 Introduction

In recent years, Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have produced realistic results on the text-to-image generation task (Xu et al. 2018; Zhang et al. 2017, 2018a; Tao et al. 2020; Hinz, Heinrich, and Wermter 2020; Zhang et al. 2021). Despite the success, unstructured text descriptions usually make it difficult for the generator to learn a good distribution coverage. The cross-modal gap between texts and inadequate real image samples during training presents optimization difficulties (e.g. overfitting (Hinz, Heinrich, and Wermter 2020; Li et al. 2019c), mode collapse), yielding image results with low quality and diversity, especially in the case of generating a complex scene conditioned on one-sentence description. XMC-GAN (Zhang et al. 2021) tries to bridge the gap by forcing strong text-image correspondence through a contrastive discriminator and achieves state-of-the-art results. However, as shown in Figure 1(a), it is still trained with fixed pairs labeled by datasets. To fully exploit the intrinsic data properties, long epochs and large model parameters are necessary for cross-modal contrastive loss, requiring huge training resources. This dilemma can be relieved by training the generator on large-scale visual-language datasets with massive image-text pairs provided (Sharma et al. 2018). Yet, the fine-grained image captioning efforts on such a large dataset are usually prohibitively heavy.

Our objective is to leverage the power of cross-modal search by feeding text-image pairs that are dynamically created, providing more informative and implicit guidance during training. The framework design should have following essential properties: 1) Large capacity to learn cross-modal distributions from dynamic image-text pairs; 2) Effective encoding scheme for implicit information transfer; 3) Good quality, diversity and controllability at inference time; 4) Can easily leverage samples from external datasets as guidance. We aim to learn high-quality text-to-image generation results with additional guidance of easily acquired text-to-image search results.

Several recently proposed methods (Hinz, Heinrich, and Wermter 2019, 2020; Casanova et al. 2021; Li, Torr, and Lukasiewicz 2022; Koh et al. 2021) also focus on generating high-quality images using additional information. For example, IC-GAN (Casanova et al. 2021) extends and improves unconditional GANs by modeling the neighborhood distribution of an instance feature. Yet, jointly encoding multi-modal information in a unified training framework is not the focus of IC-GAN. Given a text description, IC-GAN requires an inference-time CLIP-guided (Radford et al. 2021) noise optimization for image generation. OP-GAN (Hinz, Heinrich, and Wermter 2020) simultaneously learns the layout information provided by individual objects and generates an image conditioned on both layouts and text. However, OP-GAN requires multiple training stages and more fine-gained object bounding boxes so it is not efficient. MemoryGAN (Li, Torr, and Lukasiewicz 2022) builds a retrieval image memory bank and then selectively feeds image features for guidance at each stage of the network. However, even with both global and region features extracted, it still requires multi-stage discriminators (Xu et al. 2018) to generate high-resolution images (e.g. 256×256), suggesting the memory features are not fully exploited by the encoding stage.

In this paper, we take a unified view of implicit visual guidance and generative objectives and develop a new text-to-image generation framework, as shown in Figure 1. To be specific, we first conduct offline cross-modal search to build up candidate image retrieval results. A hypernetwork architecture is used in the text-conditioned encoding layer to predict the weights updating and effectively transfer visual information (e.g. content, layout) from retrieval results



(a) Difference between our method and most existing works.

(b) Scenes generated by our method. ***left***: for the *upper* reference with close view of multiple players in the field, generations also present a close view of multiple players; for the *upper* reference with distant view of single player and the whole field, our generations present similar layouts, with audience included. ***right***: for the *upper* reference with large portion of green vegetables, our generations consistently produce 'pizza' images with similar ingredients; for the *bottom* reference with less greens, the generations present a similar content.

Figure 1: Main idea of the proposed method. (a) Difference between our method and others. As highlighted in red, our method leverages the additional information from retrieval image to dynamically create image-text pairs, which produces more diversified joint distribution and enables better training with additional, easy-acquired visual conditions and guidance. Without any inference-time latent optimization, the model can directly generate high-quality images with diverse and complex scenes using the text description (b) Even with noise vector fixed (each column), the diversity can still be maintained with much better controllability (each row), as demonstrated by different retrieval references with rich visual information (e.g. styles, layouts, colors and contents). (*best view in color.*)

into the latent representation, hence improving the controllability of both generator and discriminator training via back-propagation. During inference, given a text description, the generation diversity can be ensured via not only random noise vectors, but also candidate visual feature sampling in a more controllable manner. We summarize our contributions as two-fold:

- **Unified Training.** Our training framework simultaneously encodes rich information from text description and cross-modal search results in a unified procedure with carefully designed visual-text conditional mapping layer and implicit visual guidance loss. In addition, as shown in Figure 1 and 2, our framework randomly samples candidate images with respect to text description. It has the flexibility to incorporate any new external image source.
- **Better quality, controllability and diversity.** The generator trained by our method can generate high-quality images adhere to semantic information in text descriptions. Instead of merely relying on noise sampling, a more controllable and diversified inference is reached by accommodating external features that share similar cross-modal similarity as reference. The model achieves excellent quantitative and qualitative results under common evaluation protocols.

We demonstrate these advantages through comparing experimental results with recent GAN-based methods on CUB (Wah et al. 2011) and the challenging COCO (Lin et al. 2014b) datasets.

2 Related Work

Direct text-guided image generation. (Xu et al. 2018; Zhang et al. 2017, 2018a; Tao et al. 2020; Hinz, Heinrich, and Wermter 2020; Zhang et al. 2021; Zhu et al. 2019) are proposed by only using the text descriptions as conditions for GANs. For example, StackGAN (Zhang et al. 2017, 2018a) progressively generates images in a coarse-to-fine manner. (Xu et al. 2018) improves (Zhang et al. 2018a) with cross-modal attention mechanism. DM-GAN (Zhu et al. 2019) proposes a dynamic memory module with both reading and writing gates to refine image content. XMC-GAN (Zhang et al. 2021) explores the multimodal contastive losses in a simple one-stage GAN framework and achieves the state-of-the-art performance in terms of image quality and image-text alignment. However, it still learns the joint distribution over fixed and insufficient text-image pairs provided by datasets.

GANs with additional information. (Hinz, Heinrich, and Wermter 2019, 2020; Casanova et al. 2021; Li, Torr, and Lukasiewicz 2022; Koh et al. 2021) use additional information for generating high-quality images. For example, Recently proposed IC-GAN (Casanova et al. 2021) provides extensions to unconditional GANs by feeding additional instance feature's neighbors for better distribution learning. OP-GAN (Hinz, Heinrich, and Wermter 2019) generates images conditioned on both layouts and text simultaneously, followed by a separate layout learning scheme from individual objects. MemoryGAN (Li, Torr, and Lukasiewicz 2022) builds a retrieval image memory banks, then selectively feed image features at each stage of the network. Although (Li, Torr, and Lukasiewicz 2022) and our algorithm share the same spirit of utilizing image retrieval, ours doesn't require the multi-stage discriminators (Li, Torr, and Lukasiewicz 2022; Xu et al. 2018) to generate high-resolution images.

Joint vison-language representation. Vision and language (VL) methods (Gu et al. 2017; Karpathy and Li 2015; Kiros, Salakhutdinov, and Zemel 2014; Gu et al. 2018; Wang, Li, and Lazebnik 2016; Lu et al. 2019; Tan and Bansal 2019; Chen et al. 2020b; Li et al. 2019b; Gu et al. 2020) are representatives that embrace multi-modal information for many computer vision tasks, such as image captioning and cross-modal retrieval. Such methods aim at mapping text and images into a common space, where semantic similarity across different modalities can be learned by ranking-based contrastive losses. DAMSM (Xu et al. 2018) computes the similarity between images and captions, which has been widely used in text-to-image generation methods (Xu et al. 2018; Hinz, Heinrich, and Wermter 2020; Zhu et al. 2019) for addi-

tional and fine-grained feedback. CLIP (Radford et al. 2021) provides more powerful joint representation, which is pre-trained by scaling up the simple language and image proxy task. It achieves great success in inference-time generation re-ranking (Ramesh et al. 2021) or text-driven image manipulation (Patashnik et al. 2021). Our method conducts the offline cross-modal search using image-text representations from (Xu et al. 2018), and has the flexibility to incorporate with other joint vison-language pre-trainng models.

Hypernetwork and its application in image synthesis. Hypernetwork (Ha, Dai, and Le 2017) is proposed to predict the model parameters for a target network and has been used in image synthesis tasks (Skorokhodov, Ignatyev, and Elhoseiny 2020; Anokhin et al. 2020; HAYDAROV et al. 2022; Chiang et al. 2022). (Skorokhodov, Ignatyev, and Elhoseiny 2020) presents impressive ability in unconditional image generation by modulating the generators using hypernetworks. HyperCGAN (HAYDAROV et al. 2022) extends these methods to text-to-image generation task by utilizing the text-conditional hypernetwork-based image generator. (Chiang et al. 2022) learns an implicit representation of the 3D scene with the neural radiance fields model (NeRF) (Mildenhall et al. 2020), in which a hypernetwork is used to transfer the style information. Our method also uses representative modulation power of hypernetwork but only applies it to the encoding scheme for transferring the text-conditional visual information (e.g. content, layout) into latent domain.

3 Method

Figure 2 shows the overall architecture for the proposed training framework. The system is composed of two schemes: offline cross-modal search and unified training with dynamically created image-text pairs. The cross-modal search is designed to build up an offline mapping between captions and the image database with high semantic correlations. The mapping helps to feed random data streaming continuously and efficiently into the training stage. The training scheme combines the generative power of StyleGAN architectures (Karras, Laine, and Aila 2019; Karras et al. 2020) with additional data samples provided by text and corresponding images in an end-to-end manner. The original mapping layer in StyleGAN lacks the ability of encoding semantic information from higher-level concepts into intermediate style representation (e.g. text) during training. We address this issue by introducing a new encoding scheme fitting for different types of input. The encoding is guided by both generative objectives and implicit visual guidance measured using simple feature distances. After completing the unified training, the generator can be directly used to generate images adhere to characteristics described by text and image data.

3.1 Cross-modal Image Search

To use rich visual information for better diversity and controllability during training, we build up a candidate image database which are highly correlated with text descriptions via cross-modal image search. We first embed the representation of images and captions into a common space and

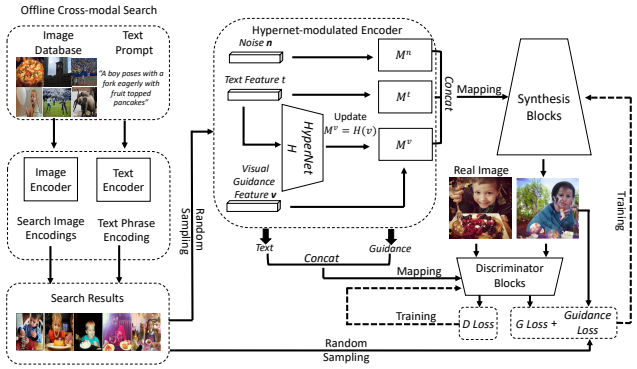


Figure 2: System overview: The proposed framework consists of two components *left*: offline cross-modal search and *right*: unified training accommodating multi-modal data stream. The cross-modal is designed to build up an offline mapping between captions and image data with high semantic correlation in the embedding space. Dynamic image-text pairs are fed through random sampling to the novel hypernet-modulated visual-text encoder, which are instantiated separately for both generator and discriminator. To be specific, hypernetworks take input of text information to predict the weights of visual encoding module $\boldsymbol{M}^v$, which transfer the visual information from visual guidance features into the latent representation. All modules are end-to-end trainable, which are optimized by the combination of guidance loss and GAN loss via back propagation.

then use the cosine similarity as ranking metrics. We denote the text set $D_C = \{C_i\}_{N_C}$ with N_C captions and image database as $D_I = \{I_j\}_{N_I}$ with N_I images. We generate the visual feature vectors using image encoder E_I and text features using text encoder E_C in the common feature space:

$$\boldsymbol{v}_i = E_I(I_j, \Theta_{E_I}) \tag{1}$$

$$\boldsymbol{t}_j = E_C(C_i, \Theta_{E_C}) \tag{2}$$

where Θ_{E_I} and Θ_{E_C} are the pre-trained weights of the image and text encoder, respectively. We build up a similarity score matrix $S = \{s_{ij}\}_{N_C \times N_I}$, where each element reflects the semantic similarity between a query text feature $\boldsymbol{t}_i$ and a candidate image feature $\boldsymbol{v}_j$. We thus formulate the image-caption cosine similarity score as:

$$s_{ij} = \frac{\boldsymbol{t}_i \cdot \boldsymbol{v}_j}{|\boldsymbol{t}_i||\boldsymbol{v}_j|} \tag{3}$$

Thus we can easily build the mappings $\Gamma = \{\gamma_{ik}\}_{N_C \times K}$ from the ith query to search results by recording the corresponding index of visual features with largest K similarity scores. The offline mappings are beneficial for continuous and efficient data streaming during training. More importantly, comparing with fixed image-text pairs provided by the training dataset itself, it yields much wider variety of pairs for GANs distribution learning.

3.2 HyperNetwork modulated Visual-Text Encoding

We start by developing an extension to StyleGAN2 (Karras et al. 2020) for modeling joint distributions of multimodal data. One simple design is adding two separate linear embedding layers M^t and M^v which directly encodes images and text features as follows:

$$\boldsymbol{v}_e = M^v(\boldsymbol{v}, \Phi_{M^v}) \tag{4}$$

$$\boldsymbol{t}_e = M^v(\boldsymbol{t}, \Phi_{M^t}) \tag{5}$$

where $\boldsymbol{v}$, $\boldsymbol{t}$ are the pre-trained image and text representation. Φ_{M^v} and Φ_{M^t} are trainable parameters. Then $\boldsymbol{v}_e$ and $\boldsymbol{t}_e$ concatenated with the noise embedding are encoded to latent domain by additional mapping layers of generator. The encoding scheme is similarly instantiated for the discriminator except that only text-visual feature are mapped to latent codes as conditions. However, the separate embedding layers fail to further encode the image and text correlation, which is a bottleneck in transferring the text-conditioned visual information into the joint latent domain. Moreover, joint encoding multimodal data without any constraint may also result in uncontrollable distribution coverage for the generator. One solution is to directly minimize the distance of v_e and t_e. However, the constraint is too strong for such simple linear layers, especially when v and t already present the semantic-correlation in the pretrained cross-modal common space. This may lead to a trivial solution and prevents the latter mapping layers to learn useful joint latent codes.

To address this issue, we use the hypernetwork modulation to transfer the text-conditional visual information (e.g. layout) into latent domain. Instead of directly optimizing Φ_{M^v}, we first encode $\boldsymbol{t}$ with a hypernetwork $\mathcal{H}$ into a feature vector, which is reshaped as a trainable parameter of M^v. We formulates the weights prediction by modifying Eq. 4 as:

$$\boldsymbol{v}_e = M^v(\boldsymbol{v}, \Phi_{M^v} = \texttt{reshaped}(\mathcal{H}(\boldsymbol{t}))) \tag{6}$$

where $\mathcal{H}$ takes text information as input and predicts the weights updating of $\boldsymbol{M}^v$, i.e. Φ_{M^v}. In our experiments, we report the performance with hypernetwork modulation and also conduct investigation on the direct optimization.

3.3 Optimization with Implicit Visual Guidance

We train the generator G and discriminator D driven by the combination of generative objectives and implicit visual guidance from retrieval images. We first accommodate the conventional generator loss with multimodal condition (i.e., additional information of caption and corresponding search images):

$$\begin{aligned}\mathcal{L}_{gen} \quad = \quad & \mathbb{E}_{z\sim p(z), k_1\sim\Gamma_i}[\log(1 - \\ & D(G(z, \boldsymbol{t}_e^{(i)}, \boldsymbol{v}_e^{(k_1)}), \boldsymbol{t}_e^{(i)}, \boldsymbol{v}_e^{(k_1)})]\end{aligned} \tag{7}$$

where z is sampled from normal distribution $N(0, I)$. Γ_i is the offline mapping for the ith query caption. We randomly sample k_1 to dynamically formulate an image. t_e^i and $v_e^{k_1}$ are derived using Eq. 5 and Eq. 4 and are trained together with the proposed hypernetwork-modulated encoding scheme in an end-to-end manner.

Table 1: FID Comparisons on CUB dataset.

Methods	FID ↓
Attn-GAN (Xu et al. 2018)	23.98
ControlGAN (Li et al. 2019a)	13.92
DM-GAN (Zhu et al. 2019)	16.09
DF-GAN (Tao et al. 2020)	14.81
Memory-GAN (Li, Torr, and Lukasiewicz 2022)	<u>10.49</u>
Ours	**5.65**

To implicitly exploit the visual information, we use the visual guidance loss measured using the distance between the generation result and the reference image:

$$\mathcal{L}_{guide} = \mathbb{E}_{z\sim p(z), k_1, k_2\sim\Gamma_i}||E_I(G(z, \boldsymbol{t}_e^{(i)}, \boldsymbol{v}_e^{(k_1)})) - \boldsymbol{v}_e^{(k_2)}|| \tag{8}$$

where E_I is a pre-trained image encoder used in our cross-modal search system. Our final loss function for generator is

$$\mathcal{L}_G = \mathcal{L}_{gen} + \lambda\mathcal{L}_{guide} \tag{9}$$

where we set λ as 1.0 in the implementation. We similarly adapt the discriminator by incorporating caption and image searching and minimize

$$\begin{aligned}\mathcal{L}_D = \quad & -\mathbb{E}_{k_1\sim\Gamma_i}[\log D(I^{(i)}, \boldsymbol{t}_e^{(i)}, \boldsymbol{v}_e^{(k_1)})] - \\ & \mathbb{E}_{\hat{x}\sim p(G), k_1\sim\Gamma_i}[\log(1 - D(\hat{I}^{(i)}, \boldsymbol{t}_e^{(i)}, \boldsymbol{v}_e^{(k_1)})]\end{aligned} \tag{10}$$

where $\hat{I}^{(i)}$ is the generation result from G. We then alternatively optimize G and D under both conditional and implicit control in the unified training framework.

4 Experiments

4.1 Experimental Setup

We evaluate our model on CUB (Wah et al. 2011) and COCO-2014 (Lin et al. 2014a) datasets. We preprocess CUB following (Xu et al. 2018), train on 8,855 samples and test on 2933 samples, with 10 captions for each. For COCO, we use 80k training samples and 40k testing samples, each annotated with 5 captions following (Xu et al. 2018).

Evaluation Metrics. Following existing works, we generate 30,000 images with randomly selected unseen text descriptions for validation, We use Inception Score (IS) (Salimans et al. 2016) and Fréchet Inception Distance (FID) (Heusel et al. 2017) for generation quality assessment. In addition to image quality metrics, we also evaluate the semantic alignment between text and generation image. We adopt Semantic Object Accuracy (SOA) (Hinz, Heinrich, and Wermter 2020) to validate whether objects mentioned in the text are recognizable by a YOLO detector (Redmon et al. 2016). We report both the SOA-C (measured by the percentage of images per class detects the given object) and SOA-I (the percentage of images a given object is detected). We report the number of parameters of the image generator. For hierarchical generation methods which have multiple stages (e.g. shape generator) , we simply report the number of parameters of the last stage generator.

Table 2: Comparisons with the state-of-the-art methods. Note that Ours-DAMSM and Ours-CLIP refer text-image representations pre-trained by DAMSM (Xu et al. 2018) and CLIP (Radford et al. 2021), respcetively.

Methods	Params (M) ↓	IS↑	FID ↓	SOA-C ↑	SOA-I ↑
Attn-GAN (Xu et al. 2018)	13	23.61 ± 0.21	33.10 ± 0.11	25.88	39.01
ControlGAN (Li et al. 2019a)	-	24.06 ± 0.60	-	-	-
Obj-GAN (Li et al. 2019c)	34	24.09 ± 0.28	36.52 ± 0.13	27.14	41.24
DM-GAN (Zhu et al. 2019)	21	$\mathbf{32.32 \pm 0.23}$	27.34 ± 0.11	33.44	48.03
OP-GAN (Hinz, Heinrich, and Wermter 2020)	18	27.88 ± 0.12	24.70 ± 0.09	35.85	50.47
DF-GAN (Tao et al. 2020)	**12**	-	19.32	-	-
HyperCGAN (HAYDAROV et al. 2022)	-	21.05	20.81	-	-
Memory-GAN (Li, Torr, and Lukasiewicz 2022)	-	-	19.47	-	-
XMC-GAN (Zhang et al. 2021) ($96\times ch$)	90	30.45	9.33	**50.94**	**71.33**
XMC-GAN (Zhang et al. 2021) ($64\times ch$)	43	30.66	11.93	39.85	59.78
Ours-DAMSM	25	29.33 ± 0.20	$\mathbf{9.13 \pm 0.07}$	35.53	50.66
Ours-CLIP	25	29.02 ± 0.23	9.23 ± 0.05	43.75	63.23

Implementation Details. For consistency and fair comparison with most existing text-to-image generation methods, we adopt DAMSM (Xu et al. 2018) pretrained image encoder and text encoder. On COCO dataset, we also utilize stronger multimodal representations from CLIP (Radford et al. 2021). During training, we choose top $K = 5$ retrieval images for random selection. During inference, we select the reference image from the training set instead of the validation one for fair comparison. We use the same set of data augmentation (i.e. random crop and random horizontal flip) with previous works. For generation models, we use extensions of StyleGAN2 architectures following (Casanova et al. 2021). M^v and M^t are both single fully connected layers that encode 256-dim visual or text features into 128-dim for both G and D. The Hypernetwork architecture is implemented as Multilayer perceptrons (MLP) (Collobert and Bengio 2004), which includes single 64-dim hidden layer with ReLU activation. It encodes 256-dim text feature into a large vector, which is reshaped as 256×128 for M^v weights updating. We train the models on two 48GB-Nvidia-A40 GPUs for resolutions 256×256, with learning rate of $3\mathrm{e}{-3}$ for both G and D. We report the results at 300 epochs when FID doesn't improve during training.

Figure 3: Generation results on COCO. For XMC-GAN, we take samples from original paper (Zhang et al. 2021).

4.2 CUB Results

As shown in Table 1, our method significantly improves FID to 5.65, comparing with the leading methods on the CUB dataset. Our method also outperforms Memory-GAN method by a large margin, which suggests that ours utilizes the retrieval results in a more effective and efficient way, even without multi-stage discriminators.

4.3 COCO Results

Quantitative results. We compare our method with representative GAN-based text-to-image generation methods using various metrics. As shown in Table 2, we achieve better FID of 9.13 than the state-of-the-art method XMC-GAN with up to $3.5\times$ fewer generator parameters. It also validates our main idea: under carefully designed guidance, enlarging diversity of visual-semantic joint distribution is beneficial to the learning of high-quality image generation. Our method also achieves comparable IS with XMC-GAN. DM-GAN (Zhu et al. 2019) achieves better IS than both XMC-GAN and ours and can generate more realistic images. This observation suggests that FID is a more robust metric and IS has limitation in capturing diversity and quality. (Hinz, Heinrich, and Wermter 2020).

For SOA-C and SOA-I, ours-DAMSM achieves comparable results with OP-GAN. This demonstrates that compared with additional individual boxes, the easily acquired retrieval samples in our unified training framework presents equivalent capability in region-level semantic preservation. XMC-GAN outperforms all methods by a large margin, showing strong text-alignment ability learnt by text-image contrastive loss, which also suggests the separately pre-trained language model in large-scale (i.e. BERT (Devlin et al. 2019)) is advantageous over the simple LSTM encoder in DAMSM. We also replace the DAMSM with a pre-trained CLIP model to incorporate more powerful joint representations. As shown in Table 2, ours-CLIP drastically improves SOA-C and SOA-I and outperforms XMC-GAN($64 \times ch$) even with less generator parameters.

Qualitative Results. In Figure 3, we compare our model with state-of-the-art text-to-image generation method XMC-GAN at resolution 256 on COCO captions. For XMC-GAN, we directly use the generation results from the original paper. Our model consistently produces more realistic images,

especially for human pose and face generation, which are very challenging in complex scenes.

Analysis on Diversity and Controllability. The dynamic image-text data during training provides not only additional visual guidance, but also a more diversified joint distribution for the generator to learn a good coverage over. During testing, the sample diversity can be realized by varying both the noise vector and retrieval images. We validate this by conducting experiments on COCO under two different settings: (1) Generation using the same caption and reference image, but with varying noise vectors. Figure 4 shows that our method is able to generate diverse images using different noises. (2) Generation using the same caption and noise vector, but with varying reference image, as shown in Figure 5. We also measure the diversity by calculating the average similarity with $d = \sum_{caps} \sum_{samples} f_{dis}(x_i, x_j)/(C * N)$, where C and N denotes the number of captions and samples. For f_{dis}, we use both L_2 and the perceptual distance measured by LPIPS (Zhang et al. 2018b). With both varying noise vectors, we show that further varying reference helps our method increase the distance by 53.2% and 67.8% in terms of L_2 and LPIPS, respectively. Note that both quantitative and qualitative diversity results help us check whether the network simply learns a nearly identity mapping from retrieval samples (i.e., a simple copy-paste operation) or not, which is important but missed by the retrieval-based method MemoryGAN (Li, Torr, and Lukasiewicz 2022). Given the diversity of our generations based on either noise or retrieval samples, we conclude that our method learns good distribution instead of a trivial solution. In Figure 6, we also observe that the retrieval guidance transfers visual information to generation results in a clear controllable manner, e.g. the backgrounds colors, human poses, room layout, and directions, etc.

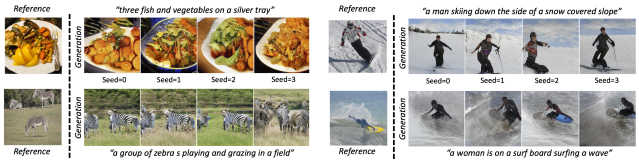
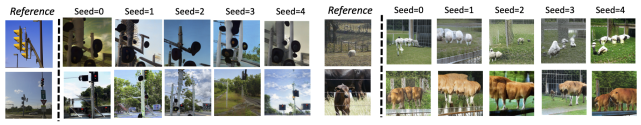


Figure 4: Generation results on COCO with varying noise vectors while fixing retrieval sample.

Figure 5: Generation results on COCO with varying retrieval images while fixing noise.

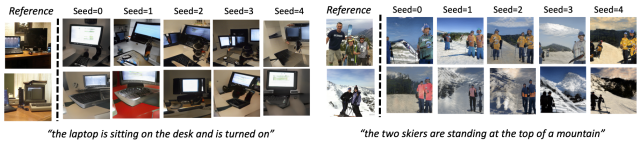
4.4 Ablations

We evaluate each component by analyzing their effects on two baselines: AttnGAN and text-conditional StyleGAN2. AttnGAN is a representative multi-stage framework adopted by many text-to-image generation methods. As such, improvements made on it may easily generalize to more. For



(a) ***left***: for the *upper* cases, our generations present close view of traffic lights, which follows the guidance of reference image with sky background; for the *bottom* reference, generations also present a similar layout that traffic lights are in the street view. ***right***: The caption mentions 'wildlife' without specifying the species. *upper* reference presents 'sheep', which is transferred to generations, showing a group of sheep on grassy area; In contrast, given *bottom* reference of cattle, the generations produce images of cattle.

(b) ***left***: both *upper* and *bottom* cases present a consistent camera view under the guidance of corresponding kitchen images. The layout information in the kitchens is transferred to the generation results. ***right***: *upper* presents fewer passengers in the boat while the *bottom* shows multiple passengers.



(c) ***left***: *upper* reference provides the information of two screens, which is also shown in generations results, producing multiple screens on a desk; In contrast, *bottom* only shows single screen. ***right***: *upper* presents three skiers on a mountain while the *bottom* shows two.

Figure 6: More qualitative results on COCO varying noise and reference, with clear controllability.

StyleGAN2 architectures, we set up the simple baseline by extending it into text-conditional version, based on which we show each component woven together into a unified framework successfully.

Cross-modal search. We build offline cross-modal image search for both baselines. As shown in Table 3, both Attn-GAN-Ret and Ours-Ret FID become worse, suggesting that diversify the joint multimodal distribution without control or feedback my confuse the generator learning. Note that the FID degradation in AttnGAN is smaller than StyleGAN2 baseline, suggesting that the additional DAMSM loss can still serves as a guidance between image and text.

Different types of Visual Guidance Loss. The implicit visual guidance, which directly applies to output of generator, are varying as L1 or contrastive loss. As shown in in Table 3, the FID gets consistently improved with both variants, compared with baselines. We also observe that contrastive loss works better for Attn-GAN while does the opposite for ours. One reason is that the DAMSM loss can be treated

Table 3: We evaluate the different components on two baselines: Attn-GAN (Xu et al. 2018) and Text-conditional StyleGAN2.

Methods	Retrieval	HyperNet	Visual Guidance	IS↑	FID ↓
Attn-GAN (Xu et al. 2018)	✗	✗	✗	23.61 (+0.00)	33.10 (-0.00)
Attn-GAN-Ret	✓	✗	✗	24.40 (+0.79)	34.67 (+1.57)
Attn-GAN Ret-L1	✓	✗	L_1	25.47 (+1.76)	27.30 (-5.80)
Attn-GAN-Ret-Contrast	✓	✗	$Contrastive$	**27.73 (+4.12)**	**21.18 (-11.92)**
Text-cond StyleGAN2	✗	✗	✗	20.96 (+0.00)	18.08 (-0.00)
Ours-Ret	✓	✗	✗	21.01 (+0.05)	24.86 (+6.78)
Ours-Ret-L1	✓	✗	L_1	27.42 (+6.46)	13.87 (-4.21)
Ours-Ret-Contrast	✓	✗	$Contrastive$	28.76 (+7.80)	16.35 (-1.73)
Ours-Ret-Hyper-Contrast	✓	✓	$Contrastive$	28.43 (+7.47)	10.07 (-8.01)
Ours-Ret-Hyper-L1 (best)	✓	✓	$L1$	**29.33 (+8.37)**	**9.13 (-8.95)**

as a variant of visual-text contrasive loss, with image-image contrastive samples introduced, the generator can learn both intra- and inter-modal information simultaneously. Another reason is the smaller batch size in our baseline makes it difficult to learn from contrastive samples (Chen et al. 2020a)

Hypernetwork modulation. One issue with Ours-Ret-L1 and Ours-Ret-Contastive is that, the diversity cannot be provided by retrieval samples, i.e., the visual information from retrieval samples cannot be effectively transferred to the latent representation for guidance. We observe that, during the optimization of separate image and text encoding scheme, the weights of the image encoder tend to be suppressed, i.e., instead of learning from diversified joint distribution, it degrades to the simpler one. We use the hypernetwork to modulate the image encoding layer to address this issue, enabling effective visual information transfer. Figure 5, Figure 6 and Table 3 show diverse generation results and improved FID, demonstrating the effectiveness of hypernetwork modulated encoding scheme.

4.5 Enhanced Controllability with Run-time Latent Optimization

Our method is trained in a unified framework and can directly produce generation results with guidance text and image inputs during inference, without any additional optimization (e.g. re-ranking, noise optimization). Benefit from the diversified joint visual-text distribution provided by dynamic pairs via cross-modal search, the guidance visual information has been successfully learnt and transferred, thus presenting the controllability property *w.r.t.* retrieval guidance images. We further show this controllablity can be enhanced via run-time latent optimization (Nikitko 2019; Abdal, Qin, and Wonka 2019).

Implementation Details. Given a text and reference image pair, we randomly sample 10k images with varying noise vectors, from which we calculate the average and standard deviation of W latent space. We optimize latent code w_{opt} using the perceptual loss (Zhang et al. 2018b) between generation result $G_{synth}(w_{opt})$ and reference image I_{ref}, i.e. $L_{percept}(G_{synth}(w_{opt}), I_{ref})$. We use Adam (Kingma and Ba 2015), with $lr = 0.02$, $(\beta_1, \beta_2) = (0.9, 0.999)$ for 300 iterations for each image.

Results. As shown in Figure 7, the initial generation (0th iteration) from the original latent space (with averaging 10k samples) and reference images may have large distinctions. After 30 iterations, the style (e.g. tone, color, saturation) starts matching with the guidance. For longer iterations, more detailed information, especially the content and layout can be learnt.

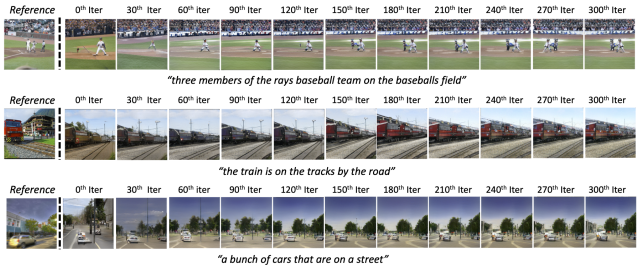


Figure 7: As optimization iterations increases, the network gradually produces similar tones, saturation, layouts, etc.

5 Conclusion

We propose a new text-to-image generation method that produces high-quaility and diversity images, with controllability provided by easy acquired retrieval images. Though our method has several components, it is not a collection of orthogonal innovations. Rather, these components are designed and woven together for a high-level vision: improving generator by training over diversified image-text joint distribution by utilizing retrieval images. To facilitate the controllable and effective visual information transfer, we propose the implicit visual guidance and hypernetwork modulated encoding scheme in a unified training framework. Quantitative and qualitative results on CUB and COCO, together with thorough ablations and analysis demonstrate the effectiveness of our method. We also show the enhanced controllability by adopting run-time latent optimization. For the current system, retrieval samples are processed as feature vectors for storage and efficiency consideration, trading off with capacity of visual information. Future work can be done by incorporating with local visual information (e.g. regions), external database for better generations.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? In *ICCV*.

Anokhin, I.; Demochkin, K.; Khakhulin, T.; Sterkin, G.; Lempitsky, V.; and Korzhenkov, D. 2020. Image Generators with Conditionally-Independent Pixel Synthesis. *arXiv preprint arXiv:2011.13775*.

Casanova, A.; Careil, M.; Verbeek, J.; Drozdzal, M.; and Romero-Soriano, A. 2021. Instance-Conditioned GAN. In *NeurIPS)*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. *CoRR*, abs/2002.05709.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*.

Chiang, P.-Z.; Tsai, M.-S.; Tseng, H.-Y.; Lai, W.-S.; and Chiu, W.-C. 2022. Stylizing 3D Scene via Implicit Representation and HyperNetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Collobert, R.; and Bengio, S. 2004. Links between perceptrons, MLPs and SVMs. In *ICML*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*.

Gu, J.; Cai, J.; Joty, S. R.; Niu, L.; and Wang, G. 2018. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval With Generative Models. In *CVPR*.

Gu, J.; Kuen, J.; Joty, S.; Cai, J.; Morariu, V.; Zhao, H.; and Sun, T. 2020. Self-Supervised Relationship Probing. *NeurIPS*, 33.

Gu, J.; Wang, G.; Cai, J.; and Chen, T. 2017. An Empirical Study of Language CNN for Image Captioning. In *ICCV*.

Ha, D.; Dai, A. M.; and Le, Q. V. 2017. HyperNetworks. In *ICLR*.

HAYDAROV, K.; Muhamed, A.; Lazarevic, J.; Skorokhodov, I.; and Elhoseiny, M. 2022. HyperCGAN: Text-to-Image Synthesis with HyperNet-Modulated Conditional Generative Adversarial Networks.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*.

Hinz, T.; Heinrich, S.; and Wermter, S. 2019. Generating Multiple Objects at Spatially Distinct Locations. In *ICLR*.

Hinz, T.; Heinrich, S.; and Wermter, S. 2020. Semantic Object Accuracy for Generative Text-to-Image Synthesis. *TPAMI*.

Karpathy, A.; and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *CoRR*, abs/1411.2539.

Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Text-to-Image Generation Grounded by Fine-Grained User Attention. *WACV*.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. H. S. 2019a. Controllable Text-to-Image Generation. In *NeurIPS*.

Li, B.; Torr, P.; and Lukasiewicz, T. 2022. Memory-Driven Text-to-Image Generation.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019b. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *Arxiv*.

Li, W.; Zhang, P.; Zhang, L.; Huang, Q.; He, X.; Lyu, S.; and Gao, J. 2019c. Object-driven Text-to-Image Synthesis via Adversarial Training. In *CVPR*.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014a. Microsoft COCO: Common Objects in Context. In *ECCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014b. Microsoft COCO: Common Objects in Context. In *ECCV*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Nikitko, D. 2019. stylegan-encoder. https://github.com/Puzer/stylegan-encoder.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *ICML*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In Meila, M.; and Zhang, T., eds., *ICML*.

Redmon, J.; Divvala, S. K.; Girshick, R. B.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*.

Salimans, T.; Goodfellow, I. J.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *NeurIPS*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Gurevych, I.; and Miyao, Y., eds., *ACL*.

Skorokhodov, I.; Ignatyev, S.; and Elhoseiny, M. 2020. Adversarial Generation of Continuous Images. *arXiv preprint arXiv:2011.12026*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*.

Tao, M.; Tang, H.; Wu, S.; Sebe, N.; Wu, F.; and Jing, X. 2020. DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis. *CoRR*, abs/2008.05865.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *CVPR*.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *CVPR*.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *CVPR*.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. 2017. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In *ICCV*.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2018a. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *TPAMI*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018b. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *CVPR*.