Generative models (e.g., GANs and diffusion models) learn the underlying data distribution in an unsupervised manner. However, many applications of interest require sampling from a specific region of the generative model’s output space or evenly over a range of characteristics. To allow efficient sampling in these scenarios, we propose Generative Visual Prompt (PromptGen), a framework for distributional control over pre-trained generative models by incorporating knowledge of arbitrary off-the-shelf models. PromptGen defines control as an energy-based model (EBM) and samples images in a feed-forward manner by approximating the EBM with invertible neural networks, avoiding optimization at inference. We demonstrate how PromptGen can control several generative models (e.g., StyleGAN2, StyleNeRF, diffusion autoencoder, and NVAE) using various off-the-shelf models: (1) with the CLIP model, PromptGen can sample images guided by text, (2) with image classifiers, PromptGen can de-bias generative models across a set of attributes, and (3) with inverse graphics models, PromptGen can sample images of the same identity in different poses. (4) Finally, PromptGen reveals that the CLIP model shows “reporting bias” when used as control, and PromptGen can further de-bias this controlled distribution in an iterative manner.1

1Introduction

Generative models learn the underlying high-dimensional data distribution and have achieved promising performance on image synthesis [4, 73, 35, 21]. Though being well praised, they still face two main criticisms. First, since generative models are typically trained in an unsupervised way, they lack controllability, meaning that it is unclear how to sample from a specific region of the space. Second, generative models are prone to inherit the imbalance and bias of training data [58, 30]. For instance, StyleGAN2 is more likely to produce images of white individuals, see Figure 1(e). Previous works have studied these challenges separately, and typical methods include editing of “style” codes [22, 66, 30] and explicit conditions [41, 67]. However, these methods are either model-dependent (i.e., requiring a well-structured style space) or label-intensive (i.e., requiring all training samples to be labeled for explicit conditions), limiting their generality and practical use.

To address the above challenges, this paper advocates a unified formulation, distributional control of generative models, which enables controllability by incorporating the knowledge of off-the-shelf models (e.g., CLIP [56], classifiers, or inverse graphics models [16]). Based on this unified view, we propose to learn distributions in the latent space of a pre-trained generative model while keeping the pre-trained weights fixed. Given its conceptual similarity to prompt learning [40, 83, 84, 29], we term our framework as Generative Visual Prompt (PromptGen). PromptGen requires no training data, and the only supervision comes from off-the-shelf models that help define the control. Specifically, PromptGen allows the user to define controls using an energy-based model (EBM) approximated
Figure 1: PromptGen uses different off-the-shelf models (e.g., CLIP [56], inverse graphics models, and classifiers) to control the output distribution of a pre-trained and fixed generative model (e.g., StyleGAN2). Colored boxes (e.g., the blue CLIP) indicate the control. See the text for details.

by an invertible neural network (INN). Unlike methods that require optimization at inference in EBM sampling [47, 14, 43, 50], PromptGen samples images in a feed-forward manner, which is highly efficient. Moreover, PromptGen stands alone at inference, meaning that we can discard the off-the-shelf models, which define the control, after training.

We illustrate the benefits of PromptGen with several experiments. Figure 1 demonstrates our main findings with StyleGAN2 trained on FFHQ [34] without any labels, while we also show results for StyleNeRF [21], diffusion autoencoder [55], and NVAE [73] in Section 4. Figure 1(a) illustrates how we can sample from StyleGAN2 based on text descriptions such as “a photo of a baby”. Figure 1(b) shows that PromptGen can leverage a race classifier (potentially trained on a different dataset) to sample uniformly across all races, de-biasing the pre-trained StyleGAN2. Figure 1(d) illustrates that PromptGen can generate images of the same identity in different poses, guided by a pose regressor.

Finally, it is worth pointing out that PromptGen not only offers generality for algorithmic design and modularity for control composition, but also enables iterative controls when some controls are contingent on others. For instance, one may train a text-controlled distribution and then de-bias this distribution. To achieve this, we view the composition of the INN and the generative model as a new “generative model” to be controlled. Figure 1(f) illustrates that PromptGen reveals “reporting bias” of the CLIP model [56], where “without makeup” is – perhaps surprisingly – positively correlated with female, and Figure 1(c) shows that PromptGen can further mitigate this bias with iterative control.

Table 1: Comparison between methods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arbitrary control (e.g., CLIP)</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Low-dimensional latent space</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Stands alone at inference</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Feed-forward (i.e., no inference-time optim.)</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Iterative distributional control</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
</tbody>
</table>

2 Related Work

Over the past few years, generative models have gained the ability to generate images with high visual quality. A few of the most widely used methods include generative adversarial networks (GANs) [18], VAE [38], invertible neural networks [11, 12], and diffusion models [24, 69]. In particular, generative models trained on large amounts of unlabeled data, e.g., BigGAN [4], StyleGANs [34, 35, 33, 62], Glow [37], and diffusion models [49, 10], achieve promising image synthesis results.

Despite their success, controllability and de-biasing are still two fundamental challenges generative models face. For controllability, existing methods include explicit conditioning at training [41, 67]
or local editing of the learned representation, e.g., “style” codes [1, 54, 44, 66, 78]. For de-biasing, existing methods include local editing of “style” codes [58, 30] and importance sampling for either training [20] or inference [28]. Existing works study these problems separately, each requiring a specific design for the task studied. On the contrary, PromptGen is a unified framework for arbitrary controls defined by off-the-shelf models. The benefits of unifying tasks have been shown by a recent trend of works in multiple research areas across vision and language [57, 46, 61, 79, 59]. Moreover, one can fine-tune generative models [17] for domain adaptation, which is orthogonal to PromptGen: since PromptGen maps a generative model to another generative model (Algorithm 1), fine-tuning can be applied before or after PromptGen training. We leave this exploration to future studies.

Previous methods usually sample from EBM [39] with Markov Chain Monte Carlo (MCMC) [71, 74, 14, 19, 13, 43]. Among them, plug-and-play generative models (PPGMs) [47] and LACE [50] define latent-space EBMs. However, MCMC requires inference-time optimization, which is inefficient and requires the off-the-shelf models to be available at inference; this is also the criticism for diffusion models, regardless of being gradient-guided [49, 65] or not [24, 25, 48]. In contrast, PromptGen achieves efficient, feed-forward sampling. Table 1 shows a comparison with previous methods. Our INN training is similar to that proposed by [52] to sample from physical systems, which is later used by [76] to solve inverse problems with two composed INNs. Notably, [52] and [76] do not leverage a low-dimensional latent space, and [76] requires training a separate model for each image sample. In this paper, we use INN to model arbitrary EBMs for various generative models.

3 Method

Figure 2 illustrates our PromptGen framework. To begin, the user selects a pre-trained generative model $G$. PromptGen then lets the user specify a control $C$ as an energy-based model (EBM) $p(z|C)$. We train an INN $f_\theta$ to approximate $p(z|C)$. PromptGen views the functional composition $G \circ f_\theta$ as a new generative model and can perform iterative control. Algorithm 1 describes the overall procedure.

Algorithm 1: Generative Visual Prompt (PromptGen)

Input: Generative model $G : \mathbb{R}^d \rightarrow \mathcal{X}$

repeat

1. Input: control $C$ of the current iteration
2. Define an EBM $p(z|C)$ for $C$ (Section 3.1)
3. Train an INN $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ to approximate $p(z|C)$ (Section 3.2)
4. $G \leftarrow G \circ f_\theta$

until user stops the iteration

return $G : \mathbb{R}^d \rightarrow \mathcal{X}$

3.1 Latent-Space EBM for Distributional Control

The plug-and-play generative model [47] was first proposed to use a latent-space EBM for controllable image synthesis. If a fixed generative model is used, then theoretically, any image-space EBM can be viewed as a latent-space EBM, as shown by [19] and [50]. We define the latent-space EBM following similar formulation as [47, 19, 50], with some new energy functions. We then extend the formulation to incorporate the moment constraint [7], which was adopted for language modeling [36], but unlike [36], we define the moment constraint in the latent space to accommodate generative vision models.
We define a control $C$ as $M$ independent properties $\{y_1, \ldots, y_M\}$, e.g., $y_1$ can be a text description and $y_2$ can be an attribute. The controllability can be defined as the EBM (detailed in Appendix B.1)

$$p(x|C) = \frac{p_x(x)e^{-E_C(x)}}{Z_X}, \quad E_C(x) = \sum_{i=1}^{M} \lambda_i E_i(x, y_i), \quad Z_X = \int x' p_x(x')e^{-E_C(x')} dx',$$

which reweights the image prior $p_x(x)$ with energy $E_C(x)$, where images with smaller energy are preferred. Using a pre-trained generative model $G : \mathbb{R}^d \rightarrow \mathcal{X}$ that maps a latent code $z$ to an image $x$, the image prior $p_x(x)$ is defined by sampling a latent code $z$ from $p_z = \mathcal{N}(0, I)$ and mapping it to $x = G(z)$. Appendix B.2 shows that this EBM is equivalent to the latent-space EBM

$$p(z|C) = \frac{p_z(z)e^{-E_C(G(z))}}{Z}, \quad E_C(x) = \sum_{i=1}^{M} \lambda_i E_i(x, y_i), \quad Z = \int z' p_z(z')e^{-E_C(G(z'))} dz'.$$

Latent-space EBM allows us to use any off-the-shelf model to specify the control. The following are some examples that are discussed in this paper (explained in Appendix B.1):

**Classifier energy:** Given a classifier $P(\cdot|x)$ and the target class $a$ that we want to sample images from, we define the classifier energy as $E_{\text{classifier}}(x, a) = -\log P(a|x)$.

**CLIP energy:** Using the CLIP model [56], we define the CLIP energy as the cosine distance between the embeddings of the image and the text $t$, averaged over $L$ differentiable augmentations [82, 44]:

$$E_{\text{CLIP}}(x, t) = \frac{1}{L} \sum_{i=1}^{L} \left(1 - \cos \langle \text{CLIP}_\text{img}(\text{DiffAug}_i(x)), \text{CLIP}_\text{text}(t) \rangle \right).$$

**Inverse graphics energy:** Given an inverse graphics model, $f_{X \rightarrow P}$, which infers image parameters (e.g., pose and expression), and the target parameters $\rho$, we define the inverse graphics energy as

$$E_{\text{inv-graphics}}(x, \rho) = d(\langle f_{X \rightarrow P}(x), \rho \rangle)^2,$$

where $d(\cdot, \cdot)$ is the geodesic distance between the inferred parameters and the target parameters.

**Moment constraint:** Some controls cannot be directly defined by off-the-shelf models, and the moment constraint [7, 36] is one of them. Given a mapping $\gamma : \mathcal{X} \rightarrow \mathbb{R}^K$ (e.g., $\gamma$ can be a classifier that outputs the probability simplex), the moment constraint defines the target distribution $p(x|C)$ as

$$p(x|C) = \arg\min_{p(x|C)} \mathbb{D}_{KL}(p(x|C) \parallel p_x(x)), \quad \text{s.t. } E_{x \sim p(z|C)}[\gamma(x)] = \mu,$$

where $\mu$ is the user-specified vector. For example, if we want to generate images that are uniformly distributed across races, we may use a race classifier as $\gamma$ and define $\mu = (|A|^{-1}, \ldots, |A|^{-1})$ where $A$ is the set of races. In this paper, we generalize the moment constraint to the latent space, and approximate the above objective as (detailed in Appendix B.6):

$$p(z|C) = \frac{p_z(z) \exp \left(\beta^T \gamma(G(z))\right)}{Z}, \quad Z = \int z' p_z(z') \exp \left(\beta^T \gamma(G(z'))\right) dz',$$

$$\beta = \arg\min_{\beta} \mathbb{E}_{z^{(1)}, \ldots, z^{(N)} \sim p_z, x^{(j)} = G(z^{(j)})} \left\| \sum_{j=1}^{N} \exp (\beta^T \gamma(x^{(j)})) \gamma(x^{(j)}) \right\|_2^2.$$
Figure 3: Illustration for Algorithm 2. In the forward pass (solid curves), we sample \( \epsilon \sim \mathcal{N}(0, I) \), map \( \epsilon \) to latent code \( z = f_\theta(\epsilon) \) with an INN \( f_\theta \), and map \( z \) to an image \( x = G(z) \) with a fixed generative model \( G \). Dashed curves show the gradients. Details are provided in Section 3.2.

**Algorithm 2:** Approximating Latent-Space EBM with INN

```
while not converged do
    1. Sample \( \epsilon \sim \mathcal{N}(0, I) \)
    2. Map \( \epsilon \) to latent code \( z = f_\theta(\epsilon) \)
    3. Map \( z \) to an image \( x = G(z) \)
    4. Optimize \( \theta \) with gradient \( \nabla_\theta \left( -\log |\det(\frac{\partial f_\theta}{\partial \epsilon})| - \log p_z(z) + E_C(x) \right) \)
```

Based on these properties of INN, we can rewrite our KL divergence objective \( \mathbb{D}_{KL}(p_\theta(z) || p(z|C)) \) as (full derivations in Appendix B.3) the following form:

\[
\mathbb{D}_{KL}(p_\theta(z) || p(z|C)) = \mathbb{E}_{z \sim p_\theta(z), x = G(z)} \left[ \log \frac{p_\theta(z)}{p_z(z)e^{-E_C(x)/Z}} \right] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), z = f_\theta(\epsilon), x = G(z)} \left[ -\log |\det(\frac{\partial f_\theta}{\partial \epsilon})| - \log p_z(z) + E_C(x) \right] - \mathbb{H}_{\mathcal{N}(0, I)} + \log Z.
\]  

(9)

Since \( \mathbb{H}_{\mathcal{N}(0, I)} \) and \( \log Z \) are independent of \( \theta \), our training objective becomes

\[
\arg \min_\theta \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), z = f_\theta(\epsilon), x = G(z)} \left[ -\log |\det(\frac{\partial f_\theta}{\partial \epsilon})| - \log p_z(z) + E_C(x) \right].
\]  

(10)

Figure 3 gives an illustration of our process, and Algorithm 2 describes the algorithmic details.

**PromptGen in a class-embedding space** (Figure 5(e)) Previous works [4, 62] have shown that class conditioning boosts generative models’ performances on ImageNet [60]. Specifically, class-conditioned generative models map a latent code \( z \) and a class embedding \( y \) to \( x = G(z, y) \). To extend PromptGen to these models, we train an INN \( h_\theta \) to map \( \xi \sim \mathcal{N}(\mu, \sigma^2 I) \) to \( y = h_\theta(\xi) \), where \( \mu \) and \( \sigma \) are the mean and standard deviation of \( G \)’s class embeddings. The motivation for defining the distribution of \( \xi \) as \( \mathcal{N}(\mu, \sigma^2 I) \) but not \( \mathcal{N}(0, I) \) is that we want the learned INN \( h_\theta \) to be volume-preserving, which is easier to train. The training objective is to minimize \( \mathbb{D}_{KL}(p_\theta(z, y) || p(z, y|C)) \), which is equivalent to (full derivations in Appendix B.4)

\[
\arg \min_\theta \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), \xi \sim \mathcal{N}(\mu, \sigma^2 I), z = f_\theta(\epsilon), y = h_\theta(\xi), x = G(z, y)} \left[ -\log |\det(\frac{\partial f_\theta}{\partial \epsilon})| - \log p_z(z) - \log |\det(\frac{\partial h_\theta}{\partial \xi})| - \log p_y(y) + E_C(x) \right],
\]  

(11)

where \( p_y(y) = \mathcal{N}(y|\mu, \sigma^2 I) \) is the estimated class-embedding distribution.

**PromptGen with conditional INN** To generalize the control to continuous values, e.g., the scene parameters \( \rho \) in Eq. (4), we condition the INN on \( \rho \). The condition is modeled by replacing \( z = f_\theta(\epsilon) \) with \( z = f_\theta(\epsilon, \rho) \). Space limited, we provide details and derivations in Appendix B.5. This extension results in a similar architecture to StyleFlow [1], but StyleFlow [1] uses MLE training on image-label pairs and is only applicable to explicit condition, which is not capable of modeling EBMs. When the condition is modeled by the equivariant operations proposed by [77], PromptGen also satisfies the homomorphism property [77] in the latent space.
Figure 4: Image synthesis based on text description, guided by the CLIP model. As the pre-trained generative model, we use StyleGAN2 trained on FFHQ 1024\(^2\) [34] with truncation \(\psi = 0.7\). The text description used in this experiment is a photo of a baby. StyleCLIP requires optimization at inference, while StyleGAN2-NADA and PromptGen do not. All images are resized for visualization.

Figure 5: PromptGen is applicable to different generative models in various domains. Figure 5(a), Figure 5(d), and Figure 5(b) are PromptGen applied to StyleNeRF [21], diffusion autoencoder [55], and NVAE [73]. Figure 5(c) is PromptGen applied to StyleGAN2 [35] trained on AFHQ-Cats [6]. Figure 5(e) is the extension to the embedding space of BigGAN [4] on ImageNet [60]. All images are resized for visualization. See Appendix C for results on more datasets and text descriptions.

4 Experiments

This section describes the experimental validation of PromptGen. See Appendix A for experiments on synthetic data and Appendix C, D, E, and F for additional experiments on images and 3D meshes. Additional experimental details are provided in Appendix B.8.

4.1 Image Synthesis based on Text Description

This experiment illustrates the capability of PromptGen to sample images from a generative model driven by a text description \(t\) using the pre-trained CLIP model [56] (the ViT-B/32 version). We used the CLIP energy from Eq. (3), with text descriptions such as a photo of a baby.

Figure 4 shows a comparison between PromptGen and two previous CLIP-guided image generation methods, StyleCLIP [54] and StyleGAN2-NADA [17]. We observe that PromptGen generates diverse and high-quality images of babies, while StyleCLIP struggles in controllability and image quality, and StyleGAN2-NADA generates baby-like adults.\(^2\) These results show that (1) locally editing the latent code (i.e., StyleCLIP) is not always an effective method for controlling generative models (e.g., not all images’ latent code can be locally edited into a baby); (2) domain adaptation (i.e., StyleGAN-NADA) is not effective in seeking modes in a generative model. Figure 5(a), Figure 5(b), and Figure 5(d) show how PromptGen applies to StyleNeRF [21], NVAE [73], and diffusion autoencoder (DiffAE; [55]). Although these generative models are all trained on FFHQ, baby images sampled from them have distinct characteristics. Figure 5(c) shows the ability of PromptGen to generate cats of a particular species. Figure 5(e) shows the extension (Section 3.2) of our PromptGen to the class-embedding space of BigGAN [4], a class-conditional GAN trained on ImageNet [60]. We observe that PromptGen helps BigGAN generate images with complex text descriptions, which are out of BigGAN’s training image distribution; however, the diversity seems to be limited in this extension.

\(^2\)Since we care about distributions, none of the images in this paper are cherry- or lemon-picked.
Figure 6: With the moment constraint, PromptGen de-biases StyleGAN2 (on FFHQ and MetFaces 1024, truncation $\psi = 0.7$). See Appendix D for image samples and Table 5 for quantitative results.

Table 2: Comparison with baselines for de-biasing binary attributes and their correlations. Following [30], we use classifiers on CelebA [45]. Baseline performances are copied from [30]. PromptGen has competitive performance, even in the cases where FairStyle achieves nearly perfect performance.

<table>
<thead>
<tr>
<th>FFHQ (binary attributes)</th>
<th>$D_{KL}(\mathbb{R}^{\psi})$</th>
<th>$D_{KL}(\mathbb{R}^{\psi})$</th>
<th>$D_{KL}(\mathbb{R}^{\psi})$</th>
<th>$D_{KL}(\mathbb{R}^{\psi})$</th>
<th>$D_{KL}(\mathbb{R}^{\psi})$</th>
<th>$D_{KL}(\mathbb{R}^{\psi})$</th>
<th>$D_{KL}(\mathbb{R}^{\psi})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFHQ (real data)</td>
<td>0.015</td>
<td>0.186</td>
<td>0.246</td>
<td>0.355</td>
<td>0.242</td>
<td>0.180</td>
<td>0.279</td>
</tr>
<tr>
<td>StyleGAN2 [35]</td>
<td>0.018</td>
<td>0.180</td>
<td>0.246</td>
<td>0.355</td>
<td>0.242</td>
<td>0.180</td>
<td>0.279</td>
</tr>
<tr>
<td>StyleFlow [1]</td>
<td>0.023</td>
<td>0.061</td>
<td>0.214</td>
<td>0.162</td>
<td>0.121</td>
<td>0.214</td>
<td>0.162</td>
</tr>
<tr>
<td>FairGen [70]</td>
<td>$4.21 \times 10^{-4}$</td>
<td>$7.07 \times 10^{-4}$</td>
<td>$0.0373$</td>
<td>$0.0330$</td>
<td>$0.00185$</td>
<td>$0.0373$</td>
<td>$0.0330$</td>
</tr>
<tr>
<td>FairStyle [30]</td>
<td>$3.20 \times 10^{-7}$</td>
<td>0</td>
<td>$0.0257$</td>
<td>$0.0157$</td>
<td>$0.000241$</td>
<td>$0.0257$</td>
<td>$0.0157$</td>
</tr>
<tr>
<td>PromptGen (ours)</td>
<td>$1.71 \times 10^{-5}$</td>
<td>$1.72 \times 10^{-5}$</td>
<td>0.0008</td>
<td>0.000558</td>
<td>0.000415</td>
<td>0.000628</td>
<td></td>
</tr>
</tbody>
</table>

4.2 De-Biasing Pre-Trained Generative Models

An important problem in generative models is to generate fair distributions w.r.t a set of attributes of interest. For instance, Figure 6(a) shows that StyleGAN2 generates images with bias across races and ages. PromptGen de-biases StyleGAN2 models trained on FFHQ 1024 and MetFaces 1024 [32] in terms of categorical attributes, using the moment constraint defined in Eq. (5) and Eq. (6). As control, we used a classifier trained on FairFace 224 [31] as $\gamma$. We defined $\mu = (|A|^{-1}, \ldots, |A|^{-1})$, where $A$ is the set of races. Similar to the energy weights $\lambda_i$ defined in Eq. (2), we propose to rescale the trained $\beta$ as $\lambda \beta$. Figure 6 shows that PromptGen de-biases the race and age effectively.

Existing de-biasing baselines consider binary attributes [30]. For a fair comparison with them, we adopted their setting to use binary classifiers trained on CelebA [45] for de-biasing and evaluation. Since classifiers trained on CelebA also suffer from the spurious correlation between attributes, we did not use the moment constraint for this experiment. Instead, since PromptGen allows conditional image generation with the classifier energy, we generated the same number of samples conditioned on each attribute or attribute combination. Table 2 shows that PromptGen has competitive performance on de-biasing for attributes and attribute combinations.

4.3 Pose-Guided Face Synthesis

With an inverse graphics model, PromptGen can control the pose of faces generated by StyleGAN2. We used the DECA model [16], which infers the parameters of FLAME [42], a parametric facial graphics model. We set $\rho \in \text{SO}(3)$ as FLAME’s three neck poses and used the conditional INN extension introduced in Section 3.2. To enable generating different poses of the same identity (ID), we propose ID energy using the IR-SE50 model [8]. Specifically, given a canonical pose $\rho_0$, we define $z_0 = f_\rho(\epsilon, \rho_0)$, and the ID energy is defined as (detailed in Appendix B.7)

$$E_{ID}(x_0, x) = 1 - \cos \langle R(x_0), R(x) \rangle, \quad x_0 = G(z_0), x = G(z),$$

where $R$ is the IR-SE50 model [8] that computes face embeddings. Figure 7 shows that PromptGen generates faces of the same ID in different poses, even without being explicitly trained with poses as conditions. We computed the FID score [23] of each model using Clean-FID [53]; following [35],
Figure 7: Using inverse graphics model DECA [16] and scene parameters \( \rho_{\text{pose}} \in \text{SO}(3) \), PromptGen controls the pose of StyleGAN2 while preserving the identity. All images are resized for visualization.

Table 3: Pose-controlled face generation. In this experiment, PromptGen uses StyleGAN2 as the pre-trained generative model. Results for GRAF [64], pi-GAN [5], GIRAFFE [51], and StyleNeRF [21] are from [21]; results for DiscoFaceGAN (DFG) [9] and GAN-Control [67] are from [67].

<table>
<thead>
<tr>
<th>Resolution</th>
<th>GRAF</th>
<th>pi-GAN</th>
<th>GIRAFFE</th>
<th>StyleNeRF</th>
<th>DFG</th>
<th>GAN-Control</th>
<th>PromptGen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pose</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FID↓</td>
<td>3</td>
<td>71</td>
<td>85</td>
<td>35</td>
<td>8</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>Dist. w/ same ID↓</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.83</td>
<td>0.68</td>
<td>0.45</td>
</tr>
<tr>
<td>Dist. w/ diff. ID</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.73</td>
<td>1.90</td>
<td>1.37</td>
</tr>
</tbody>
</table>

we did not use the truncation trick when computing the FID score. Table 3 shows that PromptGen outperforms existing models in terms of the FID score. Following [67], we then reported the average IR-SE50 [8] embedding distances for images with the same ID and with different IDs. Results show that PromptGen achieves the best ID preservation, with a slight sacrifice of ID diversity.

4.4 Decomposing Complex Control via Energy Composition

Energy composition \( E_C(x) = \sum_{i=1}^{M} \lambda_i E_i(x, y_i) \) allows us to decompose controls into simple ones. We show that \( y_1 = \text{photo of a bald black man with beard} \) (\( \lambda_1 = 6000 \)) is not successful (Figure 8(a)); by decomposing it as \( y_1 = \text{photo of a bald man} \) (\( \lambda_1 = 1500 \)), \( y_2 = \text{photo of a black man} \) (\( \lambda_1 = 3000 \)), and \( y_3 = \text{photo of a man with beard} \) (\( \lambda_1 = 1500 \)), we have better control (Figure 8(b)).

Figure 8: Decomposing complex controls (e.g., photo of a bald black man with beard) into simpler ones improves control performance. See details in Section 4.4.

4.5 Iterative Distributional Control via Functional Composition

This section discusses an interesting bias that PromptGen reveals about the CLIP model. Figure 9(a) shows images generated by PromptGen (with StyleGAN2) with the CLIP model and the text description a photo of a person without makeup, where more females are generated than males. This bias should not be attributed to image pre-training data since images contain a bias in the opposite direction, i.e., men are less likely to have makeup. We argue that this bias should be explained by CLIP having learned a “reporting bias” in vision-language pre-training data: people are more likely to say “a person without makeup” when the person is a female (detailed analysis in Appendix F).

Besides revealing the above “reporting bias”, PromptGen can also mitigate this bias via an iterative control, enabled by the functional composition in Algorithm 1. Specifically, in the second iteration, we de-biased the gender distribution of \( G \circ f_\theta \) instead of \( G \), where \( f_\theta \) is the INN learned for the text control. For de-biasing, we used the moment constraint with \( \hat{\beta} \) trained for \( G \circ f_\theta \). Figure 9(b) shows that females and males are uniformly distributed after the moment constraint in the second iteration.
Figure 9: A curious case of the CLIP model: with description a photo of a person without makeup, PromptGen generates more female images than male, showing that CLIP learns a “reporting bias”. Besides revealing this “reporting bias”, PromptGen can also mitigate this bias via an iterative control, allowing us to de-bias the text-controlled distribution. All images are resized for visualization.

4.6 Inference Latency

We compared our PromptGen and the plug-and-play generative model (PPGM) [47] in terms of the inference latency. PPGM uses Langevin dynamics [74] to optimize over the latent space at inference, while PromptGen samples images in a feed-forward manner. In this experiment, we used PPGM and our PromptGen to approximate the same EBM with CLIP energy. Inference times were estimated on an NVIDIA RTX A4000 GPU. Table 4 shows that our PromptGen has the highest performance and efficiency. Compared to our PromptGen, PPGM requires $100 \times$ inference time to achieve comparable results. Moreover, learning $f_{\theta}$ in PromptGen requires training-time optimization, and this amortized optimization is useful when one wants to reuse a controlled distribution many times.

Table 4: Comparison between PromptGen and PPGM, when approximating the same EBM with CLIP energy. $n$ is the number of inference-time optimization steps used by PPGM. All models used StyleGAN2 as the pre-trained generative model. The text descriptions are a photo of a {baby, boy, girl}.

<table>
<thead>
<tr>
<th></th>
<th>PPGM ($n = 10$)</th>
<th>PPGM ($n = 50$)</th>
<th>PromptGen (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP energy (baby) $\downarrow$</td>
<td>0.7327</td>
<td>0.7134</td>
<td>0.7038</td>
</tr>
<tr>
<td>CLIP energy (girl) $\downarrow$</td>
<td>0.7257</td>
<td>0.7184</td>
<td>0.7199</td>
</tr>
<tr>
<td>CLIP energy (boy) $\downarrow$</td>
<td>0.7263</td>
<td>0.7114</td>
<td>0.7081</td>
</tr>
<tr>
<td>Inference time per sample (sec.) $\downarrow$</td>
<td>4.4</td>
<td>21.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Back-propagation through CLIP at inference</td>
<td>10</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>When is it equal to the EBM?</td>
<td>( n \to \infty )</td>
<td>( n \to \infty )</td>
<td>( \mathbb{D}<em>{KL}(p</em>{\theta}(z)</td>
</tr>
</tbody>
</table>

5 Conclusions and Future Work

This paper proposes PromptGen, a unified framework to learn latent distributions for distributional control of pre-trained generative models. PromptGen leverages the knowledge of various off-the-shelf models and, unlike previous methods, it does not require the availability of these models at inference. PromptGen can sample images in a feed-forward manner, which is more efficient than methods that require optimization at inference. PromptGen offers the generality for algorithmic design and modularity for control composition, and it also enables iterative controls. Experiments validate that PromptGen applies to various generative models (StyleGAN2, StyleNeRF, NVAE), control types (continuous, discrete, and moment constraint), off-the-shelf models (CLIP, classifiers, and inverse graphics models), and data domains (faces, churches, animals, ImageNet, and landscapes).

Limitation and future work: We provide an error analysis in Appendix F and a discussion on societal impact in Appendix G. PromptGen is restricted by the pre-trained generative model’s coverage [3, 27]. PromptGen focuses on mode-seeking and -reweighting instead of domain adaptation, and it is possible to combine PromptGen and domain adaptation of generative models [17]. Also, PromptGen depends on the off-the-shelf models that provide knowledge about the control (Appendix F). In subsequent work, it can be beneficial to explore how to learn energy functions [15, 26] (besides our moment constraint used for de-biasing), which may provide fine-grained control with less bias.
Acknowledgments and Disclosure of Funding

The authors would like to thank the anonymous reviewers, Zoltán Ádám Milacski, Jianchun Chen, and Shubhra Aich for their valuable feedback on drafts of this paper.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
      In Section 5 and Appendix F.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
      In Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
      In Section 3 and Appendix B.
   (b) Did you include complete proofs of all theoretical results? [Yes]
      In Section 3 and Appendix B.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
      We will make our code publicly available on GitHub, with instruction for each experiment. We made great efforts to ensure that most experiments will be run in one line of command.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
      In Appendix B.8; other training details will be included in the code release.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
      In Appendix B.8.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
      We have cited the original papers.
   (b) Did you mention the license of the assets? [Yes]
      License will be included in the code.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
      We do not have new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes]
      All data we use are published datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
      All data we use are published datasets.

5. If you used crowdfunding or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
      We do not have user study.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
      We do not have user study.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
      We do not have user study.
A PromptGen on Synthetic Data

We demonstrate the behavior of our PromptGen with two-dimensional synthetic data, using GAN as the generative model. This synthetic experiment is illustrated in Figure 10. Specifically, we created a “real” distribution that is a Gaussian mixture distribution biased towards one mode. We derive a closed-form fair classifier based on this distribution, detailed in the purple block in Figure 10. When a GAN is trained to approximate this distribution, its outputs are biased towards the over-represented mode, as shown by Figure 10(b). Figure 10(c) illustrates the controllability experiment: using the fair classifier as control, PromptGen learns a distribution concentrated in a specific region of the output space. Figure 10(d) illustrates the de-biasing experiment: using the moment constraint (Section 3.1 and Section 4.2), PromptGen upweights the under-represented regions of the output space.

B Method Details and Derivations

B.1 Controllability as EBM

In Section 3, we defined a control $C$ as $M$ independent properties $\{y_1, \ldots, y_M\}$. For example, $y_1$ can be a text description, and $y_2$ can be an attribute. In this part, we first elaborate on why controllability can be formed as EBMs in Eq. (1). We then provide concrete examples of the distributions derived from different energy functions defined in Section 3.1.

We denote the image prior as $p_\theta(x)$, the only distribution that can be estimated from data when labels for the control are not provided during generative model pre-training. Given the control $C = \{y_1, \ldots, y_M\}$, we resort to Bayes’ theorem to rewrite the conditional distribution $p(x|C)$ as

$$p(x|C) \propto p_\theta(x)p(C|x)$$

(13)

$$= p_\theta(x)p(y_1|x) \prod_{i=2}^{M} p(y_i|x, y_{<i})$$

(14)

$$= p_\theta(x) \prod_{i=1}^{M} p(y_i|x) \quad \text{(independence assumption)}.$$  

(15)

For Eq. (15) to be well-defined, we need to define $p(y_i|x)$ for each $y_i$. In order to incorporate the knowledge of arbitrary off-the-self models (besides image classifiers), we define each $p(y_i|x)$ as

$$p(y_i|x) = \frac{\exp(-\lambda_i E_i(x, y_i))}{Z_i}, \quad Z_i = \int_{y'_i} \exp(-\lambda_i E_i(x, y'_i))dy'_i.$$  

(16)
Eq. (16) says that \( p(y_i | x) \) is proportional to the exponential of an energy function \( E_i(x, y_i) \), where \( y_i \) with lower energy has higher density or mass. Note that Eq. (16) defines a distribution over all possible values of \( y_i \) instead of all possible values of \( x \). Combining Eq. (15) and Eq. (16), we have

\[
p(x | C) = \frac{p_x(x)e^{-E_C(x)}}{Z_X}, \quad E_C(x) = \sum_{i=1}^{M} \lambda_i E_i(x, y_i), \quad Z_X = \int_{x'} p_x(x')e^{-E_C(x')} dx',
\]

which is the same equation as Eq. (1).

In the following, we use Eq. (16) to derive the distributions from the classifier energy, CLIP energy, and inverse graphics energy defined in Section 3.1.

**Classifier energy:** Given a classifier \( P(\cdot | x) \) and the target class \( a \), we define the classifier energy as \( E_{\text{classifier}}(x, a) = -\log P(a | x) \). Using Eq. (16), we arrive at:

\[
p_{\text{classifier}}(a | x) = \frac{\exp(\lambda_{\text{classifier}} \log P(a | x))}{\sum_{a'} \exp(\lambda_{\text{classifier}} \log P(a' | x))} = \frac{P(a | x)^{\lambda_{\text{classifier}}}}{\sum_{a'} P(a' | x)^{\lambda_{\text{classifier}}}},
\]

which is equivalent to a temperature-adjusted distribution of the original classifier.

**CLIP energy:** Using Eq. (16), the CLIP energy in Eq. (3) is equivalent to

\[
p_{\text{CLIP}}(t | x) \propto \exp \left( -\frac{\lambda_{\text{CLIP}}}{L} \sum_{i=1}^{L} \left( 1 - \cos \left( \text{CLIP}_{\text{img}}(\text{DiffAug}_i(x), \text{CLIP}_{\text{ext}}(t)) \right) \right) \right).
\]

**Inverse graphics energy:** Using Eq. (16), the inverse graphics energy in Eq. (4) is equivalent to

\[
p_{\text{inv-graphics}}(\rho | x) \propto \exp \left( -\lambda_{\text{inv-graphics}} \delta(f_X \rightarrow p(x), \rho)^2 \right).
\]

If the geodesic distance \( \delta(\cdot, \cdot) \) is the Euclidean distance, then \( p_{\text{inv-graphics}}(\rho | x) \) is a Gaussian distribution whose mean is \( \rho \) and whose variance depends on the hyperparameter \( \lambda_{\text{inv-graphics}} \); if the geodesic distance \( \delta(\cdot, \cdot) \) is the spherical distance (e.g., the distance between pose parameters defined on a unit sphere), then \( p_{\text{inv-graphics}}(\rho | x) \) is a vMF distribution.

**B.2 Equivalence between Image-Space EBM and Latent-Space EBM**

**Proposition 1.** Define \( x \sim p_x(x) \) as \( z \sim p_z(z), x = G(z) \) and \( p(x | C) \) as \( z \sim p(z | C), x = G(z) \), where \( p(z | C) \) is defined as the following EBM:

\[
p(z | C) = \frac{p_z(z)e^{-E_C(G(z))}}{Z}, \quad E_C(x) = \sum_{i=1}^{M} \lambda_i E_i(x, y_i), \quad Z = \int_{z'} p_z(z')e^{-E_C(G(z'))} dz'.
\]

We have

\[
p(x | C) = \frac{p_x(x)e^{-E_C(x)}}{Z_X}, \quad Z_X = \int_{x'} p_x(x')e^{-E_C(x')} dx'.
\]

In spirit, our proof follows the proof in [50], which follows [19]. The difference between our proof and that in [50] is that we derive \( p(x | C) \) from \( p(z | C) \) while they derived \( p(z | C) \) from \( p(x | C) \).

**Proof.** Based on Lemma 1 in [19] and Lemma 1 in [50], \( p(z | C) \) is equivalent to rejection sampling with proposal distribution \( p_z(z) \) and acceptance probability

\[
r(z) = \frac{e^{-E_C(G(z))}}{M_C \cdot Z}, \quad \text{where } \forall z, M_C > \frac{e^{-E_C(G(z))}}{Z}.
\]

Since \( p(x | C) \) is defined as \( z \sim p_z(z | C), x = G(z) \), \( p(x | C) \) is equivalent to rejection sampling with proposal distribution \( p_z(z), x = G(z) \) and acceptance probability

\[
r(x) = \frac{e^{-E_C(x)}}{M_C \cdot Z}.
\]
Note that the above proposal distribution $z \sim p_z(z), x = G(z)$ is the same as $p_x(x)$ by definition. Based on Lemma 1 in [19] and Lemma 1 in [50], we arrive at:

$$p(x|C) = \frac{p_x(x) r(x)}{\mathbb{E}_{x' \sim p_x(x')} [r(x')]}$$

(25)

$$= \frac{p_x(x) e^{-E_C(x)} / (M_C \cdot Z)}{\mathbb{E}_{x' \sim p_x(x')} [e^{-E_C(x')}/(M_C \cdot Z)]]}$$

(26)

$$= \frac{p_x(x) e^{-E_C(x)}}{\mathbb{E}_{x' \sim p_x(x')} [e^{-E_C(x')}]}$$

(27)

$$= \frac{p_x(x) e^{-E_C(x)}}{Z_X}, \quad Z_X = \int_{x'} p_x(x') e^{-E_C(x')} \, dx'.$$

(28)

### B.3 Derivation of Eq. (9): Approximating EBM with INN

The full derivation of Eq. (9) is given by:

$$\mathbb{D}_{KL}(p_\theta(z)||p(z|C))$$

$$= \mathbb{E}_{z \sim p_\theta(z)} \left[ \log \frac{p_\theta(z)}{p(z|C)} \right]$$

$$= \mathbb{E}_{z \sim p_\theta(z), x = G(z)} \left[ \log \frac{p_\theta(z)}{p_z(z) e^{-E_C(x)/Z}} \right]$$

$$= \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), z = f_\theta(\epsilon), x = G(z)} \left[ \log \mathcal{N}(\epsilon|0, I) - \log \det \left( \frac{\partial f_\theta}{\partial \epsilon} \right) \right]$$

$$- \log p_z(z) + E_C(x) + \log Z$$

(29)

$$= \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), z = f_\theta(\epsilon), x = G(z)} \left[ - \log \det \left( \frac{\partial f_\theta}{\partial \epsilon} \right) \right] - \log p_z(z)$$

$$+ E_C(x) - \mathbb{H}_{\mathcal{N}(0, I)} + \log Z.$$

### B.4 Derivation of Eq. (11): Extension to Generative Models with a Class-Embedding Space

The full derivation of Eq. (11) is given by:

$$\mathbb{D}_{KL}(p_\theta(z, y)||p(z, y|C))$$

$$= \mathbb{E}_{(z, y) \sim p_\theta(z, y)} \left[ \log \frac{p_\theta(z, y)}{p(z, y|C)} \right]$$

$$= \mathbb{E}_{(z, y) \sim p_\theta(z, y), x = G(z, y)} \left[ \log \frac{p_\theta(z, y)}{p_z(z, y) e^{-E_C(x)/Z}} \right]$$

(30)

$$= \mathbb{E}_{\xi \sim \mathcal{N}(\mu, \sigma^2 I), \epsilon \sim \mathcal{N}(0, I), z = f_\theta(\epsilon), y = h_\theta(\xi), x = G(z, y)} \left[ \log \mathcal{N}(\epsilon|0, I) - \log \det \left( \frac{\partial f_\theta}{\partial \epsilon} \right) \right]$$

$$+ \log \mathcal{N}(\xi|\mu, \sigma^2 I) - \log \det \left( \frac{\partial h_\theta}{\partial \xi} \right) - \log p_z(z) - \log p_y(y) + E_C(x) + \log Z$$

$$= \mathbb{E}_{\xi \sim \mathcal{N}(\mu, \sigma^2 I), \epsilon \sim \mathcal{N}(0, I), z = f_\theta(\epsilon), y = h_\theta(\xi), x = G(z, y)} \left[ - \log \det \left( \frac{\partial f_\theta}{\partial \epsilon} \right) \right] - \log \det \left( \frac{\partial h_\theta}{\partial \xi} \right)$$

$$- \log p_z(z) - \log p_y(y) + E_C(x) - \mathbb{H}_{\mathcal{N}(0, I)} - \mathbb{H}_{\mathcal{N}(\mu, \sigma^2 I)} + \log Z.$$
Algorithm 3: Extension of Algorithm 2 to Conditional INN

\[\text{while not converged do}\]
\[\begin{align*}
1. \ & \text{Sample } \epsilon \sim \mathcal{N}(0, I), \rho \sim p_{\rho}(\rho) \\
2. \ & \text{Map } \epsilon \text{ to latent code } z = f_\theta(\epsilon, \rho) \\
3. \ & \text{Map } z \text{ to an image } x = G(z) \\
4. \ & \text{Optimize } \theta \text{ with gradient } \nabla_\theta \left( - \log |\det(\frac{\partial f_\theta}{\partial \epsilon})| - \log p_z(z) + E_{C_\mu}(x) \right)
\end{align*}\]

B.5 Extension to Conditional INN

Algorithm 3 illustrates how we extend the training algorithm (Algorithm 2) to conditional INN, which helps generalize to controls specified by continuous values \(\rho\). The training objective becomes

\[
\arg \min_\theta \mathbb{E}_{\rho \sim p_\rho(\rho)} \left[ \mathcal{D}_{KL}(p_\theta(z|\rho) \| p(z|\mathcal{C}_\rho)) \right],
\]

where \(\mathcal{C}_\rho\) means that the control is specified by value \(\rho\). The derivation of Algorithm 3 is given by:

\[
\begin{align*}
\mathbb{E}_{\rho \sim p_\rho(\rho)} \left[ \mathcal{D}_{KL}(p_\theta(z|\rho) \| p(z|\mathcal{C}_\rho)) \right] \\
= \mathbb{E}_{\rho \sim p_\rho(\rho), z \sim p_\rho(z|\rho)} \left[ \log \frac{p_\theta(z|\rho)}{p(z|\mathcal{C}_\rho)} \right] \\
= \mathbb{E}_{\rho \sim p_\rho(\rho), z \sim p_\rho(z|\rho), x = G(z)} \left[ \log \frac{p_\theta(z|\rho)}{p(z)} \right. \\
& \left. \times e^{-E_{C_\rho}(x)/Z} \right] \\
= \mathbb{E}_{\rho \sim p_\rho(\rho), \epsilon \sim \mathcal{N}(0, I), z = f_\theta(\epsilon, \rho), x = G(z)} \left[ \log \mathcal{N}(\epsilon|0, I) - \log |\det(\frac{\partial f_\theta}{\partial \epsilon})| \right. \\
& \left. - \log p_z(z) + E_{C_\mu}(x) + \log Z \right] \\
= \mathbb{E}_{\rho \sim p_\rho(\rho), \epsilon \sim \mathcal{N}(0, I), z = f_\theta(\epsilon, \rho), x = G(z)} \left[ - \log |\det(\frac{\partial f_\theta}{\partial \epsilon})| - \log p_z(z) \\
& + E_{C_\mu}(x) - \mathbb{H}_{\mathcal{N}(0, I)} + \log Z \right]
\end{align*}
\]

B.6 Derivation for Latent-Space Moment Constraint

Given a mapping \(\gamma : \mathcal{X} \rightarrow \mathbb{R}^K\), moment constraint [7, 36] defines the target distribution \(p(x|C)\) as

\[
p(x|C) = \underset{p(x|C)}{\arg \min} \mathbb{D}_{KL}(p(x|C) \| p_x(x)) \quad \text{s.t.} \quad \mathbb{E}_{x \sim p(x|C)}[\gamma(x)] = \mu,
\]

(33)

Deviation from the pre-trained distribution

Moment constraint

where \(\mu\) is the user-specified constraint. Examples are provided in the main text, and we omit them here for brevity. [7] showed that Eq. (33) can be approximated to an arbitrary precision by

\[
p(x|C) \propto p_x(x) \exp \left( \hat{\beta}^\top \gamma(x) \right),
\]

(34)

where \(\hat{\beta}\) needs to be computed. [36] estimates \(\hat{\beta}\) by solving the following regression problem:

\[
\hat{\beta} = \underset{\beta}{\arg \min} \mathbb{E}_{x^{(1)}, \ldots, x^{(N)} \sim p_x(x)} \left\| \sum_{j=1}^N \exp \left( \beta^\top \gamma(x^{(j)}) \right) \gamma(x^{(j)}) \right\|_2^2.
\]

(35)

In [36], sampling \(x \sim p_x(x)\) is straightforward since they focus on autoregressive language models. In the context of latent-variable vision generative models, where \(x \sim p_x(x)\) is implicitly defined as \(z \sim p_z(z), x = G(z)\), Eq. (35) is equivalent to

\[
\hat{\beta} = \underset{\beta}{\arg \min} \mathbb{E}_{z^{(1)}, \ldots, z^{(N)} \sim p_z(z), x = G(z)} \left\| \sum_{j=1}^N \exp \left( \beta^\top \gamma(x^{(j)}), x^{(j)} \right) \gamma(x^{(j)}) \right\|_2^2.
\]

(36)
Finally, based on Proposition 1, we can generalize the moment constraint to the latent space as

$$p(z|C) = \frac{p_z(z) \exp \left( \hat{\beta}^T \gamma(G(z)) \right)}{Z}, \quad Z = \int_{z'} p_z(z') \exp \left( \hat{\beta}^T \gamma(G(z')) \right) dz'.$$

(37)

Note that Eq. (36) and Eq. (37) are verbatim copies of Eq. (7) and Eq. (6), respectively.

**Algorithm 4:** Extension of Algorithm 3 with ID Energy (Section 4.3)

while not converged do

1. Sample $\epsilon \sim N(0, I), \rho \sim p_\rho(\rho)$
2. Map $\epsilon$ to latent codes $z = f_\theta(\epsilon, \rho)$ and $z_0 = f_\theta(\epsilon, \rho_0)$
3. Map $z$ to an image $x = G(z)$ and $z_0$ to an image $x_0 = G(z_0)$
4. Optimize $\theta$ with gradient

$$\nabla_\theta \left( -\log |\det(\frac{\partial f_\theta}{\partial \epsilon})| - \log p_z(z) + E_{C_\rho}(x) + \lambda_{ID} E_{ID}(x_0, x) \right)$$

B.7 Inverse Graphics Control with Identity Energy

This subsection provides more detail on how we use an inverse graphics model, DECA [16], to control the pose of faces generated by StyleGAN2 trained on FFHQ 1024$^2$, as introduced in Section 4.3. Recall that we use the conditional INN extension, detailed in Algorithm 3, for this experiment. To enable generating different poses of the same identity, we add additional identity energy using the IR-SE50 model [8]. The training algorithm is detailed in Algorithm 4. Specifically, we generate a canonical latent code $z_0 = f_\theta(\epsilon, \rho_0)$ for each $\epsilon$, where $\rho_0$ is the canonical pose. The latent code $z$ and the canonical latent code $z_0$ are mapped to the image $x$ and the canonical image $x_0$. The identity energy in Eq. (12), copied below, encourages the embeddings of the two images to be similar:

$$E_{ID}(x_0, x) = 1 - \cos \langle R(x_0), R(x) \rangle, \quad x_0 = G(z_0), x = G(z).$$

(38)

B.8 Experimental Details

Our INN architecture contains 8 blocks. Each block consists of a soft permutation of channels [2], an affine coupling layer [12], and an ActNorm layer [37]. In the affine coupling layer, we model the sub-network as an MLP with one hidden layer, where the hidden dimension is 256 and the non-linear activation is LeakyReLU$_{0.1}$.

For the prior distribution $p_\rho(\rho)$ of the pose parameter $\rho$ in the inverse graphics experiments (Section 4.3), we sample the $x$-axis and $y$-axis rotations relative to a canonical pose $\rho_0$, whose $x$-axis and $y$-axis rotations are 0.3 and 0, respectively. The relative pose’s $x$-axis and $y$-axis rotations are uniformly sampled from $[-(1/9)\pi, (1/9)\pi]$.

Each experiment was run on an NVIDIA RTX A4000 GPU (with 16G memory). Our code implementation is based on the PyTorch framework. Our code can be found at https://github.com/ChenWu98/Generative-Visual-Prompt.
Figure 11: Additional results for image synthesis based on text description, guided by the CLIP model. As the backbone, we use StyleGAN2 trained on Landscape-HQ with truncation \( \psi = 0.7 \). All images are 256x256 and resized for visualization.

Table 5: Quantitative results for de-biasing categorical attributes (Figure 6). See details in Appendix D. PromptGen de-biases StyleGAN2 in terms of race, age, and gender.

<table>
<thead>
<tr>
<th></th>
<th>FFHQ</th>
<th>MetFaces</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D_{\text{face}}^{\text{KL}} ↓</td>
<td>D_{\text{age}}^{\text{KL}} ↓</td>
</tr>
<tr>
<td>StyleGAN2</td>
<td>0.860</td>
<td>0.597</td>
</tr>
<tr>
<td>PromptGen (ours; ( \lambda = 1 ))</td>
<td>0.286</td>
<td>0.357</td>
</tr>
<tr>
<td>PromptGen (ours; ( \lambda = 2 ))</td>
<td>\textbf{0.099}</td>
<td>\textbf{0.172}</td>
</tr>
</tbody>
</table>

C Additional Results for Image Synthesis based on Text Description

This section presents additional results of PromptGen for image synthesis from text, using the CLIP model as control. As the pre-trained generative model, we explore StyleGAN2 trained on AFHQ-Cats [6], LSUN-Churches [81], FFHQ [34], and Landscape-HQ [68]. Figure 11 and Figure 12 present more results of PromptGen with diverse text descriptions.

D Additional Results for De-Biasing Generative Models

Table 5 provides quantitative results of de-biasing StyleGAN2 across categorical attributes (Figure 6). Specifically, we used the pre-trained classifier provided by FairFace [31] to classify the attributes of the generated images. We then report the KL divergence between the attribute distribution of the generated images and the uniform distribution. Figure 13 and Figure 14 visualize the de-biasing results. We report the KL divergence between the generated distribution and the uniform distribution. Interestingly, on MetFaces, de-biasing the race results in more sculptures, and we postulate that the reason is that almost all paintings and sketches in MetFaces are for white individuals.
Figure 12: Additional results for image synthesis based on text description, guided by the CLIP model. As the pre-trained generative model, we use StyleGAN2 trained on FFHQ, AFHQ-Cats and LSUN-Churches datasets. The captions are the text descriptions given to the CLIP model.
Figure 13: Using our moment constraint, PromptGen de-biases the racial and age distributions of StyleGAN2 trained on FFHQ with truncation $\psi = 0.7$. All synthesized images are $1024^2$ in resolution and resliced for visualization. We fixed the random seed for PromptGen, so please zoom in to see detailed differences between images.
Figure 14: Using the moment constraint, PromptGen de-biases the racial, age, and gender distributions of StyleGAN2 trained on MetFaces with truncation $\psi = 0.7$. All images are $1024^2$ and resized for visualization. Interestingly, de-biasing the race results in more sculptures, and we postulate that the reason is that almost all paintings and sketches in MetFaces are for white individuals. We fixed the random seed for PromptGen, so we recommend zooming in to see differences between images.
Figure 15: PromptGen with intra-sample signed-distance constraints that generate 3D face models with varying distances in the mouth, eye openness, nose-tip stretch, and lip forward.

E Experiments on Generative 3D Face Models

3D face modeling has been an active area of research that has recently gained major interest due to applications in virtual humans, deep faces, and digital actors. Existing 3D deep learning generative models build 3D compact representations of shape and appearance capable of modeling non-linearity (e.g., scatter effects, specularities) that is necessary for generating photo-realistic faces. Providing a generative model with the capability of generating 3D faces with a particular geometry (e.g., a specific distance between eyes or nose length) is particularly useful for technical artists when creating new characters. However, this problem remains unaddressed, partially due to the lack of publicly available 3D databases labeled with such constraints. We refer to this capability as the intra-sample constraint in 3D generative models, and this section shows how our PromptGen framework is able to achieve this by defining a new energy function based on signed distances.

We conducted the experiments on the FaceScape dataset [80], a large-scale 3D human face dataset consisting of 16,940 topologically uniformly registered 3D face meshes along with high-quality textures. The dataset contains 847 identities with 20 expressions per identity, and each mesh consists of 26,317 vertices of which 68 are 3D landmark vertices covering eyes, nose, mouth, jaw, and
eyebrows. With some abuse of notation, we denote a 3D face mesh as \( \mathbf{x} \in \mathbb{R}^{3V} \) (recall that \( \mathbf{x} \) is used to represent an image in other parts of this paper), where \( V = 26,317 \) is the number of vertices. Given an index \( l \) of a vertex (e.g., one of the 68 landmark vertices), the 3D coordinate of this vertex is defined as \( \mathbf{x}[l] \in \mathbb{R}^3 \). We pre-train a generative mesh model \( G \) that maps each latent code \( z \sim \mathcal{N}(0, I) \) to a 3D mesh \( \mathbf{x} \). PromptGen allows us to obtain geometric control on the 3D meshes. Specifically, given two landmark vertex indices \( l_1 \) and \( l_2 \), we define an intra-sample constraint that enforces the signed distance (along a given direction) between the two vertices to be \( s \). To this effect, we define the signed-distance energy as

\[
E_{\text{signed-distance}}(\mathbf{x}) = |d(v_1, v_2) - s|, \quad v_1 = \mathbf{x}[l_1], \ v_2 = \mathbf{x}[l_2]
\]  

(39)

where \( d(\cdot, \cdot) \) is the signed-distance (along a given direction), and \( |\cdot| \) is the absolute value.

Figure 15 presents several examples of 3D mesh control, in which random textures are applied to the generated meshes. It shows that PromptGen successfully controls the generative 3D mesh model in terms of the mouth stretch, eye openness, nose-tip stretch, and lip forward.

F Error Analysis

As we discussed in Section 5, the controlled distribution depends on (1) the pre-trained generative model’s coverage and (2) the off-the-shelf models used for the control. In this section, we provide some examples of PromptGen failing to generate satisfactory images using the CLIP energy. For the pre-trained generative model, we use StyleGAN2 trained on FFHQ, AFHQ-Wild, or LSUN-Churches.
Our first observation is that the CLIP model sometimes fails in modeling linguistic negation. For instance, the text description photo of a man without beard results in a distribution of photos of men with beard (Figure 16(a)). Meanwhile, Figure 16(b) shows that CLIP is capable of modeling the negation of having makeup, but with the “reporting bias” discussed in Section 4.5. Moreover, CLIP seems to have difficulty in numerical reasoning, and gaining control over the count of specific objects in a scene tends to be unsuccessful. We showed this by specifying photo of a church with three windows, which did not result in the desired specification (Figure 16(c)).

To gain a deeper understanding of the above failures and biases, in Figure 17, we provide some image samples from the LAION-400M dataset [63] with CLIP retrieval, using the same text descriptions as Figure 16. CLIP retrieval is based on the CLIP embedding similarity between the web images and text descriptions, while the original text below each individual image is not used. We observe an impressive consistency between CLIP retrieval and PromptGen: in both Figure 16(a) and Figure 17(a), most images have a beard; in both Figure 16(b) and Figure 17(b), all images are female; in both Figure 16(c) and Figure 17(c), some images do not have exactly three windows. This consistency suggests that the failures and biases in Figure 16 should be mostly attributed to the CLIP model rather than to our PromptGen algorithm. We believe that our observation sheds light on the intricacy of contrastive multi-modal (vision-language) pre-training, which is worthy of being further investigated.

Figure 17: For a deeper understanding of the failures and biases illustrated in Figure 16, we provide some image samples from the LAION-400M dataset [63] with CLIP retrieval, using the same text descriptions as Figure 16. CLIP retrieval is based on the CLIP embedding similarity between the web images and text descriptions, while the original text below each individual image is not used. In Figure 17(a), most images have a beard; in Figure 17(b), all images are female; in Figure 17(c), some images do not have exactly three windows. These observations are consistent with those in Figure 16.
Another observation is that the control tends to fail when the text description requires sampling from low-density regions of the pre-trained generative model’s output space. In other words, the control usually fails if the pre-trained generative model does not cover the mode we are trying to gain control over. For example, images faithful to photo of a person yawning and photo of a baby with long hair are not commonly observed in the FFHQ dataset and, hence, these two text descriptions result in degeneration (Figure 18(a)) or weird images (Figure 18(b)). Another example is photo of an animal from the side, which is not commonly observed in the AFHQ-Wild dataset, and Figure 18(c) shows that the generated images fail to follow this description. Even when the control is successful (e.g., when a complex description is decomposed in Figure 8(b)), sampling from low-density regions results in limited diversity (e.g., the backgrounds in Figure 8(b) look similar to each other).

Finally, we also ran some failure controls using PPGM, showing that PromptGen and PPGM reveal similar failure cases of pre-trained generative models and CLIP. Results are shown in Figure 19.

**Figure 18:** When the pre-trained generative model fails in covering certain modes required by the text description, unsatisfactory outputs are produced. In this figure, we show several text descriptions that require sampling from low-density regions of the pre-trained generative model’s output space.

**Figure 19:** Failure modes revealed by PromptGen (Figure 16 and Figure 18) also hold for PPGM.
G Societal Impact

With the improvements in generative models for images, DeepFake technology [75, 72] has become more accessible. Like any new technology, it is a double-edged sword, and it is crucial to research and comprehend the possible advantages and disadvantages of generative models for society.

On the positive side, we show that PromptGen can be used to de-bias pre-trained generative models and to reveal biases learned by text-image models (like CLIP), indicating that PromptGen might be a useful tool for fair AI if used appropriately. The efficient inference provided by PromptGen also helps reduce the computational expense, which has a positive impact on the environment. Better controllability, however, unavoidably makes it simpler to synthesize targeted pictures, which might have detrimental social effects in creating deceptive media (e.g., DeepFakes) or privacy leaks (e.g., identity-conditioned human face synthesis). To battle these cases, there are current technologies that can detect fake media effectively, and we expect companies and users to use these technologies to distinguish what is real from fake. We encourage practitioners to consider these risks when using PromptGen to develop systems.