Unite and Conquer: Plug & Play Multi-Modal Synthesis using Diffusion Models

Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara and Vishal M. Patel
Johns Hopkins University, Baltimore, MD,USA
{ngopala2, wbandar1 and vpatel36}@jhu.edu

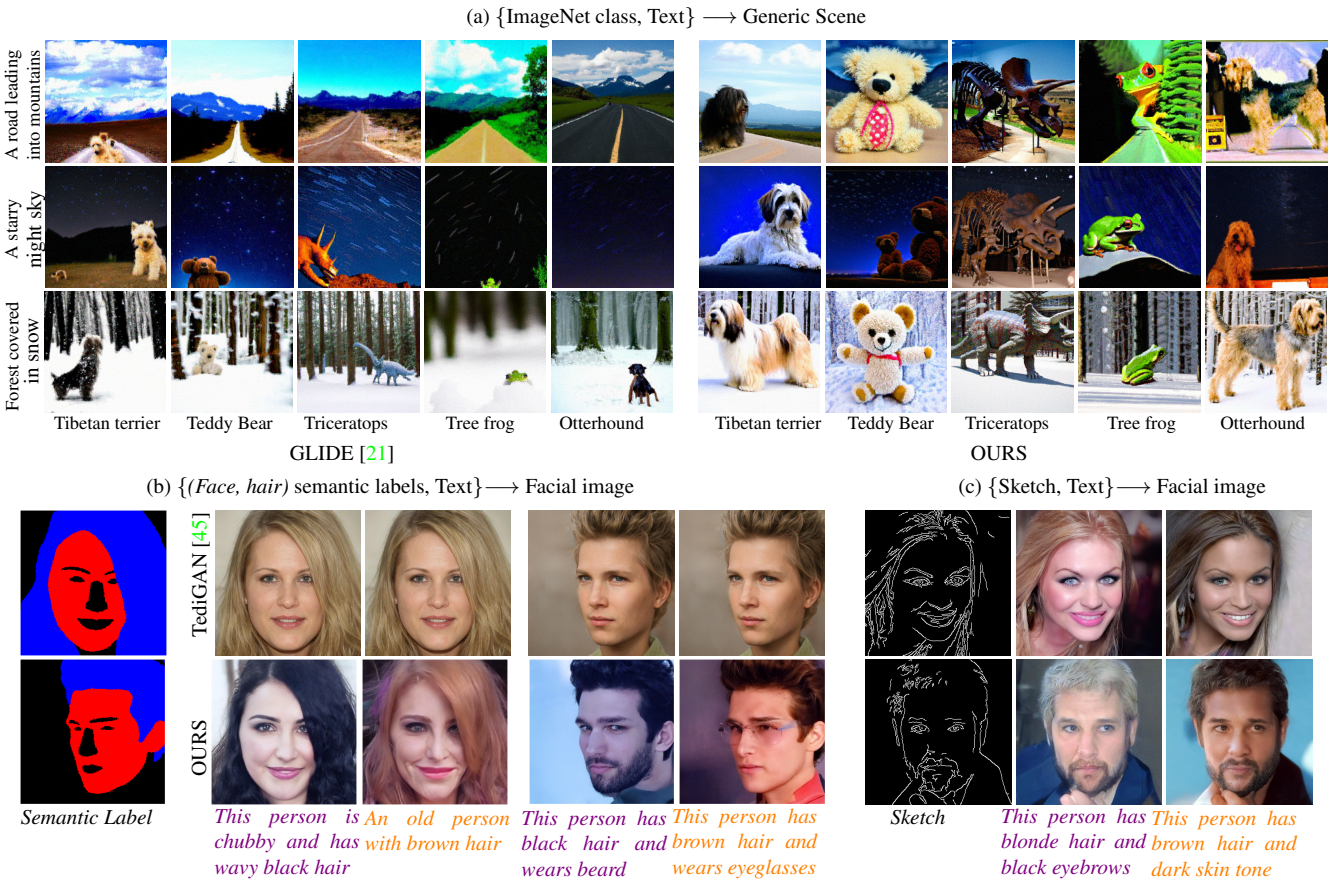

Figure 1. **Applications of our method: (a)** Cross-dataset multimodal generation (ImageNet and GLIDE [21]) using two off-the-shelf diffusion models. We bring in new novel classes to a predefined background. We consider five rare classes from the ImageNet classes, and three different text prompts as shown in the figure. For GLIDE [21], we create a new text prompt by adding an extra sentence utilizing *"and a {} in the photo"*. For GLIDE [21], we sample five images and show the results having the maximum value of CLIP [24] correlation value to the text prompt. (b) Multimodal Face generation with three modalities (hair segmentation map, skin segmentation map, and text). (c) Multimodal Generation with two modalities. The text prompt for (b) and (c) is a user-defined text prompt with multiple attributes.

Abstract

Generating photos satisfying multiple constraints finds broad utility in the content creation industry. A key hurdle to accomplishing this task is the need for paired data consisting of all modalities (i.e., constraints) and their corresponding output. Moreover, existing methods need retraining using paired data across all modalities to introduce a new condition. This paper proposes a solution to this problem based on denoising diffusion probabilistic models (DDPMs). Our motivation for choosing diffusion models over other generative models comes from the flexible internal structure of diffusion models. Since each sampling step in the DDPM follows a Gaussian distribution, we show that there exists a closed-form solution for generating an image given various constraints. Our method can unite multiple diffusion models trained on multiple sub-tasks and conquer the combined task through our proposed sampling

strategy. We also introduce a novel reliability parameter that allows using different off-the-shelf diffusion models trained across various datasets during sampling time alone to guide it to the desired outcome satisfying multiple constraints. We perform experiments on various standard multimodal tasks to demonstrate the effectiveness of our approach. More details can be found at: https://nithingk.github.io/projectpages/Multidiff

1. Introduction

Today's entertainment industry is rapidly investing in content creation tasks [12, 22]. Studios and companies working on games or animated movies find various applications of photos/videos satisfying multiple characteristics (or constraints) simultaneously. However, creating such photos is time-consuming and requires a lot of manual labor. This era of content creation has led to some exciting and valuable works like Stable Diffusion [28], Dall.E-2 [26], Imagen [29] and multiple other works that can create photorealistic images using text prompts. All of these methods belong to the broad field of conditional image generation [25, 33]. This process is equivalent to sampling a point from the multi-dimensional space $P(z|x)$ and can be mathematically expressed as:

$$\hat{z} \sim P(z|x), \tag{1}$$

where $\hat{z}$ denotes the image to be generated based on a condition x. The task of image synthesis becomes more restricted when the number of conditions increases, but it also happens according to the user's expectations. Several previous works have attempted to solve the conditional generation problem using generative models, such as VAEs [18,25] and Generative Adversarial Networks (GANs) [7, 34]. However, most of these methods use only one constraint. In terms of image generation quality, the GAN-based methods outperform VAE-based counterparts. Furthermore, different strategies for conditioning GANs have been proposed in the literature. Among them, the text conditional GANs [3,27,41,45] embed conditional feature into the features from the initial layer through adaptive normalization scheme. For the case of image-level conditions such as a sketches or semantic labels, the conditional image is also the input to the discriminator and is embedded with an adaptive normalization scheme [22, 31, 40, 42]. Hence, a GAN-based method for multimodal generation has multiple architectural constraints [11]

A major challenge in training generative models for multimodal image synthesis is the need for paired data containing multiple modalities [12, 32, 44]. This is one of the main reasons why most existing models restrict themselves to one or two modalities [32, 44]. Few works use more than two domain variant modalities for multimodal generation [11, 45]. These methods can perform high-resolution

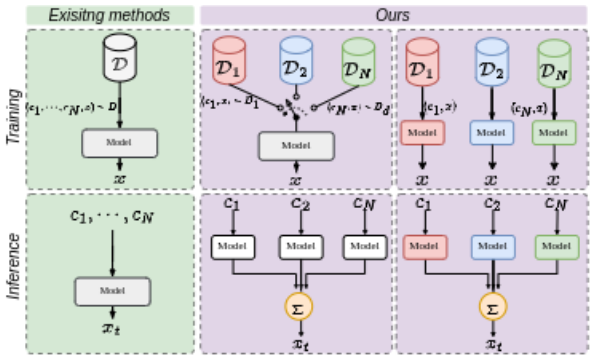

Figure 2. **An illustration of the difference between the existing multimodal generation approaches [45] and the proposed approach.** Existing multimodal methods require training on paired data across all modalities. In contrast, we present two ways that can be used for training: (1) Train with data pairs belonging to different modalities one at a time, and (2) Train only for the additional modalities using a separate diffusion model in case existing models are available for the remaining modalities. During sampling, we forward pass for each conditioning strategy independently and combine their corresponding outputs, hence preserving the different conditions.

image synthesis and require training with paired data across different domains to achieve good results. But to increase the number of modalities, the models need to be retrained; thus they do not scale easily. Recently, Shi *et al*. [32] proposed a weakly supervised VAE-based multimodal generation method without paired data from all modalities. The model performs well when trained with sparse data. However, if we need to increase the number of modalities, the model needs to be retrained; therefore, it is not scalable. Scalable multimodal generation is an area that has not been properly explored because of the difficulty in obtaining the large amounts of data needed to train models for the generative process.

Recently diffusion models have outperformed other generative models in the task of image generation [5, 9]. This is due to the ability of diffusion models to perform exact sampling from very complex distributions [33]. A unique quality of the diffusion models compared to other generative processes is that the model performs generation through a tractable Markovian process, which happens over many time steps. The output at each timestep is easily accessible. Therefore, the model is more flexible than other generative models, and this form of generation allows manipulation of images by adjusting latents [1, 5, 23]. Various techniques have used this interesting property of diffusion models for low-level vision tasks such as image editing [1, 17], image inpainting [20], image super-resolution [4], and image restoration problems [16].

In this paper, we exploit this flexible property of the denoising diffusion probabilistic models and use it to design a solution to multimodal image generation problems with-

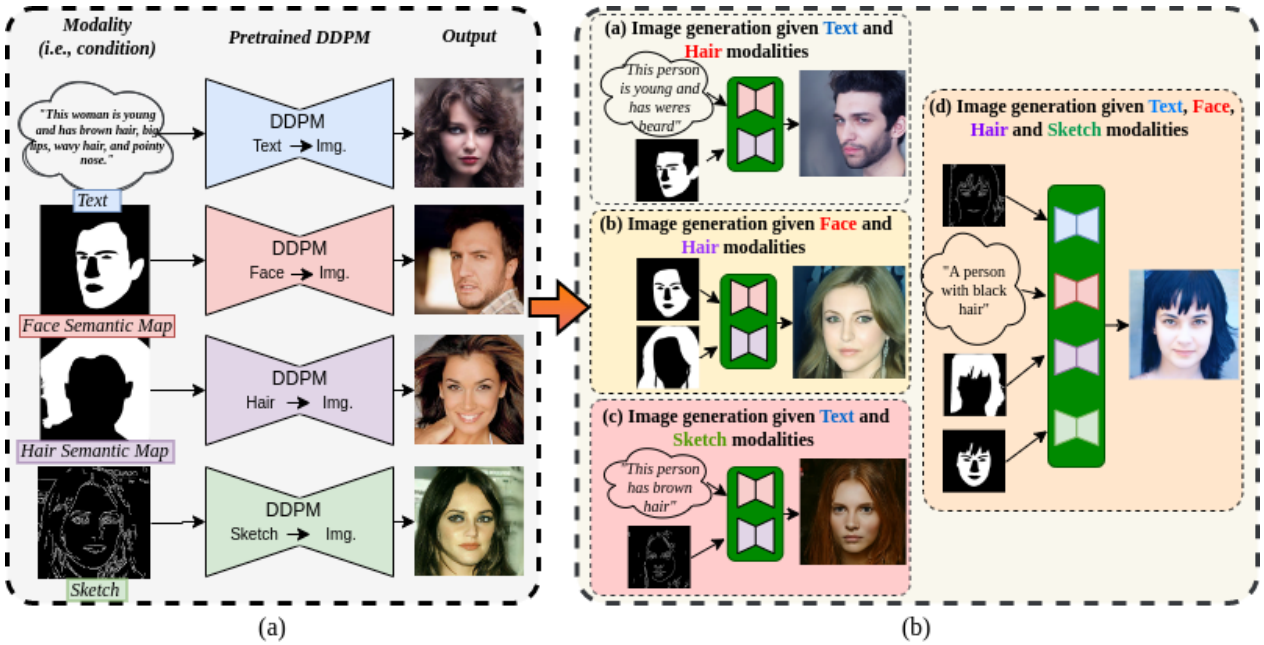

Figure 3. **An illustration of our proposed approach.** *During training,* we use diffusion models trained across multiple datasets (we can either train a single model that supports multiple different conditional strategies one at a time or multiple models). *During Inference,* we sample using the proposed approach and condition them using different modalities at the same time.

out explicitly retraining the network with paired data across all modalities. Figure 2 depicts the comparison between existing methods and our proposed method. Current approaches face a major challenge: the inability to combine models trained across different datasets during inference time [9, 19]. In contrast, our work allows users flexibility during training and can also use off-the-shelf models for multi-conditioning, providing greater flexibility when using diffusion models for multimodal synthesis task. Figure 1 visualizes some applications of our proposed approach. As shown in Figure 1-(a), we use two open-source models [5, 21] for generic scene creation. Using these two models, we can bring new novel categories into an image (e.g. Otterhound: the rarest breed of dog). We also illustrate the results showing multimodal face generation, where we use a model trained to utilize different modalities from different datasets. As it can be seen in 1-(b) and (c), our work can leverage models trained across different datasets and combine them for multi-conditional synthesis during sampling. We evaluate the performance of our method for the task of multimodal synthesis using the only existing multimodal dataset [45] for face generation where we condition based on semantic labels and text attributes. We also evaluate our method based on the quality of generic scene generation.

The main contributions of this paper are summarized as follows:

- We propose a diffusion-based solution for image generation under the presence of multimodal priors.
- We tackle the problem of need for paired data for multimodal synthesis by deriving upon the flexible property of diffusion models.
- Unlike existing methods, our method is easily scalable and can be incorporated with off-the-shelf models to add additional constraints.

2. Related Work

In this section, we describe the existing works on Conditional Image generation using GANs and multimodal image generation using GANs

2.1. Conditional Image Generation

The earliest methods for conditional image generation are based on non-parametric models [46]. However, these models often result in unrealistic images. On the other hand, deep learning-based generative models produce faster and better-quality images. Within deep learning based techniques, multiple methods have been proposed in literature for performing conditional image generation, where the images are generated conditioned on different kinds of input data. have proposed methods where images are generated based on a text prompts [27]. [22] et. al proposed a method for conditioning semantic label inputs based on a spatially-adaptive normalization scheme. [31, 40] Multiple conditional GAN-based techniques also tackle the problem where the conditioning happens from image-level semantics like thermal image [13, 42].

2.2. Multimodal Image generation

Recently multimodal image synthesis has gained significant attention [11,32,37,38,44,45,48], where these method attempts to learn the posterior distribution of an image when conditioned on the prior joint distribution of all the different modalities. The approaches [32, 37, 38, 44, 48] follow a variational auto-encoder-based solution, where all the input modalities are first processed through their respective encoders to obtain the mean and variance of the underlying Gaussian distributions which are combined using the product of experts in the latent space. Finally, the image is generated by sampling using the new posterior mean and variance. Some methods using GANs for multimodal image synthesis have also gained recent attention. Huang et al [11] perform multimodal image synthesis using introduces a new Local-Global Adaptive Instance normalization to combine the modalities. Huang et al [11] also use the product of experts theory to combine encoded feature vectors and use a feature decoder to obtain the final output. TediGAN [45] uses a StyleGAN [15] based framework where the different visual-linguistic modalities are combined in the feature space and decoded to obtain the final output. TediGAN allows user-defined image manipulation according to the input of other modalities.

3. Proposed Method

3.1. Langevian score based sampling from a diffusion process

An alternate interpretation of denoising diffusion probabilistic models is denoising score-based approach [36], where the sampling during inference is performed using stochastic gradient langevian dynamics [43]. Here a network is used to compute the score representing the gradient of likelihood of the data. The sampling during the reverse timestep can be represented by [36]:

$$z_{t-1} \leftarrow \frac{1}{\sqrt{1-\beta_t}}\left(z_t - \beta_t s_\theta(z_t, x, t)\right) + \sigma_t^2 \boldsymbol{\eta}, \quad (2)$$

where $\boldsymbol{\eta} = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, z_t is the sample at timestep t and x is the condition. The score value $s_\theta(\cdot)$ is given by,

$$s_\theta(z_t, t) = \nabla_x \log P(z_t|x) = \frac{\epsilon_\theta(z_t, x, t)}{\sqrt{1-\bar{\alpha}_t}}, \quad (3)$$

where ϵ_θ is the output from the denoising network.

3.2. Multimodal conditioning using diffusion models

In regular conditional denoising diffusion models [30], the input image (i.e., the condition) is concatenated with the sampled noise when passing through the network. When multiple modalities are present, the trivial solution of finding the conditional distribution is by concatenating all the N modalities with the noisy image. However, if we want to improve the functionality of the trained network by adding a new modality, the whole model needs to be retrained with all $N+1$ modalities. Instead, we propose an alternative way to achieve this goal. Let (z, x_i) denote a point in the space of the images of a particular domain and $p(z|x_i)$ denote the distribution of the predicted image z based on the modality x_i. Let $\mathbf{X} = \{x_1, , x_2, ..x_N\}$. Let the distribution of the image conditioned on all modalities be denoted by $P(z|\mathbf{X})$ and the distribution of the image conditioning on the individual modalities be $P(z|x_i)$. Assuming that all the modalities are statistically independent,

$$P(z|\mathbf{X}) = \frac{P(z)}{P(\mathbf{X})}\prod_{i=1}^{N} P(x_i|z) = KP(z)\frac{\prod_{i=1}^{N} P(z|x_i)}{\prod_{i=1}^{N} P(z)}, \quad (4)$$

where K is a term which is independent from z_t. Assuming the individual distributions $P(z|x_i)$ and $P(z)$ follow a Gaussian distribution, the distribution $P(z|\mathbf{X})$ will also follow a Gaussian distribution. Please note that here we consider a graphical model starting from x_i and pointing towards z. The "explaining away" effect does not happen since x_i are independent inputs. Now, let's assume that N diffusion models are trained to generate samples from the distributions $P(z|x_i)$ conditioned on each modality x_i separately. We have N modalities from where the unconditional distribution could be computed. However, how good each model can model the unconditional distribution is not certain, hence we utilize the generalized product of experts rule [2] to compute the effective unconditional density as

$$P(z) = \prod_{i=1}^{N} P_{\delta_i}^{a_i}(z|\phi), \quad (5)$$

where a_i is the confidence factor of each individual distribution with a null condition to modelling the overall unconditional density. To preserve the effective variance so that the reverse diffusion process still holds, we set the constraint $\sum_{i=1}^{N} a_i = 1$. As mentioned in Section 3.1, we can use stochastic gradient Langvein sampling based sampling to sample from the conditional distribution $P(z|\mathbf{X})$. Please note that we are imposing an assumption that the diffusion process for each of the individual modalities have the same variance schedule (a more generalized derivation when the variance schedules are not equal is provided in the supplementary document). Hence the score-based derivations are valid and the effective diffusion process has the same variance schedule

$$\nabla_{z_t} \log(z_t|\mathbf{X}) =$$

$$\nabla_{z_t} \log\left((\prod_{i=1}^{N} P_{\delta_i}^{a_i}(z_t|\phi))\frac{\prod_{i=1}^{N} P_{\delta_i}(z_t|x_i)}{\prod_{j=1}^{N}\prod_{i=1}^{N} P_{\delta_i}^{a_i}(z_t|\phi)}\right) =$$

$$\sum_{i=1}^{N}\left(\nabla_{z_t} \log P_{\delta_i}(z_t|x_i) - \sum_{j\neq i} a_j \nabla_{z_t} \log P_{\delta_j}(z_t|\phi)\right) \quad (6)$$

where δ_i denotes the parameters of the individual distribution densities and ϕ denotes the null condition. b_{ij} denotes

the confidence of a paramteric model to estimate the unconditional density of another model. Hence the effective score when conditioned on all the modalities can be represented in terms of scores of the individual conditional distribution as well as the score of the unconditional model. Hence the effective score s_c is given by:

$$s_c = \frac{\epsilon_c}{\sqrt{1-\bar{\alpha}_t}},$$

$$\epsilon_c = \epsilon_\theta(z_t, \mathbf{X}, t) = \sum_{i=1}^{N} a_i \epsilon_i(z_t, \phi, t) + \sum_{i=1}^{N}\left(\epsilon_i(z_t, x_i, t) - \sum_{j=i}^{N} a_j \epsilon_j(z_t, \phi, t)\right), \quad (7)$$

where $\epsilon_\theta(z_t, x_i, t)$ denotes the output prediction of the individual conditional networks and $\epsilon_\theta(z_t, t)$ is the prediction of the unconditional network. After computing the effective score, sampling could be performed using equation by,

$$z_{t-1} \leftarrow \frac{1}{\sqrt{1-\beta_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_c\right) + \sigma_t^2 \boldsymbol{\eta}, \quad (8)$$

where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. An additional parameter for stringer conditions can be incorporated to (7) using Generalized product of experts [2] (proof in supplementary). Hence giving importance to some modalities over the others by partial weighting to the scores estimated by each modality as follows:

$$\epsilon_c = \sum_{i=1}^{N} w_i \epsilon_i(z_t, x_i, t) - (\sum_{i=1}^{N} w_i - 1)\sum_{j=1}^{N} a_j \epsilon_j(z_t, \phi, t). \quad (9)$$

$$, w_i \geq 1$$

Recently Ho *et al.* [10] proposed a method for sampling from a particular conditional distribution without the need of an explicit classifier as used by previous methods [5]. According to this method, given a model trained for modeling a conditional distribution $p(z|c)$, and with the same model trained for an modelling an unconditional distribution $p(z)$, The effective score for generating samples could be obtained using,

$$\hat{\epsilon}(z_t, c, t) = (1+w)\cdot\epsilon(z_t, c, t) - w\cdot\epsilon(z_t, t), \quad (10)$$

where w is a scalar. By comparing equations (9) and (10), we can see that equation (10), is the case of multimodal conditioning using diffusion models with all modalities being the same.

4. Experiments

In this section we describe in detail the experiments performed and reason out the choice on experimental setup. We consider different multimodal settings for our network and evaluate the performance quantitatively for and multimodal image generation on the CelebA and FFHQ datasets. For multimodal image semantics to face generation, we choose networks that can perform semantic labels to face generation retrain them for scratch for the different scenarios considered. As detailed in earlier sections, one major challenge of multimodal image generation is the lack of paired data across all modalities. To extend existing approaches for the case of multimodal generation, two approaches could be followed. The first is to take a model trained for a particular conditioning modality and finetune it for another modality. In the second approach, the training corpus becomes the combination of the datasets with individual modalities and different iterations could model training sees a different modality and its corresponding image. We re-train the exisitng methods with both these settings and compare the performance on multimodal face generation. We also evaluate the performance of our method with existing compositional models.

4.1. Multimodal Face Generation

To create a criteria to quantitatively evaluate the existing works and our method for multimodal generation, we follow the works [44] where complementary information comes from semantic labels of different portions of the same image. To make the problem even more generic, during training we choose the semantic labels for different regions from different datasets. For multimodal semantics to face generation, we use the CelebA-HQ dataset [14] and the FFHQ dataset [15]. We choose the semantic labels of hair from the CelebA dataset and the semantic labels of skin from the FFHQ dataset for the training process. The semantic labels for the skin and the hair are obtained using the face parser released by Zheng et al. [49]. For training our dataset, we choose 27,000 images from the CelebA-HQ dataset and 27,000 images from the FFHQ dataset, and train the corresponding individual diffusion models with these semantic labels as well as attributes. We test our method using 3,000 images chosen from the CelebA-HQ dataset and 3,000 images from the FFHQ dataset. During evaluations, we utilize the semantic labels of hair as well as skin from the same image to illustrate how well our method is able to generalize. As for the evaluations metrics we choose the FID score [8] and the LPIPS score [47] to evaluate the sample diversity and quality of facial images generated. To evaluate the structural similarity of the produced result with the actual image we use the SSIM score. Finally to see how close the input semantic labels are to the ones produced by our method, we obtain the parsed masks of the generated images and compute the mean intersection over union over all classes (mIoU) and F1 precision scores of the semantic images obtained from the reconstructed image and the input semantic image and report the mean for all testing images. We set the value of $a_i = 1/N$ and $w_i = 1$ for all the experiments. Here N denotes the number of modalities used. For training the sketch to face model, we utilized Canny egde detector [6] to extract the edges. For text to face generation, we utilize the FFHQ dataset and trained a model based on

Figure 4. **Qualitative results for cross dataset multimodal generation** (ImageNet and CompGen [19]) using two off-the-shelf diffusion models.

Figure 5. **Qualitative comparisons for semantic to face generation.** In this case, a single model is trained by alternating different input datasets across different iterations. During Inference time all the modalities are taken from a single dataset and the proposed sampling technique is used.

Table 1. **Quantitative results for multimodal semantic labels to face generation on CelebA dataset.** The combined training based strategy and fine-tuning based training strategy are shown in the corresponding sections. (↑)/ (↓) represents higher/ lower the metric lower the metrics, the better respectively.

Type	Method	FID↓	LPIPS↓	SSIM↑	mIoU↑	F1 ↑
Fine-Tuning	SPADE [22]	131.91	0.638	0.316	0.605	0.694
	OASIS [31]	118.67	0.624	0.318	0.579	0.663
	PIX2PIXHD [42]	153.19	0.611	0.282	0.716	0.819
	INADE [40]	125.43	0.632	0.279	0.881	0.932
Combined	SPADE [22]	89.29	**0.523**	0.376	0.890	0.936
	OASIS [31]	71.20	0.571	0.323	0.792	0.870
	PIX2PIXHD [42]	73.32	0.512	0.373	0.872	0.925
	INADE [40]	54.27	0.552	0.332	0.887	0.933
	TediGAN [45]	69.51	0.4823	0.417	0.834	0.905
	OURS	**26.09**	0.519	**0.416**	**0.911**	**0.948**

the extracted face embeddings using FARL [49]. The reader is referred to the supplementary material for more details on the implementation of the individual multimodal face generation sub-parts.

4.2. Generic Scenes Generation

To show that our method is generalized and could combine existing works to make their generation process more powerful, we use the text-to-image generation-based diffusion model released by GLIDE and an ImageNet class conditional generation model. We perform a combined multimodal generation task where we choose the GLIDE model to decide the background in the image and use the Imagenet model to bring specific objects to the image. During evaluations, we generate images using text prompts that contain a scene information as well as an ImageNet object using an and conditioning. We generate 500 such images for all methods and evaluate how close the images are to the text prompts using the average clip distance between the image embeddings and the embeddings of the input text prompt. We make use of nonreference quality metrics NIQE to evaluate the method. We also present the accuracy using the state-of-the-art Imagenet classifier [39] to detect whether the class is present in the image. We set $w_i = 5$ for all the experiments.

4.3. Analysis and Discussion

Semantic label to face generation. The quantitative results for semantic face generation on the FFHQ and CelebA datasets can be found in Tables 2 and 1, respectively. As the choice of comparison methods, we use the current state-of-the-art method for semantic face generation TediGAN [45] and several recently introduced semantic to face generation methods [22, 31, 40, 42]. The fine-tuning-based multimodal

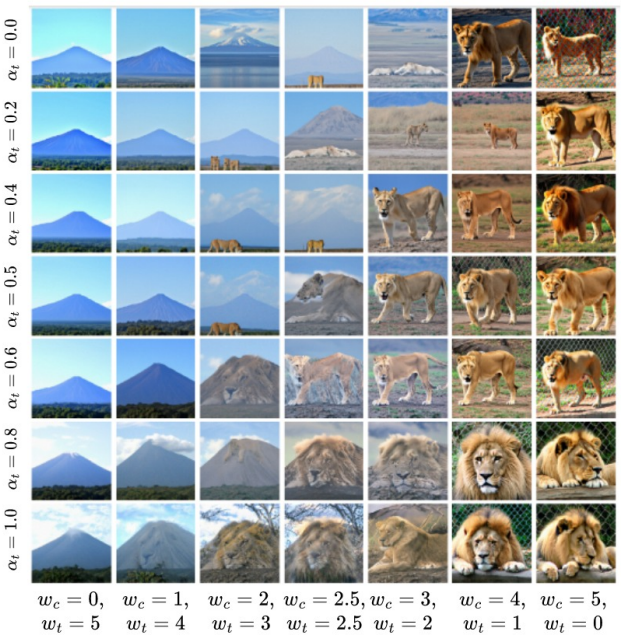

Figure 6. **The role of reliability factor for multimodal generation.** The text is "A mountain". ImageNet class is "Lion". Here,α_t denotes the factor at which the unconditional model for text is weighed. w_c, w_t denotes weights for diffusion model generating based on class and text, respectively.

Type	Method	FID↓	LPIPS↓	SSIM↑	MIoU↑	F1 ↑
Fine-Tuning	SPADE	91.77	0.624	0.340	0.649	0.728
	OASIS	99.94	0.634	0.320	0.625	0.704
	PIX2PIXHD	223.85	0.670	0.265	0.666	0.775
	INADE	134.11	0.656	0.274	0.869	0.922
Combined	SPADE	75.73	**0.545**	0.373	0.876	0.921
	OASIS	75.57	0.593	0.331	0.786	0.863
	PIX2PIXHD	138.30	0.560	0.363	0.840	0.898
	INADE	47.40	0.574	0.334	0.862	0.910
	TediGAN [45]	125.27	0.545	0.409	0.813	0.887
	OURS	**33.60**	0.542	**0.421**	**0.919**	**0.950**

Table 2. Quantitative results for multimodal semantic labels to face generation on FFHQ dataset

generation and the combined unpaired training-based results are shown separately. TediGAN [45] has been trained with paired data across all modalities since it supports such a provision. From Tables 2, and 1 we can see that all the methods fail to produce reasonable results when trained in the finetuning-based strategies. This can be seen in the high FID scores, low SSIM, and parsed mask accuracy metrics. The alternating training strategy produces reasonably good results, and improve all the evaluation metrics improve. But training this way introduces dataset-specific bias because of which the quality of the existing semantic-to-face generation techniques deteriorates when used on an independent test set during testing.

Generic Scenes Generation. Table 3 shows the quantitative comparison for the proposed method. As we can see, the accuracy of the object being in the image is low for

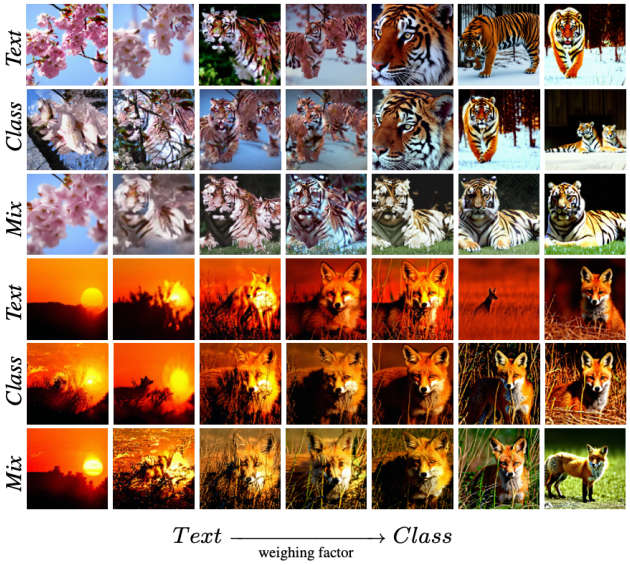

Figure 7. **Interpolation across multiple modalities.** Here, {*class* →*Text*} denotes to where the unconditional model comes from.

Method	Modality	NIQE(↓)	Clip(↑)	Acc(↑)
GLIDE [21]	*Uni*	6.64	**0.317**	0.3
CompGen [19]	*Uni*	5.71	0.282	22.4
Ours	*Multimodal*	**5.34**	0.286	**87.0**

Table 3. **Comparison for generic scenes creation.** Here we consider three different reliability values $\{0, 0.5, 1\}$ and report the best possible value.Acc denotes ImageNet classification accuracy

GLIDE [21] because it focuses on regions on the text which are more easier to generate. One could always make the case that the text model fails because of its weak robustness to doctored text prompts. Hence we perform a comparison with compositional unimodal generation [19] against a multimodal scenario where we utilize information across datasets. This analysis can be seen in Table 3. we evaluate the performances on 500 generated images using the CLIP score, NIQE score and ImageNet classification accuracy. As can be seen multimodal generation has its advantages that it can introduce new novel classes to the image hence being more accurate and can generate realistic images leading to better metrics.

How to choose a good reliability factor? When we have multiple models trained on different datasets, and conditioning needs to be applied based on both of these models. There exist two scenarios. In the case of independent attributes like a face semantic mask and hair semantic map, the reliability factors doesn't affect the composite image since each attribute could be added without affecting the others performance. The next possible case is of non independent attributes, where a blend of both images is a possible solution like in 6, here the reliability factor depends on how much of each image is desired by the user. For an

Label Combined OURS

Figure 8. A visualization of generated images with combined conditional generation (uni-modal) with ours (multi-modal).

Semantics	Modality	FID	SSIM	mIoU	F1
Face	*Uni*	27.66	0.373	0.842	0.907
Face+Hair	*Uni*	61.62	0.376	0.872	0.922
Ours	*Multi*	26.09	0.416	0.911	0.948

Table 4. Analyzing advantage of the proposed sampling method over normal training and regular sampling techniques.

illustration, we refer the reader to Figure 6. Here we utilize the GLIDE model and the ImageNet generation model [5]. We utilize 100 steps of deterministic DDIM [35] with same random initial noise for all settings. The vertical values to the left denotes increasing reliability factor for the text model. When reliability factor equals to zero, the top most row, the class model is used as the unconditional model and bottom row shows the case when the text model is used as the unconditional model. Since the text model has more generation capability, we can see that it creates much better naturalistic series of images compared to the ImageNet generation model [5] which cannot to create artistic scenes.

Why is a reliable score better than a direct solution? One straightforward solution to the case of multi-modal generation is to use equation (10) and consider the most powerful model as the unconditional model. But this scenario is a subset of our reliable mean solution. As one can imagine, this formulation puts a strict bias towards the space of data points of one of the models over the other. Whereas using a reliable mean, this specific bias can be negated and more user defined control is possible. For example, as seen in Fig 6, one can obtain a user defined mix of how much each modality should be mixed using the reiable mean. Moreover, In Fig 7, row 2, we can see that distorted cherry blossoms are created when the ImageNet class conditional model is treated as the stronger model. The text model used is a very powerful model [21]. Hence it is able to model class specific points accurately and a mix of the unconditional densities of both the models can perform valid interpoaltions between both modalities.

5. Ablation studies

To show the effectiveness of the proposed multimodal strategy over a normal method trained unimodally, we retrain a diffusion model that can take one or more conditioning simultaneously. We analyze the FID score and the SSIM scores on the images and their corresponding ground truths, and the F1 score and mIoU score between the original parsed maps vs the reconstructed parsed maps. We perform ablation over three different scenarios: with using one modality when both conditioning modality is given at a time and our proposed sampling strategy. As we can see from Table 4, there is a significant boost to the performance using the specified sampling strategy. This shows that the multimodal sampling strategy enforces stronger conditioning and produces better-quality results. Figure 8 shows the results corresponding to the combined conditioning strategy and our proposed method. As can be seen from this figure8, when regular diffusion-based inference time sampling is performed from a model trained with different modalities across different datasets, the sampling procedure generates unrealistic results. In contrast, we are able to generate much more realistic results.

6. Limitations and Future scope

One limitation of our method is that the dimension of the latent space modelled by the diffusion models is required to be the same. If the number of channels differ for the latent variable z_t, then this method cannot be utilized. This is the reason why we could not utilize stable diffusion model [28] for our experiments. Another scenario when our model can fail is when one model is asked to create a sample that could be easily generated and the other model is asked for a harder class. Thirdly, if contradictory information is given as input across modalities, our method fails to produce the desired output. More visual results corresponding to this condition are given in the supplementary document. This problem can be easily alleviated by using different reliability weights to the different conditioning modalities and giving more weight to the most desired conditioning modality as can be seen from Fig. 6.

7. Conclusion

In this paper, we propose one of the first methods that can perform multimodal generation using individual models trained for multiple sub-tasks. The multimodal generation is enabled by a newly proposed formulation utilizing a generalized product of experts. We introduce a new reliability parameter that allows user-defined control while performing multimodal mixing of correlated modalities. We briefly discuss the design choices of the reliability parameter for different applications. The proposed sampling significantly boosts the performance of multimodal modal generation using diffusion models compared to the sampling using a unimodal network. We show results on various multimodal tasks with trained as well as publically available off-the-shelf models to show the effectiveness of our method.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[2] Yanshuai Cao and David J Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014. 4, 5

[3] Zhangling Chen, Ce Wang, Huaming Wu, Kun Shang, and Jun Wang. Dmgan: Discriminative metric-based generative adversarial networks. *Knowledge-Based Systems*, 192:105370, 2020. 2

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 2

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 5, 8

[6] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern recognition*, 34(3):721–725, 2001. 5

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 6

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[11] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. *arXiv preprint arXiv:2112.05130*, 2021. 2, 4

[12] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *European Conference on Computer Vision*, pages 91–109. Springer, 2022. 2

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4, 5

[16] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 2

[17] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[19] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022. 3, 6, 7

[20] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[21] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 3, 7, 8

[22] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3, 6

[23] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[25] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

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

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2, 3

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 8

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

[30] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021. 4

[31] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations*, 2021. 2, 3, 6

[32] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 4

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

[34] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 2

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 8

[36] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[37] Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal elbo. *arXiv preprint arXiv:2105.02470*, 2021. 4

[38] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016. 4

[39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[40] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7962–7971, 2021. 2, 3, 6

[41] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[42] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3, 6

[43] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011. 4

[44] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 4, 5

[45] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021. 1, 2, 3, 4, 6, 7

[46] Xiaohui Zeng, Raquel Urtasun, Richard Zemel, Sanja Fidler, and Renjie Liao. Np-draw: A non-parametric structured latent variable model for image generation. *arXiv preprint arXiv:2106.13435*, 2021. 3

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[48] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. M6-ufc: Unifying multi-modal controls for conditional image synthesis. *arXiv preprint arXiv:2105.14211*, 2021. 4

[49] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. *arXiv preprint arXiv:2112.03109*, 2021. 5, 6