MultiGen: Zero-shot Image Generation from Multi-modal Prompts

Zhi-Fan Wu, Lianghua Huang, Wei Wang, Yanheng Wei, and Yu Liu

Alibaba Group
wuzhifan.wzf@gmail.com,
{xuangen.hlh, ww413411, yanheng.wyh, ly103369}@alibaba-inc.com

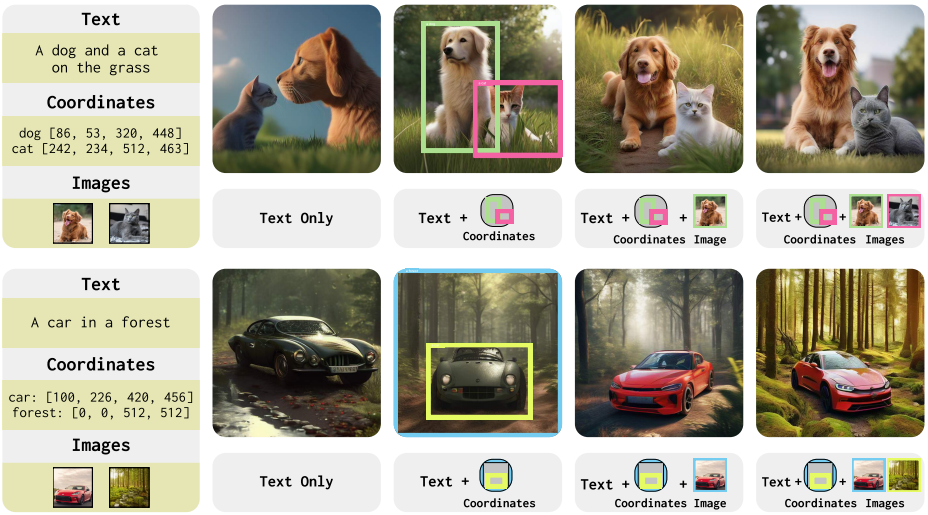

Fig. 1: Zero-shot image generation from multi-modal prompts. MultiGen can generate images by either using text alone (first column) or by combining text with coordinates (second column) and *object-level* images (third and fourth column) *without any model tuning*. This empowers the control of the image generation process through multi-object multi-modal prompts.

Abstract. The field of text-to-image generation has witnessed substantial advancements in the preceding years, allowing the generation of high-quality images based solely on text prompts. However, accurately describing objects through text alone is challenging, necessitating the integration of additional modalities like coordinates and images for more precise image generation. Existing methods often require fine-tuning or only support using single object as the constraint, leaving the zero-shot image generation from multi-object multi-modal prompts as an unresolved challenge. In this paper, we propose MultiGen, a novel method designed

to address this problem. Given an image-text pair, we obtain object-level text, coordinates and images, and integrate the information into an "augmented token" for each object. The augmented tokens serve as additional conditions and are trained alongside text prompts in the diffusion model, enabling our model to handle multi-object multi-modal prompts. To manage the absence of modalities during inference, we leverage a coordinate model and a feature model to generate object-level coordinates and features based on text prompts. Consequently, our method can generate images from text prompts alone or from various combinations of multi-modal prompts. Through extensive qualitative and quantitative experiments, we demonstrate that our method not only outperforms existing methods but also enables a wide range of tasks.

Keywords: Image generation · Multi-modal prompt · Customization

1 Introduction

The field of image generation, particularly in text-to-image synthesis, has experienced notable advancements in recent years [11,23,26,31]. Existing methods can generate high-quality images based on given text prompts [26, 28, 31]. However, text has inherent limitations in accurately describing objects, which restricts text-to-image models from generating user-specified objects [9, 29]. Therefore, other modalities need to be introduced as supplements for precise image generation. Among them, coordinates and images are two important modalities, corresponding to positions and attributes of objects in an image.

Previous studies have made efforts to incorporate modalities other than text into image generation. Customized generation methods often fine-tune pre-trained models using a few object-specific images [9, 14, 29], but this process is time-consuming due to parameter updates through back propagation. Additionally, storing extra parameters for each object increases the overall storage overhead. Recently, some methods explore to introduce image modality in a zero-shot manner to alleviate these challenges [4, 5, 15]. However, these methods typically only support using a single object as the constraint, limiting their applicability. Furthermore, they often only focus on the fixed combination of text and image modalities, lacking flexibility. The challenge of utilizing multi-object multi-modal prompts for zero-shot image generation remains unresolved.

In this paper, we propose a new method called MultiGen to address this challenge. During the training process, given a pair of image and text, we obtain object-level information including the text, coordinate, and image for each object through off-the-shelf open-set detection methods. Subsequently, we integrate the information into an augmented token for each object. These augmented tokens serve as additional conditions and are jointly trained with the text prompts in the diffusion model. By leveraging the augmented tokens, the diffusion model can accurately generate objects within an image. This approach enables our model to effectively handle multi-object multi-modal prompts in a zero-shot manner.

However, during the inference process, there may arise challenges related to missing modalities, such as scenarios where only text modality is provided or

when multi-modal input is available for only some objects. To address these missing modalities, we propose to generate these modalities through generation models. By leveraging the object-level text as conditions, we generate corresponding coordinates and image features for each object. This enables us to complete the missing modalities in situations where only text modality is available. As a result, MultiGen can support generation solely from text or from a combination of text, coordinates, and image modalities. This overcomes the limitation of some existing methods that rely on complete modalities for generation.

We conduct extensive experiments across various scenarios to validate the promising effectiveness of MultiGen in addressing zero-shot image generation from multi-modal prompts. Our qualitative experiments demonstrate that MultiGen successfully generates images using various combinations of multi-modal prompts, while enabling multiple generation applications. Importantly, our model exhibits successful image generation when a relatively large number of objects are provided. Furthermore, our quantitative experiments highlight the superior performance of our method in zero-shot subject-driven generation as well as text-to-image generation.

2 Related Work

Text-to-Image Generation. Text-to-image generation [13,23,26,28,31,32] has made unprecedented progress in recent years. These methods have demonstrated strong capabilities in generating high-quality images based on text prompts. These advancements are mainly based on diffusion models [6, 11, 35, 37] and auto-regressive methods [3, 27]. Among them, models based on diffusion have received significant attention due to high-quality images they generate. Despite these methods being able to generate high-quality, realistic images, they often only support using text prompts and are unable to provide more precise control for image generation through other modalities.

Multi-modal and Customized Generation. Many recent works have extended text-to-image models to include multi-modal and customized generation. Some studies have introduced additional conditions to enhance control capabilities [12, 17, 22, 41, 44]. These methods often rely on the provision of extra modalities or conditions, and cannot generate images when the conditions are missing. Customized generation methods [9,20,29] typically involve an optimization process to enable the model to learn to generate new objects. Although these methods produce high-quality results, they need lengthy training periods and involve additional storage expenses to facilitate the generation of new objects. Some methods enable zero-shot generation by incorporating image modality into the prompts [4,5,15,34]. However, these methods are limited to generating single objects and can only accept combined inputs of text and image modalities. Our method supports image generation leveraging object-level text, coordinates, and image modalities. Additionally, it can accommodate arbitrary combinations of modalities. This makes MultiGen an effective and versatile method.

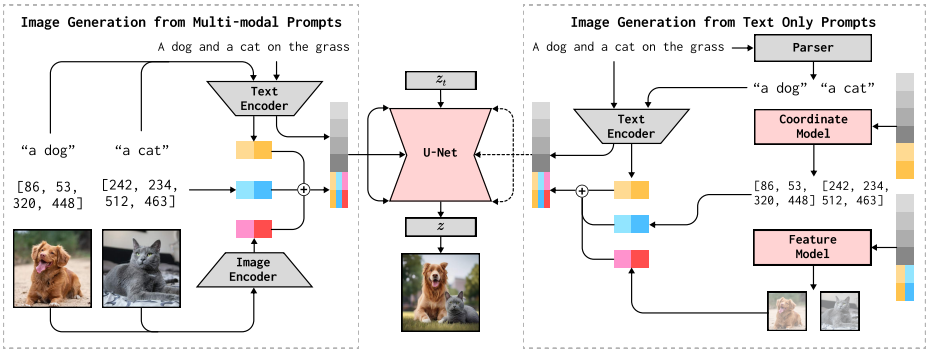

Fig. 2: Overall pipeline of MultiGen, which is for generating images from multi-modal prompts. Our method can handle various combinations of different modalities. When provided with inputs (Sec. 3.1), MultiGen utilizes the multimodal information to obtain augmented tokens for each object, which are used to generate images. In cases where some modalities are missing or only textual modality is provided (Sec. 3.2), MultiGen utilizes the coordinate model to generate coordinates and the feature model to generate image features, in order to compensate for the missing modalities and obtain augmented tokens, which are then used to generate images.

3 Method

We propose MultiGen, an image generation model based on the diffusion process that supports multi-modal prompts by integrating augmented tokens. Given an image-text pair, we obtain object-level multi-modal information including text, coordinates, and images. We construct augmented tokens for each object by incorporating multi-modal information and train them alongside the text prompts as conditions, enabling our model to generate images from multi-object multi-modal prompts in a zero-shot manner.

Moreover, during the inference process, there may be cases where certain modalities are absence, such as only text prompts or incomplete multi-modal prompts are provided. To address this, we propose to utilize generative models to generate the missing modalities. By leveraging the object-level texts extracted from the NLP parser, we generate coordinates and image features for each object to complete the missing modalities. Consequently, our model is capable of generating images by combining various modalities in a flexible manner.

In the following part, we describe the components of MultiGen in detail.

3.1 Image Generation with Augmented Tokens

Diffusion Models. We provide a concise overview of diffusion models, which belong to the category of generative model. Diffusion models convert Gaussian noise into a learned data distribution through an iterative denoising procedure. In addition to generating images of a distribution unconditionally, diffusion models

can also be conditional, such as generating images based on text or low-resolution images. Typically, the mean square error is used as the denoising objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\mathbf{c},\epsilon,t}(||\epsilon - \epsilon_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \epsilon, \mathbf{c})||_2^2), \tag{1}$$

where $\mathbf{x}_0$ is the training data, $\mathbf{c}$ is the optional condition, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the additive Gaussian noise. $t \sim \mathcal{U}([0, 1])$, and α_t, σ_t are scalar functions of t. ϵ_θ represents the diffusion model with learnable parameters θ. When the $\mathbf{x}_0$ is substituted as $z_0 = \mathcal{E}(x_0)$, where $\mathcal{E}$ is a compression model such as Autoencoder, the diffusion process is trained on a latent space. Compared with high-dimensional pixel space, the generative model in latent space pays more attention to important semantic information, and is trained in a lower dimension, so the computational cost is more efficient. In this paper, for the purpose of saving computing resources, we train our model on a pre-trained latent diffusion model. For conditional data sampling from diffusion models, a popular approach is to use classifier-free guidance to adjust the predicted noise:

$$\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) = \omega \epsilon_\theta(\mathbf{x}_t, \mathbf{c}) + (1 - \omega)\epsilon_\theta(\mathbf{x}_t), \tag{2}$$

where $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$, and ω is a guidance weight.

Multi-Modal Augmented Token. Text-to-image diffusion models usually use word embedding $\mathbf{c}_w$ extracted from text prompt p as condition: $\mathbf{c} = [\mathbf{c}_w]$, where $\mathbf{c}_w = f_{\text{text}}(p)$, and f_{text} is a language model. However, the text has a high level of abstraction and limited information density, making it difficult to describe objects accurately. Therefore, we introduce two additional modalities to describe objects, i.e., images and coordinates.

We design our method based on the following principles. Firstly, the multi-modal information of the same object should be bound together to avoid affecting other objects or the global condition. Secondly, we aim to maintain the original structure of the pre-trained text-to-image model as much as possible, enabling the integration of extensions to the existing model.

We introduce augmented tokens to achieve the required properties. Each token contains object-level text, coordinate, and image information at the same time, and is used to describe an object in the generated image. We obtain the augmented token a_i for object i in the following way:

$$a_i = f^g_{\text{text}}(p_i) + \text{MLP}_{\text{img}}(f^g_{\text{img}}(x_i)) + \text{MLP}_{\text{crd}}(r_i), \tag{3}$$

where $f^g_{\text{text}}(\cdot)$ and $f^g_{\text{img}}(\cdot)$ represent extracting global embedding from a language model and image model, respectively. p_i is object-level text of object i, x_i is the cropped image corresponding to p_i in image $\mathbf{x}$, and r_i is the 4-dimensional coordinates of x_i in the image $\mathbf{x}$. MLP_{img} and MLP_{crd} are two feed-forward neural networks for mapping image features and coordinates into the same space as text. Each MLP consists of two linear layers with a GeLU [10] activation function in between.

For all objects involved in the image, we obtain such augmented tokens $\mathbf{c}_a = [a_1, ..., a_n]$, where n is the number of objects in the image. Then, augmented

tokens are appended to the text embeddings: $\mathbf{c} = [\mathbf{c}_w, \mathbf{c}_a]$, and serve as the sampling condition of the diffusion model for image generation. Our method does not require changing the architecture of the diffusion model but only changes the input, maintaining the availability of techniques for building on top of the diffusion model.

Training Data. In order to train the model, we need to get the object-level data required for augmented tokens. Specifically, given an image-text pair $(\mathbf{x}, \mathbf{p})$, we need to detect the object i in image $\mathbf{x}$ and its corresponding coordinates, to construct the object-level multi-modal data (x_i, p_i, r_i). Such data can be obtained through off-the-shelf open-set detection methods. Here, we use Grounding DINO [19] to process approximately 100M of data offline. We perform NMS post-processing on the detected data, set the threshold to 0.7, and limit the number of detected objects to a maximum of 15. For cases where the number of objects is less than 15, we pad augmented tokens with zero.

3.2 Handling the Missing Modalities

In the process of inference, acquiring multi-modal information for all objects can pose challenges. For instance, there may be situations where only text prompts are available, or only certain objects have their multi-modal information provided, such as when users desire to merge real objects with generated ones in an image. Therefore, we propose the utilization of generative models to generate the missing modalities including coordinates and image features, enabling the flexible combinations of multiple modalities.

Object Extraction from Text Prompts. When given a text prompt, the task is to extract the mentioned object from it. Various methods can be employed for object extraction in text, such as constructing a constituency tree to identify noun phrases within the prompts or utilizing a large language model for in-context learning to extract objects. In this case, we utilize the constituency tree to identify objects in text prompts due to its high accuracy and minimal resource requirements. Consequently, we obtain the object-level text: $\mathbf{p}_o = [p_1, ..., p_n] = f_{\text{parser}}(p)$. For the sake of notation simplicity, we will also represent the text embedding mapped by $f^g_{\text{text}}(\cdot)$ as $\mathbf{p}_o$ in the following discussion.

Coordinate Generation. When coordinates are not provided, we generate them for augmented tokens to indicate positions of objects. Intuitively, with the text prompt of the image and object-level text, we have sufficient information to generate the coordinates of objects. In this case, we employ a diffusion model to generate the coordinates for each object. Formally, given $\mathbf{c}_r = [\mathbf{c}_w, \mathbf{p}_o]$, the goal is to generate reasonable $\mathbf{r} = [r_1, ..., r_n]$ for objects in the image. Thus, the training objective can be written as

$$\mathcal{L}_{\text{crd}} = \mathbb{E}_{\mathbf{r}_0, \mathbf{c}_r, \epsilon, t}(||\epsilon - \epsilon_{\theta_{\text{crd}}}(\alpha_t \mathbf{r}_0 + \sigma_t \epsilon, \mathbf{c}_r)||_2^2). \tag{4}$$

Here, the neural network backbone $\epsilon_{\theta_{\text{crd}}}$ is implemented as a Transformer model. It should be noted that since r_i and p_i have a one-to-one correspondence, to enable the network to learn this relationship, we add learnable position embedding

in these tokens. During the training process, we convert absolute coordinate values into relative coordinate values to facilitate generation at various resolutions.
Image Feature Generation. In the case that the object-level images are not given, we generate the corresponding image features based on object-level texts and coordinates of objects, to complete the missing image feature modality, and maintain the alignment among augmented tokens. In Sec. **??**, we verified the importance of incorporating image features through ablation study. Given $\mathbf{c}_e = [\mathbf{c}_w, \mathbf{p}_o, \mathbf{r}]$ as condition, our aim is to generate image embeddings $\mathbf{e} = [e_1, ..., e_n]$ for objects in the image. We leverage the following objective:

$$\mathcal{L}_{\text{feat}} = \mathbb{E}_{\mathbf{e}_0, \mathbf{c}_e, \epsilon, t}(||\epsilon - \epsilon_{\theta_{\text{feat}}}(\alpha_t \mathbf{e}_0 + \sigma_t \epsilon, \mathbf{c}_e)||_2^2), \tag{5}$$

where $\epsilon_{\theta_{\text{feat}}}$ is also a network based on Transformer. For the conditions p_o and r, we follow the form of the augmented token and add $\mathbf{r}$ to $\mathbf{p}_o$ after MLP mapping as the extra condition in addition to $\mathbf{c}_w$. Since there exists a one-to-one correspondence between e_i and the condition token, we also incorporate learnable position embeddings for them.

4 Experiments

In this section, we first introduce the detailed setup of experiments, then present the qualitative and quantitative experimental results. Finally, we conduct detailed ablation studies on MultiGen.

4.1 Experimental Details

Model and Training Details. We employ Stable Diffusion 1.5 [28] as the image generation diffusion model. The batch size is set to 640. We use AdamW [21] as the optimizer, with a learning rate of 1e-4, and training is performed for a total of 100K steps using the cosine scheduler.

The coordinate model and feature model are based on the Transformer architecture [7, 40]. Each model consists of 24 transformer blocks, with 32 attention heads and the dimension of 2048. The total parameter count for each model is approximately 1.2B. Similarly, AdamW is used as the optimizer for the training of these models, with a learning rate of 1e-4, and the training is conducted for 100K steps from scratch. The overall batch size is set to 4096.

To extract text features, we utilize CLIP ViT-L/14 [25] as the text encoder. For image feature extraction, we employ SigLIP ViT-SO400M-384 [43] as the encoder, and for comparison, CLIP ViT-L/14-336 [25] is used.
Datasets Details. We train our model using a combination of internal datasets and public datasets, which collectively consist of approximately 100M images. The public datasets used in our training include ImageNet21K [30], WebVision [16], and a filtered version of the LAION dataset [33]. To ensure data quality, we remove duplicates, low-resolution images, and those that potentially contain harmful content from the LAION dataset.

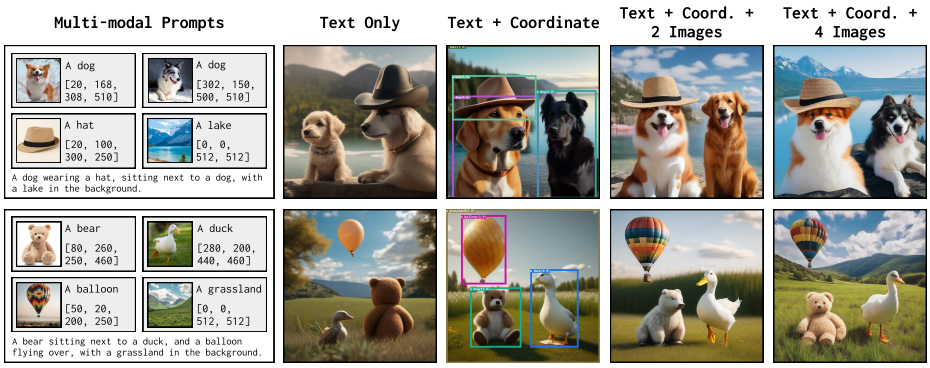

Fig. 3: Qualitative results of MultiGen under different multi-modal prompt combinations. MultiGen supports various combinations of modalities as inputs, including but not limited to: (1) text only, (2) text and coordinates, (3) text, coordinates, and images of some objects, (4) text, coordinates, and images of all objects. Even when provided with 4 object images, MultiGen still generates images successfully.

4.2 Qualitative Results

In this section, we present qualitative results. Due to limited space, we have included the full results in the supplementary material.

Zero-shot Image Generation from Multi-modal Prompts. In Fig. 3, we present the main qualitative results of MultiGen. The results demonstrate the capability of our method in handling different combinations of multimodal prompts. Our method supports flexible combinations of text modality with coordinate and image modalities. Here, we validate four different scenarios, including generating images by (1) using only text, (2) using text and coordinates, (3) using text, coordinates, and images of some objects, and (4) using text, coordinates, and images of all objects. The results indicate that our method is capable of handling a variety of modality combinations effectively. This makes our approach applicable to a wide range of tasks, from simple text-to-image generation to complex tasks which involve placing both real and generated objects at specified locations within a generated image. Notably, even when given images of multiple objects, our method is able to generate images effectively while preserving the original characteristics of the objects.

Comparison of Generation from Single-object Prompts. We compare the image generation capabilities of various methods when provided with a single image and a text prompt. We compare MultiGen with InstructPix2Pix [1], BLIP-Diffusion [15] and Kosmos-G [24] and present the results in Fig. 4. MultiGen displays several advantages. Firstly, MultiGen is capable of seamlessly blending objects with the background mentioned in text prompts, without being affected by the background of the original image. In contrast, we find that InstructPix2Pix and Kosmos-G tends to preserve the background of the original image while disregarding the text prompt. Although BLIP-Diffusion performs slightly better, it occasionally faces similar challenges. Secondly, MultiGen exhibits a stronger

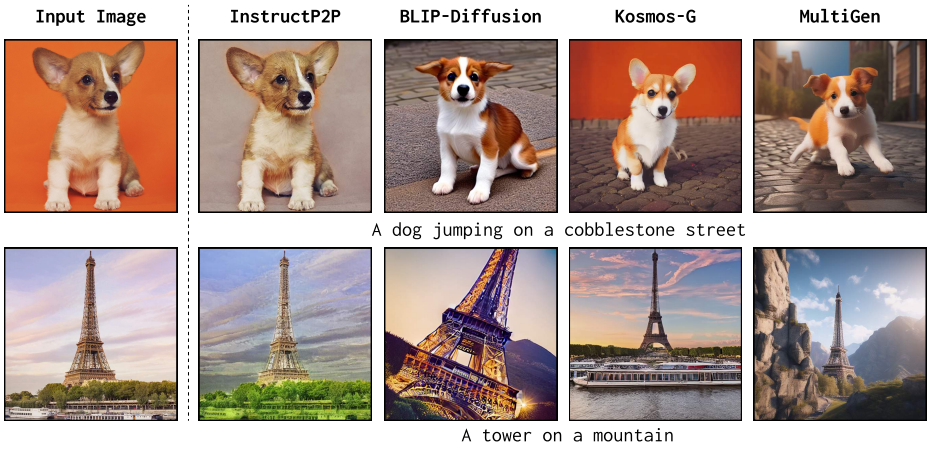

Fig. 4: Qualitative comparison of zero-shot image generation results from *single-object* multi-modal prompts. We compare MultiGen (fifth column) with InstructP2P [1] (second column), BLIP-Diffusion [15] (third column) and Kosmos-G [24] (fourth column). While ensuring object fidelity, MultiGen can make changes to objects and backgrounds according to the requirements of the text prompt.

ability to make reasonable modifications to objects based on the instructions given in the text prompt. For instance, it successfully generates a jumping dog or a cat in a nurse suit, whereas other methods struggle with more complex text prompts. Lastly, MultiGen also demonstrates an advantage in preserving the fidelity of objects.

Comparison of Generation from Multi-object Prompts. We compare the capabilities of image generation when provided with multiple images and the text prompt. We compare MultiGen with Kosmos-G [24] and Emu2 [39], and present the results in Fig. 5. It is clear that MultiGen exhibits significant advantages. Firstly, MultiGen does not omit any of the objects when combining multiple objects. In contrast, Kosmos-G and Emu2 often fail to include all objects in the scene. For example, when prompted with "A bear next to a bird", our method accurately depicts both the bear and the bird in the image, while Kosmos-G and Emu2 overlook the bird. Similarly, with the prompt "An architecture under the planet", MultiGen retains the planet instead of discarding it, unlike Kosmos-G. Secondly, MultiGen effectively blends the backgrounds of two images. For instance, when prompted with "An architecture under the planet", MultiGen seamlessly integrates the architecture and the planet, whereas the image generated by Emu2 still contains remnants of the background from the original architecture image. Lastly, the results produced by MultiGen appear more coherent with the prompts. Overall, MultiGen exhibits a significant advantage in comprehending multi-modal prompts when compared to other existing methods.

Zero-shot Image Customization. Fig. 6 showcases the diverse zero-shot image customization capabilities of MultiGen, utilizing multi-modal prompts. The

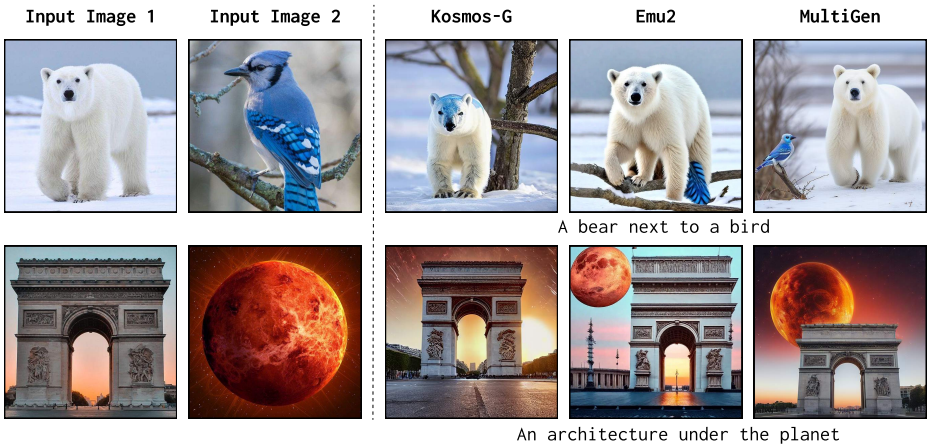

Fig. 5: Qualitative comparison of zero-shot image generation results from *multi-object* multi-modal prompts. We compare MultiGen (fifth column) with Kosmos-G [24] (third column) and Emu2 [39] (fourth column). When combining images of multiple objects, it is evident that MultiGen does not result in missing objects. Additionally, it excels in seamlessly fusing images with backgrounds.

first row exemplifies the ability in artistic stylization, i.e., transforming a provided object image into various styles as dictated by textual descriptions. The second row showcases its attribute modification capability, enabling alterations of object attributes like color and texture via textual descriptions. In the third row, we show MultiGen is able to re-contextualization, situating the provided object within different contexts. The final row reveals its skill in accessorization, allowing the incorporation of a variety of accessories such as hats, clothing, and glasses, into the given object. Collectively, these results demonstrate the ability of MultiGen to effectively integrate information from different modalities.

Zero-shot Style Transfer. Since Multigen inherits the original architecture of stable diffusion, our method can be directly integrated with ControlNet. Such combination can provide additional structure information to MultiGen, which enables zero-shot style transfer. In our approach, we utilize PidiNet [38] to extract the sketch of the structure reference image, which serves as the input to ControlNet and obtain the structure information. Given such structure information provided by ControlNet, we can integrate the features of the reference image. At this point, the style of the generated image is guided by the feature embedding provided to MultiGen. We set the coordinates of the reference image to be the same size as the generated image. Consequently, the reference image effectively guides the structure image in generating the desired style. In Fig. 7, we present the outcomes of zero-shot style transfer facilitated by MultiGen. The results demonstrate that MultiGen can effectively perform zero-shot style transfer, enabling the structure image to acquire the style of a specified reference image without any additional training.

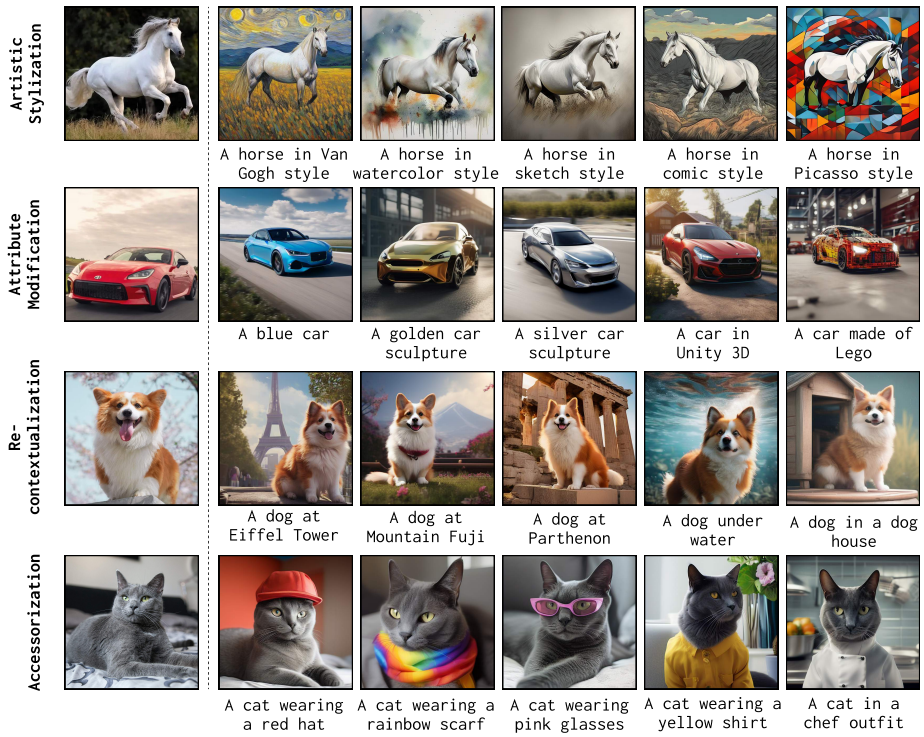

Fig. 6: Zero-shot image customization with MultiGen. Our method supports artistic stylization, attribute modification, re-contextualization, and accessorization through multi-modal prompts in a zero-shot manner.

4.3 Quantitative Results

Results on DreamBooth Benchmark. We compare MultiGen with Texual Inversion, DreamBooth, Re-Imagen, and BLIP-Diffusion on DreamBooth benchmark. This dataset contains 30 subjects in total, each with 4 to 7 images. There are 25 prompt templates, and we generate 4 images for each subject-prompt pair, resulting in 3000 images in total. Since MultiGen is zero-shot and accepts one input image for each object, we pick one image from the 4 to 7 images provided by each subject as description image and extract features. We use a classifier-free guidance scale of 4.0 and 50 DDIM [36] inference steps for sampling.

We follow the setting in DreamBooth [29] and report DINO [2], CLIP-I, and CLIP-T as evaluation metrics. DINO and CLIP-I reflect the subject fidelity, i.e., the preservation of subject details in generated images. While CLIP-T measures the prompt fidelity, i.e., the image-text alignment.

As shown in Tab. 1, our method outperforms other zero-shot generation methods, including Re-Imagen and BLIP-Diffusion. It is worth noting that our method surpasses Textual Inversion without any optimization and with only a

Fig. 7: MultiGen enables zero-shot style transfer. In the first column, we present the structure images, and in the first row, we display the reference images. We are able to transfer the style without any training.

single image input. This indicates that our method exhibits excellent zero-shot multi-modal generation capabilities.

Results on MS-COCO Benchmark. Our method is capable of supporting diverse combinations of multimodal prompts, including image generation from text prompts. To evaluate the effectiveness of our method in generating images based on text, we conduct experiments on the MS-COCO 2014 validation set [18]. Following the approach of previous studies [28, 31], we randomly sample 30,000 captions for image generation. We set the classifier-free guidance scale to 2.0 and utilize DDIM sampling for 50 steps. Notably, our zero-shot FID achieves 9.84 as shown in Tab. 2, surpassing the performance of the base model SD v1.5 that we employed. This showcases the strong performance of our method in the text-to-image task while also supporting multi-modal prompts.

4.4 Ablation Studies

In this section, we validate the impact of image feature encoders and image feature extraction methods on the generation based on image conditions. We also analyze the capabilities of the image feature model and the coordinate model, demonstrating their efficacy in tackling the issue of missing modalities. Due to space limitations, we include this part of work in the supplementary materials.

Different Image Feature Encoders. We observe that the choice of image encoders for feature extraction significantly impacts the fidelity and quality of

Table 1: Quantitative comparisons on Dreambooth Benchmark. MultiGen achieves the best zero-shot generation performance.

Methods	DINO↑	CLIP-I↑	CLIP-T↑
Real Images (Oracle)	0.774	0.885	-
Tuning-based generation			
Textual Inversion [9]	0.569	0.780	0.255
DreamBooth [29]	0.668	0.803	0.305
BLIP-Diffusion [15]	0.670	0.805	0.302
Zero-shot generation			
Re-Imagen [5]	0.600	0.740	0.270
BLIP-Diffusion [15]	0.594	0.779	0.300
MultiGen	**0.615**	**0.780**	**0.308**

Table 2: Zero-shot text-to-image generation FID on MS-COCO validation set. 30,000 samples are randomly sampled for evaluation.

Methods	FID↓
GLIDE [23]	12.24
Make-A-Scene [8]	11.84
DALL-E 2 [26]	10.39
Imagen [31]	7.27
Parti [42]	7.23
SD v1.5 [28]	9.93
MultiGen	**9.84**

image generation. In the main experiments, we utilize SigLIP ViT-SO400M-384, a CLIP-based model with 400M parameters that achieved an accuracy of 83.2% on the ImageNet-1k validation set, to extract features. For comparison purposes, we train another image generation diffusion model using CLIP ViT-L/14-336, which achieves a zero-shot accuracy of 76.2% on the ImageNet-1k validation set.

Despite similar parameter numbers, SigLIP demonstrates superior representation ability, resulting in improved generation outcomes. In Fig. 8, we showcase the generated results using SigLIP and CLIP for image feature extraction in row (c) and row (d), respectively. It is evident that the model utilizing SigLIP features better preserves the attributes of objects in the original image, including shape, color, material, appearance, etc. This highlights the importance of using an image encoder with a more effective representation space as a crucial factor in enhancing image generation from multi-modal prompts.

Different Image Feature Extraction Methods. Another factor that affects the generation results is the way to extract object-level image features. In general, there are two methods to extract these features. The first involves obtaining the coordinates of objects in the image, subsequently cropping the image based on these coordinates to yield object-level images, and finally extracting features from these images. The alternative approach involves extracting the complete feature map of the original image and then employing feature pooling techniques to obtain regional features. For example, RoIAlign can be used to pool regional features from the feature map of the full image through interpolation. In the main experiments, we leverage the first method, e.g., we directly crop the image according to the coordinates to obtain the object-level images, then resize them to the input size of the image encoder for feature extraction.

We compare the two methods and the results are shown in Fig. 8. It can be seen that the regional features extracted using RoIAlign cannot accurately restore the attributes of the original object as showing in row(e). The obvious deviations include color, shape, material, and texture compare with the origi-

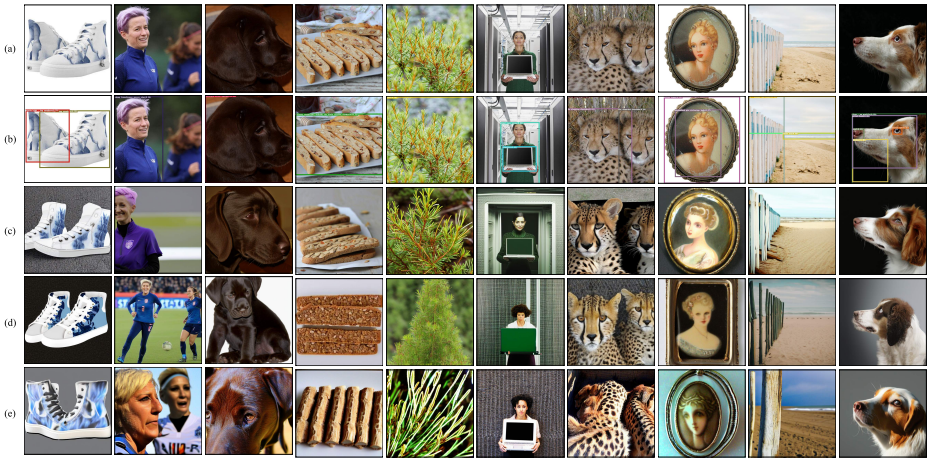

Fig. 8: Ablation studies of image encoders and image feature extraction methods. (a) Randomly sampled original images. (b) Object detection is performed using open-set detection to identify objects in the images along with their corresponding texts. (c) Images are generated based on the provided object-level images, texts, and coordinates, where image features are extracted by SigLIP-SO400M-384. (d) Image features are extracted by CLIP ViT-L/14-336. (e) Images are generated using regional features obtained through RoIAlign instead of cropping-based feature extraction.

nal objects. We speculate that this is due to the inherent problem of CLIP's local features, as previous research has shown that its local features contain a considerable amount of noisy activations. Therefore, extracting local features from the feature map can lead to attribute confusion, resulting in unsatisfactory generation results. On the other hand, directly extracting global features from object-level images is a better approach as it avoids the problem of inaccurate local features in CLIP.

5 Conclusion

In this paper, we present MultiGen, a diffusion-based model capable of generating images from multi-object multi-modal prompts. By creating augmented tokens that combine text, image, and coordinate modalities, we enable the model to generate images based on fine-grained multi-modal prompts, thus enhancing the capability of image generation. Our model can generate missing modalities through the text modality, allowing our approach to support the flexible combination of multiple modal inputs. We consider MultiGen as a significant advancement in zero-shot image generation from multi-modal prompts. Moving forward, we aim to expand the incorporation of modalities and explore wider applications based on the proposed MultiGen framework.

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. arXiv:2211.09800 (2022)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9650–9660 (2021)
3. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M., Murphy, K.P., Freeman, W.T., Rubinstein, M., Li, Y., Krishnan, D.: Muse: Text-to-image generation via masked generative transformers. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 4055–4075 (2023)
4. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. arXiv:2304.00186 (2023)
5. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. arXiv:2209.14491 (2022)
6. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems 34 (NeurIPS). pp. 8780–8794 (2021)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations (ICLR) (2021)
8. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision (ECCV) (2022)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv:2208.01618 (2022)
10. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv:1606.08415 (2016)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems 33 (NeurIPS). pp. 6840–6851 (2020)
12. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 13753–13773 (2023)
13. Kang, M., Zhu, J., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10124–10134 (2023)
14. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. arXiv:2212.04488 (2022)
15. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. arXiv:2305.14720 (2023)
16. Li, W., Wang, L., Li, W., Agustsson, E., Gool, L.V.: Webvision database: Visual learning and understanding from web data. arXiv:1708.02862 (2017)
17. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: GLIGEN: open-set grounded text-to-image generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22511–22521 (2023)

18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
19. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. arXiv:2303.05499 (2023)
20. Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones: Concept neurons in diffusion models for customized generation. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 21548–21566 (2023)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 5th International Conference on Learning Representations (ICLR) (2017)
22. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv:2302.08453 (2023)
23. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: Proceedings of the 39th International Conference on Machine Learning (ICML). pp. 16784–16804 (2022)
24. Pan, X., Dong, L., Huang, S., Peng, Z., Chen, W., Wei, F.: Kosmos-G: Generating images in context with multimodal large language models. arXiv:2310.02992 (2023)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125 (2022)
27. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8821–8831 (2021)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
29. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv:2208.12242 (2022)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
31. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv:2205.11487 (2022)
32. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 30105–30118 (2023)
33. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv:2111.02114 (2021)

34. Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., Taigman, Y.: knn-diffusion: Image generation via large-scale retrieval. In: 11st International Conference on Learning Representations (ICLR) (2023)
35. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning (ICML). vol. 37, pp. 2256–2265 (2015)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: 9th International Conference on Learning Representations (ICLR) (2021)
37. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Advances in Neural Information Processing Systems 32 (NeurIPS). pp. 11895–11907 (2019)
38. Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., Pietikäinen, M., Liu, L.: Pixel difference networks for efficient edge detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 5097–5107 (2021)
39. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. arXiv:2312.13286 (2023)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NeurIPS). pp. 5998–6008 (2017)
41. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., Wang, L.: Reco: Region-controlled text-to-image generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14246–14255 (2023)
42. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research (2022)
43. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. arXiv:2303.15343 (2023)
44. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv:2302.05543 (2023)