# MMoT: Mixture-of-Modality-Tokens Transformer for Composed Multimodal Conditional Image Synthesis

Jianbin Zheng[1†], Daqing Liu[2†], Chaoyue Wang[2*], Minghui Hu[3], Zuopeng Yang[4], Changxing Ding[1*] and Dacheng Tao[2]

[1*]School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510006, China.
[2*]JD Explore Academy, Beijing, 100176, China.
[3]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, Singapore.
[4]Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240, China.

*Corresponding author(s). E-mail(s): chaoyue.wang@outlook.com; chxding@scut.edu.cn;
Contributing authors: jabir.zheng@outlook.com; liudq.ustc@gmail.com; e200008@e.ntu.edu.sg; yzpeng@sjtu.edu.cn; dacheng.tao@gmail.com;
†These authors contributed equally to this work.

**Abstract**

Existing multimodal conditional image synthesis (MCIS) methods generate images conditioned on any combinations of various modalities that require all of them must be exactly conformed, hindering the synthesis controllability and leaving the potential of cross-modality under-exploited. To this end, we propose to generate images conditioned on the compositions of multimodal control signals, where modalities are imperfectly complementary, *i.e.*, composed multimodal conditional image synthesis (CMCIS). Specifically, we observe two challenging issues of the proposed CMCIS task, *i.e.*, the modality coordination problem and the modality imbalance problem. To tackle these issues, we introduce a Mixture-of-Modality-Tokens Transformer (MMoT) that adaptively fuses fine-grained multimodal control signals, a multimodal balanced training loss to stabilize the optimization of each modality, and a multimodal sampling guidance to balance the strength of each modality control signal. Comprehensive experimental results demonstrate that MMoT achieves superior performance on both unimodal conditional image synthesis (UCIS) and MCIS tasks with high-quality and faithful image synthesis on complex multimodal conditions. The project website is available at https://jabir-zheng.github.io/MMoT.

**Keywords:** Image synthesis, Multimodal conditions, Transformer, Modality coordination, Modality imbalance

## 1 Introduction

With the maturity of image synthesis quality, we start focusing on improving its controllability to generate specific images as we expected. To control the image content, various control signals from different modalities have been proposed, including texts (Zhou et al., 2021), sketches (T.-C. Wang et al., 2018), segmentation masks (T.-C. Wang et al., 2018), and bounding box layouts (Sun & Wu, 2021), where each of them has its own advantages, *e.g.*, texts usually describe attributes of objects

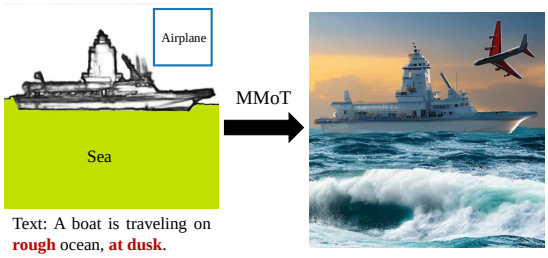



**Fig. 1** CMCIS relaxes the stringent requirements for inputs. The input can be the composition of multiple complementary modalities (*e.g.*, text, sketch, segmentation mask, and bounding boxes), and our MMoT can generate reasonable results leveraging all inputs.

and style of images, segmentation masks depict the environmental contexts, and sketches/layouts are able to control the position and size of each object.

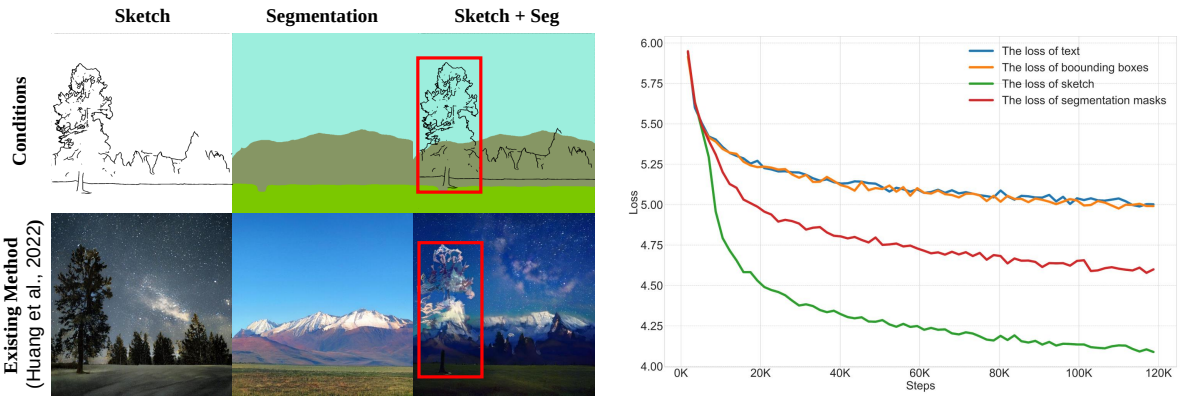
The pioneering methods mostly focus on unimodal conditional image synthesis (**UCIS**) that generates images with a single unimodal control signal, *e.g.*, text-to-image (Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022; Ramesh et al., 2021; Saharia et al., 2022; Yu et al., 2022; Zhou et al., 2021), sketch-to-image (Esser, Rombach, & Ommer, 2021; Park, Liu, Wang, & Zhu, 2019; T. Wang et al., 2022; T.-C. Wang et al., 2018), segmentation-to-image (Esser et al., 2021; Park et al., 2019; Sushko et al., 2022; T. Wang et al., 2022; T.-C. Wang et al., 2018), and layout-to-image (S. He et al., 2021; Li, Wu, Koh, Tang, & Sun, 2021; Sun & Wu, 2021; Sylvain, Zhang, Bengio, Hjelm, & Sharma, 2021; Z. Yang, Liu, Wang, Yang, & Tao, 2022; Zhao, Yin, Meng, & Sigal, 2020). Despite the great progress they have achieved, those methods fail to fully utilize the information from different modalities, thus hindering their controllability and applications in real scenarios. To this end, recent works (X. Huang, Mallya, Wang, & Liu, 2022; Li et al., 2022; Z. Zhang et al., 2021) propose to generate images conditioned on any combinations of various modalities, *e.g.*, sketch+segmentation or text+segmentation, termed multimodal conditional image synthesis (**MCIS**). However, the current MCIS strictly requires every unimodal signal must be *exactly* conformed with each other, leaving the potential of cross-modality underexploited. Additionally, it is unfriendly to the majority of users without a professional painting background.

In this paper, we propose a more challenging task, namely Composed Multimodal Conditional Image Synthesis (**CMCIS**), which allows each unimodal signal to be imperfectly complementary and the final expectation of images to be composed of all modality signals. For example, as illustrated in Figure 1, the user can generate a desired image by using a variety of different modalities to describe different components of the scene, *i.e.*, drawing a sketch of the boat, filling the sea with a segmentation mask, using a bounding box to decide the size and location of the airplane, and the text giving high-level semantic information such as dusk.

Unfortunately, existing MCIS methods, no matter GAN-based (X. Huang et al., 2022), Transformer-based (Li et al., 2022; Z. Zhang et al., 2021), or diffusion-based (L. Huang et al., 2023), are not able to handle the new challenging CMCIS task. As illustrated in Figure 2, we observe two main issues of existing MCIS methods: **(i)** the modality coordination problem due to the non-adaptive fusion on fine-grained information across multiple modalities, *e.g.*, the tree incorrectly composed with the texture of the mountain in Figure 2 (a); and **(ii)** the modality imbalance problem that caused by the imbalanced distribution of each modality in datasets, *i.e.*, different modalities tend to converge at different rates in Figure 2 (b).

Specifically, the *modality coordination problem* arises from the imperfectly complementary of input multi-modality signals, where each image region may involve a different combination of modality (*e.g.*, sky by segmentation masks, tree by sketches+segmentation masks in Figure 2 (a)), *i.e.*, the proposed CMCIS task. Such imperfection demands dynamic coordination of modalities (*i.e.*, adaptive fusion) to adapt to varying image regions. The existing methods have limitations in effectively capturing the fine-grained coordination among modalities adapted to different regions (*i.e.*, non-adaptive fusion). They either represent each modality as a single latent vector, thereby sacrificing the regional specificity of modality coordination, or take all modality signals as input then simultaneously learn cross-modal and cross-region interactions, making it challenging to achieve fine-grained adaptation.

And the *modality imbalance problem* arises due to the varying levels of information density exhibited by different modalities, *e.g.*, text tokens being




**(a) Modality coordination problem** **(b) Modality imbalance problem**

**Fig. 2** (a) Modality coordination problem: incorrect coordination of multiple modalities, *e.g.*, the tree incorrectly composed with the mountain. (b) Modality imbalance problem: various modalities tend to converge at different rates and converge to different endpoints.

high density, while segmentation tokens being lower. Such differences over the whole dataset lead to imbalanced distribution of each modality. During training, the imbalance further affects convergence difficulty (K. He et al., 2022), which manifests as different modalities converging at different rates and to different endpoints (*e.g.*, sketch converges faster than other modalities in Figure 2 (b) because it describes more detailed conditional information and is therefore easier to optimize). However, existing methods treat each modality equally thus suffering from the issue of modality imbalance, particularly in the proposed CMCIS task.

To tackle the former modality coordination problem, we propose the Mixture-of-Modality-Tokens Transformer (MMoT) to fully exploit the cooperativity across modalities. Specifically, MMoT uses multiple encoders to model the intra-modal interaction. Then, modality-specific cross-attention is adopted to inject multimodal conditional information into the decoder. Finally, the key module multistage token-mixer adaptively fuses multimodal conditioning information with the masked cross-attention mechanism. To tackle the modality imbalance problem, we propose a multimodal balanced loss to adaptively control the optimization of each modality during the training phase, as well as a multimodal sampling guidance during the sampling phase to control the influences of different modalities and introduce divergence maps to the sampling process to realize more spatially coordinated generation.

To the best of our knowledge, we are the first to focus on specific challenges of the CMCIS task, and our contributions are summarized as follows:

- We propose a new challenging task, namely Composed Multimodal Conditional Image Synthesis, which allows users to input various control signals that are not perfectly complementary.
- We propose the Mixture-of-Modality-Tokens Transformer (MMoT) for CMCIS, which adaptively fuses fine-grained conditional signals across different modalities.
- We introduce the multimodal balanced training loss and the divergence-driven sampling guidance to alleviate the imbalance problem between multiple modalities in CMCIS.
- The proposed MMoT accomplishes high-quality image synthesis conditioned on complex compositions of multiple modalities and achieves new state-of-the-art performance on COCO-stuff (Caesar, Uijlings, & Ferrari, 2018; Lin et al., 2014) and LHQ (Skorokhodov, Sotnikov, & Elhoseiny, 2021).

## 2 Related Work

### 2.1 Unimodal Conditional Image Synthesis

Deep generative models are a family of techniques in which deep neural networks are

trained to simulate the distribution of training data. (Bond-Taylor, Leach, Long, & Willcocks, 2021). There are a variety of generative models have been proposed, such as energy-based models (LeCun, Chopra, Hadsell, Ranzato, & Huang, 2006), normalizing flows (Kobyzev, Prince, & Brubaker, 2020; Papamakarios, Nalisnick, Rezende, Mohamed, & Lakshminarayanan, 2021), variational autoencoders (VAEs) (Kingma & Welling, 2014; Sohn, Lee, & Yan, 2015), generative adversarial networks (GANs) (Goodfellow et al., 2020; Mirza & Osindero, 2014; C. Yang, Shen, & Zhou, 2021), generative image transformers (GITs)(Chang, Zhang, Jiang, Liu, & Freeman, 2022; Esser et al., 2021) and denoising diffusion models (Dhariwal & Nichol, 2021; Ho, Jain, & Abbeel, 2020). Generative models typically make trade-offs in quality, sampling speed, and diversity.

GIT is one of the more popular models of late, especially for uni-modal conditional generation (Gafni et al., 2022; Yu et al., 2022) as the advance in discretizing multi-modalities and the powerful sequence modeling capabilities. Recently, large-scale text-to-image generative models (Ramesh et al., 2022; Saharia et al., 2022; Yu et al., 2022) have made explosive processes and achieved unprecedented superior results.

Traditional GITs (M. Chen et al., 2020; Child, Gray, Radford, & Sutskever, 2019; N. Parmar et al., 2018) treat image synthesis as a "pixel-by-pixel" autoregressive sequence generation task with the help of the self-attention mechanism (Vaswani et al., 2017). However, as the computation requirement is highly correlated with the sequence length, sampling a high-resolution image may be a challenging endeavor. The proposed Vector Quantised model (Van Den Oord, Vinyals, et al., 2017) significantly reduces the processing burden and enables the sampling of high-resolution images based on GITs (Esser et al., 2021; Ramesh et al., 2021). And as the scale of the model increases, so does the ability to generate the model (Yu et al., 2022).

The design of model architecture is another point of interest. Instead of using decoder-only language models, recent research (Wu, Liang, Ji, et al., 2022; Yu et al., 2022) employs an encoder-decoder transformer for conditional image synthesis and achieves promising results. Our work proposed a novel encoder-decoder-based architecture that decouples intra-modal interaction and fusion.

## 2.2 Multimodal Conditional Image Synthesis

Multimodal conditional image synthesis has attracted increasing attention recently. Representative work includes M6-UFC (Z. Zhang et al., 2021) and PoE-GAN (X. Huang et al., 2022).

M6-UFC is a Bert-based framework based on the two-stage image synthesis method (M. Chen et al., 2020; Esser et al., 2021; Ramesh et al., 2021; Razavi, Van den Oord, & Vinyals, 2019; Van Den Oord et al., 2017). In M6-UFC, the multimodal conditional inputs and generated image are transformed into a sequence of tokens to be processed by the unidirectional Transformer decoder (Vaswani et al., 2017). The advantages of M6-UFC are that it unifies various modalities in a universal form and can thus easily extend to more guidance modalities. However, it employs concatenation to combine multimodal user inputs, and intra-modal interaction is interlaced with fusion; as a result, it may struggle with handling missing modalities (X. Huang et al., 2022; Ma, Ren, Zhao, Testuggine, & Peng, 2022).

PoE-GAN is GANs based method with the multiscale projection discriminator (Liu, Yin, Shao, Wang, et al., 2019; Miyato & Koyama, 2018; T.-C. Wang et al., 2018). In PoE-GAN, conditional information from multiple modalities is first encoded into a unified latent space and then fused using product-of-experts modeling (Hinton, 2002). The advantages of PoE-GAN are that it decouples intra-modal interaction and fusion, and is more robust to missing modalities. However, conditional GANs are known to be susceptible to mode collapse (Isola, Zhu, Zhou, & Efros, 2017; Odena, Olah, & Shlens, 2017) and spatial information is lost in the latent space of PoE-GAN. MMoT inherits the advantages of the above two works and is able to produce reasonable outputs even with counterfactual condition modalities.

Recently, there have been some methods (L. Huang et al., 2023; Mou et al., 2023; L. Zhang & Agrawala, 2023) that attempt to introduce more control signals to large-scale Text-to-Image pre-trained models (*e.g.*, Stable diffusion (Rombach, Blattmann, Lorenz, Esser, &

Ommer, 2022)) to achieve multimodal conditional image generation. However, such methods are text-centric, so it is difficult to leverage the impact of various imperfectly complementary modality in the case of our proposed CMIS task.

# 3 Method

The goal of CMCIS is to train a single model that approximates image distributions conditional on any combination of feasible modalities. Mathematically, the objective is to learn $p(x|\mathcal{C})$ when given a dataset of images $x$ paired with a conditional input $\mathcal{C}$ which consists of $M$ different modalities $\mathcal{C} \subseteq \{c_1, c_2, \ldots, c_M\}$.

Our model follows an encoder-decoder structure, consisting of $M$ encoders for modeling intra-modal interaction and a common decoder for fusion among multi-modalities, as shown in Figure 3. Our proposed methods are described in detail later, comprised of 1) describing the two-stage transformer-based framework, 2) introducing the Mixture-of-Modality-Token transformer with a multistage token-mixer module, 3) training with multimodal balanced loss, and 4) sampling with divergence-driven guidance.

## 3.1 Preliminary

In this section, we review the two-stage transformer-based framework (M. Chen et al., 2020; Esser et al., 2021; Ramesh et al., 2021; Razavi et al., 2019; Van Den Oord et al., 2017) for image synthesis. At the first stage, a convolutional model consisting of an encoder $E$ and a decoder $D$ is learned, such that, an image $x \in \mathbb{R}^{3\times H\times W}$ can be represented as a collection of code-words $z_{\mathbf{q}} \in \mathbb{R}^{n_z\times h\times w}$ with code from a learned, discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K}$, where $n_z$ is the dimensionality of codes, $z_k \in \mathbb{R}^{n_z}$ is the $k^{th}$ code-word, and $K$ is the number of code-words. More precisely, the convolutional encoder $E$ first encodes the $x$ as $\hat{z} = E(x) \in \mathbb{R}^{n_z\times h\times w}$ and then an element-wise quantization $\mathbf{q}(\cdot)$ is applied to each spatial element $\hat{z}_{ij} \in \mathbb{R}^{n_z}$ to obtain the closest discrete code-work $z_k$, *i.e.*, $\mathbf{q}(\hat{z}_{ij}) = \arg\min_{z_k\in\mathcal{Z}} \|\hat{z}_{ij} - z_k\|$. At last, a given image $x$ can be approximated by a convolutional decoder $D$, *i.e.*, $\hat{x} = D(z_{\mathbf{q}})$. Overall, the first stage can be denoted by:

$$z_{\mathbf{q}} = \mathbf{q}(E(x)), \hat{x} = D(z_{\mathbf{q}}), \tag{1}$$

the convolutional model and the codebook can be trained via the loss function:

$$\mathcal{L}_{\mathcal{VQ}}(E, D, \mathcal{Z}) = \|x - \hat{x}\|^2 + \|\mathrm{sg}[E(x)] - z_{\mathbf{q}}\|_2^2 + \|\mathrm{sg}[z_{\mathbf{q}}] - E(x)\|_2^2, \tag{2}$$

where sg[·] denotes the stop-gradient operation.

At the second stage, the quantized encoding of an image $x$ can be transformed into a sequence of code-word index $s \in \{0, \ldots, |\mathcal{Z}| - 1\}^{h\times w}$. Therefore, image generation can be formulated as autoregressive sequence generation modeled with a transformer. The transformer learns to predict the distribution $p(s_i|s_{<i})$ of possible next token $s_i$ given preceding tokens $s_{<i}$. Then the likelihood of the full representation can be computed as $p(s) = \prod_i p(s_i|s_{<i})$. The transformer can be trained by maximizing the log-likelihood:

$$\mathcal{L}_T = \mathbb{E}_{x\sim p(x)}\left[-\log p(s)\right]. \tag{3}$$

## 3.2 Mixture-of-Modality-Tokens Transformer

We propose a novel framework, namely Mixture-of-Modality-Tokens Transformer (MMoT) for adaptive fusion among multimodal information.

### 3.2.1 Modality representation in unified-form

In this paper, we consider four commonly used input conditions, including text, segmentation mask, sketch, and bounding box layout. For visual modalities, *i.e.*, image, segmentation masks, and sketch, we quantize it with the help of discrete VAE, *e.g.*, VQGAN (Esser et al., 2021). For bounding boxes layout, we follow the same tokenization method as in (Jahn, Rombach, & Ommer, 2021). And CLIP (Radford et al., 2021) is our solution for tokenizing text modalities.

Specifically, to unify the embeddings of different modalities including image, segmentation masks, sketch, bounding boxes, and text, we treat all of them as language-like tokens and embed the tokens to the latent space $U_m \in \mathbb{R}^{l\times d}$, where $m$ is

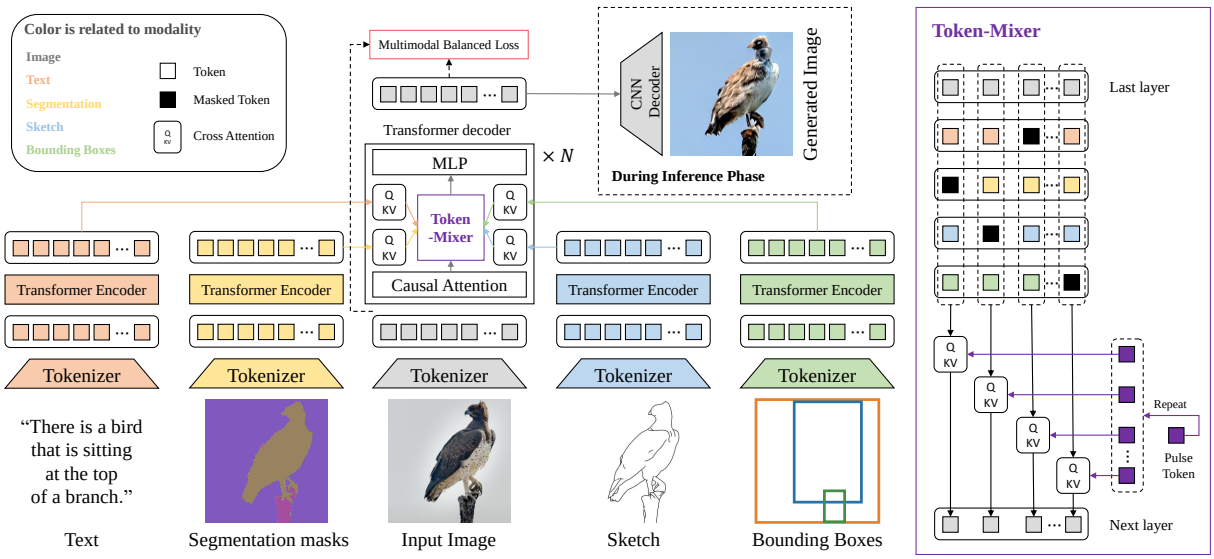



**Fig. 3** The overview of the Mixture-of-Modality-Tokens Transformer for CMCIS task. Given an image and multiple modalities, including text, segmentation masks, sketch, and bounding boxes, we tokenize them into discrete tokens with different tokenizers, and then (a) model intra-modal interaction with modality-specific encoders; (b) inject multimodal conditioning information into the decoder with modality-specific cross-attention; (c) adaptively fuse conditional signals via the multistage token-mixer. We train MMoT with multimodal balanced loss and sample with divergence-driven multimodal guidance.

used to distinguish different modalities, $l$ denotes the number of tokens and $d$ means the dimension of each embedding.

Given an image $\mathcal{X} \in \mathbb{R}^{3\times H_I\times W_I}$, we quantize it with the help of a discrete VAE, e.g., VQGAN (Esser et al., 2021). Specifically, the discrete VAE encodes the image to a latent space, and maps the latent features to the closest discrete tokens from the codebook $\mathcal{Z}_{image} = \{z_k\}_{k=1}^{K}$ with $K$ entries. The quantized representations are in the format of $\mathbb{N}_{image}^{\{h_i,w_i\}}$, where $\mathbb{N}_{image}$ is the set of integers in $[0, 1, \ldots, K)$, and $h_i, w_i$ are the dimension of quantized representations. The segmentation masks $\mathcal{S} \in \mathbb{R}^{H_S\times W_S}$ and the sketch $\mathcal{K} \in \mathbb{R}^{3\times H_K\times W_K}$ can be tokenized in the same pipeline but with different pre-trained discrete VAE models. Such content of segmentation masks and sketch in the raw space can be represented as $s \in \mathbb{N}_{seg}^{\{h_s,w_s\}}$ and $k \in \mathbb{N}_{sketch}^{\{h_k,w_k\}}$ with different codebook $\mathbb{N}_{seg} \in [0, 1, \ldots, K')$ and $\mathbb{N}_{sketch} \in [0, 1, \ldots, K^{\dagger})$, respectively.

Given a bounding boxes layout consists of a set of objects of their positions and category classes, we directly tokenize it into sequential object tokens $b = \{(o_i, p_i)_{i=1}^{N_B}\}$ with $N_B$ objects, where $o_i$ denotes the $i^{th}$ object's category, and $p_i = [tl_i, br_i]$ represents its top-left and bottom-right corner positions.

For image, segmentation masks, sketch, and bounding boxes, we feed their discrete codes to respective embedding layers to get the continuous representations.

As for text modalities, CLIP (Radford et al., 2021) is our solution. A transformer-based encoder is used to represent the discrete token sequence into a high dimensional vector once the sentence is tokenized with Byte-Pair Encoding (BPE). Then we use this vector as the final representation in the form of $\mathbb{R}^{\{1,d\}}$, which means the text embedding only occupies one token as the input to our MMoT module. For datasets that lack corresponding textual descriptions, we use image embedding as the pseudo-text representation from CLIP as a replacement.

In general, the representations of images can be denoted as $X \in \mathbb{R}^{l_{image}\times d}$ and the representations of the conditional modality $m$ can be denoted in the same dimension as $C_m \in \mathbb{R}^{l_m\times d}$ but with different length of tokens $l_m$, and as we discussed, the text modality contributes only one token, where $l_{text} = 1$.

### 3.2.2 Attention mechanism

We recall the attention mechanism since it is an important means for MMoT to achieve interaction and fusion. The attention mechanism draws

global dependencies between them with Query-Key-Value (QKV) model, where queries $Q = W_Q X$, keys $K = W_K C$ and values $V = W_V C$ with the learnable weights $W$. Thus the attention function can be defined as:

$$\text{Attn}(X, C) = softmax(\frac{QK^T}{\sqrt{D}})V. \quad (4)$$

It is worth noting that with the given condition $C$ that differs from the input $X$, the attention mechanism is widely known as cross-attention $\text{CA}(X, C)$. If the condition information $C$ is the same as the input $X$, the attention mechanism can be expressed as $\text{Attn}(X, X)$, which is known as self-attention $\text{SA}(X)$.

#### 3.2.3 Modeling intra-modal interaction with modality-specific encoders

To model the intra-modal interaction for various modalities, we introduce modality-specific encoders that project input features into an intermediate representation.

Each modality encoder consists of $N$ self-attention layers. Given the input features $C_m^e(n-1) \in \mathbb{R}^{L_m \times D}$ of modality $m$ in $(n-1)^{th}$ layer, the output $C_m^e(n) \in \mathbb{R}^{L_m \times D}$ of the $n^{th}$ encoder layer can be calculated as:

$$C_m^e(n) = \text{SA}[C_m^e(n-1)]. \quad (5)$$

#### 3.2.4 Injecting multimodal conditioning information with modality-specific cross attention

In order to inject multimodal conditional information into the decoder, we use modality-specific cross-attention to fuse the image feature with each modality feature.

Specifically, given the input image features $X(n-1)$ of the $(n-1)^{th}$ decoder layer and the output $C_m^e(N)$ of the encoder's last layer, the output $C_m^d(n) \in \mathbb{R}^{L_{image} \times D}$ for $n^{th}$ decoder layer is given by:

$$\begin{aligned} X(n) &= \text{SA}[X(n-1)], \\ C_m^d(n) &= \text{CA}[X(n), C_m^e(N)]. \end{aligned} \quad (6)$$

#### 3.2.5 Adaptive fusion with multistage token-mixer

After applying cross-attention, the multistage token-mixer is proposed to fuse the modality tokens which contain conditional information related to a specific modality. A special `[PULSE]` token $P$ within token-mixer is introduced to adaptively estimate the combination weights (*i.e.*, attention scores) of each modality token and fuse them with the masked cross-attention mechanism. The combination-weight maps in Figure 9 (c) show that `[PULSE]` token can effectively evaluate the influences of different modalities in different decoder layers.

Specifically, with the output $X(n)$ of the $n^{th}$ self-attention layer and a set of outputs of the $n^{th}$ cross-attention layer, we adapt a multistage token-mixer to fuse the conditional information from different modalities and then feed the fusion features to the subsequent decoder layers:

$$X^{\dagger}(n) = \text{Mixer}[X(n), \boldsymbol{C}^d(n)], \quad (7)$$

where $\boldsymbol{C}$ is the stack of all conditional modalities, *i.e.*, $\boldsymbol{C} \in \mathbb{R}^{M \times L_{image} \times D}$.

The Mixer function can be defined as:

$$\text{Mixer}(X, \boldsymbol{C}) := softmax(\frac{PF^T}{\sqrt{D}})F, \quad (8)$$

where we concatenate the latent representation of image $X(n)$ with the stack representation $\boldsymbol{C}$ of $M$ modalities along the modality dimension to form $F \in \mathbb{R}^{L_{image} \times (M+1) \times D}$, while its transpose can be denoted as $F^T \in \mathbb{R}^{L_{image} \times D \times (M+1)}$. $P \in \mathbb{R}^{L_{image} \times 1 \times D}$ is the `[PULSE]` token. Noted that the random masks will be applied in the Mixer function during the training phase, which serve as modality dropouts to handle missing-modality cases during inference.

### 3.3 Multimodal Balanced Loss

To train the MMoT, we propose $\mathcal{L}_{mmb}$, an improved cross-entropy loss named multimodal balanced loss, for sequential prediction tasks to realize balanced optimization among different modalities:

$$\mathcal{L}_{mmb} = \mathbb{E}_{s \sim p(s)} \mathbb{E}_{x, \mathcal{C}_s \sim p(x, \mathcal{C}_s)} \left[ -\log p(x|\mathcal{C}_s) \right], \quad (9)$$

where $p(s)$ is the probability of the occurrence of subset $\mathcal{C}_s$. In order to adaptively control the optimization of each subset, we set:

$$p(s) = -\log p(x|\mathcal{C}_s) \Big/ \sum_{k=1}^{2^M} \left[-\log p(x|\mathcal{C}_k)\right]. \tag{10}$$

As the joint conditional distribution in Eq. (9) is able to be represented as the product of sequential conditional distributions in an auto-regressive process, we can have:

$$p(x|\mathcal{C}_s) = \prod_i p(x_i|x_{<i}, \mathcal{C}_s), \tag{11}$$

where $i$ is the index in a token sequence. In the simplest case, $p(s) = 1/2^M$, which means that any subset $\mathcal{C}_s$ of input modalities has the same chance to appear in the forward process, would cause an imbalanced multimodal optimization because different subsets contain different levels of control information. It is therefore necessary to introduce a parameter to indicate the strength of the given conditions. Intuitively, $p(x|\mathcal{C}_s)$ indicates how easy it is to optimize this subset, so that when $p(s)$ proportional to $[-\log p(x|\mathcal{C}_s)]$ can make MMoT focus on the subsets that are more difficult to optimize. We will show the effectiveness of multimodal balanced loss in ablation experiments.

### 3.4 Divergence-driven Sampling Guidance

During inference, we propose multimodal guidance for the CMCIS task to balance the influences of various control signals. Assuming that all the input conditional modalities are statistically independent and using $M'$ to denote the number of modalities in subset $\mathcal{C}_s$, the sequential conditional distribution $p(x_i|x_{<i}, \mathcal{C}_s)$ in Eq. (11) can be rewritten as follows:

$$\begin{aligned} p(x_i|x_{<i}, \mathcal{C}_s) &= p(x_i|x_{<i}) \left[\frac{p(\mathcal{C}_s|x_i, x_{<i})}{p(\mathcal{C}_s|x_{<i})}\right] \\ &= p(x_i|x_{<i}) \prod_{m=1}^{M'} \left[\frac{p(c_m|x_i, x_{<i})}{p(c_m|x_{<i})}\right]. \end{aligned} \tag{12}$$

Inspired by the influence of the conditioning signal can be amplified by the guidance scale (Dhariwal & Nichol, 2021), we use $\lambda_m$ to control the influence of the $m^{th}$ conditioning signals:

$$\begin{aligned} p_\lambda(x_i|x_{<i}, \mathcal{C}_s) &\propto p(x_i|x_{<i}) \prod_{m=1}^{M'} \left[\frac{p(c_m|x_i, x_{<i})}{p(c_m|x_{<i})}\right]^{\lambda_m} \\ &= p(x_i|x_{<i}) \prod_{m=1}^{M'} \left[\frac{p(x_i|x_{<i}, c_m)}{p(x_i|x_{<i}, \emptyset)}\right]^{\lambda_m}. \end{aligned} \tag{13}$$

For mitigation of computation, we denote Eq. (13) to the log in logarithmic form:

$$\begin{aligned} \log p_\lambda(x_i|x_{<i}, \mathcal{C}_s) = \log p(x_i|x_{<i}) + \\ \sum_{m=1}^{M'} \lambda_m \left[\log p(x_i|x_{<i}, c_m) - \log p(x_i|x_{<i}, \emptyset)\right], \end{aligned}$$

we can then synchronously generate multiple parallel token streams: token streams conditioned on different modalities including empty input, and apply multimodal guidance on logit scores:

$$\begin{aligned} \mathrm{p}^{uncon} &= \mathrm{TL}(x|\emptyset), \\ \mathrm{p}_m^{con} &= \mathrm{TL}(x|c_m), \\ \mathrm{p} &= \mathrm{p}^{uncon} + \sum_{m=1}^{M'} \lambda_m (\mathrm{p}_m^{con} - \mathrm{p}^{uncon}), \end{aligned} \tag{14}$$

where the function $\mathrm{TL}(x|c)$ computes the logits outputted by MMoT decoder when conditioned on $c$, $\emptyset$ means the null condition for classifier free, $\mathrm{p}^{uncon}$ are logits scores obtained by unconditional token stream, $\mathrm{p}_m^{con}$ are logits scores obtained by conditional token stream of modality $m$, p are the multimodal guided logits score, and $\lambda_m$ is the guidance scale relevant to the corresponding modality.

In addition, based on an observation that the Jensen–Shannon Divergence (JSD) between the unconditional logits and conditional logits contains rich semantic information (Figure 9 (b)), we use JS divergence to decide the multimodal guidance scale:

$$\lambda_m \propto \mathrm{JSD}(\mathrm{p}_m^{con} - \mathrm{p}^{uncon}).$$

It is worth noting that the suggested multimodal guidance can not only increase sample quality but, more crucially, lead to more spatially coordinated images.

# 4 Experiments

In this section, we evaluate the quality and diversity of our versatile MMoT in synthesizing images under various conditional modalities or compositions of them. The generated images with a set of conditional input modalities show that MMoT can carry out effective interaction and fusion of multimodal information (Sect. 4.3), and superior results with extensive conditional image synthesis methods state that MMoT is robust to all kinds of modalities (Sect. 4.2). We conduct ablation studies to validate the effectiveness of different modules of MMoT (Sect. 4.4), and we provide some important insights about how MMoT realizes interaction and fusion via several visualizations (Sect. 4.5).

We performed our experiments on two datasets: COCO-Stuff (Caesar et al., 2018; Lin et al., 2014) and LHQ (Skorokhodov et al., 2021). COCO-Stuff is a derivative work of the COCO dataset, which contains dense pixel-level and instance-level annotations including text descriptions, segmentation maps, bounding boxes, and so on. LHQ is a dataset of 90k nature landscapes but without any annotations, so we use pseudo-labeling methods to obtain text, segmentation masks, and sketch annotations. More details about the datasets are in 4.1.1.

## 4.1 Experimental Setup

### 4.1.1 Datasets

We evaluate the proposed MMoT on COCO-Stuff (Caesar et al., 2018; Lin et al., 2014) and LHQ (Skorokhodov et al., 2021). All input modalities are obtained from either human annotations or pseudo-labeling methods. And for fair comparisons with PoE-GAN, the same pseudo-labeling methods were used in our approach. We describe details about each dataset in the following.

**COCO-Stuff** is an expansion of the Microsoft Common Objects in Context (MSCOCO) dataset (Lin et al., 2014), which includes 91 stuff classes and 80 object classes. It contains 123,287 images of complex scenes, including 118,287 training images and 5,000 test images. All images are randomly cropped to 256×256 in our main experiments and ablation studies.

Annotations (*i.e.*, text, segmentation mask, sketch, and bounding box layout) for each image are obtained from either human annotations or pseudo-labeling methods: **(i)** In COCO-Stuff, each image has 5 *text* captions, we use CLIP text encoder to extract a high dimension vector per caption. **(ii)** We direct use the *segmentation mask* provided in COCO-Stuff. **(iii)** To obtain the *sketch* annotation, we first detect the edge per image with the HED (Xie & Tu, 2015) and then simplify the rough sketch with the sketch cleanup process (Simo-Serra, Iizuka, Sasaki, & Ishikawa, 2016). **(iv)** We use the bounding boxes and labels provided in COCO-stuff as the ground truth *bounding box layout*.

**LHQ** is a dataset containing 90,000 high-resolution landscape images crawled and preprocessed from Unsplash and Flickr. The dataset is randomly split into an 86,400 training set and a 3,600 test set, and all images are randomly cropped to 256×256 in our main experiments.

Since the vanilla dataset does not come with any manual annotations, annotations (*i.e.*, text, segmentation mask, and sketch) are obtained from pseudo-labeling methods: **(i)** For the *text* annotation, we use the pre-trained CLIP image encoder to extract a feature vector as the pseudo text embedding. **(ii)** DeepLab-v2 (L.-C. Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017) was used to produce pseudo *segmentation mask* annotation. **(iii)** HED (Xie & Tu, 2015) followed by the sketch cleanup process (Simo-Serra et al., 2016) was adopted to annotate each image with a *sketch* map.

### 4.1.2 Evaluation metrics

For different conditional image synthesis tasks, we use different metrics to evaluate the generation performance over all existing methods and our proposed MMoT. They are Inception Score (IS)(Salimans et al., 2016), Frechet Inception Distance (FID)(Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017) and Clean-FID(G. Parmar, Zhang, & Zhu, 2021). IS and FID are the most commonly used metrics to evaluate the quality and diversity of generated images. Clean-FID is an improved version of FID.

**Inception Score (IS)** measures the quality of generated images by computing the expected KL-divergence between the marginal class distribution over all generated images and the conditional distribution for a particular generated image, using

the class probability predicted by the Inception Net. This metric is expected to capture both the fidelity and diversity of generated images.

**Frechet Inception Distance (FID)** measures the similarity between the embedding feature of generated and real images. This is achieved by fitting the embedding features into a multivariate Gaussian distribution and computing their Frechet distance.

**Clean-FID.** Under the hood, computing FID contains several subtle implementation decisions, notably image resizing, quantization, and formatting. Any inconsistency in the steps leads to results that are no longer comparable to other methods. The resize operation and the image quantization/compression are especially impactful. To facilitate an easy comparison, (G. Parmar et al., 2021) propose an easy-to-use library, *i.e.*, clean-fid, which is more suitable for benchmarking due to its reported benefits over previous implementations of FID.

### 4.1.3 Hyper-parameters

The number of encoder layers in the MMoT is 12, while the number of decoder layers is 24. We use this asymmetrical structure because the encoder is mainly used to extract features, while the decoder is responsible for more complex mapping, *i.e.*, converting the input features into the final image output, and this design can greatly reduce the number of parameters. Encoders for different modalities have the same number of layers. Both encoders and decoders share an architecture of 12 attention heads and an embedding dimension of 768. All training images in COCO-Stuff and LHQ are randomly cropped to $256 \times 256$ in our main experiments and ablation studies. We use the AdamW (Loshchilov & Hutter, 2017) optimizer with $\beta_1 = 0.9, \beta_2 = 0.95$ and the weight decay is set to be 0.01. We use a batch size of 64 for training all our models and set the learning rate to be 4.5e-6, and all the models are trained for 300 epochs on 8 A100 GPUs.

## 4.2 Comparisons with Existing Methods

We compare MMoT with one of the state-of-the-art MCIS methods PoE-GAN (X. Huang et al., 2022) and also with a wide range of UCIS approaches in the unimodal setting. Since M6-UFC (Z. Zhang et al., 2021) and Composer (L. Huang et al., 2023) are performed on different MCIS datasets and their codes are unavailable till our submission, we did not make direct comparisons with them.

**Table 1** Comparison on COCO-Stuff (256×256)

(a) Text to Image

| Method | FID ↓ | Clean-FID ↓ |
|---|---|---|
| - | - | - |
| - | - | - |
| DF-GAN | - | 45.2 |
| DM-GAN+CL | - | 29.9 |
| VQGAN+T* | 27.8 | 28.1 |
| PoE-GAN | - | 20.5 |
| *MMoT(Ours)* | **17.9** | **17.8** |

(b) Bounding boxes to Image

| Method | FID ↓ | Inception Score ↑ |
|---|---|---|
| LostGAN-V2 | 42.6 | 18.0±0.50 |
| OC-GAN | 41.7 | 17.8±0.00 |
| VQGAN+T | 33.7 | - |
| LAMA | 31.1 | - |
| Context-L2I | 29.56 | 18.57±0.54 |
| TwFA | 22.1 | 24.3±1.04 |
| *MMoT* | **19.2** | **26.7±0.50** |

(c) Segmentation masks to Image

| Method | FID ↓ | Clean-FID ↓ |
|---|---|---|
| pix2pixHD | 111.5 | - |
| SPADE | 22.6 | 22.1 |
| VQGAN+T | 22.4 | 21.6 |
| OASIS | 17.0 | 19.2 |
| PITI | 15.8 | - |
| PoE-GAN | - | 15.8 |
| *MMoT* | **12.7** | **12.9** |

(d) Sketch to Image

| Method | FID ↓ | Clean-FID ↓ |
|---|---|---|
| - | - | - |
| pix2pixHD* | 44.4 | 46.2 |
| SPADE* | 78.8 | 80.3 |
| VQGAN+T* | 33.9 | 34.4 |
| PITI | **20.3** | - |
| PoE-GAN | - | 25.5 |
| *MMoT* | 23.1 | **23.9** |

(e) All†

| Clean-FID ↓ |
|---|
| - |
| - |
| - |
| - |
| - |
| 13.6 |
| **12.6** |

We evaluate models conditioned on different modalities (*i.e.*, text, bounding boxes, segmentation masks, sketch). The best scores are highlighted in bold and the second best ones are underlined. For fair comparisons, all the results are taken from the relative papers. '-' means the related value is unavailable in their papers. '*' denotes results on samples from retrained models with the official implementation. All† means image synthesis conditioned on text+segmentation masks+sketch.

### 4.2.1 Text to image synthesis

Text-to-Image is designed to render a realistic image from a text description, which is a rather challenging task that dominates image generation. It is also a cross-modal generation task, which requires the model to be able to generate images that meet people's expectations based on understanding the objects and their relationships described in the text. For text-to-image synthesis, we compare with DF-GAN (Tao et al., 2020), DM-GAN+CL (Ye, Yang, Takac, Sunderraman, & Ji, 2021), VQGAN+T (Esser et al., 2021) and PoE-GAN (X. Huang et al., 2022) on COCO-Stuff. Since text annotation is not available in LHQ, we compare with VQGAN+T (Esser et al., 2021), MaskGIT (Chang et al., 2022), and NUWA-Infinity (Wu, Liang, Hu, et al., 2022) in the language-free setting, *i.e.*, language annotations are unavailable and we take the pseudo-labelling texts as condition. The quantitative results of


**Fig. 4** Qualitative comparison of text-to-image synthesis results on COCO-Stuff. More results are demonstrated in A.1

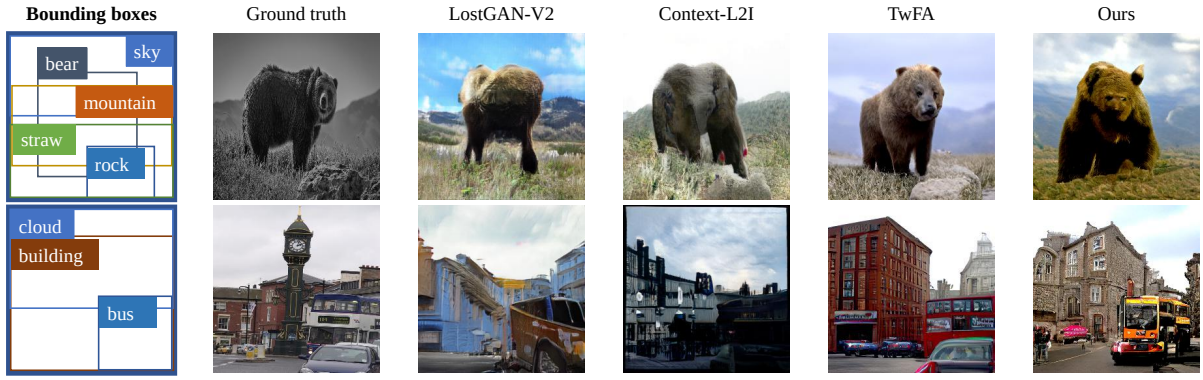



**Fig. 5** Qualitative comparison of bounding boxes-to-image synthesis results on COCO-Stuff. More results are demonstrated in A.1

text-to-image synthesis on COCO-Stuff and LHQ are reported in Table 1 (a) and Table 2 (a), respectively. In Figure 4, we also provide several qualitative comparisons. MMoT has comparable performance, both quantitatively and qualitatively.

### 4.2.2 Bounding boxes to image synthesis

Bounding boxes to image generation aims to generate photo-realistic conditioned on specified layouts which consists of a set of object bounding boxes and corresponding categories. Compared with text-to-image synthesis, such a layout provides a simple sketch of the image, which makes the generation more user-friendly and controllable, but this also reduces the diversity of generated images to some extent. For bounding-boxes-to-image generation, we compare with LostGAN-V2 (Sun & Wu, 2021), OC-GAN (Sylvain et al., 2021), VQGAN+T (Esser et al., 2021), LAMA (Li et al., 2021), Context-L2I (S. He et al., 2021) and TwFA (Z. Yang et al., 2022). The performance of several methods for bounding boxes to image synthesis on COCO-Stuff is evaluated quantitatively and qualitatively. The quantitative results are reported in Table 1 (b), while several qualitative comparisons are shown in Figure 5. The evaluations demonstrate that MMoT achieves better performance compared to the other methods.

### 4.2.3 Segmentation to image synthesis

The goal of segmentation to image synthesis is to generate a full-color image from a grayscale segmentation mask, where each pixel in the mask corresponds to a specific object or region in the image. For segmentation masks-to-image synthesis, we compare with pix2pixHD (T.-C. Wang et al., 2018), SPADE (Park et al., 2019), VQGAN+T (Esser et al., 2021), OASIS (Sushko et al., 2022), PITI (T. Wang et al., 2022) and


**Fig. 6** Qualitative comparison of segmentation-to-image synthesis results on COCO-Stuff. More results are demonstrated in A.1

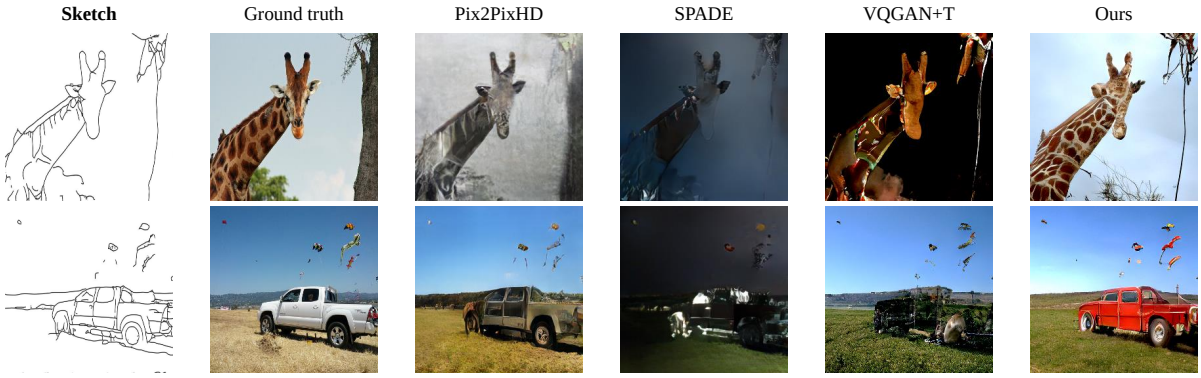



**Fig. 7** Qualitative comparison of sketch-to-image synthesis results on COCO-Stuff. More results are demonstrated in A.1

PoE-GAN (X. Huang et al., 2022). Table 1 (c) and Table 2 (b) present the quantitative results of segmentation-to-image synthesis on COCO-Stuff and LHQ, respectively. Figure 6 provides several qualitative comparisons. Both the quantitative and qualitative evaluations show that MMoT outperforms the other methods.

### 4.2.4 Sketch to image synthesis

The same as segmentation to image synthesis, converting a sketch to an image is also an image-to-image translation task that involves creating an image from a rough, hand-drawn sketch or line drawing. This task can be challenging because sketches often lack detail, texture, and color information, and require a model to infer these missing details to generate a realistic image. For sketch-to-image generation, we compare with pix2pixHD (T.-C. Wang et al., 2018), SPADE (Park et al., 2019), VQGAN+T (Esser et al., 2021), PITI (T. Wang et al., 2022) and PoE-GAN (X. Huang et al., 2022). Table 1 (d) and Table 2 (c) show the quantitative results of sketch-to-image synthesis on COCO-Stuff and LHQ, respectively. Additionally, Figure 7 presents several qualitative comparisons. The evaluations indicate that MMoT performs better than the other methods, both in terms of quantitative metrics and qualitative comparisons.

Sect. 4.2.1 to Sect.4.2.2 demonstrates MMoT's unimodal conditional image generation capabilities. Surprisingly, MMoT achieves comparable performance with unimodal methods specifically designed for that modality on both datasets. In Figure 4 to Figure 7, we found that MMoT is robust to different input modalities and can produce photo-realistic images of refiner textures (*e.g.*, pasta, bus, and building), clearer structures (*e.g.*, giraffe, bear, and car), and more reasonable interactions (*e.g.*, the reflection on the bus windows, the car's shadow on the ground). It is worth

**Table 2** Comparison on LHQ (256×256)

| (a) Text to Image | | | |
|---|---|---|---|
| **Method** | FID ↓ | Clean-FID ↓ | Inception Score ↑ |
| VQGAN+T | 12.71 | 12.78 | 4.61 |
| MaskGIT† | 24.33 | - | 4.61 |
| NUWA-Infinity† | **9.71** | - | **4.98** |
| *MMoT(Ours)* | 11.38 | **11.46** | 4.94 ± 0.19 |

| (b) Segmentation masks to Image | | | |
|---|---|---|---|
| **Method** | FID ↓ | Clean-FID ↓ | Inception Score ↑ |
| pix2pixHD | 36.52 | 41.00 | 3.42±0.10 |
| SPADE | 25.47 | 26.70 | 3.71±0.12 |
| VQGAN+T | 14.92 | 15.00 | 4.42±0.11 |
| *MMoT* | **11.87** | **11.98** | **4.43 ± 0.13** |

| (c) Sketch to Image | | | |
|---|---|---|---|
| **Method** | FID ↓ | Clean-FID ↓ | Inception Score ↑ |
| pix2pixHD | 32.19 | 36.23 | 3.51±0.13 |
| SPADE | 35.83 | 36.35 | 2.98±0.07 |
| VQGAN+T | 13.91 | 13.9 | 4.13 ± 0.21 |
| *MMoT* | **11.68** | **11.76** | **4.26 ± 0.09** |

We evaluate models conditioned on different modalities (*i.e.*, text, segmentation masks, sketch). The best scores are highlighted in bold and the second best ones are underlined. All the results are on samples from retrained models with the official implementation. '†' denotes results taken from the relative papers which trained on LHQC with 1024×1024 resolution. '-' means the related value is unavailable in their papers.

noting that, unlike unimodal conditional synthesis models, MMoT supports the combination of many different types of inputs.

#### 4.2.5 Multimodal conditional image synthesis

As mentioned earlier, unimodal conditional image generation supports only one type of conditioning information. To make the generation more flexible and controllable, multimodal conditional image generation can synthesize images conditioned on multiple types of modality inputs. As illustrated in Table 1 (e), we also obtain better results than PoE-GAN when conditioned on All† (*i.e.*, text + segmentation masks + sketch). It is worth noting that, since PoEGAN is trained with only three modalities, we also test our MMoT on these modalities. It could be a little bit unfair but has some reference significance.

Overall, when conditioned on a single modality, the superiority of the proposed MMoT is validated on both quantitative metrics and qualitative visual comparison. The improved performance of unimodal conditional image synthesis indicates the robustness of the multistage token-mixer and the balanced optimization of each modality. In addition, MMoT also outperforms previous state-of-art MCIS method PoE-GAN when conditioned on multimodal conditional inputs.

### 4.3 Image Synthesis Conditioned on the Compositions of Inputs

The most exciting ability of MMoT is that it can synthesize imagery images according to the compositions of a set of input modalities.

In Figure 8, we show some multimodal conditional samples generated by our MMoT. As illustrated in Figure 8 (a), using segmentation masks to give the coarse layout of semantic classes (*e.g.*, sky, trees, and grass) is easy but it is impossible to specify the color of an object, however, with the help of text, the color of clouds can be defined. Similarly, sketch allow us to describe the shape (*e.g.*, ridges of mountains) and texture (*e.g.*, ripples of water) of an object with simple strokes, but only under the control of the text can the state (*e.g.*, snow or stone mountain, frozen lake or dirty river) and category (*e.g.*, river, sand or grassland) of the object be given. Moreover, as shown in Figure 8 (c) and (d), MMoT can synthesize realistic and diverse images when conditioned on the compositions of segmentation masks and sketch/bounding box layout.

### 4.4 Ablation Study

In Table 3, we analyze the importance of different components of MMoT. The settings in both studies are similar to the comparison with existing methods in Table 1. The baseline model is the vanilla encoder-decoder transformer, which simply concatenates the features outputted by respective encoders while performing modality dropout. Each row corresponds to a model trained with the additional element.

Compared with the baseline, when adding the multistage token-mixer, the performances of unimodal conditions and multiple conditions have gained significant improvement, which indicates the robustness on missing-modality and the effectiveness of the fusion of the multistage token-mixer. And since multimodal balanced loss can

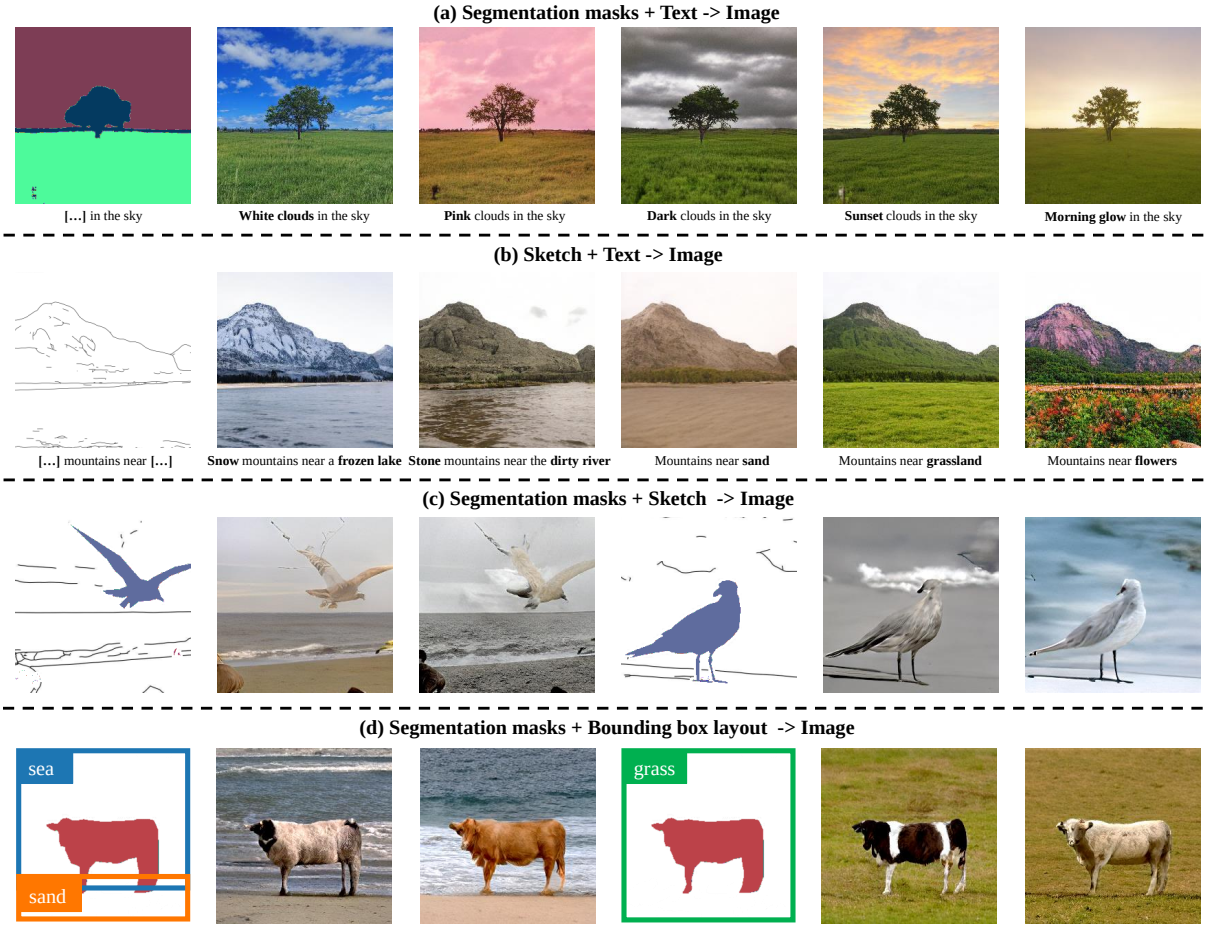



**Fig. 8** Samples of composed multimodal conditional image synthesis generated by MMoT. We show compositions of different modalities: (a) segmentation masks and text; (b) sketch and text; (c) segmentation masks and sketch; (d) segmentation masks and bounding box layout. MMoT can synthesize reasonable images leveraging conditional information conveyed by different modalities. More results are demonstrated in A.2

**Table 3** Ablation on COCO-Stuff (256×256)

| | Text | | Bounding boxes | | Segmentation masks | | Sketch | | All | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | FID ↓ | Clean-FID ↓ | FID ↓ | Inception Score ↑ | FID ↓ | Clean-FID ↓ | FID ↓ | Clean-FID ↓ | FID ↓ | Clean-FID ↓ |
| Base | 55.93 | 56.27 | 26.27 | 22.50±0.67 | 20.07 | 20.43 | 32.91 | 33.73 | 13.43 | 13.68 |
| + Multistage Token-Mixer | 31.19 | 31.40 | 21.36 | 24.18±0.59 | 14.52 | 14.77 | 23.75 | 24.37 | 12.79 | 13.04 |
| + Multimodal Balanced Loss | 23.04 | 23.32 | 20.65 | 24.90±0.57 | 13.24 | 13.43 | **23.08** | **23.70** | 12.23 | 12.44 |
| + Multimodal Guidance | **17.91** | **17.83** | **19.24** | **26.67±0.50** | **12.73** | **12.91** | 23.42 | 23.93 | **11.75** | **11.73** |

The best scores are highlighted in bold and the second best ones are underlined.

facilitate the optimization of each modality, unimodal conditional image synthesis achieved further improvement, especially for text-to-image generation. We note that multimodal guidance is useful for text-to-image synthesis but not essential for each input condition.

## 4.5 Qualitative Analysis

In Figure 9, we show some visualizations of the underlying process, including cross-attention maps, divergence maps, and combination-weight maps.

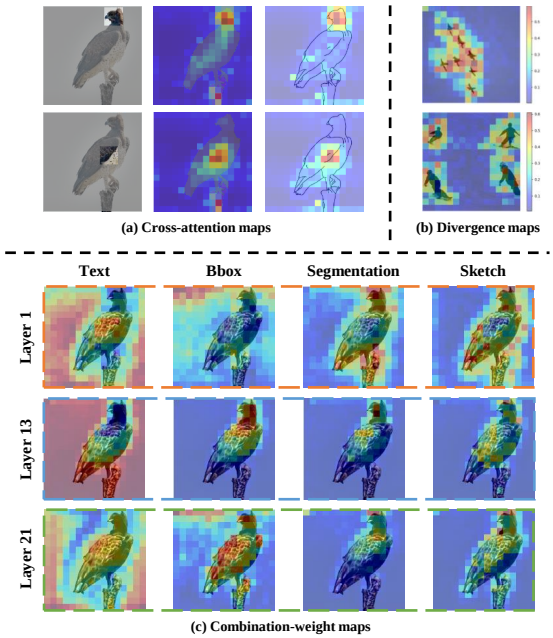



**Fig. 9** Visualizations of cross attention maps, divergence maps, and combination weight maps.

#### 4.5.1 Cross-attention maps

Part (a) in Figure 9 is the average attention score maps across all transformer decoder layers. We show the cross-attention map conditioned on segmentation masks and sketches. It is noted that when generating the specific areas of an image (*e.g.*, the head or thorax of a bird), high responsiveness is observed in the corresponding areas of the segmentation map or sketch.

#### 4.5.2 Divergence map

Part (b) in Figure 9 is the distribution of the JS divergence calculated between unimodal conditional logits and unconditional logits over different generated images, which contain rich semantic relations reflecting the influence of different conditions. So injecting the divergence map as composition prior to controlling the value of the guidance scale during the sampling process can lead to more spatially coordinated images.

#### 4.5.3 Combination-weight maps

The key to effective fusion is to calculate accurate combination weights, which represent the influences of each input modality. The combination weights in MMoT are related to attention scores calculated in Eq. (8). Part (c) in Figure 9 shows the combination-weight maps of individual modalities in layers 1, 13, and 21, which show that the modality at different semantic levels contributes to different layers, *i.e.*, text and bbox have more contributions in the higher layer, while segmentation and sketch play a role in all layers, especially in the earlier layer. It also illustrates that the special token `[PULSE]` within the multistage token-mixer can adaptively detect the influences of modality tokens across the generated image.

## 5 Conclusions

In this paper, we focus on the challenging Composed Multimodal Conditional Image Synthesis (CMCIS) task and propose a novel Mixture-of-Modality-Tokens Transformer (MMoT). Towards two severe issues of CMCIS, *i.e.*, modality coordination and imbalance problems, we introduce a multistage token-mixer, multimodal balanced loss, and divergence-driven sampling guidance to fully exploit the cooperativity across different modalities. Extensive experiments on the COCO-Stuff and LHQ datasets demonstrate that the proposed MMoT successfully generates high-quality and faithful images conditioned on composed multimodal signals, and achieves superior performance over most existing UCIS and MCIS models.

**Limitations.** Since our framework is based on the autoregressive Transformer, it suffers from limited inference speed. In the future, we will attempt to adapt the proposed module and training/testing schemes to other deep neural frameworks and hope to produce more realistic, high-quality, and diverse results.

**Broader Impacts.** The proposed composed multimodal image synthesis offers unprecedented fine-grained generation capabilities, which lead to both positive and negative societal impacts. Multimodal control signals as input for synthesis greatly improve the flexibility of user interaction and ease the use of deep generative models. However, the increasing generation capabilities also make it easier to synthesize desired images for malicious purposes, *i.e.*, the misuse of fake or nefarious information. In the future, sufficient guardrails, access control, and detection techniques are encouraged to minimize the risk of misuse.

**Data Availability Statement.** The data that support the findings of this study are available from the corresponding author upon reasonable request.

# Appendix A More Qualitative Results

## A.1 Qualitative Comparisons with UCIS Models

In Figure A1 to A4, we show additional qualitative comparisons with a wide range of UCIS models when conditioned on text, segmentation mask, sketch, and bounding boxes, respectively. Competitive visual results compared with UCIS models specially designed for a single modality indicate MMoT is robust to different modalities.

## A.2 Qualitative CMCIS Examples

Figure A5 to A9 show that MMoT can generate high-quality, faithful, and diverse images when conditioned on complex compositions of two or three different modalities.

In Figure A5 to A7, we also show more visual comparisons with PoE-GAN when conditioned on compositions of text+segmentation mask, text+sketch, and segmentation mask+sketch, respectively. The modality coordination problem and the modality imbalance problem are common in MCIS models when conditioned on complex multimodal conditions. In contrast, MMoT addresses both issues and can synthesize high-quality and faithful images.

**Modality coordination problem.** The modality coordination problem is caused by the non-adaptive fusion of fine-grained information across multiple modalities. As illustrated in Figure A7, when PoE-GAN synthesizes an image, the generated contents from the sketch condition are incorrectly composed with the generated contents from the segmentation mask condition.

**Modality imbalance problem.** The modality imbalance problem is caused by the imbalanced distribution of each modality in datasets. As illustrated in Figure A5 and A6, PoE-GAN tends to ignore text inputs when generating images.

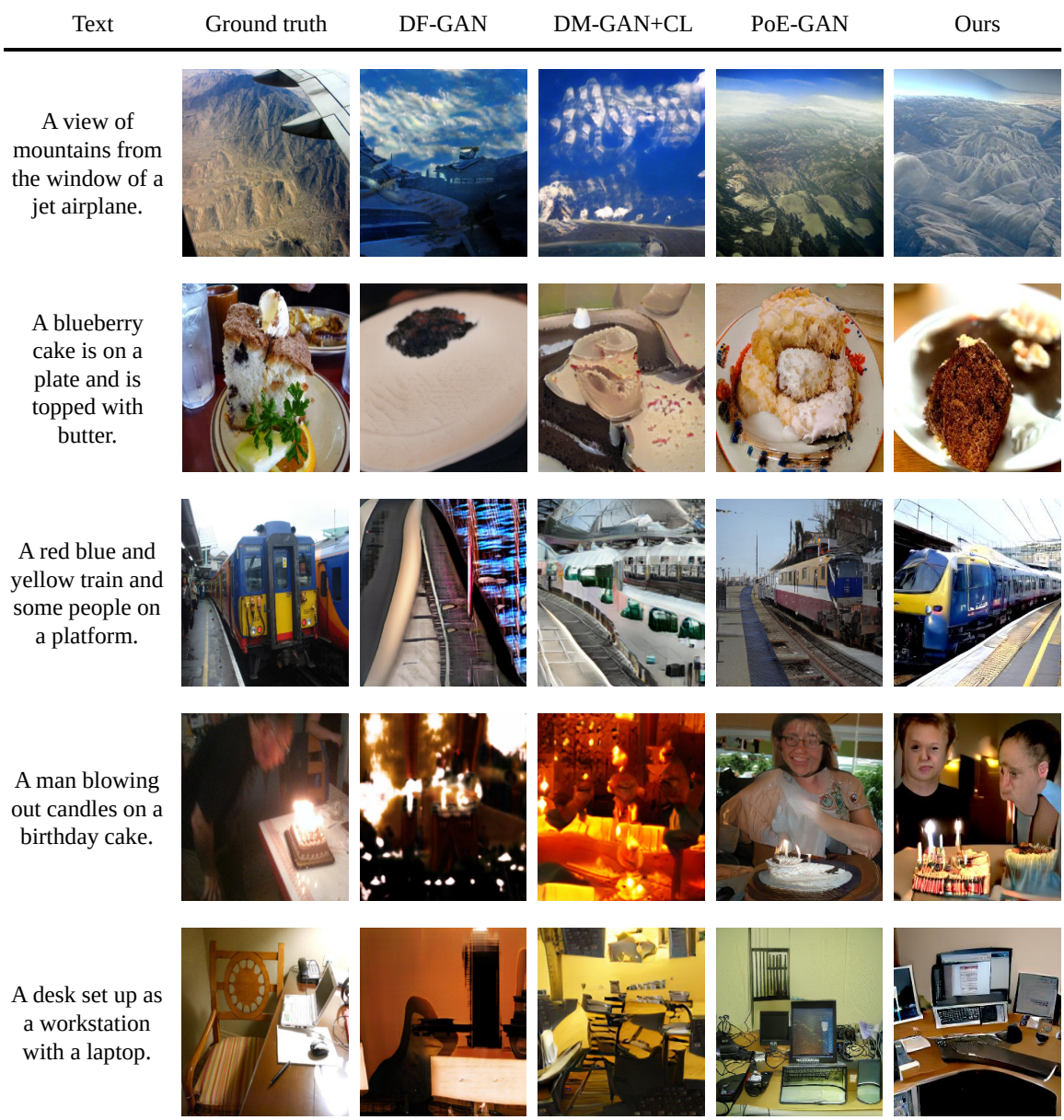



**Fig. A1** Additional qualitative comparison of text-to-image synthesis on COCO-Stuff.


**Fig. A2** Additional qualitative comparison of segmentation mask-to-image synthesis on COCO-Stuff.

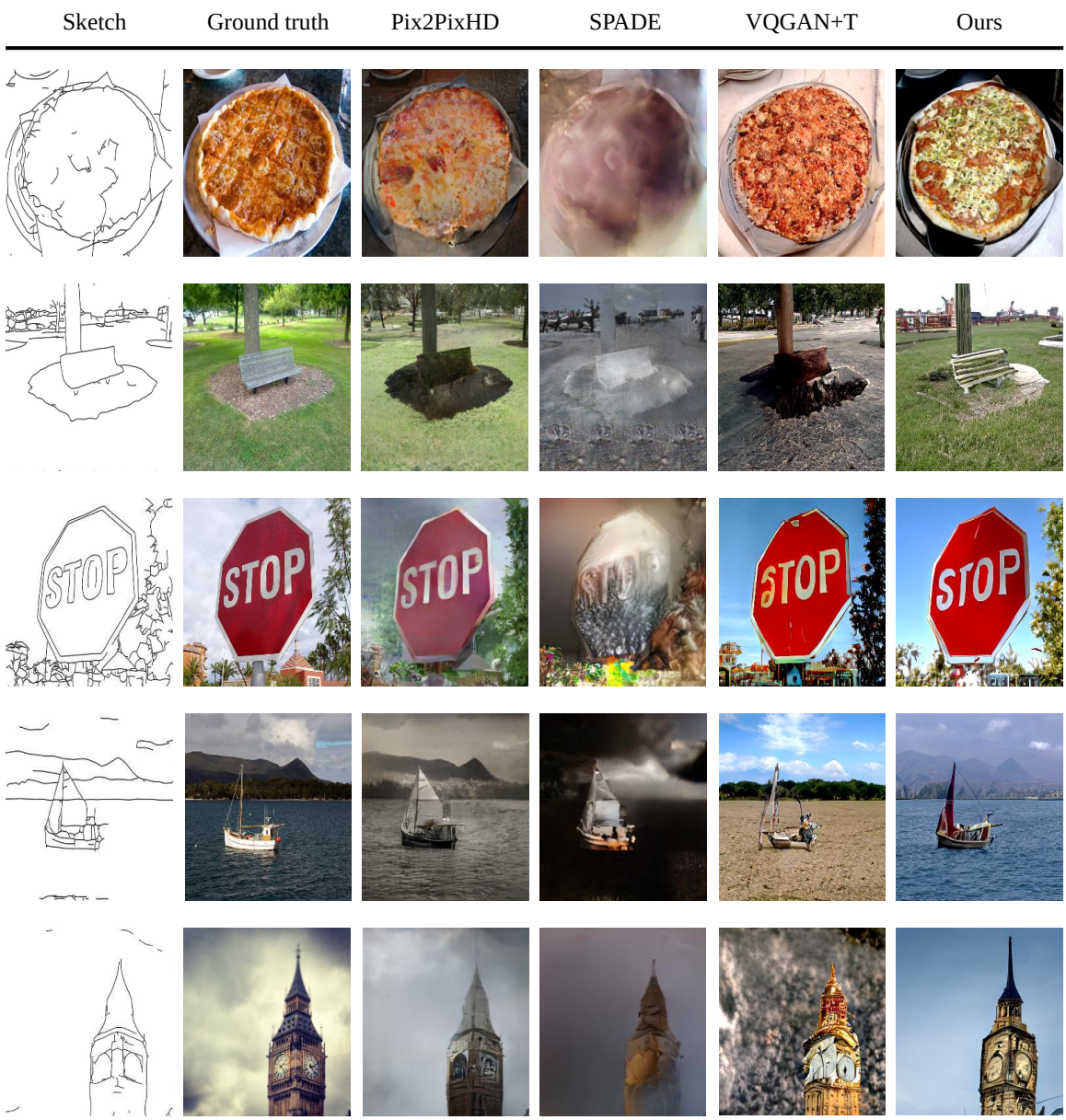



**Fig. A3** Additional qualitative comparison of sketch-to-image synthesis on COCO-Stuff.


**Fig. A4** Additional qualitative comparison of bounding boxes-to-image synthesis on COCO-Stuff.

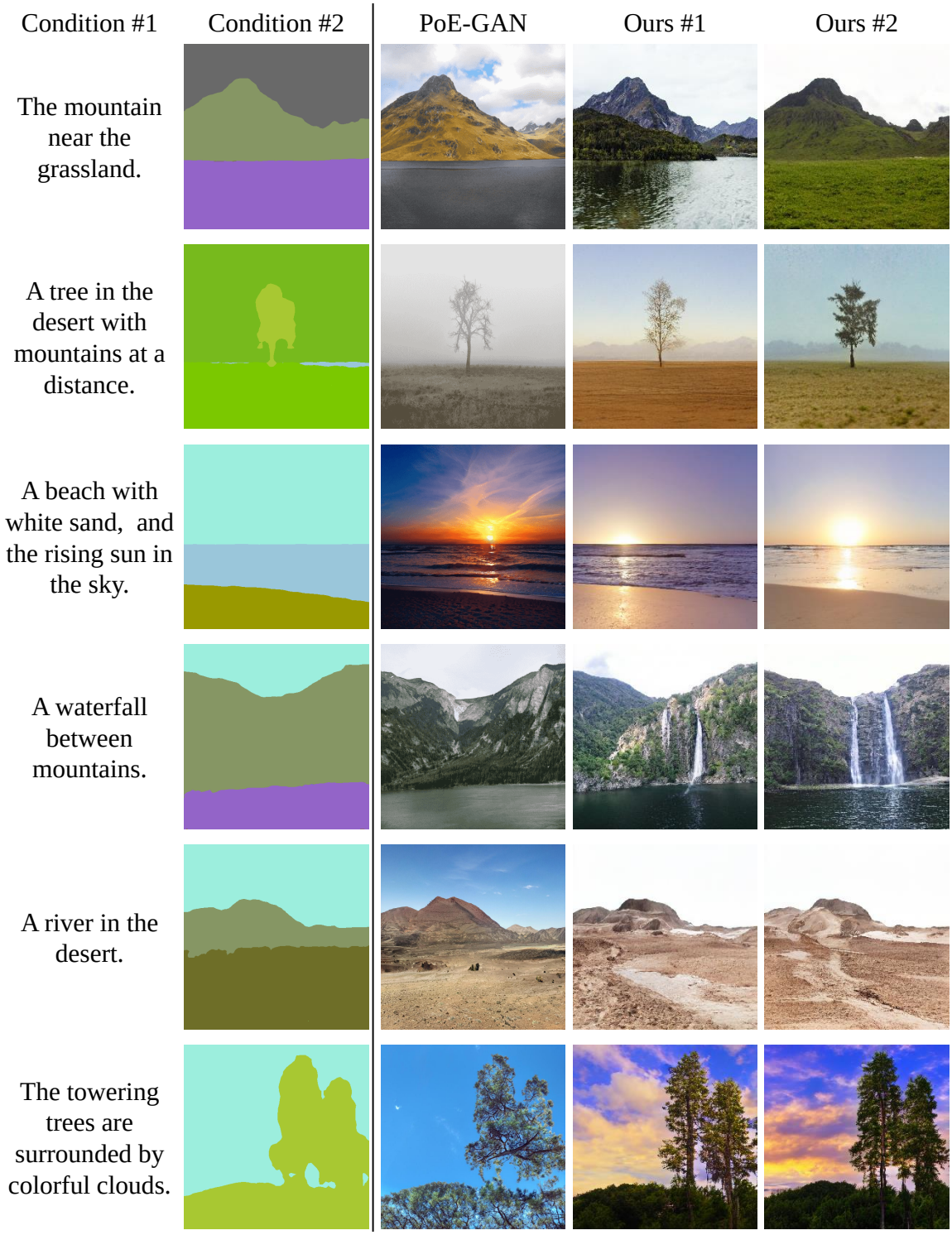



**Fig. A5** Examples of composed multimodal conditional image synthesis when conditioned on text and segmentation mask. From left to right: text, segmentation, a random sample from PoE-GAN, and two random samples from our MMoT. PoE-GAN always struggles with the modality imbalance problem. In contrast, MMoT can balance the information of the two modalities to synthesize images.

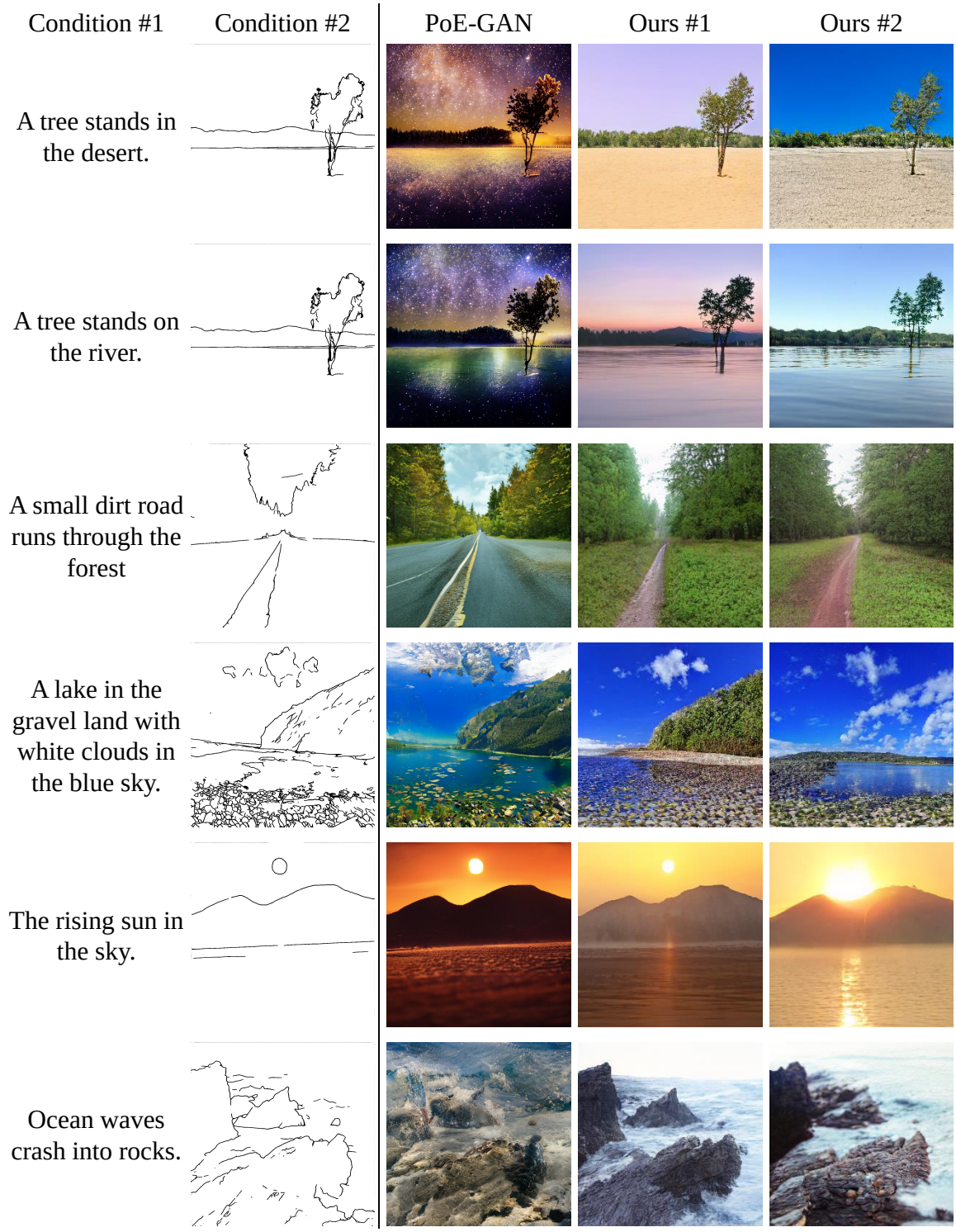



**Fig. A6** Examples of composed multimodal conditional image synthesis when conditioned on text and sketch. From left to right: text, sketch, a random sample from PoE-GAN, and two random samples from our MMoT. PoE-GAN always struggles with the modality imbalance problem. In contrast, MMoT can balance the information of the two modalities to synthesize images.

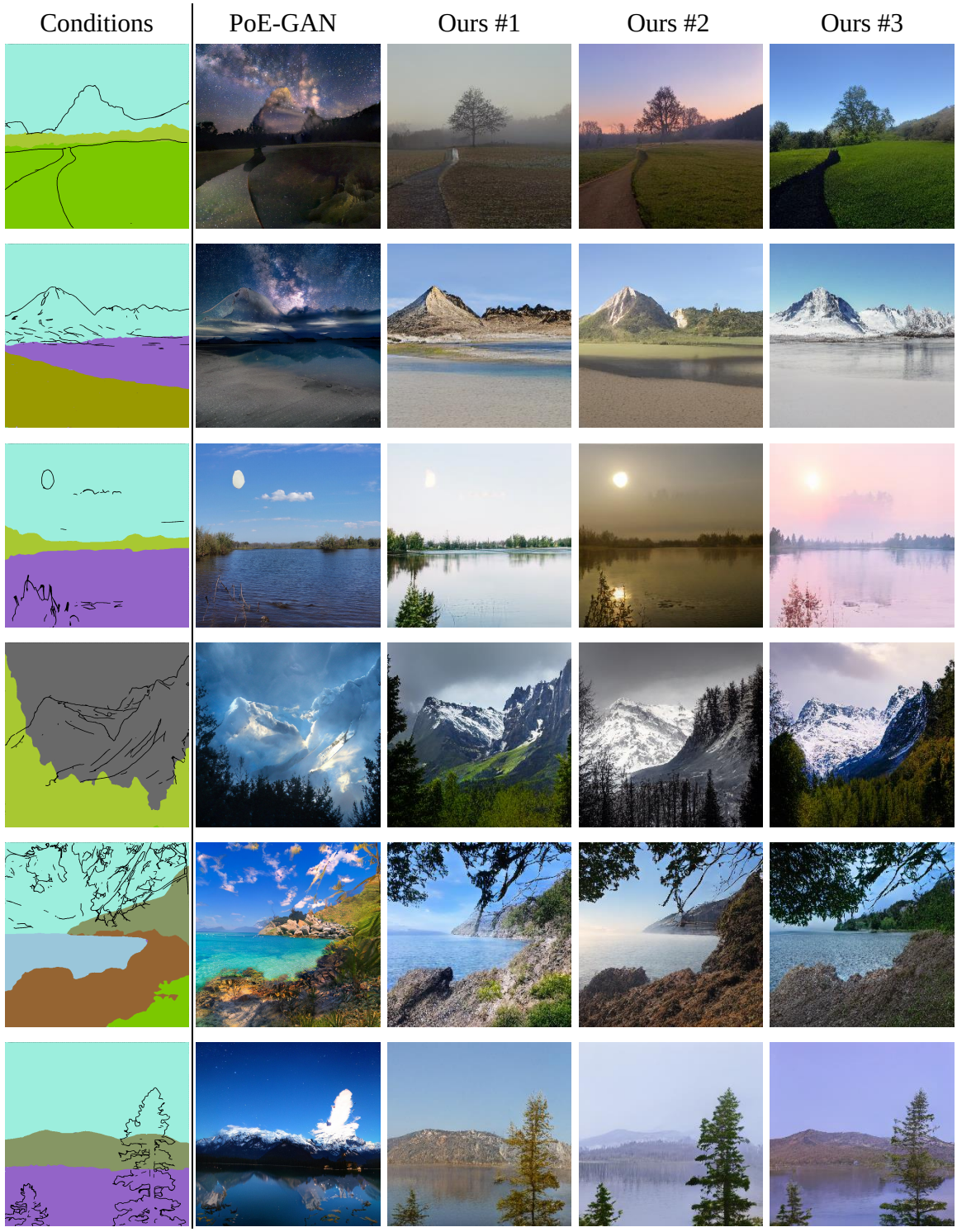



**Fig. A7** Examples of composed multimodal conditional image synthesis when conditioned on segmentation and sketch. From left to right: segmentation mask, sketch, a random sample from PoE-GAN, and three random samples from our MMoT. PoE-GAN always struggles with the modality coordination problem. In contrast, MMoT can generate more spatially coordinated images.

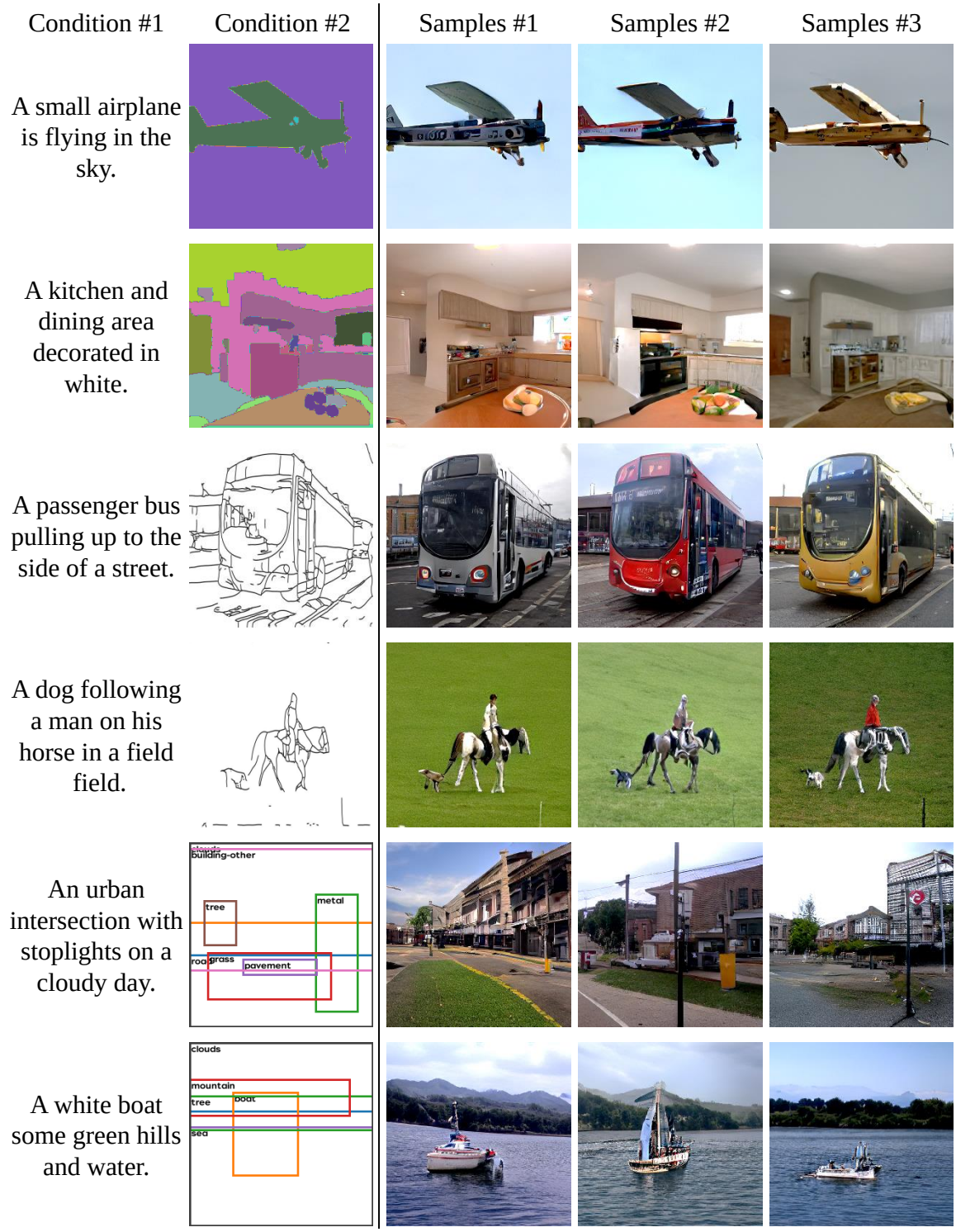



**Fig. A8** Examples of composed multimodal conditional image synthesis. We show three random samples from MMoT conditioned on compositions of different modalities (from top to bottom: text+segmentation mask, text+sketch, and text+bounding boxes).

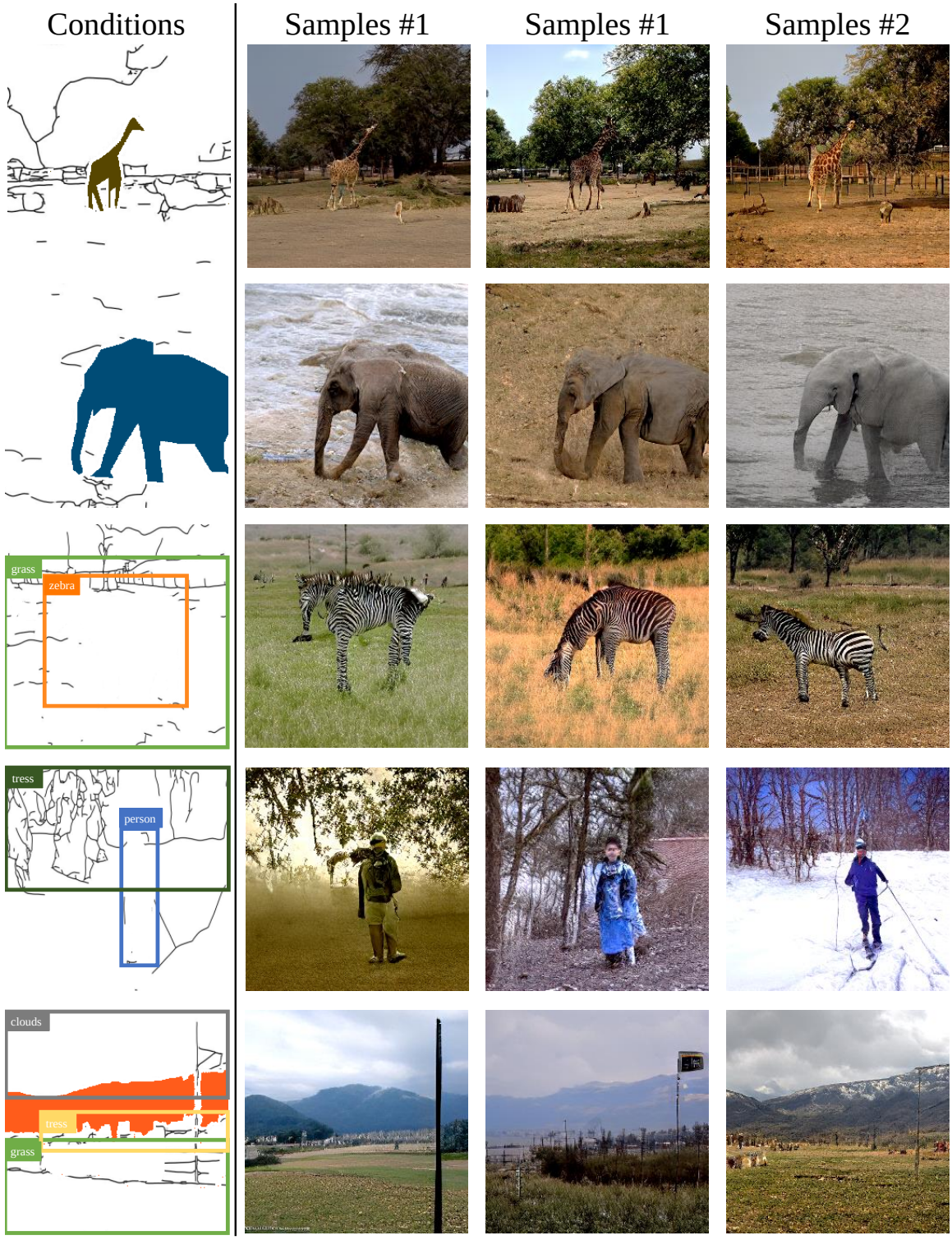



**Fig. A9** Examples of composed multimodal conditional image synthesis. We show three random samples from MMoT conditioned on compositions of different modalities (from top to bottom: segmentation mask+sketch, bounding boxes+sketch, and segmentation mask+sketch+bounding boxes).

# References


Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G. (2021). Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *Transactions on Pattern Analysis and Machine Intelligence*.

Caesar, H., Uijlings, J., Ferrari, V. (2018). Cocostuff: Thing and stuff classes in context. *Conference on computer vision and pattern recognition* (pp. 1209–1218).

Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T. (2022). Maskgit: Masked generative image transformer. *Conference on computer vision and pattern recognition* (pp. 11315–11325).

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I. (2020). Generative pretraining from pixels. *International conference on machine learning* (pp. 1691–1703).

Child, R., Gray, S., Radford, A., Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, *34*, 8780–8794.

Esser, P., Rombach, R., Ommer, B. (2021). Taming transformers for high-resolution image synthesis. *Conference on computer vision and pattern recognition* (pp. 12873–12883).

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y. (2022). Make-a-scene: Scene-based text-to-image generation with human priors. *European conference on computer vision* (pp. 89–106).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Conference on computer vision and pattern recognition* (pp. 16000–16009).

He, S., Liao, W., Yang, M.Y., Yang, Y., Song, Y.-Z., Rosenhahn, B., Xiang, T. (2021). Context-aware layout to image generation with enhanced object appearance. *Conference on computer vision and pattern recognition* (pp. 15049–15058).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, *14*(8), 1771–1800.

Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J. (2023). Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Huang, X., Mallya, A., Wang, T.-C., Liu, M.-Y. (2022). Multimodal conditional image synthesis with product-of-experts gans. *European conference on computer vision* (pp. 91–109).

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. *Conference on computer vision and pattern recognition* (pp. 1125–1134).

Jahn, M., Rombach, R., Ommer, B. (2021). High-resolution complex scene synthesis with transformers. *Conference on computer vision and pattern recognition workshop*.

Kingma, D.P., & Welling, M. (2014). Auto-encoding variational bayes. *International conference on learning representations*.

Kobyzev, I., Prince, S.J., Brubaker, M.A. (2020). Normalizing flows: An introduction and review of current methods. *Transactions on Pattern Analysis and Machine Intelligence*, *43*(11), 3964–3979.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, *1*(0).

Li, Z., Wu, J., Koh, I., Tang, Y., Sun, L. (2021). Image synthesis from layout with locality-aware mask adaption. *International conference on computer vision* (pp. 13819–13828).

Li, Z., Zhou, H., Bai, S., Li, P., Zhou, C., Yang, H. (2022). M6-fashion: High-fidelity multi-modal image generation and editing. *arXiv preprint arXiv:2205.11705*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C.L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision* (pp. 740–755).

Liu, X., Yin, G., Shao, J., Wang, X., et al. (2019). Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *Advances in neural information processing systems*, *32*.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *International conference on learning representations*.

Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X. (2022). Are multimodal transformers robust to missing modality? *Conference on computer vision and pattern recognition* (pp. 18177–18186).

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *Advances in neural information processing systems workshop*.

Miyato, T., & Koyama, M. (2018). cgans with projection discriminator. *International conference on learning representations.*

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X. (2023). T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Odena, A., Olah, C., Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. *International conference on machine learning* (pp. 2642–2651).

Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, *22*(1), 2617–2680.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y. (2019). Semantic image synthesis with

spatially-adaptive normalization. *Conference on computer vision and pattern recognition* (pp. 2337–2346).

Parmar, G., Zhang, R., Zhu, J.-Y. (2021). On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D. (2018). Image transformer. *International conference on machine learning* (pp. 4055–4064).

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning* (pp. 8748–8763).

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021). Zero-shot text-to-image generation. *International conference on machine learning* (pp. 8821–8831).

Razavi, A., Van den Oord, A., Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, *32*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Conference on computer vision and pattern recognition* (pp. 10684–10695).

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., ... others (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, *35*, 36479–36494.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, *29*.

Simo-Serra, E., Iizuka, S., Sasaki, K., Ishikawa, H. (2016). Learning to simplify: fully convolutional networks for rough sketch cleanup. *ACM Transactions on Graphics*, *35*(4), 1–11.

Skorokhodov, I., Sotnikov, G., Elhoseiny, M. (2021). Aligning latent and image spaces to connect the unconnectable. *International conference on computer vision* (pp. 14144–14153).

Sohn, K., Lee, H., Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, *28*.

Sun, W., & Wu, T. (2021). Learning layout and style reconfigurable gans for controllable image synthesis. *Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5070–5087.

Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A. (2022). Oasis: Only adversarial supervision for semantic image synthesis. *International Journal of Computer Vision*, *130*(12), 2903–2923.

Sylvain, T., Zhang, P., Bengio, Y., Hjelm, R.D., Sharma, S. (2021). Object-centric image generation from layouts. *Aaai conference on artificial intelligence* (Vol. 35, pp. 2647–2655).

Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X.-Y., Wu, F., Bao, B. (2020). Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint*

*arXiv:2008.05865*.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, *30*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F. (2022). Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. *Conference on computer vision and pattern recognition* (pp. 8798–8807).

Wu, C., Liang, J., Hu, X., Gan, Z., Wang, J., Wang, L., ... Duan, N. (2022). Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *arXiv preprint arXiv:2207.09814*.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N. (2022). Nüwa: Visual synthesis pre-training for neural visual world creation. *European conference on computer vision* (pp. 720–736).

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. *International conference on computer vision* (pp. 1395–1403).

Yang, C., Shen, Y., Zhou, B. (2021). Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, *129*, 1451–1466.

Yang, Z., Liu, D., Wang, C., Yang, J., Tao, D. (2022). Modeling image composition for complex scene generation. *Conference on computer vision and pattern recognition* (pp. 7764–7773).

Ye, H., Yang, X., Takac, M., Sunderraman, R., Ji, S. (2021). Improving text-to-image synthesis using contrastive learning. *British Machine Vision Conference*.

Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., ... others (2022). Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Zhang, L., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, Z., Ma, J., Zhou, C., Men, R., Li, Z., Ding, M., ... Yang, H. (2021). Ufc-bert: Unifying multi-modal controls for conditional image synthesis. *Advances in neural information processing systems*, *34*, 27196–27208.

Zhao, B., Yin, W., Meng, L., Sigal, L. (2020). Layout2image: Image generation from layout. *International Journal of Computer Vision*, *128*, 2418–2435.

Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., ... Sun, T. (2021). Lafite: Towards language-free training for text-to-image generation. *Conference on computer vision and pattern recognition*.