Trace Controlled Text to Image Generation

Kun Yan[1], Lei Ji[2], Chenfei Wu[2], Jianmin Bao[2], Ming Zhou[3], Nan Duan[2], and Shuai Ma[1]

[1] SKLSDE Lab, Beihang University, Beijing, China
{kunyan,mashuai}@buaa.edu.cn
[2] Microsoft Research Asia, Beijing, China
{leiji,chewu,jianbao,nanduan}@microsoft.com
[3] Langboat Technology, Beijing, China
zhouming@chuangxin.com

Abstract. Text to Image generation is a fundamental and inevitable challenging task for visual linguistic modeling. The recent surge in this area such as DALL·E has shown breathtaking technical breakthroughs, however, it still lacks a precise control of the **spatial** relation corresponding to semantic text. To tackle this problem, mouse trace paired with text provides an interactive way, in which users can describe the imagined image with natural language while drawing traces to locate those they want. However, this brings the challenges of both controllability and compositionality of the generation. Motivated by this, we propose a Trace Controlled Text to Image Generation model (TCTIG), which takes trace as a bridge between semantic concepts and spatial conditions. Moreover, we propose a set of new technique to enhance the controllability and compositionality of generation, including trace guided re-weighting loss (TGR) and semantic aligned augmentation (SAA). In addition, we establish a solid benchmark for the trace-controlled text-to-image generation task, and introduce several new metrics to evaluate both the controllability and compositionality of the model. Upon that, we demonstrate TCTIG's superior performance and further present the fruitful qualitative analysis of our model.

Keywords: Controllable Text to Image Generation, Mouse Trace, Diffusion Decoder

1 Introduction

Human beings always imagine about how a different world could look like even it may not ever exists. For humans, imagining is a kind of creation, a kind of deconstruction and reconstruction of the external environment. Artists can easily "create" the imagined images precisely on what, where as well the actual size, color, shape etc. However, it is quite challenging to ask the current AI agent to create the images human imagined due to 1) lack of natural interaction between human and AI agents and 2) the technique barrier of decoupling semantics concepts and then composing them into image. Although the emerging research

works on text to image generation has made great progress, it is still far from satisfaction.

Controllable image generation aims to create image with fine-grained conditions and plays a key role in both research and industrial applications. These advanced works can be categorized into three types depending on the controllable signals as *layout-to-image*, *text-to-image*, and *multimodal-to-image*. We argue that only layout or text is not enough for precise image generation. Although layout can represent the spatial relation accurately, *layout-to-image* works rely on predefined object categories and are infeasible in describing open-domain objects, attribute and relations. Additionally, *text-to-image* works rely only on text to describe the image, which are complicated to describe the detailed texture and spatial etc.

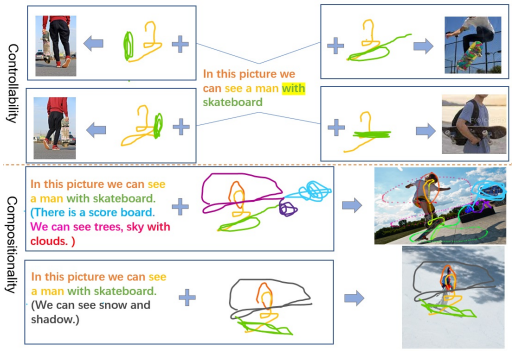


Fig. 1: The task showcases. Both controllability and compositionality are crucial to image generation.

To tackle these limitations, *multimodal-to-image*[21,20,41] generation empowers both text and layout as multimodal inputs. In this paper, we mainly investigate this task, specifically, taking text and mouse traces for image generation, as traces provide a more natural and interactive way than layouts to ground the text into the corresponding position of the image. **Also, in many advanced HCI scenarios, our fingers and eye-sight movement is also trace alike instead of layout alike.** Localized Narrative[29] is such a dataset, where each image is labeled with the detailed narration while simultaneously hovering their mouse pointer over the region they are describing. Thus each word in the narration is grounded into a sequence of traces-points in the image. Upon this, the correspondence between mouse traces and text for an image is the key for trace-controlled text to image generation.

There are two key characteristics for the trace-controlled text to image generation: 1) the narration is open domain natural language description with various types of objects, attributes, and relations; 2) the traces and text are manually grounded and aligned. Figure 1 presents several showcases of the tasks, and demonstrates the two major challenges. On one hand, **controllability** consists of

both semantic condition and spatial condition, specifically object-level, sentence-level semantics, and spatial correspondence. As shown in the case, the text "man with skateboard" correlates to various appearances with different poses between the man and the skateboard, while each appearance can be specified by the mouse traces. On the other hand, **compositionality** is the generalization capability of different combination between context objects as well as various scales. In the showcase, "man with skateboard" coupled with different contexts of "tree" or "snow" aims to generate distinct correlated background.

To deal with these challenges, we first propose a novel mouse trace controlled text to image generation (TCTIG) model with two-stage training of visual discrete tokenization as the first stage and visual generation model as the second stage. In details, we design a module regarding trace as the bridge between the semantic concepts in the natural language description and spatial conditions in the image together with a trace guided re-weighting loss to improve the controllability. Besides, the segment-aligned augmentation strategy are promoted to enrich the compositionality. Furthermore, the existing evaluation metrics only focus on image quality or semantic relevance, we introduce a new evaluation metric Spatial Semantic CLIP Score (SSCP) to evaluate both the controllability and compositionality of the model. Lastly, experimental results on the LN-COCO dataset[28] present the effectiveness of our method on all these existing and the new SSCP evaluation metrics.

2 Related Work

2.1 Text to Image Generation

The text-to-image generation works can be divided into two categorizes: end-to-end and two-stage methods. Previous methods mainly adopt the conditional generative adversarial networks(GAN) [34,38,30,40]. The literature [10,1] presents a review for adversarial text-to-image generation. Recent works investigated the Denoising Diffusion Model[23,12,19], which demonstrate the capability of high quality image generation and beat GAN based method[6]. Although the diffusion models are capable of generating photorealistic images, it still cannot compose complex semantic and needs several pass of re-editing to make up all needed semantics as mentioned in [23]. Another research direction is the two-stage method, which is a new paradigm for pretraining with web-scale image and text pairs. They demonstrated the pretrained models are effective in generalizing high semantically related open-domain images, which first adopt VQVAE[24]/VAGAN[8] to tokenize images into discrete tokens and then generates these visual tokens for further decoding to real images[32,7,41]. However, these methods always produce blurred images with artifacts more or less. Motivated by the superior performance of these models, we propose a novel two-stage model with carefully designed module for trace as another controlled signal. Specifically, our model takes advantages of both high-quality capability of diffusion model and semantics relevance of the two-stage model.

2.2 Conditioned Image Generation

Besides the text-controlled image generation, there are also other conditions including explicit conditions like a semantic mask[5,26,9], layout[42,36], scene graph[17], trace[20], and implicit conditions like style [18,27]. Recently hybrid conditions for image[21,20,41] generation are emerging in the research community by combining text with layout/style inputs. *TRECS[20] is the only work for text and mouse traces to image generation, which retrieved and composed a mask for a further mask-to-image generation.* However, TRECS heavily relied on the object/thing categories predefined in the COCO-STUFF[3] dataset and is hard to scale to open-domain text with fine-grained attributes and long-tail objects. Besides, the grounding of the description to the trace is another challenge for generating both semantically and spatially matched images. To deal with these, we propose a novel trace-controlled text-to-image generation (TCTIG) model by leveraging the advantage of open-domain text-to-image methods, and a semantically aligned regularization for grounding.

2.3 Trace related image Tasks

Trace is a natural way of spatial input for humans to interact with images for image-related tasks. With the release of the Localized narrative [28] dataset, several trace related image tasks are proposed including image and trace to caption generation [28,39,22], image and text to trace generation[22], image to caption with trace generation [22], trace and text to image generation [20], multimodal queries for image retrieval [4], as well as Panoptic narrative grounding [11]. Those tasks elicit more challenges for this research direction, such as fine-grained semantics and dense grounding, and further prompt many real applications for multimodal cognition. In this work, we mainly study the problem of trace and text to image generation, one of the most challenging task.

3 Method

3.1 Preliminaries

Problem Formulation The task is defined as, given text and mouse trace, output the synthesized image. The text is defined as a sequence of word tokens and the trace is defined as a sequence of tracepoints. We adopt the two-stage training strategy. The *first* stage is to tokenize the image into a sequence of discrete visual tokens. The *second* stage is to train our trace controlled text to image generation (TCTIG) model for conditioned image generation by using these visual tokens.

3.2 Visual Tokenization

Specifically, Vector Quantized Variational Autoencoder (VQVAE)[24] learns discrete representations for image and then reconstruct the image by these discrete tokens with a typical encoder-decoder framework. In this paper, we directly adopt the pretrained VQGAN[9], which improves VQVAE by adversarial training.

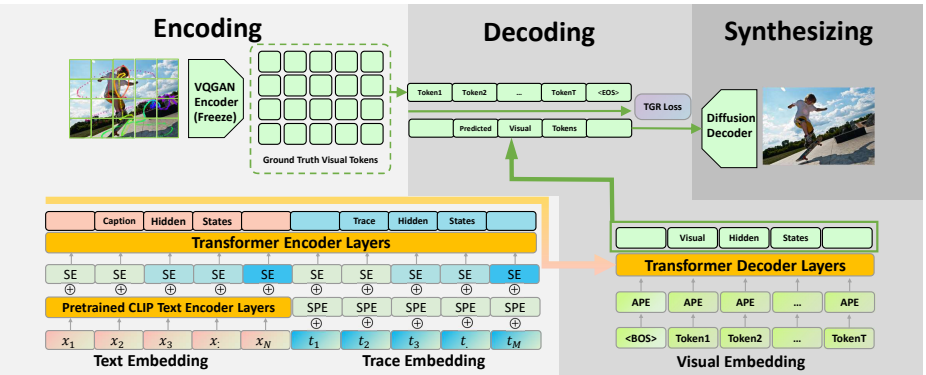


Fig. 2: Model Architecture Overview. "SE" is Segment Embedding, "SPE" is Sinusoidal Positional Embedding, and "APE" is Axial Positional Embedding.

3.3 TCTIG model

Our TCTIG model consists of three major modules: encoding, decoding, and generation. The encoding and decoding process is streamlined with a transformer-based encoder-decoder, which mainly employs multi-head self-attention and cross attention mechanism [37]. Here, we highlight several task-oriented modifications.

Encoding

Text Preprocessing The text description sequence is $\boldsymbol{X} = \{x_1, \ldots, x_l\}$, in which x_i is the i-th token and l is the text sequence length.

Trace Preprocessing The raw trace input is a sequence of trace-point coordinates with timestamps. We segment the trace-point sequences uniformly by the same time window τ, and then each trace segment is converted to its minimal bounding rectangle. Every bounding rectangle can be represented by a 5D vector which contains normalized coordinates of the top-left and bottom-right corners, and the area ratio with respect to the whole image. We denote the trace input as $\boldsymbol{T} = \{t_1, \ldots, t_M\}$, where $t_j \in \mathbb{R}^5$, M is the number of bounding rectangles.

Text-Trace Encoder The text $\boldsymbol{X}$ and traces $\boldsymbol{T}$ are embedded separately and then concatenated together as a single input sequence feeding into a transformer encoder.

– **Text Embedding**: Each text token x_i is embedded into $\hat{x}_i \in \mathbb{R}^d$. Then, we employ relative positional embeddings and learnable segment embeddings to represent token-level and sentence-level semantics repectively. The relative positional embeddings o_i, more specifically, are Sinusoidal Positional Embeddings (SPE) [37] to capture the temporal order of the text. For Segment Embedding (SE) s_i, every token in the same sentence shares the same

embedding to differentiate different sentences. The final text embedding $\tilde{\boldsymbol{X}} = \{\tilde{x}_1, \ldots, \tilde{x}_l\}$, where $\tilde{x}_i = \hat{x}_i + o_i + s_i$.

- **Trace embedding**: We project each 5D vector t_j into a spatial embedding $\hat{t}_j \in \mathbb{R}^d$, where d is the embedding size across the model. Trace is taken as the grounding between text to image. Besides the spatial embedding for image grounding, the trace also encodes the same positional embedding SPE o_j and segment embedding SE s_j as text for text grounding. Thus trace is regarded as a bridge for semantic text and spatial image. The final trace embedding is $\tilde{\boldsymbol{T}} = \{\tilde{t}_1, \ldots, \tilde{t}_M\}$, where $\tilde{t}_j = \hat{t}_j + o_j + s_j$.
- **Encoder**: The embeddings are concatenated together and input to the transformer encoder for further processing. In order to leverage the pretrained model, we initiate our text encoder with the CLIP [31] weights.

$$[\bar{X}; \bar{T}] = \text{Transformer}([CLIP_{text}(\tilde{\boldsymbol{X}}); \tilde{\boldsymbol{T}}]). \tag{1}$$

Decoding

Visual Tokens We use VQGAN[9] with a discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^K$ to encode every target image into a sequence of discrete image tokens $V = [v_1, \ldots, v_{h\times w}] \in \{0, \ldots, |\mathcal{Z}| - 1\}^{h\times w}$. Specifically, we adopt the VQGAN model pretrained on ImageNet with codebook size $K = 1024$, down sample factor $f = 16$ [4]. Thus, for images with size of 256×256, the corresponding visual token length is $N = h \times w = 256$, where $h = 256/f, w = 256/f$.

Visual Token Decoder Visual token decoder combines text and trace information using cross attention connected to the hidden states of Text-Trace Encoder's last layer. Each visual token v_i is embedded into $\hat{v}_i \in \mathbb{R}^d$. In order to align the spatial relation between trace and image, the input for decoder combines the visual embedding with the axial positional embedding (APE) [16]. The position embedding p_i is a linear projection of column and row axis of each visual token in the image. The final input visual embedding is $\tilde{\boldsymbol{V}} = \{\tilde{v}_1, \ldots, \tilde{v}_N\}$, where $\tilde{v}_i = \hat{v}_i + p_i$.

$$\hat{V} = \text{Transformer}(\tilde{\boldsymbol{V}}, [\hat{X}; \hat{T}]). \tag{2}$$

A cross-entropy generation loss $\mathcal{L}_{gen}$ is then computed with the logits transformed from the last decoder layer's hidden states and ground truth visual token ids.

$$\mathcal{L}_{gen} = - \mathop{\mathbb{E}}_{\hat{v}_i \sim \hat{V}} \log p \left(\hat{v}_i \mid \hat{V}_{<i}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{T}}; \boldsymbol{\theta} \right). \tag{3}$$

[4] The checkpoint and model config can be found at https://heibox.uni-heidelberg.de/d/8088892a516d4e3baf92/

Trace Guided Re-weighting(TGR) Loss During self-regression training process, traditional models treat all N target image tokens equally without discrimination. But naturally the tokens corresponding to user-described objects should gain more attention and are usually hard to learn due to its high-frequency information density. In order to help the model focus on important details and better ground the description to the image, we develop a trace guided re-weighting loss (named as TGR) on top of traditional cross-entropy loss. Given a sequence of trace boxes $\boldsymbol{T} = \{t_1, \ldots, t_M\}$ defined as above, we first calculate the center coordinates of each boxes, we denote the x coordinates of those centers as $\boldsymbol{T}^x = \{t_1^x, \ldots, t_M^x\}$ and the y coordinates of those centers as $\boldsymbol{T}^y = \{t_1^y, \ldots, t_M^y\}$, we calculate each trace-box's corresponding token position id $P = \{pid_1, \ldots, pid_M\}$ where

$$pid_i = \lfloor t_i^x * w \rfloor + \lfloor t_i^y * h \rfloor * w \tag{4}$$

We then calculate the frequency distribution of each position ids within every image

$$\boldsymbol{D} = \frac{bincount(P)}{|P|} \tag{5}$$

The re-weighted loss function can be formulated as:

$$\mathcal{L}_{TGR} = -\frac{\mathop{\mathbb{E}}_{\hat{v}_i \sim \hat{V}, d_i \sim \boldsymbol{D}} (1 + \alpha * d_i) \log p\left(\hat{v}_i \mid \hat{V}_{<i}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{T}}; \boldsymbol{\theta}\right)}{\mathbb{Z}} \tag{6}$$

Where α is a learnable parameter to fade the weight into cross-entropy loss and $\mathbb{Z} = 1 + \frac{\alpha * \sum \boldsymbol{D}}{|\boldsymbol{D}|}$ is a scaling factor to keep the loss scale comparable to traditional cross entropy loss.

Synthesizing One option of synthesizing is directly adopting the pretrained VQ-GAN decoder to synthesize the image I given the decoded discrete visual tokens V. But we find the VQ-GAN decoder brings its limitation on reconstruction quality to the whole pipeline in practical experiments. Even using ground truth tokens, the reconstruction result still has non-negligible artifacts.

In recent days, diffusion models [6] is shown to be effective in generating high-quality images and beat GANs on image synthesizing. In our work, we find using latent embedding extracted from VQ-code book as additional conditions to train an guided diffusion model dramatically improves the reconstruction quality. Specifically, we adopt similar framework as in [6]. To guided the model generate images represented by latent code, we concatenate the VQ-GAN latent code embeddings of each image to the U-Net bottle neck during diffusion steps and further train a VQ-latent conditioned diffusion model (Please find the training details in supplementary materials). Thus, at inference stage, we feed the transformer generated V's corresponding code book embeddings to the diffusion decoder, and get a high quality image I.

Training the transformer on the tokens from the VQ-GAN encoder allows us to allocate its modeling capacity to the low-frequency information that makes images visually recognizable to us. And diffusion decoder plays a good role to reconstruct high frequency, long-tailed information. By carefully combining those two methods, we reach a sweet point between efficiency and quality.

3.4 Segment Aligned Augmentation(SAA)

Traces, as a type of control signal, should be highly sensible by the model. Not only the spatial location but also the scale are essential to model. That is, when moving a trace segment spatially, the correlated semantic on the image should move along the same direction. Besides, when we enlarge or shrink the scale of a trace segment, the corresponding object should be enlarged or shrunk accordingly. An intuitive way to enhance this controllability is by employing a data-centric strategy. We design a so-called segment aligned augmentation to create more spatial-aware training cases dynamically while keeping correct alignment between text, trace, and image target.

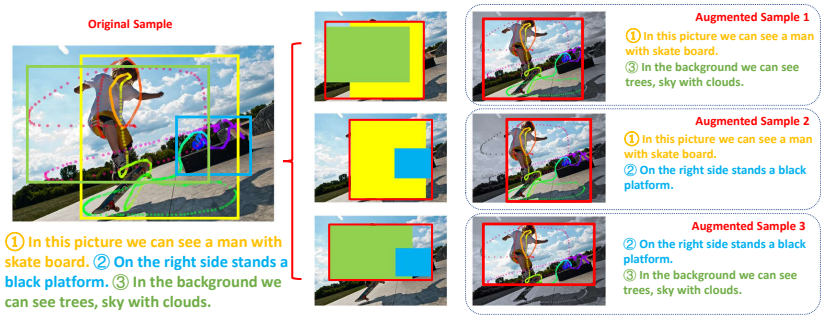


Fig. 3: Segment Aligned Augmentation.

As illustrated in Figure 3, given a sample of the dataset, we first segment the caption paragraph into sentences and its corresponding traces into several pieces. For every piece of trace segment, we get its minimal bounding box. Next, we randomly sample several trace segments and reunite the corresponding bounding boxes, and thus their outer bounding forms a new cropping window. We crop the original image along those new cropping windows, and re-calculate the coordinates of the corresponding trace segments since the relative position of every tracepoint has been changed. Finally, we collect the selected sentences, re-scaled trace segments, and cropped images as a new augmented sample.

By applying this strategy, the relative scale and positions of visual objects as well as trace coordinates encoding are dynamically changed but kept aligned all along. Our experiments demonstrate this approach is crucial for preventing the model from falling into trivial biases and it also improves the overall performance significantly.

4 Experiments

4.1 Dataset

We use the annotated COCO subset of Localized Narratives[28] to evaluate our method, which is called **LN-COCO** for short. Each image has one or several pairs of the captioning paragraph and corresponding mouse traces. There are 134,272 samples in the training set and 8,573 in the validation set.

4.2 Quantitative Results

Dataset	Method	Fidelity		Relevance		Controllability
		$IS \uparrow$	$FID \downarrow$	$SOA-I \uparrow$	$SOA-C \uparrow$	$SSCP \uparrow$
COCO-Caption	DALL·E [32]	17.8	28	-	-	-
	CogView [7]	18.2	27.1	-	-	-
LN-COCO	Real Images	35.70	0.48	0.6608	0.6739	1.00
	AttnGAN [38]	20.80	51.80	-	-	-
	TRECS [20]	**21.3**	48.7	***0.3523**	***0.3288**	*0.862
	TCTIG(-Trace)	10.70	75.34	0.1021	0.0698	0.727
	TCTIG	17.59	**11.94**	0.2658	0.1787	**0.973**

Table 1: Main Results. * indicates the results evaluated by us and missed in the original paper, - means to remove the module.

In this section, we investigate how our approach quantitatively compares to existing models and further establish metrics to assess the characteristic performance of trace-controlled image synthesis task. To measure the key components of the task including"image generation", "text to image" and "trace controlled", we evaluate the model's performance from the following three perspectives:

- **Fidelity** of synthesized images
- **Semantic relevance** between text conditions and generated images
- **Controllability** of semantic arrangement and spatial composition empowered by traces

The metrics and results are presented in Table 1 and are discussed in the following paragraphs.

Fidelity

Inception Score(IS) As described in [35], Inception Score compares each image's label distribution with the whole set marginal image label distribution, encouraging models to synthesize distinguished and diverse samples.

While IS has been shown to correlate with human judgments of generated image quality, it is likely less informative as it overfits easily and can be manipulated to achieve much higher scores using simple tricks [2,40]. The other limitation of IS is that it is designed for images with salient objects which can be classified into those predefined categories. When it comes to complex scenes such as most cases in LN-COCO, the score will intrinsically drop with a large margin. To follow the traditions, we still present the IS score of our method but do not take it as a major concern.

Fréchet Inception Distance (FID) was first proposed by [14], which measures feature distribution distance between real and fake images.

From Table 1, we can see that our method achieves the best FID on trace-controlled text-to-image generation tasks. To the best of our knowledge, our approach is the first to apply the token-based two-stage image synthesis paradigm to Localized Narratives. In addition, we make a rough comparison with two-stage baseline methods on text-to-image generation including DALL·E [32] and CogView [7], which conduct evaluations on COCO-Caption. The image distribution of LN-COCO and COCO-Caption is merely identical, thus the FID comparison between our method on LN-COCO and theirs on COCO-Caption is reasonable. Compared with those models, our model is several times smaller(750M parameters compared to 3B of Cogview, 12B of DALL·E), our training data is limited (130K samples compared to 30M of Cogview and 250M of DALL·E), and the task is more challenging (Narratives are four times longer than MS-COCO captions on average with fine-grained description). From the results, we can find that although our method without trace (TCTIG(-Trace)) performs worse, our method with trace (TCTIG) achieves the lowest FID of 11.94. Thus, we demonstrate **trace is a powerful and informative signal and our design to inject trace** to build text-to-image generative models is effective and efficient.

Semantic Relevance

Semantic Object Accuracy (SOA) SOA is a score introduced by [15] to evaluate the semantic relevance of generative text-to-image models. Given the captions of the MS-COCO dataset, first generate images from the model to be evaluated. Then a pre-trained object detector YOLOv3 [33] detects whether the generated image contains the objects specified in the caption. In this paper, to adapt the SOA score calculation on LN-COCO, we use the same keywords and exclusive exceptions defined by [15] to filter all captions of the LN-COCO validation set. The metadata and statistics of each label is listed in the Appendix A.

Conclusion We calculate the SOA score of TRECS [20] by using their released generated samples on LN-COCO. We also conduct an evaluation on reconstructed images by VQGAN decoder given ground truth images tokens. We can see that even using ground truth tokens to reconstruct, the result images still cannot surpass the TRECS, a semantic mask-based generative method, on SOA metric. TRECS are trained with exactly the same categories of objects with YOLOv3, this is beneficial to SOA scores but harmful to **generalizability**. What's more, since VQGAN cannot ensure the object shape is always kept after quantization, this indicates that YOLOv3 detects objects highly relying on shapes. The result of the reconstructed image sets an upper bound on the SOA metric for the future visual token-based method. This paper provides solid baselines for the trace-controlled visual-token-based generative method and demonstrates a way to reduce the gap to the upper bound. While to raise the aforementioned upper bound closer to real images, the image tokenization technique should be further improved to keep the shape of visual objects as efficiently as possible.

Controllability

Spatial Semantic CLIP Score (SSCP) Although the SOA can evaluate whether the semantic concepts in the text are depicted as an object in generated images, it still lacks detailed controllability evaluation. For instance, whether the objects are at the right place, whether those objects have proper relations, whether the descriptive word such as color, texture, and postures are correctly presented. All those aspects can not be differentiated by the SOA score but play a critical role in the assemblage of our visual world. Recently, CLIPScore[13] and CLIP-R-Precision[25] are effective evaluation metrics using the pre-trained CLIP[31] for specific tasks. Motivated by this, we propose a novel Spatial Semantic CLIP (SSCP) score to evaluate the controllability perspective of this task.

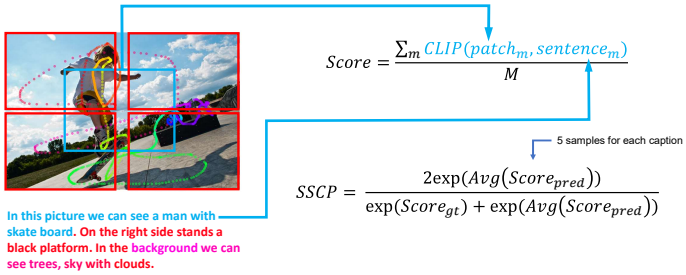


Fig. 4: SSCP Score Calculation

Conclusion The newly proposed SSCP score is to evaluate the controllability of both semantic and spatial grounding for image generation, the detailed calculation of which is presented in Figure 4. First, we crop each generated image into four corner patches and one central patches. Next, we calculate the CLIP score between each patch with the spatially aligned sentences and average scores of these patches as $Score_{pred}$. For each caption, we generate five samples and calculate the average $Score_{pred}$ of them. Finally, together with the $Score_{gt}$ calculated for the ground truth image in the same way, we can get the SSCP score for each sample. We report the SSCP of TRECS by using their publicly released generated samples on LN-COCO and our model in Table 1. Note that, according to our metric, the SSCP score for ground truth images in the validation set is 1.0. Our TCTIG model performs significantly better than TRECS on the SSCP metric(0.973 VS 0.862), which confirms that our generated samples have more accurate spatial layouts and general flexible control over descriptive semantics.

4.3 Ablation Study

Table 2 demonstrates the ablation results including trace, trace guided re-weighting loss (TGR), semantc segment augumetation (SSA) as well as the diffusion v.s. GAN synthesizing. 1) Take a closer look at the TCTIG(-Trace-SAA-TGR) which only takes text as input, we can find that the quantitative results are merely poor.

Dataset	Method		Fidelity		Relevance		Controllability
	Model	*Sythesis*	$IS \uparrow$	$FID \downarrow$	$SOA-I \uparrow$	$SOA-C \uparrow$	$SSCP \uparrow$
LN-COCO	TCTIG(-Trace-SAA-TGR)	GAN	10.70	75.34	0.1021	0.0698	0.727
	TCTIG(-SAA-TGR)	GAN	14.80	29.58	0.1320	0.0728	0.923
	TCTIG (-TGR)	GAN	15.02	26.50	0.2200	0.1565	0.955
	TCTIG	GAN	16.65	19.54	0.2593	0.1609	0.962
	GT Image	GAN	18.74	12.74	0.3501	0.3463	0.989
	TCTIG	Diffusion	**17.59**	**11.94**	**0.2658**	**0.1787**	**0.973**
	GT Image	Diffusion	21.17	8.62	0.4058	0.3636	0.994

Table 2: Ablation Results. - means to remove the module, SAA means Segment Aligned Augmentation, and TGR means semantically aligned loss.

This indicates that with limited caption image pair, the model can not learn enough knowledge about the correlation of linguistic and visual tokens. When we incorporate **traces** into the model, the performance boosts significantly(SSCP from 0.727 to 0.923). 2) And with the segment-aligned augmentation, our TCTIG (-TGR) significantly improve the performance with the **compositionality** and controllability. 3) By the semantic regularization (TGR) supervision, the grounding between the caption and image with mouse traces are learned to improve the **controllability**. 4) Lastly, the VQ based diffusion model further enhance the performance especially on the **image quality**, which achieves the best results. Those ablations are strong evidence of the effectiveness of our method.

4.4 Human Evaluation

As mentioned in [7], human evaluation is much more persuasive than these automatic evaluation metrics on text-to-image generation. We conduct a similar human evaluation through a side-by-side comparison between TRECS and TCTIG model on 1000 randomly selected images. For each case, we ask the annotators to evaluate the generated images from three perspectives including image quality, semantic relevance, and spatial grounding. Figure 5 presents the results, from which we can see that TCTIG performs better especially on spatial grounding. For semantics, we found that TRECS is good at generating predefined popular objects, while TCTIG is more general on open-domain concepts.

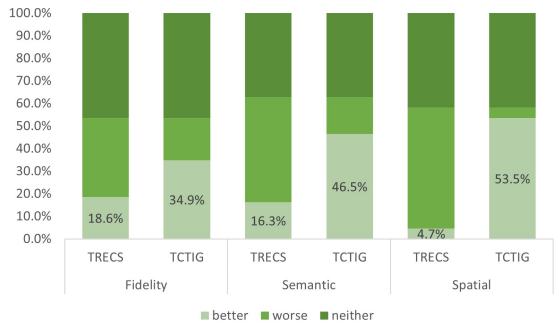
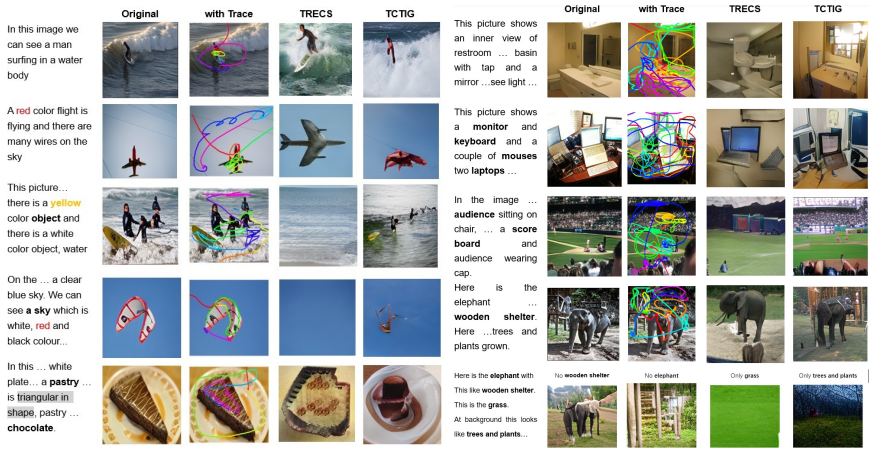


Fig. 5: Human evaluation. "neither" means both methods are equally good or poor to compare.

4.5 Qualitative Analysis

Comparison with related methods Figure 6 lists the results of these samples reported in TRECS[20], and we add our results as well. The baseline AttGAN tends to generate images with semantically relevant textures, while TRECS rely on scene mask to generate more accurate predefined objects. Our model is able to generate open-domain semantically and spatially relevant images.

Fig. 6: Qualitative study with related methods



(a) Controllability Analysis (b) Compositional Analysis

Fig. 7: Qualitative study on controllability and compositional

Controllablity Figure 7a presents the synthesized images with the relatively simple scene with fine-grained attributes and appearance descriptions. We validate the controllability on both semantics and spatial perspectives. As shown in the first case, both TRECS and our model are able to generate accurate images.

These cases present several findings. (1) As expected, with the trace, TRECS and our model are sensitive to the **spatial position**, and the generated objects are in the corresponding spatial position of the image. (2) TRECS model relies on the predefined 181 objects and things in COCO-Stuff for object generation, which is hard to deal with **open-domain** ambiguous and long-tailed objects like "object" in case 3, and "chocolate" in case 6. Our model with an open-domain text encoder is able to generate more accurate semantically related images. (3) Our model is sensitive to the **color attribute** . From cases 2 to 4, the color is explicitly mentioned in the text and the objects in the image should be painted with the right color. Our model can generate the colored objects with the trace position as the patches of the area have the corresponding color. (4) All models are infeasible to learn the **shape attribute** but prefer to ground to the shape of the trace, like case 5. Although it is mentioned "triangular" shape, both TRECS and our model generate trace-shaped objects.

Compositionality Figure 7b presents the synthesized images with composed and complex scenes. From these results, we can see that: (1) TRECS model rely on a composed mask for image generation, and thus would like to generate **composed** images by gathering all objects and lacks a smooth and consistent transition between boundaries as shown in case 1 and case 2. (2) The trace is important for our model to **ground** the objects to the corresponding spatial position in the image, e.g. "mirror ", "laptop", "shelter", and "audience" in the showcases. Our model is able to generate a similar layout to the ground truth image. (3) from the last row, we randomly select sentences from the last example to generate the corresponding image to present the **compositionality**.

5 Conclusion

In this work, we propose a Trace Controlled Text to Image Generation model (TCTIG) to provide a straightforward and natural solution tackling the controllability problem of text to image generation. We establish a solid benchmark for the trace-controlled text-to-image generation task. Upon that, we further demonstrate TCTIG's superior performance by detailed quantitative results and analyze the controllability and compositionality by qualitative studies.

6 Acknowledgement

This work is supported in part by NSFC 61925203.

References

1. Agnese, J., Herrera, J., Tao, H., Zhu, X.: A survey and taxonomy of adversarial neural networks for text-to-image synthesis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(4), e1345 (2020)

2. Barratt, S., Sharma, R.: A note on the inception score (2018)
3. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Computer vision and pattern recognition (CVPR), 2018 IEEE conference on. IEEE (2018)
4. Changpinyo, S., Pont-Tuset, J., Ferrari, V., Soricut, R.: Telling the what while pointing to the where: Multimodal queries for image retrieval. arXiv preprint arXiv:2102.04980 (2021)
5. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1511–1520 (2017)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34** (2021)
7. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. arXiv preprint arXiv:2105.13290 (2021)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis (2020)
10. Frolov, S., Hinz, T., Raue, F., Hees, J., Dengel, A.: Adversarial text-to-image synthesis: A review. arXiv preprint arXiv:2101.09983 (2021)
11. González, C., Ayobi, N., Hernandez, I., Hernández, J., Pont-Tuset, J., Arbelaez, P.: Panoptic narrative grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1364–1373 (2021)
12. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. arXiv preprint arXiv:2111.14822 (2021)
13. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
15. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. IEEE transactions on pattern analysis and machine intelligence **PP** (2020)
16. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multi-dimensional transformers (2019)
17. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1219–1228 (2018)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4396–4405 (2019). https://doi.org/10.1109/CVPR.2019.00453
19. Kim, G., Ye, J.C.: Diffusionclip: Text-guided image manipulation using diffusion models. arXiv preprint arXiv:2110.02711 (2021)
20. Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Text-to-image generation grounded by fine-grained user attention. In: Winter Conference on Applications of Computer Vision (WACV) (2021)

21. Li, B., Qi, X., Torr, P.H., Lukasiewicz, T.: Image-to-image translation with text guidance. arXiv preprint arXiv:2002.05235 (2020)
22. Meng, Z., Yu, L., Zhang, N., Berg, T.L., Damavandi, B., Singh, V., Bearman, A.: Connecting what to say with where to look by modeling human attention traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12679–12688 (2021)
23. Nichol, A., 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. arXiv preprint arXiv:2112.10741 (2021)
24. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf
25. Park, D.H., Azadi, S., Liu, X., Darrell, T., Rohrbach, A.: Benchmark for compositional text-to-image synthesis (2021)
26. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
27. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
28. Pont-Tuset, J., Uijlings, J., Changpinyo, B., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: ECCV (2020), https://arxiv.org/abs/1912.03098
29. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: ECCV (2020)
30. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1505–1514 (2019)
31. 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. arXiv preprint arXiv:2103.00020 (2021)
32. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. ArXiv **abs/2102.12092** (2021)
33. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement (2018)
34. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning. pp. 1060–1069. PMLR (2016)
35. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NIPS (2016)
36. Sun, W., Wu, T.: Image synthesis from reconfigurable layout and style. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10531–10540 (2019)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

38. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 1316–1324 (2018)
39. Yan, K., Ji, L., Luo, H., Zhou, M., Duan, N., Ma, S.: Control image captioning spatially and temporally. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2014–2025 (2021)
40. Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: CVPR (2021)
41. Zhang, Z., Ma, J., Zhou, C., Men, R., Li, Z., Ding, M., Tang, J., Zhou, J., Yang, H.: M6-ufc: Unifying multi-modal controls for conditional image synthesis. arXiv preprint arXiv:2105.14211 (2021)
42. Zhao, B., Meng, L., Yin, W., Sigal, L.: Image generation from layout. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)