Text-to-Image Generation Grounded by Fine-Grained User Attention

Jing Yu Koh,* Jason Baldridge, Honglak Lee and Yinfei Yang
Google Research
{jykoh,jridge,honglak,yinfeiy}@google.com

Abstract

Localized Narratives [28] is a dataset with detailed natural language descriptions of images paired with mouse traces that provide a sparse, fine-grained visual grounding for phrases. We propose TRECS, a sequential model that exploits this grounding to generate images. TRECS uses descriptions to retrieve segmentation masks and predict object labels aligned with mouse traces. These alignments are used to select and position masks to generate a fully covered segmentation canvas; the final image is produced by a segmentation-to-image generator using this canvas. This multi-step, retrieval-based approach outperforms existing direct text-to-image generation models on both automatic metrics and human evaluations: overall, its generated images are more photo-realistic and better match descriptions.

1. Introduction

Text-to-image synthesis goes back at least to 1983 [1]. WordsEye [6] marked a significant evolution: it was a retrieval-based system that extracted relevant 3D models from a database to depict scenes described in text. More recently, deep neural networks based on Generative Adversarial Networks (GANs) [12] have enabled end-to-end trainable photo-realistic text-to-image generation [30, 42, 39]. In contrast with these end-to-end models, others proposed hierarchical models [15, 35, 21]; in these, object bounding boxes and segmentation masks are used as intermediate representations for realistic image generation. Several other approaches also use intermediate scene graph representations [18, 41] for improving image synthesis. Others have explored dialogue-based interactions in which users incrementally provide instructions to refine and adjust generated scenes [33, 8]. This provides users with greater control by allowing them to designate the relative positions of objects in the scene. However, the language used is restricted, and the images are synthetic 3D visualizations or cartoons.

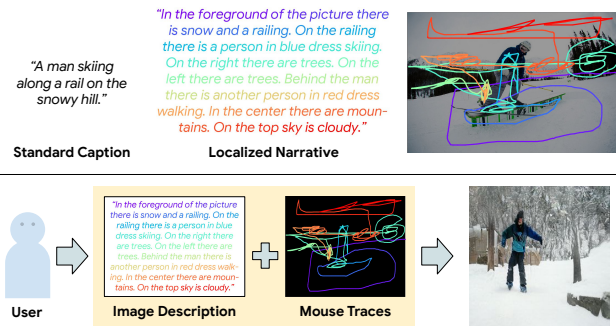*Work done as a member of the AI Residency program.

Figure 1. **[Top]**: Localized narratives comparing against standard captions. **[Bottom]**: Image synthesis from descriptions and traces.

The Localized Narratives dataset [28] provides an alternative paradigm. Instead of writing short captions, annotators scan a mouse pointer over images while describing them. They transcribe their speech, allowing the text and traces to be time-aligned (Figure 1, top). These grounded narratives support the task of *user attention grounded text-to-image generation* [28]: generate an image given a free-form narrative and aligned traces (Figure 1, bottom).

Pont-Tuset et al. (2020) [28] present a proof of concept method which naively generates images using exact matches between words and labels; we build on this core approach. Our system, TRECS (**T**ag-**Re**trieve-**C**ompose-**S**ynthesize, Fig. 2), significantly enhances the image generation process by improving how language evokes image elements and how traces inform their placement.

- A **tagger** predicts object labels for *every* word. We train a BERT [7] model on the output of a constrained Hidden Markov Model (HMM). The HMM is generated from noisy narrative-to-image alignments.

- A text-to-image dual encoder **retrieves** images with *semantically relevant* masks. This helps select contextually appropriate masks for each object (e.g. skiers instead of bikers for the *person* class in Fig. 2). One mask is chosen per trace sequence to maximize spatial overlap. This overcomes the lack of ground-truth text-to-object information and better grounds descriptions.
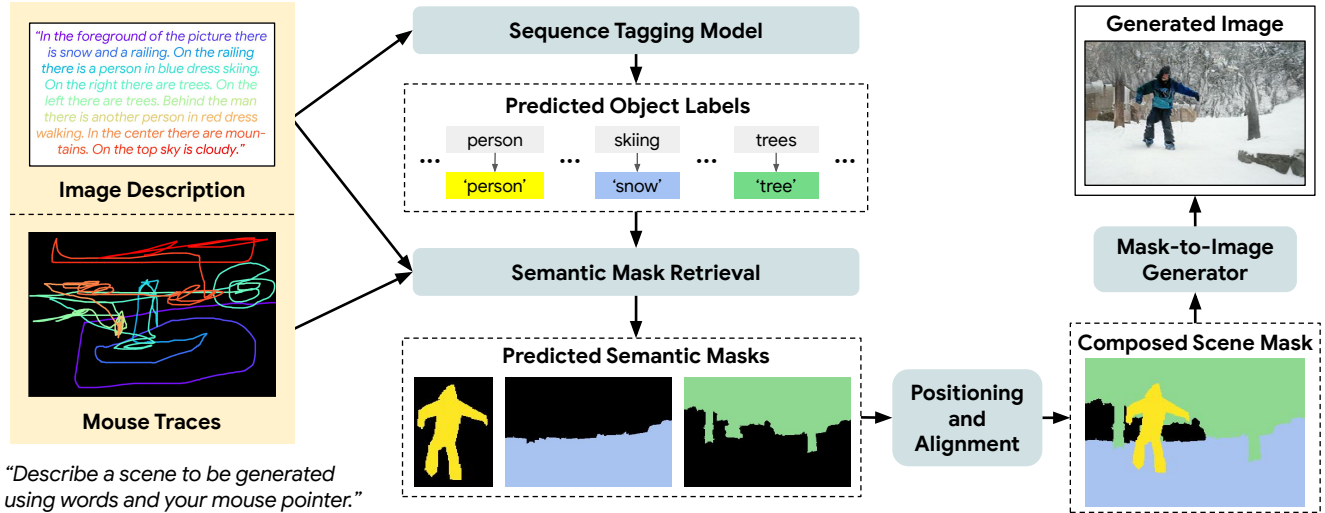
Figure 2. Multi-stage TRECS system for image synthesis using both descriptions and mouse traces.

- Selected masks are **composed** corresponding to trace order, with separate canvases for background and foreground objects (closely related to `stuff` and `thing` masks from [4]). The foreground mask is placed over the background one to create a full scene segmentation.

- Finally, a realistic image is **synthesized** by inputting the complete segmentation to mask-to-image translation models such as SPADE [27] or CC-FPSE [23].

TRECS exploits both text and mouse traces. Compared to other strategies, especially those requiring scene graphs, pointing with a mouse while talking is a more natural way for users to indicate their intent during image synthesis.

Localized Narratives are longer and more detailed than standard captions; they average 41.8 words in length, four times that of MS-COCO captions (averaging 10.5 words). This presents a major challenge for existing direct text-to-image models like AttnGAN [39]. Our multi-stage strategy better copes with the detail and captures the spatial configurations of the described objects and backgrounds. On the COCO portion of Localized Narratives (LN-COCO), TRECS beats AttnGAN on automatic metrics of image quality as well as side-by-side human evaluations of both realism and image-text alignment: TRECS is preferred 77% of the time for realism (compared to AttnGAN's 23%), and 52.4% of the time for alignment (compared to AttnGAN's 32.0%, with the remainder being ties or failures for both models). On the Open Images portion of Localized Narratives (LN-OpenImages), AttnGAN obtains better scores on automatic metrics, but TRECS's output is still judged better in human evaluations, with 77.2% of TRECS images preferred for realism (compared to AttnGAN's 22.8%), and 45.8% preferred for language alignment (compared to AttnGAN's 40.5%). We show that high quality images can

be generated for many narratives, but there is nevertheless much room for improvement on this challenging task.

Our key contributions are:

- We show, for the first time (to the best of our knowledge), viability for the very difficult task of grounded text-to-image synthesis for *narratives* (as compared to prior work on shorter captions).

- We propose TRECS, a sequential generation model that uses state-of-the-art language and vision techniques to generate high quality images that are aligned with both language and spatial mouse traces.

- We conduct both automatic and human evaluations that demonstrate the improved quality of TRECS generated images over prior state-of-the-art. Through extensive ablative studies, we identify key components of the TRECS pipeline that are essential for the user attention grounded text-to-image generation task.

2. The TRECS System

We observed that outputs from leading end-to-end text-to-image models [42, 39, 21] leave much to be desired; in particular, their generated images captured a visual gist of the descriptions but lacked well-defined objects and coherent composition. Motivated by the fact that models like SPADE [27] can produce realistic images when given gold-standard segmentation masks, [28] briefly sketch out an alternative based on mask retrieval and composition, followed by segmentation-to-image generation. We found this approach produces recognizable objects in some cases, but also often produces blank images–due largely to the limited use of the language in the narratives. Our TRECS system (Figure 2) significantly enhances this strategy by better
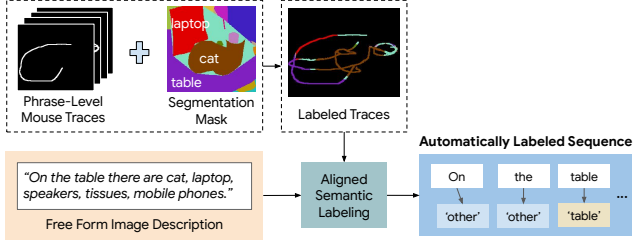
Figure 3. Overview of mouse trace sequence tagging.

modeling the relation between language and the selection and placement of visual elements.

2.1. Sequence Labeling with Pixel Semantics

Traces in Localized Narratives cover a small portion of an image's pixels, but their fine-grained alignment to the *narrative* makes them valuable indicators of the placement and scale of described items. Given both sources of information, a skilled artist could render a scene that visually captures a narrative. Datasets often used for text-to-image generation, e.g. COCO [22], Caltech-UCSD Birds [38], and Oxford Flowers-102 [25], do not contain such fine-grained descriptions. The latter two furthermore lack diversity in both descriptions and images.

TRECS exploits word-trace alignments and transition information to assign image labels to each word in training set narratives (Fig. 3). These are used to train a BERT [7] model to predict tags for new narratives. To address noise in the traces and alignments, we combine three long-standing methods to automatically refine word-object assignments: *tf-idf* weighting [24], IBM Model 1 [3], and Hidden Markov Models [29]. We use two key observations: (1) narratives mention items that are found in the images and (2) traces pass through coherent image regions and thus provide useful category transition information (e.g., *cloud* labels frequently occur next to *sky* labels). To extract semantic labels of the image, we rely on the COCO-Stuff dataset [4], which provides pixel-level semantic segmentation masks for the COCO portion of Localized Narratives.

As a starting point, we directly use word-trace alignments and gold-standard segmentation masks of each image in the training set to tag its narrative with image labels. For a given phrase and its corresponding trace, we obtain the convex hull for the trace and determine the image label that is assigned most frequently within it. This label is assigned to all words in the phrase. Using the hull rather than just the trace reduces noise, especially when annotators refer to an item by circling it with the mouse pointer (a common approach when describing objects). This produces assignments such as:

```
this/snow  picture/other   shows/snow  a/snow
person/person skiing/snow  on/snow  the/snow  snow/snow
he/snow  wore/person  a/person  helmet/person  on/person
```
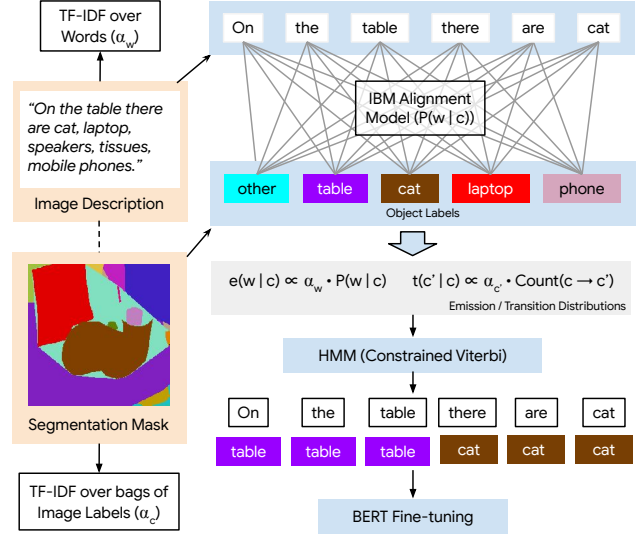

Figure 4. Semantic label refinement process.

```
his/person head/person and/other he/snow  wore/other
ski/other pads/other to/background
```

Words do not always match their assigned labels, but the overall phrase extents correspond with traces that visited image regions with those labels. However, strong word-label associations are missed, producing evidently inappropriate tags such as he/snow and ski/other.

Constrained Label Refinement We combine these assignments with several further steps to obtain cleaner assignments (Fig. 4). First, we compute term frequency-inverse document frequency (*tf-idf*) scores α_w for each word w in the vocabulary, treating each narrative as a document. We use these to reduce the influence of common, uninformative phrases (e.g. *"in the image"*). *tf-idf* scores α_c for image labels c are computed similarly. Next, we learn word-label alignments by pairing each narrative and the corresponding bag of image labels and training IBM Model 1 on this corpus. From this, we obtain translation probabilities $P(w|c)$.

We construct an HMM with these building blocks. The emission distributions are obtained by scaling the translation probabilities with the α_w scores:

$$e(w|c) \propto \alpha_w \cdot P(w|c).$$

Similarly, the transition probabilities are defined as:

$$t(c'|c) \propto \alpha_{c'} \cdot \text{Count}(c \rightarrow c')$$

where the counts are obtained from the noisy word-trace assignments discussed above. Add-1 smoothing is used for both distributions. Finally, we scale the contribution of the transitions: for a given step we use $e(w|c')t(c'|c)^{10}$, which

239

we found necessary to allow tags to flip due to peaky emissions. To assign tags to a *training* sentence (used to auto-supervise BERT as described in the next section), we constrain Viterbi decoding to only those image labels annotated on the image itself. The HMM produces more appropriate and semantically aligned tags; for example, the above narrative is tagged as:

```
this/person picture/person shows/person a/person
person/person skiing/snow on/snow the/snow snow/snow
he/person wore/person a/person helmet/person
on/person his/person head/person and/person he/person
wore/person ski/skis pads/person to/person
```

Auto-Supervised Training   When human-annotated data is scarce, a weak generative model with strong initialization can label examples for a more powerful model that generalizes better. Garrette and Baldridge (2012) [10] show that such *auto-supervised training* is effective for low-resource part-of-speech tagging. Here, we fine-tune a pretrained BERT model [7] on the constrained HMM labels. We visually inspected the sequence tags and found that this improved both tagging quality (recall that there are no gold-standard per-word image category annotations) and final image generation quality (see detailed results in Sec. 4.2).

2.2. Semantically Aligned Mask Retrieval

For a narrative we wish to generate an image for, the BERT tagger detects image classes that TRECS must include in the final generated image. We prepare a full semantic scene segmentation by: (1) identifying masks which match those detected classes, (2) are relevant to the narrative, and (3) are spatially aligned to the traces.

We train a cross-modal dual encoder to retrieve k training set images that best match the narrative, and then select COCO-Stuff masks of the detected classes from those images. For each detected class instance c_i, we extract the masks that match that class in the top k images; to satisfy spatial alignment, we select the mask m_i with the highest mean intersection over union (mIOU) with the convex hull S_i corresponding to the traces for instance i:

$$m_i = \underset{M_{c_i,j} \ \forall j \in \{1...,k\}}{\arg\max} \ \text{mIOU}(M_{c_i,j}, S_i)$$

where $M_{c_i,j}$ denotes mask instances of class c_i for the jth retrieved image.

We train an image-text dual encoder (Figure 5) [11, 5, 40] for cross-modal retrieval [9], using the model and code of Parekh et al. (2020) [26]. For text, we extract pre-trained BERT embeddings, which are passed into a 1-layer transformer with 8 attention heads and a hidden dimension of 128 units. The outputs are passed to a fully-connected layer which maps them to $\mathbb{R}^{2048}$. For images, we use a pretrained Inception v3 model [34] fine-tuned during training.
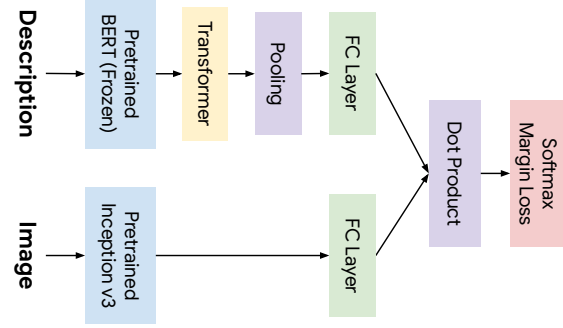

Figure 5. Dual-encoder approach for learning aligned image-text representations.

Following prior work [16, 26], we pre-train on Conceptual Captions [32] and then train on LN-COCO (the Localized Narratives portion for MS-COCO images). The model is trained by minimizing the softmax margin loss with negatives samples from the same batch [40], which simulates image-caption and caption-image retrieval in each batch.

2.3. Mask Composition

The masks selected in the second stage must be composed to represent the scene as a complete semantic segmentation (Fig. 2, bottom right) that is consistent with language and spatial descriptions. Object (`thing`) classes (e.g. *person*, *cat*, *airplane*), have specific sizes, shapes, and identifiable features, while background (`stuff`) classes (e.g. *grass*, *sky*, *water*) are amorphous [4]. `stuff` masks are also generally larger, so they often occlude `thing` masks if care is not taken. Here, we use a straightforward but effective strategy that separately proposes composed layers for `thing` (foreground) and `stuff` (background) masks and then superimposes the former over the latter.

The `thing` layer is created by centering each mask on its corresponding set of traces. They are placed in the reverse order they appear in the narrative—a good default strategy as annotators tend to describe salient objects before less notable details. This strategy works poorly for the `stuff` layer because the masks are larger and clash with each other—as such, many background scenes composed this way would contain only one or two `stuff` classes. To address this, we instead use a full semantic segmentation map from a single image—the one which has the `stuff` objects with the highest mIOU with the `stuff` hulls identified from the narrative and the corresponding traces. Because they are derived from a real image, these background `stuff` masks are semantically coherent with respect to both the narrative and each other. Finally, pixels not associated with an explicit semantic mask are assigned the class label of the closest `stuff` pixel. In addition to providing more coherent backgrounds, this also reduced the number of generated images with large blank regions.

2.4. Mask-to-Image Translation

The first three stages of TRECS construct a novel scene segmentation respecting the descriptions and traces. The last stage aims to produce a photo-realistic image given the full scene segmentation. Compared to free-form text-to-image generation, generating from a segmentation mask is much more constrained. Mask-to-image translation is an area which has seen remarkable recent progress [17, 37, 27, 23]. When given a full segmentation mask for an image, existing networks are able to generate high fidelity images.

For the final stage of TRECS, we experiment with two state-of-the-art mask-to-image generation models: SPADE [27] and CC-FPSE [23], and find that CC-FPSE works better for our use case (see Section 4.2). SPADE and CC-FPSE consist of a generator and a discriminator that are trained adversarially with the GAN framework [12]. The CC-FPSE generator learns a mask-to-image mapping using conditional convolution kernels; the weights of each kernel are predicted based on the mask layout. This gives the model explicit control over the generation process depending on the labels at each spatial location. The CC-FPSE discriminator incorporates multi-scale feature pyramids, which promote higher fidelity details and textures.

Scene segmentation masks created by TRECS are composed from multiple masks retrieved from different images and placed according to the narrative's traces, creating a mismatch with the gold standard segmentation masks that SPADE and CC-FPSE are trained on. Nevertheless, we find that this strategy works well compared to direct text-to-image generation, as we show in Section 4.1.

3. Evaluation Metrics

Image Quality We rely foremost on side-by-side human evaluations. In each case, a human evaluator is presented with output from two competing models for the same narrative and is *forced* to pick which image is more photo-realistic. (Presentation order is randomized for each judgment.) Each image is rated by 5 independent annotators to allow for a majority vote (reducing variation) and provide nuanced breakdowns. See the Appendix for details.

Inception Score (IS) [31] is a widely-used automated metric. To compute IS, the predictions of a model (usually Inception v3 [34] pretrained on ImageNet) are obtained for a *set* of images, and the distribution of its predictions is measured. Higher IS is achieved both when generated images denote clear objects and when predictions are diverse.

Fréchet Inception Distance (FID) [14] improves on IS by comparing real and generated examples. Two multivariate Gaussians are fit to the Inception outputs of the real and generated samples; the FID score is then the Fréchet distance between the two Gaussians. Lower FID indicates greater similarity between the real and generated distributions.

Image-Language Alignment Generated images should also match the description. [15] assessed this fit by running a generated image through an image captioning model and computing BLEU, METEOR and CIDER scores of the generated caption given the original description. They showed that this approach correlated with human judgments. Nevertheless, this is a very indirect measure as it involves two generative models *and* automated text similarity measures. More importantly, as shown in [28], generating long *narratives* from the image alone (not including mouse traces) rather than short captions is a much harder problem, which makes caption-based measures less reliable. Hence, in this paper, we focus our evaluation for image-language alignment on human evaluations.

Similar to the human evaluation for image quality, we present generated images from two models, but also including the narrative they were conditioned on. The evaluators are asked to select the image that is more closely aligned to the narrative. Because at times neither image is a good match, evaluators can also select *Neither*.

4. Experimental Results

Using automatic evaluations and human judgments, we compare TRECS's performance with existing text-to-image generation models. We evaluate all models on the COCO validation set of Localized Narratives (LN-COCO) and on a held out test set of Open Images data [20] that is covered by Localized Narratives (LN-OpenImages). The latter provides a stronger test of model generalization. We also perform several ablations and variations to better understand the impact of choices for each of TRECS's stages.

4.1. Main Results

Many models have been proposed for text-to-image synthesis [15, 39, 21, 36]. We compare closely to AttnGAN [39], as we observed that the pretrained version produced better images for LN-COCO compared to others (e.g. Obj-GAN [21], see Table 1). For fair comparison with TRECS, we fine-tuned AttnGAN on LN-COCO's training set. Note, however, that AttnGAN uses only the narratives and not the traces—giving TRECS an advantage with respect to available inputs. Incorporating traces into end-to-end models like AttnGAN is non-trivial and worth exploring in future.

Image Quality Qualitatively, TRECS's images are crisper and more realistic, as seen in cherry-picked (Fig. 8) and random (Fig. 9) examples. AttnGAN tends to produce textures and blobs that are semantically relevant (e.g. giraffe patterns) but do not represent clear objects. This is confirmed in side-by-side human evaluation of image quality (Fig. 6): for 77.0% of 1000 LN-COCO narratives, the TRECS image was preferred to the AttnGAN image. The same preference was found for the LN-OpenImages test set. We attribute the improvement in image quality to our staged
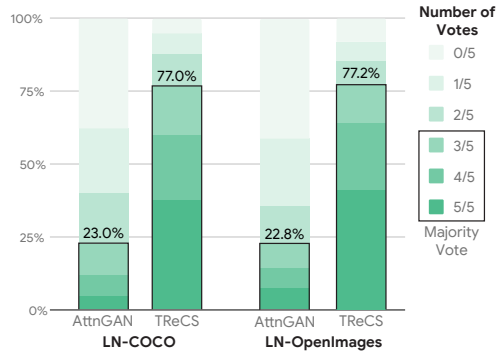
Figure 6. Human evaluation of image quality on LN-COCO validation set and LN-OpenImages test set. Models were fine-tuned on the LN-COCO training set. Of the decisions with 5/5 votes (indicating unanimous preference), TRECS was selected 88.3% of the time on LN-COCO, compared to 11.7% for AttnGAN. On LN-OpenImages, TRECS was selected unanimously 84.1%, compared to 15.9% for AttnGAN.
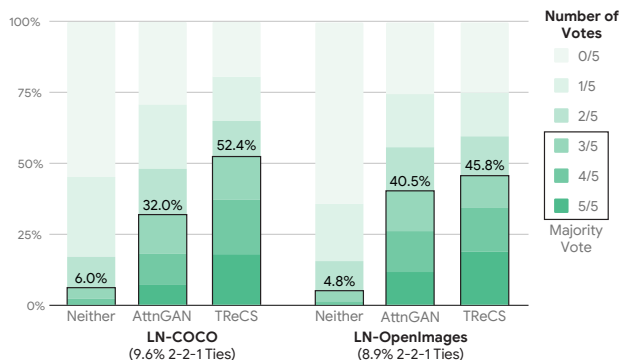

Figure 7. Human evaluation of image-text alignment on LN-COCO validation and LN-OpenImages test sets. Models were fine-tuned on the LN-COCO training set. Of the decisions with 5/5 votes (indicating unanimous preference), TRECS was selected 71.0% of the time on LN-COCO, compared to 27.8% for AttnGAN. On LN-OpenImages, TRECS was selected unanimously 61.4%, compared to 37.9% for AttnGAN.

approach and use of traces. By composing a scene mask instead of direct generation, TRECS explicitly imposes object shapes when composing scenes, and subsequently take advantages of strong mask-to-image generation models to create high fidelity images.

Table 1 shows IS and FID scores for both models on both datasets. Clearly, fine-tuning AttnGAN on LN-COCO makes a large improvement (59.4 FID → 51.8 FID), so all other AttnGAN results in this paper are for the fine-tuned version. On LN-OpenImages, TRECS and fine-tuned AttnGAN swap leadership on these measures: e.g., TRECS is 3.1 FID points better on LN-COCO and 5.3 FID points worse on LN-OpenImages. We find that these metrics pro-

Dataset	Method	IS ↑	FID ↓
LN-COCO	Obj-GAN[†]	16.5	66.5
	AttnGAN[†]	17.4	59.4
	AttnGAN	20.8	51.8
	TRECS	21.3	48.7
LN-OpenImages	AttnGAN	15.3	56.6
	TRECS	14.7	61.9

Table 1. Image quality scores on LN-COCO validation and LN-OpenImages test sets. ↑ (↓) indicates that a higher (lower) number is better performance. † indicates models pretrained on the original COCO, but not fine-tuned on the LN-COCO training set.

vide valuable feedback while developing models, but stress that human evaluation provides the better measure of generated image quality. Note that it is possible to optimize for IS directly and create incomprehensible or adversarial images that nonetheless achieve IS as high as 900 [2].

Image-Text Alignment TRECS also outperforms AttnGAN on human evaluations of image-text alignment (see Fig. 7). Of 1000 LN-COCO held out images, 52.4% were chosen by human raters as being better aligned to the given narrative, compared to 32.0% of AttnGAN images. We observe a similar trend on LN-OpenImages, with TRECS winning 45.8% of contests to AttnGAN's 40.5%.

TRECS also has distinctly better performance when considering cases for which there is full agreement (i.e. decisions where 5/5 voters selected either model, or neither). For these images, TRECS is selected 71.0% of the time, as compared to 27.8% for AttnGAN (the remaining being *Neither*). Similarly, on LN-OpenImages, TRECS is selected unanimously 61.4% of the time, as compared to 37.9% for AttnGAN, indicating a clear preference for TRECS images when evaluating for text-alignment.

TRECS's superior performance may be due to its ability to handle longer free form descriptions. Narratives are much longer (average of 41.8 words) than MS-COCO captions (average of 10.5 words). Also, the narratives are transcriptions of free-form speech and incorporate filler words used in everyday speech. This data presents a challenge for existing text-to-image synthesis models, which were originally created to handle concise and clean captions. By explicitly assigning image labels to words, performing cross-modal matching during mask retrieval, and composing foreground and background's separately, TRECS captures the full range of described objects more effectively.

4.2. Ablations

Sequence Labeling Using HMM tags (Section 2.1), we fine-tuned an uncased BERT-Large model [7]. Weights are optimized using Adam [19] with a learning rate of 1e-5.
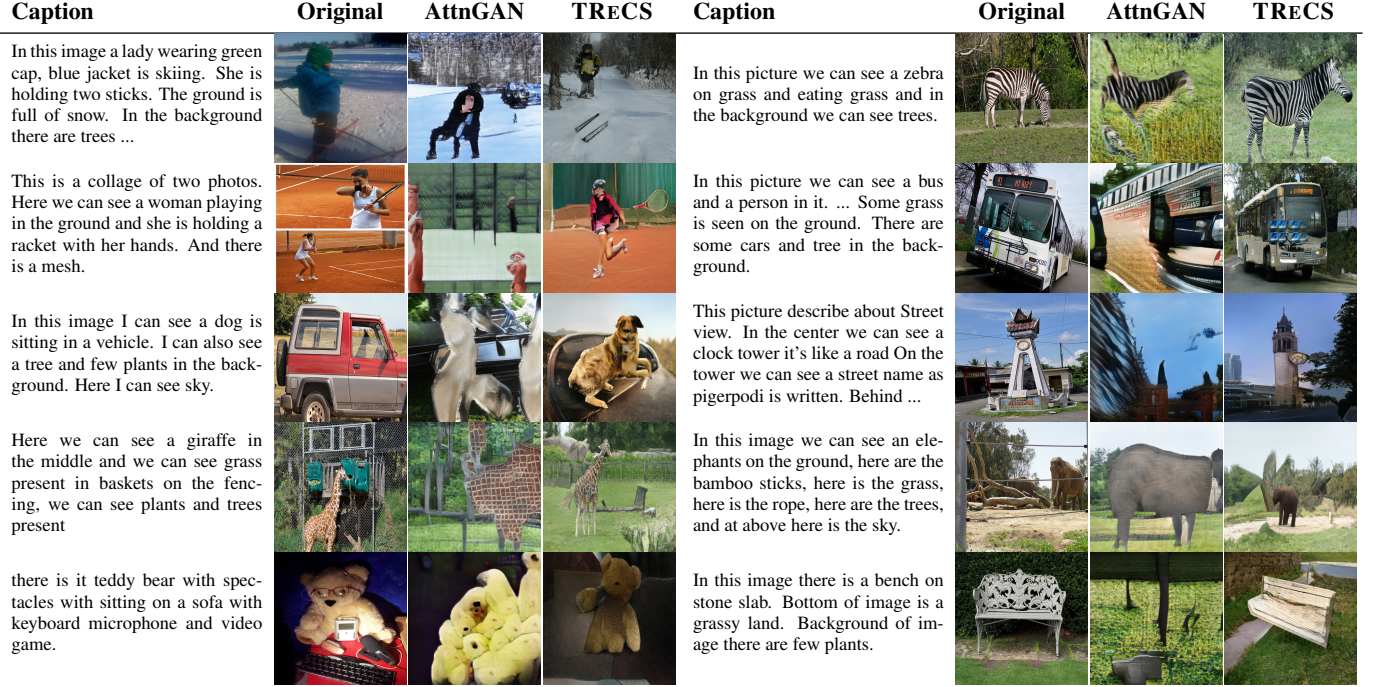
Caption	Original	AttnGAN	TRECS	Caption	Original	AttnGAN	TRECS
In this image a lady wearing green cap, blue jacket is skiing. She is holding two sticks. The ground is full of snow. In the background there are trees ...				In this picture we can see a zebra on grass and eating grass and in the background we can see trees.			
This is a collage of two photos. Here we can see a woman playing in the ground and she is holding a racket with her hands. And there is a mesh.				In this picture we can see a bus and a person in it. ... Some grass is seen on the ground. There are some cars and tree in the background.			
In this image I can see a dog is sitting in a vehicle. I can also see a tree and few plants in the background. Here I can see sky.				This picture describe about Street view. In the center we can see a clock tower it's like a road On the tower we can see a street name as pigerpodi is written. Behind ...			
Here we can see a giraffe in the middle and we can see grass present in baskets on the fencing, we can see plants and trees present				In this image we can see an elephants on the ground, here are the bamboo sticks, here is the grass, here is the rope, here are the trees, and at above here is the sky.			
there is it teddy bear with spectacles with sitting on a sofa with keyboard microphone and video game.				In this image there is a bench on stone slab. Bottom of image is a grassy land. Background of image there are few plants.			

Figure 8. Original and generated images for cherry picked examples from LN-COCO.

Sequence Labels	Generator	IS ↑	FID ↓
BERT (raw labels)	CC-FPSE	20.0	49.9
HMM	SPADE	20.2	50.6
HMM	CC-FPSE	20.7	49.0
BERT (HMM)	SPADE	20.2	49.5
BERT (HMM)	CC-FPSE	21.3	48.7

Table 2. Ablation experiments on the validation set of LN-COCO. ↑ (↓) indicates that a higher (lower) number equates better performance. BERT (raw labels) indicate a BERT model that was trained on raw segmentation labels, as compared to the HMM processed labels. A dual-encoder with k=5 is used for mask retrieval.

k	Recall@k(%)		IS ↑	FID ↓
	C→I	I→C		
GT	-	-	22.0	44.6
1	26.3	27.0	20.3	49.0
5	53.8	52.2	21.3	48.7
10	66.8	63.1	21.0	50.9
20	78.4	74.1	20.7	52.3
50	89.9	86.6	20.0	56.5
100	95.4	92.9	18.8	60.5

Table 3. Retrieval and generation evaluation (LN-COCO) with varying k. **Labeler:** BERT; **image sequencer:** CC-FPSE. **[C→I]:** Caption to image retrieval (vice versa for **I→C**). **[GT]:** Retrieval using groundtruth masks, providing IS/FID upper bounds.

During training and inference, we set the class probabilities of the COCO-Stuff `other` and `background` tags to 0. This assists in downstream image generation, as we found that these classes were not meaningful when presented to the mask-to-image translation models.

Table 2 shows that training BERT on the output of the HMM (auto-supervision) improves image quality over training it on noisy labels (20.0/49.9 → 21.3/48.7) or using the output of the HMM itself (20.7/49.0 → 21.3/48.7). Note that the HMM tags exploit ground-truth image-level labels (and hence are not valid for testing). We observed from manually inspecting image outputs that auto-supervision improved both image quality and image-text alignment.

Mask Retrieval The retrieval model is strong (Table 3): given a query narrative, the groundtruth image is retrieved 53.8% of the time in the top 5 and 95.4% in the top 100, over 8573 images in LN-COCO's validation set. Inspecting retrieved results (e.g. Fig. 10) indicates that this underestimates performance: many retrieved images are good matches, but these narrative-image connections are not in the paired data. Ilharco et al. (2020) [16] provide human evaluations that show retrieval performance is underestimated by the available data, and Parekh et al. (2020) [26] provide new annotations that partly address this gap.

Increasing k (number of retrieved images) improves im-

Caption	Original	AttnGAN	TRECS
A person wearing jacket, helmet is on a ski board holding ski sticks. There is snow. On the back there is a banner, stand, ropeway ...			
Group of people standing and we can see kites in the air and sky with clouds. A far we can see trees. This is grass.			
In this picture we can see one boy is holding a bat and playing a game, he is keeping a cap. soundings there is ...			
In the picture there is a road on the road there are many vehicles there are many poles on the road there are many trees ...			
The image is outside of the city. In the image in middle there are few bags, on right side we can see a person standing ...			
In this image we can see three fire engines on the road. In the background there are trees, houses and sky.			
In this image i can see few benches and the ground covered with the snow. In the background i can see few ...			
In this image I can see the road ... In the back there are signal lights and the vehicles. I can see many trees, building ...			
In this image I can see a cat on a vehicle.			
This picture shows a man standing and skiing and we see blue cloudy sky and a tree and he wore a cap on his head			
This woman wore yellow t-shirt, headband, holding bottle and standing beside this yellow hydrant. On this grass ...			
This is the picture of a floor where we have some laptops, bags and an other back pack in which some things ...			

Figure 9. Original and generated images for random examples from LN-COCO.

Query Description	Retrieved Image #1	Retrieved Image #2	Retrieved Image #3
In this picture we can see food and spoon in the plate.			
In the image there is a donkey truck on the side of the road beside it there is a caution board, on back ...			
In a room people are seated on wooden chairs. In the center there is a rectangular dining table ...			

Figure 10. Dual encoder retrieval examples.

age generation quality up to a certain point, after which IS and FID drop. For LN-COCO, $k=5$ produces the best images (Table 3). This matches our intuition: as k increases, the number of less-relevant masks increases, and some of these are chosen because their shape better matches the trace hull. Conversely, when k is too small, there are insufficient semantically connected masks.

Image Generation CC-FPSE [23] generally outperforms SPADE [27] on IS/FID while keeping other choices fixed (Table 2). We manually inspected generated images from both models and found CC-FPSE also had better image-text alignment. Hence, we used CC-FPSE to synthesize the primary results in Section 4.1.

We also tried fine-tuning CC-FPSE with segmentation outputs created by our retrieval module. The model is trained on a set of original segmentation annotations and a set of noisy segmentation masks created by our retrieval module. The goal of this training process was to fine-tune the generative model to handle noise introduced in the mask retrieval process. However, this training setting did not result in any notable improvements to generation results, most likely due to the large spatial disjoint between noisy masks and real image pixels. Hence, we used the pretrained CC-FPSE model to report our final results.

5. Conclusion

As judged by people, TRECS outperforms AttnGAN in producing higher quality images that are better aligned to the descriptions. TRECS is a step towards an interface that allows a user to describe a scene while controlling size and placement naturally. Such interfaces would support various applications, allowing users to rapidly create interactive scenes without specialized software [6], assist the creative process of an artist [13], or generate realistic images that are more closely grounded to intended descriptions [33].

References

[1] Giovanni Adorni and Mauro Di Manzo. Natural language input for scene generation. In *First Conference of the European Chapter of the Association for Computational Linguistics*, Pisa, Italy, Sept. 1983. Association for Computational Linguistics.

[2] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[3] Peter F. Brown, Stephen A. Della-Pietra, Vincent J. Della-Pietra, and Robert L. Mercer. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313, 1993.

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP)*, 2019.

[6] Bob Coyne and Richard Sproat. Wordseye: an automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[8] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *British Machine Vision Conference*, 2018.

[10] Dan Garrette and Jason Baldridge. Type-supervised hidden markov models for part-of-speech tagging with incomplete tag dictionaries. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012.

[11] Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. End-to-end retrieval in continuous space. *arXiv preprint arXiv:1811.08008*, 2018.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[13] David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations*, 2018.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

[15] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[16] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China, 2019. Association for Computational Linguistics.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[18] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[19] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2018.

[21] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer, 2014.

[23] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems*, 2019.

[24] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[25] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008.

[26] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. *arXiv:2004.15020*, 2020.

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive nor-

malization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[28] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. *Proceedings of the European Conference on Computer Vision*, 2020.

[29] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[30] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, 2016.

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[33] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. Chatpainter: Improving text to image generation using dialogue. In *International Conference on Learning Representations: Workshop*, 2018.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[35] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. *arXiv preprint arXiv:1809.01110*, 2018.

[36] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6710–6719, 2019.

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

[38] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. 2010.

[39] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[40] Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[41] LI Yikang, Tao Ma, Yeqi Bai, Nan Duan, Sining Wei, and Xiaogang Wang. Pastegan: A semi-parametric method to generate image from scene graph. In *Advances in Neural Information Processing Systems*, pages 3948–3958, 2019.

[42] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.