Sequential Attention GAN for Interactive Image Editing via Dialogue

Yu Cheng [1] Zhe Gan [1] Yitong Li [2] Jingjing Liu [1] Jianfeng Gao [3]

[1] Microsoft Dynamics 365 AI Research [2]Duke University
[3]Microsoft Research

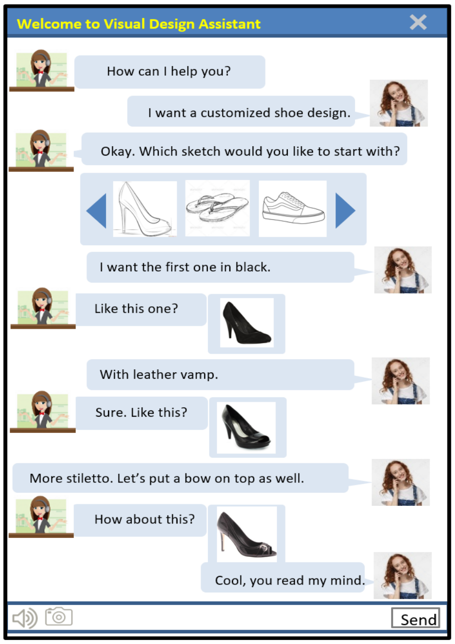
Abstract

In this paper, we introduce a new task - interactive image editing via conversational language, where users can guide an agent to edit images via multi-turn natural language dialogue. In each dialogue turn, the agent takes a source image and a natural language description from the user as the input, and generates a new image following the textual description. Two new datasets are introduced for this task, Zap-Seq and DeepFashion-Seq. We propose a novel Sequential Attention Generative Adversarial Network (SeqAttnGAN) framework, which applies a neural state tracker to encode both the source image and the textual description in each dialogue turn and generates high-quality new image consistent with both the preceding images and the dialogue context. To achieve better region-specific text-to-image generation, we also introduce an attention mechanism into the model. Experiments on the two new datasets show that the proposed SeqAttnGAN model outperforms state-of-the-art (SOTA) approaches on the dialogue-based image editing task. Detailed quantitative evaluation and user study also demonstrate that our model is more effective than SOTA baselines on image generation, in terms of both visual quality and text-to-image consistency[1].

1. Introduction

The volume of visual media has grown tremendously in recent years, which has intensified the need for professional image editing tools (e.g., Adobe Photoshop, Microsoft Photos). However, image/video editing still remains a challenging task that relies heavily on manual efforts and is time-consuming, as visual design requires not only expert artistic creativity but also iterative experimentation. An AI-powered interactive design environment is a natural direction to automate the process, where a system can automatically generate new designs following users' orders.

To reach the ultimate goal of enabling creative collaboration between designers and algorithms for better user experience, we propose a new task - interactive image editing via conversational language, where a system can generate new images by engaging with users in a multi-turn dialogue. Figure 1 illustrates such an interactive image editing system, which supports natural language communication with a user for customizing shape, color, size, texture, etc. of a visual design through natural language conversations. Users can provide feedback on intermediate results, which in turn allows the system to further refine the images. Potential application of such a system can go also beyond dialogue-based visual design to language-guided visual assistance/navigation.



Welcome to Visual Design Assistant

How can I help you?

I want a customized shoe design.

Okay. Which sketch would you like to start with?

I want the first one in black.

Like this one?

With leather vamp.

Sure. Like this?

More stiletto. Let's put a bow on top as well.

How about this?

Cool, you read my mind.

Figure 1. Example of a visual design assistant powered by interactive image editing. In each dialogue turn, the user provides natural language feedback to guide the system to modify the design. The system refines the images iteratively based on the user's feedback.

There are some related studies that explored similar tasks. For example, [3, 11, 6, 32, 26] proposed neural approaches that allow systems to take keyword-based text in-

[1] Code and datasets will be released upon acceptance.

put (e.g., object attributes) for image generation. While these paradigms are effective to some degree, they are either restricted to keyword input or single-turn setting. Only allowing keyword for user feedback inevitably constrains how much information a user can convey to the system to influence the image generation process. Furthermore, without multi-turn capability, the degree of interactive user experience with system assistance is very limited.

To solve these challenges, we propose a new conditional Generative Adversarial Network (GAN) framework, which uses an image generator to generate intermediate results, and a neural state tracker for encoding dialogue context. In each dialogue turn, the generator generates a new image by taking into account both the dialogue history and previously generated images. To fully preserve the sequential information in a dialogue session, the model is trained end-to-end with full dialogue sequences. To achieve better fine-grained image quality and coherent refinement throughout a dialogue session, the model also uses an attention mechanism and a multimodal regularizer based on an image-text matching score.

As this is a newly proposed task, we also introduce two new datasets, namely Zap-Seq and DeepFashion-Seq, which were collected via crowdsourcing in a real-world application scenario. In total, there are 8,734 dialogues in Zap-Seq and 4,820 in DeepFashion-Seq. Each dialogue consists of a sequences of images, with slight variation in design, accompanied by a sequence of textual descriptions on the difference between each pair of images. Figure 1 shows some examples of image/description sequences from both datasets.

Experiments on these two datasets show that the proposed SeqAttnGAN framework achieves better performance than state-of-the-art techniques. In particular, by incorporating dialogue history information, SeqAttnGAN is able to generate high-quality images, beating all baseline models on metrics of contextual relevance and consistency. Detailed qualitative analysis and user study also show that allowing natural language feedback in image editing task is more effective than only taking keywords or visual attributes as input, as used in previous approaches. The contributions of our work can be summarized as follows:

- We propose a new task - interactive image editing via conversational language, which allows users to provide natural language feedback for image editing, via multi-turn dialogue.
- We introduce two new datasets for this task, Zap-Seq and DeepFashion-Seq. With free-formed descriptions and diverse vocabularies, the two datasets provide reliable benchmarks for measuring multi-turn image editing models.
- We propose a new conditional GAN framework - SeqAttnGAN, which can fully utilize dialogue history to synthesize images that conform to user's iterative feedback in a sequential fashion.

2. Related Work

Image Generation and Editing Language-based image editing [6, 22] is a task designed for minimizing labor work while helping users create visual data. One big challenge is that systems should be able to understand which part of the image the user is referring to, for the current editing command. To achieve this, the system requires comprehensive understanding of both natural language and visual information. Following this thread, several studies have explored the task. Hu *et al.* [17] worked on the language-based image segmentation task, taking phrase as the input. Ramesh *et al.* [22] developed a system using simple language to modify the image, where a classification model is used to understand user intent. Wang *et al.* [34] also proposed a neural model for global image editing.

Since the introduction of GAN [13], there has been a surge of interest in image generation tasks. In the conditional GAN space, there has been some studies on generating images from either images [18] or text (e.g., captions [28], attributes [11], long-paragraph [20]). There were also studies on how to parameterize the model and training framework [24] beyond the vanilla GAN [27]. Zhang *et al.* [39] stacked several GANs for text-to-image synthesis, with different GANs to generate images of different sizes.

AttnGAN [33] proposed by Xu *et al.* embedded attention mechanism into the generator, to focus on fine-grained word-level information. Chen *et al.* [6] presented a framework targeting image segmentation and colorization with a recurrent attentive model. The FashionGAN work [32] aimed at creating new clothing on a human body based on textual descriptions. The text-adaptive GAN [26] proposed a method for manipulating images with natural language description. While these paradigms are effective, they all have certain restrictions on the user input (either pre-defined attributes or single-turn interaction), which limits the scope of image editing capability.

Dialogue-based Vision Tasks There are many similar tasks that lie in the intersection between computer vision and natural language processing, such as visual question-answering [2], visual-semantic embeddings [36], grounding phrases in image regions [29], and image-grounded conversation [25].

Most approaches have focused on end-to-end neural models based on the encoder-decoder architecture and sequence-to-sequence learning [12, 31, 4, 8]. Das *et al.* [1] proposed the visual dialogue task, where the agent aims to answer questions about images in an interactive dialogue. De Vries *et al.* [9] introduced the GuessWhat?! game, where a series of questions is asked to pinpoint a specific object in an image. However, these dialogue settings are mainly text-based, where visual feature only plays a supportive role. DeVault *et al.* [23] investigated building dialogue systems that can help users efficiently explore data through visualization. Guo *et al.* [14] introduced an agent presenting candidate images to the user and retrieving new images based on user's feedback. Another piece of related work is [3] for interactive image generation by encoding history information. Different from these studies, in our work, text information is heavily relied on for guiding the image editing process throughout each dialogue session.

3. Datasets: Zap-Seq and DeepFashion-Seq

We define the interactive image editing task as follows: in the t-th dialogue turn, the system presents a generated image $\hat{x}_t$ to the user. The user then provides a natural language feedback o_t to describe the change he/she likes to make for the desired design. The system then takes into account the user feedback and creates a new image based on the preceding image from the previous turn. This process continues iteratively until the user is satisfied with the result rendered by the system, or the maximum number of dialogue turns has been reached.

Existing image generation datasets are mostly single-turned, thus not suitable for this new sequential task. Therefore, we introduce two new datasets - Zap-Seq and DeepFashion-Seq, which were derived from two existing datasets - UT-Zap50K [38] and DeepFashion [21], respectively. UT-Zap50K contains 50,025 shoe images collected from Zappos.com, while DeepFashion contains around 290,000 clothes images from different settings (e.g., stores, street snapshots). Each image comes along with a set of reference attributes.

First, we retrieve sequences of images from the two datasets, with each sequence containing 3 to 5 images. Every pair of consecutive images are slightly different in certain attributes [40]. As a result, a total of 8,734 image sequences were extracted from UT-Zap50K and 4,820 sequences from DeepFashion.

After collecting these image sequences, the second step is to collect natural language descriptions that can capture the difference between each image pair. We resort to crowdsourcing via Amazon Mechanical Turk [5] for this data collection task. Specifically, each annotator was asked to provide a free-formed sentence to describe the difference between any two given images. Figure 2 provides some image examples with textual annotations from the turkers (more examples and the interface of the data collection task is provided in Appendix). To provide robust datasets for measurement, we also randomly select a subset of images from the two original datasets to form additional sequences, which make up to 10% of the whole datasets.

Dataset	Zap-Seq	DeepFashion-Seq
# dialogues	8,734	4,820
# turns per dialogue	3.41	3.25
# descriptions	18,497	12,765
# words per description	6.83	5.79
# unique words	973	687

Table 1. Statistics on the Zap-Seq and DeepFashion-Seq datasets.

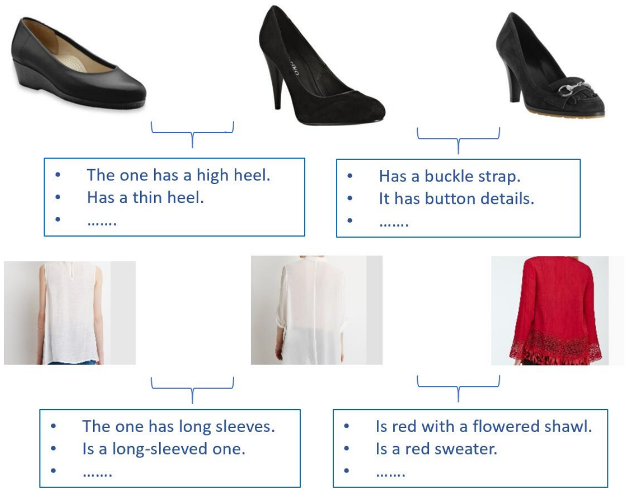


Figure 2. Examples of the collected data. Each annotator is asked to provide a natural language sentence describing the difference between two design images. The images and collected descriptions are used to form "dialogue sequences" for the task.

After manually removing wrong or duplicate annotations, we obtained a total of 18,497 descriptions for Zap-Seq and 12,765 for DeepFashion-Seq. Table 1 provides the statistics on the two datasets.

Most descriptions are concise (between 4 to 8 words), yet the vocabulary is highly diverse (943 unique words in the Zap-Seq dataset and 687 in DeepFashion-Seq). Compared with pre-defined keyword-based attributes provided in the original datasets, descriptions in the new datasets often include fine-grained details on the design. More details about the datasets (e.g., length distribution of text, phrase-type analysis) can be found in Appendix.

4. Sequential Attention GAN

For this new task, we develop a new Sequantial Attention GAN (SeqAttnGAN) model to generate a sequence of images $\hat{x}_1, \ldots, \hat{x}_T$, given an input initial image x_0, and a sequence of natural language user feedback $o_1, \ldots, o_T$. As illustrated in Figure 3, in the t-th dialogue turn ($t \geq 2$), (i) the Dialogue State Tracker fuses the current user feedback o_t with the hidden state h_{t-1} to obtain an updated hidden state h_t; (ii) the Attention Module, together with the upsampling module, fuses the word features of the current user

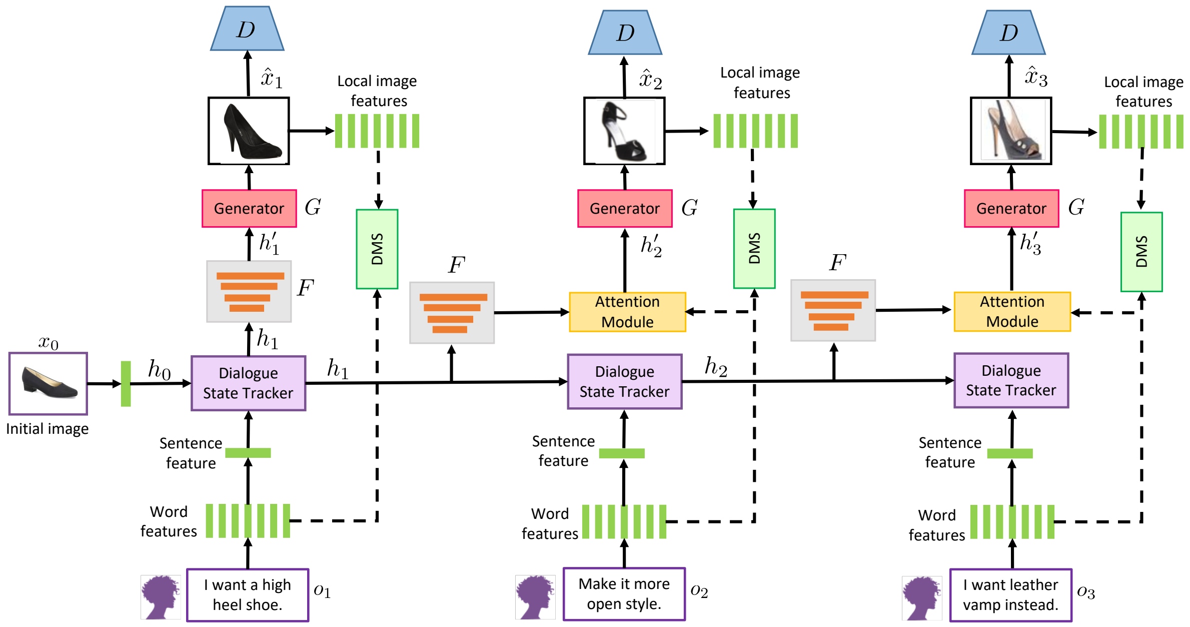


Figure 3. The framework of SeqAttnGAN. The dialogue state tracker keeps track of the contextual information that has been passed on during the sequential image editing process. The attention module absorbs dialogue context for refining different sub-regions of the image. The DMS regularizer provides a fine-grained image-text matching loss. F: Up-sampling. D: Discriminator.

input o_t with the feature map that is up-sampled from h_t, to obtain a dialogue-context-aware image feature set h'_t; (iii) the Image Generator generates the current image $\hat{x}_t$ based on h'_t. The following sub-sections introduce each individual component in detail.

4.1. Dialogue State Tracker

The dialogue state tracker is modeled as a Recurrent Neural Network (RNN) with the Gated Recurrent Unit (GRU) [7]. In the first step, the initial input image x_0 is directly encoded into a feature vector using ResNet-101 [15] pre-trained on ImageNet [10]. This image feature vector serves as our initial hidden state $h_0 \in \mathbb{R}^{d_h}$ of the state tracker. In the t-th step, the hidden state h_t is updated via:

$$h_t = \text{GRU}(h_{t-1}, o_t)\,, \tag{1}$$

$$h'_t = F_{\text{attn}}(o_t, F(h_{t-1}))\,, \tag{2}$$

$$\hat{x}_t = G(h'_t, \epsilon_t)\,, \tag{3}$$

where $F(\cdot)$ is an up-sampling module that transforms the current hidden vector $h_t \in \mathbb{R}^{d_h}$ into a feature map $F(h_t) \in \mathbb{R}^{d_{h'} \times N}$, where N is the number of sub-regions in the image. ϵ_t is a noise vector sampled in each step from a standard normal distribution, and $G(\cdot)$ is the image generator that takes h'_t and ϵ_t as inputs, to generate the image $\hat{x}_t$. h'_t is obtained via the proposed attention module F_{attn} as shown in (2).

4.2. Attention Module

To perform compositional mapping [37, 32, 33], *i.e.*, enforcing the model to produce regions and associated features that conform to the textual description, we introduce an attention module F_{attn} into our framework. Specifically, F_{attn} has two inputs: the word features $e_t \in \mathbb{R}^{d_e \times L}$ that correspond to o_t and the image features $h'_{t-1} = F(h_{t-1}) \in \mathbb{R}^{d_{h'} \times N}$. L is the number of words in a sentence, and d_e is the dimension of the word vector. A word-context vector is computed for each sub-region of the image based on its hidden features h'_{t-1}. For the i-th sub-region of the image (*i.e.*, the i-th column of h'_{t-1}), a word-context vector c_i can be obtained by learning the attention weights of every word in o_t given the i-th sub-region of the image. Finally, $F_{\text{attn}}(o_t, F(h_{t-1}))$ produces a word-context matrix $(c_0, c_1, \ldots, c_{N-1}) \in \mathbb{R}^{d_{h'} \times N}$, which is passed to the image generator G to generate an image in the t-th step.

Compared with AttnGAN [33], our model deploys the attention module in a dialogue sequence. All the dialogue turns share the same image generator, while AttnGAN has disjoint generators for different scales. Hence we name our model as *Sequential Attention GAN* (SeqAttnGAN). Following [39], the objective of SeqAttnGAN is the joint conditional-unconditional losses over the discriminator and the generator. With the supervision of x_t in the t-th dialogue turn, the loss of the generator G is defined as:

$$\mathcal{L}_G = -\frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log D_t(\hat{x}_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log D_t(\hat{x}_t, h_t)]$$

And the loss of the discriminator D is calculated by:

$$\mathcal{L}_D = -\frac{1}{2}\mathbb{E}_{x_t \sim P_{data}}[\log D(x_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log(1 - D(\hat{x}_t))]$$
$$-\frac{1}{2}\mathbb{E}_{x_t \sim P_{data}}[\log D(x_t, h_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log(1 - D(\hat{x}_t, h_t)]$$

where x_t is from the true data distribution P_{data} and $\hat{x}_t$ is from the model distribution P_G.

4.3. Deep Multimodal Similarity Regularizer

In addition to the above GAN loss, an additional image-text matching loss is introduced into SeqAttnGAN. Specifically, we adopt the Deep Attentional Multimodal Similarity Model (DAMSM) developed in [33] to: (i) maximize the utility of all the input information (such as image attributes) to boost the model performance; and (ii) regularize the model in order to stabilize the training of the image generator. DAMSM aims to match the similarity between the synthesized images and user input sentences, acting as an effective regularizer. For simplicity, we call it the DMS regularizer in our paper.

The DMS regularizer is pre-trained on our new datasets. Given a training sample $\{x_0, x_1, o_1, \ldots, x_T, o_T\}$, we first transform it into T image-text pairs. Specifically, for each $t = 1, \ldots, T$, we use x_t as the input image I_i, and then use the concatenation of the image-attribute value of x_{t-1} and the user feedback o_t as the paired text D_i. Note that here we combine image attributes and user feedback as the new text, which is different from [33, 39]. After selecting N image-text pairs, we obtain the new corpus $\{I_i, D_i\}_{i=1}^N$ for DMS training. Following [16, 33], the posterior probability of text D_i matching image I_i is defined as:

$$P(D_i|I_i) = \frac{\exp(\gamma R(I_i, D_i))}{\sum_{j=1}^{M}\exp(\gamma R(I_i, D_j))} \tag{4}$$

where γ is a smoothing factor, $R(\cdot,\cdot)$ is the word-level attention-driven image-text matching score [33] (*i.e.*, the attention weights are calculated between the sub-region of an image and each word of its corresponding text). Given a mini-batch of M pairs, the loss function for matching the images with their corresponding text is:

$$\mathcal{L}_{\text{DMS}}^{i \to d} = -\sum_{i=1}^{M} \log P(D_i|I_i) \tag{5}$$

Symmetrically, we can also define the loss function for matching textual descriptions with their corresponding images (by switching D_i and I_i). Combining these two, the pre-trained DMS regularizer is computed by:

$$\mathcal{L}_{\text{DMS}} = \mathcal{L}_{\text{DMS}}^{i \to d} + \mathcal{L}_{\text{DMS}}^{d \to i} \tag{6}$$

In summary, we use a similar idea to DAMSM in [33] to form our DMS regularizer. The training pairs are created by concatenating image attributes and user descriptions, and the image-text matching score is calculated in each step. By bringing in the discriminative power of DMS, the model can generate region-specific image features that better align with the user's text input, as well as improving the visual diversity of generated images.

By adding all the terms together, the overall objective of SeqAttnGAN is defined as:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_G + \mathcal{L}_D + \lambda\mathcal{L}_{\text{DMS}} \tag{7}$$

where λ is the hyperparameter to balance the two loss functions. $\mathcal{L}_G$, $\mathcal{L}_D$ and $\mathcal{L}_{\text{DMS}}$ are computed in each step and aggregated back to update the gradients, similar to the training of a standard GAN model.

Implementation Details Bi-directional Long Short-Term Memory (BiLSTM) [30] is used to encode the text. The mini-batch size M is set to 50. λ is set to 2 on Zap-Seq and 2.5 on DeepFashion-Seq, respectively. The setting of γ follows [33].

5. Experiments

We conduct both quantitative and qualitative evaluations to validate the effectiveness of our proposed model. Given the subjective nature of this new task, we also conduct human evaluation to compare our approach with state-of-the-art methods.

5.1. Datasets and Baselines

All the experiments are performed on the newly-collected Zap-Seq and DeepFashion-Seq datasets with the same splits: 90% images are used for training, and the model is evaluated on a held-out test set from the rest 10%. The training process uses image pairs sampled from the training set that has no overlap with the test set. We compare our approach with several baselines:

(1) **StackGAN**. The first baseline is StackGAN v1 [39] (we chose v1 due to the low resolution of images in Zap-Seq and DeepFashion-Seq). A generator is trained to generate target images with a resolution of 64 × 64 pixels, based on the reference attributes of images and the descriptions. In other components of the StackGAN architecture, all hyper-parameters and training epochs remain the same as the original.

(2) **AttnGAN**. AttnGAN [33] achieves state-of-the-art Inception Score on MS-COCO. We use AttnGAN1 to generate images at a resolution of 64 × 64 pixels. The discriminator and all the hyper-parameters stay unchanged.

(3) **LIBE**. We also used the recurrent attentive model developed in [6] for image coloring and segmentation tasks

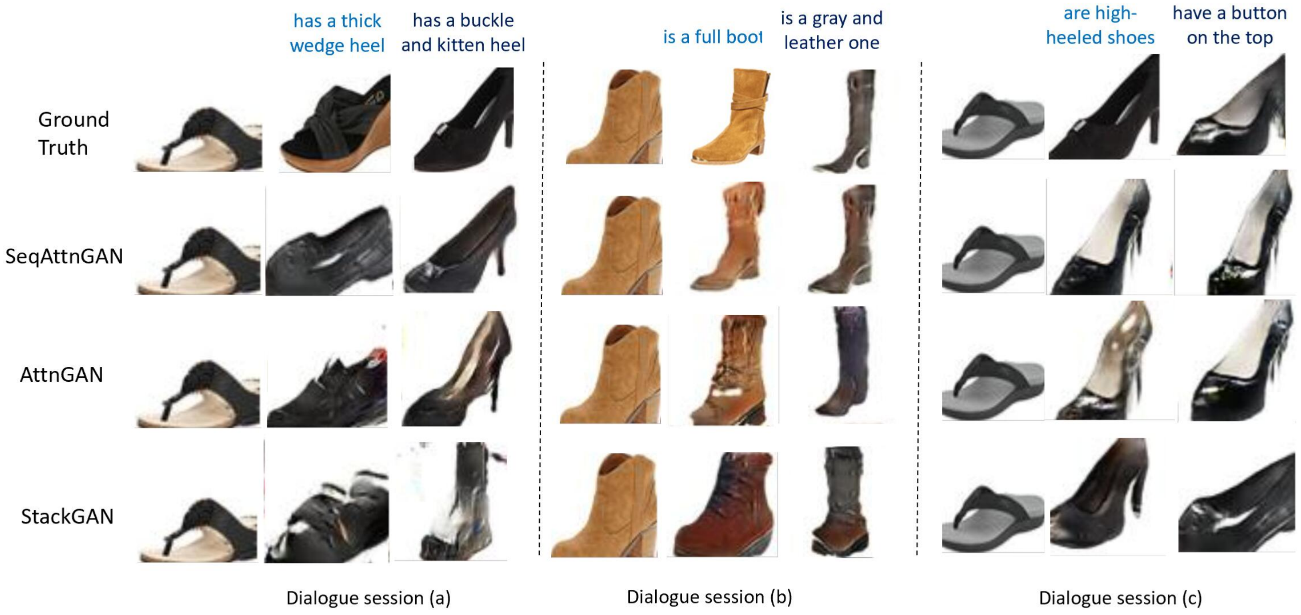


Figure 4. Examples of images generated from the given descriptions in the Zap-Seq dataset. The first row shows the ground-truth images and its reference descriptions, followed by images generated by the three approaches: SeqAttnGAN, AttnGAN and StackGAN. To save space, we only display key phrases of each description.

as another baseline. Similar to AttnGAN and StackGAN, LIBE utilizes image-caption pairs to train the model. The hyper-parameter setting and training details remain the same as in the original paper.

For training, we use bounding box information for images. Data augmentation is also performed on both datasets. Specifically, images are cropped to 64×64 and augmented with horizontal flips. For fair comparison, all models share the same structure of generator and discriminator. Text encoder is also shared. We use the Adam optimizer [19] in training. DMS only appears in the training phase. Baseline model training follows standard conditional-GAN training procedure.

5.2. Quantitative Evaluation

In this section, we provide quantitative evaluation and analysis on the two datasets. For each dialogue turn in the test set, we randomly sampled one image from each model, then calculated Inception Score (IS) and Frechet Inception Distance (FID) scores comparing each selected sample with the ground-truth image. The averaged numbers are presented in Table 2. Our SeqAttnGAN model outperforms both StackGAN and LIBE on the Zap-Seq dataset, with slightly lower performance than AttnGAN. On the DeepFashion-Seq dataset, our model achieves the best results among all the models.

Next, to evaluate whether the generated images are coherent with the input text, we measure the Structural Similarity Index (SSIM) score between generated images and ground-truth images. Table 3 summarizes the results, which shows that the generated images yielded by our model are more consistent with the ground-truth than all the baselines. This indicates that the proposed model can generate images with higher contextual coherency.

Model	Zap-Seq		DeepFashion-Seq	
	IS	FID	IS	FID
StackGAN	7.88	60.62	6.24	65.62
AttnGAN	**9.79**	**48.58**	8.28	55.76
LIBE	4.73	76.52	3.89	79.04
SeqAttnGAN	9.58	50.31	**8.41**	**53.18**

Table 2. Comparison of Inception Score (IS) and Frechet Inception Distance (FID) between our model and the baselines on the two datasets.

Dataset	StackGAN	AttnGAN	LIBE	SeqAttnGAN
Zap-Seq	0.437	0.527	0.159	**0.651**
DF-Seq	0.316	0.405	0.112	**0.498**

Table 3. Comparison of SSIM score between our model and the baselines on the two datastes. Here DF-Seq is the DeepFashion-Seq dataset.

Figure 4 and Figure 5 present a few examples comparing all the models with the ground-truth (more visualized examples are provided in Appendix E). In each example, it is observable that our model generates images more consistent with the ground-truth images and the reference descriptions than the baselines. Specifically, SeqAttnGAN can generate:

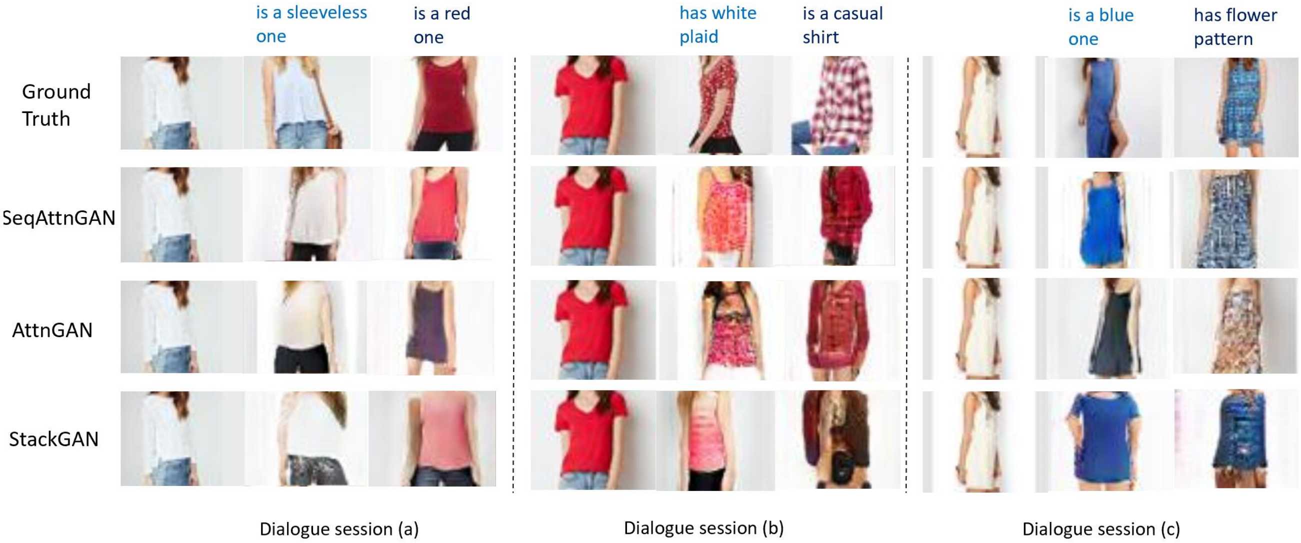


Figure 5. Examples of images generated by different methods on the DeepFashions-Seq dataset.

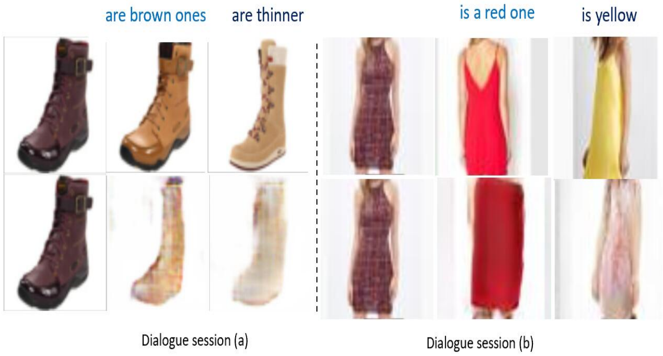


Figure 6. Examples generated by LIBE. In each dialogue session, the first row images are ground-truth images, and the second row images are generated from the model.

1) better regional changes (e.g., session (a) in Figure 4, session (c) in Figure 5); and 2) more consistent global changes on color, texture, etc. (session (b) in Figure 4, session (a) in Figure 5). Even for fine-grained features (e.g., ”kitten heel”, ”leather”, ”button”), the images generated by our model can well satisfy the requirement. AttnGAN is able to synthesize visually sharp/diverse images, but not as good as our model in terms of context relevance (i.e., the generated image does not match the textual description). StackGAN does not perform as well as our model and AttnGAN, in terms of both visual quality and content consistency (i.e., the generated image has drastic design change from the previous image). This observation is consistent with the quantitative study.

Figure 6 also provides some example images generated by LIBE. These examples show that LIBE cannot generate high quality images as the other models do, and can only realize color changes to some degree.

Model	Zap-Seq		DeepFashion-Seq	
	Vis.	Rel.	Vis.	Rel.
StackGAN	2.84$_{\pm 0.25}$	2.61$_{\pm 0.22}$	2.72$_{\pm 0.2}$	2.68$_{\pm 0.19}$
AttnGAN	2.23$_{\pm 0.22}$	2.45$_{\pm 0.21}$	2.26$_{\pm 0.24}$	2.52$_{\pm 0.18}$
SeqAttnGAN	**1.79**$_{\pm 0.17}$	**1.58**$_{\pm 0.14}$	**1.86**$_{\pm 0.19}$	**1.74**$_{\pm 0.13}$

Table 4. Results from the user study for both **vis**ual quality (Vis.) and context **rel**evance (Rel). A lower number indicates a higher rank.

5.3. Human Evaluation

We perform human evaluation via Amazon Mechanical Turk. From each dataset, we randomly sampled 200 image sequences generated by all the models, each assigned to 3 workers to label. The source model of each image is hidden from the annotators for fair comparison. The participants were asked to rank the quality of the generated image sequences based on two aspects independently: 1) consistency to the description and the source image, 2) visual quality and naturalness.

Table 4 provides the ranking comparison of SeqAttnGAN and the other two methods. For each approach, we computed the average ranking (1 is the best and 3 is the worst) and standard deviation. Results show that our approach achieves the best rank on all dimensions. This human study indicates that our solution achieves the best visual quality and image-text consistency among all the models.

Besides the crowdsoured human evaluation, we also recruited real users to interact with our system. Figure 7 shows examples of several dialogue sessions with real users. We observe that users often start the dialogue with a high-level description of main attributes (e,g., color, cat-

Model	Zap-Seq			DeepFashion-Seq		
	IS	FID	SSIM	IS	FID	SSIM
SeqAttnGAN	**9.58**	**50.31**	**0.651**	**8.41**	**53.18**	**0.498**
w/o Attn	8.52	57.19	0.548	7.58	58.15	0.433
w/o DSM	8.21	58.07	0.478	7.24	60.22	0.412

Table 5. Ablation study on using different variations of SeqAttnGAN, measured by IS, FID and SSIM.

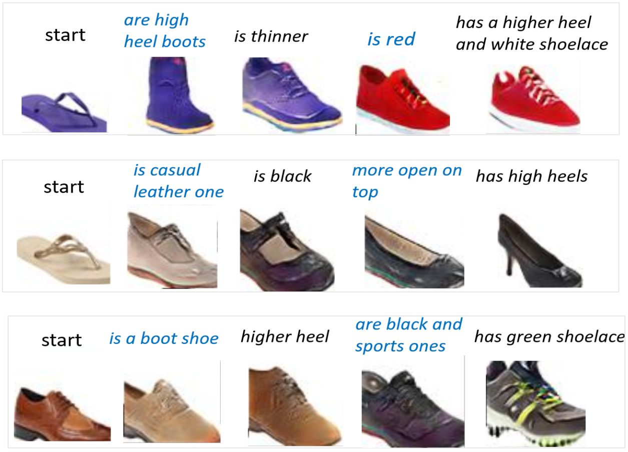


Figure 7. Example sessions of users interacting with our image editing system using SeqAttnGAN model. Each row represents an interactive dialogue between the user and our system.

egory). As the dialogue progresses, users gradually give more specific feedback on fine-grained changes. Our model is able to capture both global and subtle changes between images through multi-turn refinement, and can generate high-quality images with fine-grained attributes (e.g., white shoelace) as well as comparative descriptions (e.g., thinner, more open).

5.4. Ablation Study

We conducted an ablation study to validate the effectiveness of two main components in the proposed SeqAttnGAN model: the attention module and the DMS regularizer. We first compare the IS, FID and SSIM scores of SeqAttnGAN with/without attention and DMS. Table 5 shows the ablation results on Zap-Seq and DeepFashion-Seq, indicating that both attention and DMS can improve the model with a large margin. Figure 8 shows some examples generated by SeqAttnGAN without attention and DMS. As can be observed, the ablated systems generate images that are drastically different from previous image, losing contextual consistence, and the textual descriptions are not well reflected in the generated images either. This ablation study validates that the attention module helps improving image-and-text consistency and DMS helps stabilizing the training. Similar observation can been found on the DeepFashion-Seq dataset (details provided in Appendix C).

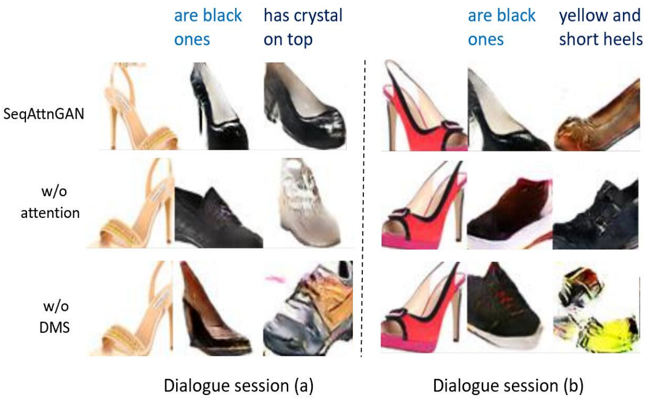


Figure 8. Examples generated by different variations of our model. The first row is from SeqAttnGAN, the second row is from SeqAttnGAN without attention, and the last row is from SeqAttnGAN without DMS.

6. Conclusion

In this paper, we present interactive image editing via dialogue, a novel task that resides in the intersection of computer vision and language. To provide benchmarks for this new task, we introduce two datasets, which contain image sequences accompanied by textual descriptions. To solve this task, we propose the SeqAttnGAN model, which can jointly model user's description and dialogue history to iteratively generate images.

Experimental results demonstrate the effectiveness of SeqAttnGAN. In both quantitative and human evolution, our approach with sequential training outperforms baseline methods that rely on pre-dened attributes or trained in a single-turn paradigm, while offering a more expressive and natural human-machine communication. Particularly, the proposed attention technique can enforce the networks to focus on specific regions of the image, and the DMS function can regularize the model to boost the rendering power.

The results are limited by the current fashion data we adopted. In future work, we plan to build a generic system for other image categories (e.g., face [35]). Currently the framework focuses on individual image synthesis. We would like to explore an approach to generating more consistent image sequences, i.e., disentangling the learned representations into attributes and other factors. In addition, understanding semantic meanings (e.g. "in front", "on side") of user feedback is also important. Finally, we plan to investigate models to support more robust natural language interactions, which requires techniques such as user intent understanding, co-reference resolution, etc.

References

[1] *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017. 2

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2

[3] R. Y. Benmalek, C. Cardie, S. J. Belongie, X. He, and J. Gao. The neural painter: Multi-turn image generation. *CoRR*, abs/1806.06183, 2018. 1, 3

[4] A. Bordes, Y.-L. Boureau, and J. Weston. Learning end-to-end goal-oriented dialog. In *ICLR*, 2017. 2

[5] M. Buhrmester, T. Kwang, and S. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6:3–5, 8 2011. 3

[6] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu. Language-based image editing with recurrent attentive models. *arXiv preprint arXiv:1711.06288*, 2017. 1, 2, 5

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop*, 2014. 4

[8] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, pages 2970–2979. IEEE Computer Society, 2017. 2

[9] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, pages 4466–4475. IEEE Computer Society, 2017. 2

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 4

[11] M. Dixit, R. Kwitt, M. Niethammer, and N. Vasconcelos. Aga: Attribute guided augmentation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jan 2017. 1, 2

[12] J. Gao, M. Galley, and L. Li. Neural approaches to conversational ai. *arXiv preprint arXiv:1809.08267*, 2018. 2

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *NIPS*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[14] X. Guo, H. Wu, Y. Cheng, S. Rennie, and R. S. Feris. Dialog-based interactive image retrieval. *CoRR*, abs/1805.00145, 2018. 3

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015. 4

[16] X. He, L. Deng, and W. Chou. Discriminative learning in sequential pattern recognition. volume 25, pages 14–36. Institute of Electrical and Electronics Engineers, Inc., September 2008. 5

[17] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. *ECCV*, 2016. 2

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[20] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao. Storygan: A sequential conditional gan for story visualization. *arXiv:1812.02784*, 2018. 2

[21] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 3

[22] R. Manuvinakurike, T. Bui, W. Chang, and K. Georgila. Conversational Image Editing: Incremental Intent Identication in a New Dialogue Task. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–295, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[23] R. Manuvinakurike, D. DeVault, and K. Georgila. Using Reinforcement Learning to Model Incrementality in a Fast-Paced Dialogue Game. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Saarbruecken Germany, Aug. 2017. SIGDIAL. 3

[24] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2

[25] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *IJCNLP*, 2017. 2

[26] S. Nam, Y. Kim, and S. J. Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *NIPS*. 2018. 1, 2

[27] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In D. Precup and Y. W. Teh, editors, *ICML*, volume 70, pages 2642–2651, 2017. 2

[28] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In M. F. Balcan and K. Q. Weinberger, editors, *ICML*, volume 48, pages 1060–1069, 2016. 2

[29] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 2

[30] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. 1997. 5

[31] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 2016. 2

[32] R. U. D. L. C. C. L. Shizhan Zhu, Sanja Fidler. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017. 1, 2, 4

[33] Q. H. H. Z. Z. G. X. H. X. H. Tao Xu, Pengchuan Zhang. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 2, 4, 5

[34] H. Wang, J. D. Williams, and S. Kang. Learning to globally edit images with textual description. *CoRR*, abs/1810.05786, 2018. 2

[35] J. Wang, Y. Cheng, and R. S. Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, 2016. 8

[36] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 2

[37] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 4

[38] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 3

[39] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2, 4, 5

[40] B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. *CVPR*, pages 6156–6164, 2017. 3