Sequential Attention GAN for Interactive Image Editing

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
{yu.cheng,zhe.gan,jingjl,jfgao}@microsoft.com,yitong.li@duke.edu

ABSTRACT

Most existing text-to-image synthesis tasks are static single-turn generation, based on pre-defined textual descriptions of images. To explore more practical and interactive real-life applications, we introduce a new task - Interactive Image Editing, where users can guide an agent to edit images via multi-turn textual commands on-the-fly. In each session, the agent takes a natural language description from the user as the input, and modifies the image generated in previous turn to a new design, following the user description. The main challenges in this sequential and interactive image generation task are two-fold: 1) contextual consistency between a generated image and the provided textual description; 2) step-by-step region-level modification to maintain visual consistency across the generated image sequence in each session. To address these challenges, we propose a novel Sequential Attention Generative Adversarial Network (SeqAttnGAN), which applies a neural state tracker to encode the previous image and the textual description in each turn of the sequence, and uses a GAN framework to generate a modified version of the image that is consistent with the preceding images and coherent with the description. To achieve better region-specific refinement, we also introduce a sequential attention mechanism into the model. To benchmark on the new task, we introduce two new datasets, Zap-Seq and DeepFashion-Seq, which contain multi-turn sessions with image-description sequences in the fashion domain. Experiments on both datasets show that the proposed SeqAttnGAN model outperforms state-of-the-art approaches on the interactive image editing task across all evaluation metrics including visual quality, image sequence coherence and text-image consistency.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

Generative Adversarial Network, Sequential Attention, Image Editing with Natural Language

ACM Reference Format:
Yu Cheng[1], Zhe Gan[1], Yitong Li[2], Jingjing Liu[1], Jianfeng Gao[3]. 2020. Sequential Attention GAN for Interactive Image Editing. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414053

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414053

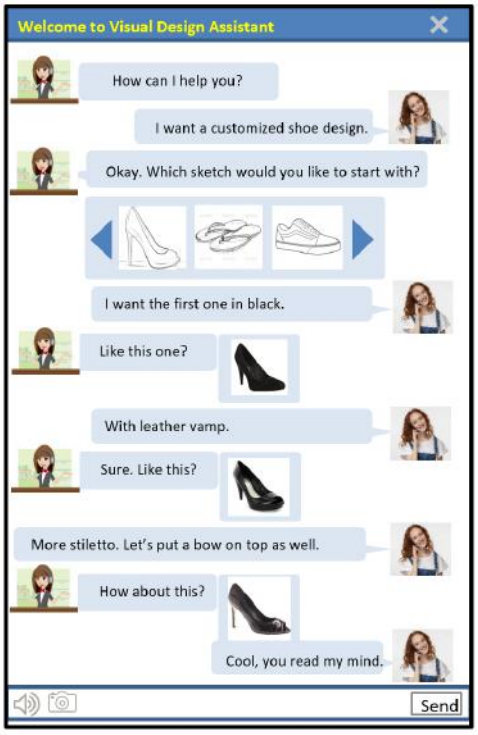

Figure 1: Example of a visual design assistant powered by interactive image editing. In each sequence turn, the user provides natural language feedback to guide the system to modify the design. The system refines the images iteratively based on the user's feedback.

1 INTRODUCTION

In recent years we have witnessed a tremendous growth in visual media, which has intensified users' needs for professional image editing tools (e.g., Adobe Photoshop, Microsoft Photos). However, image/video editing relies heavily on manual effort and is time-consuming, as visual design requires not only expert artistic creativity but also iterative experimentation through trial and error. To automate the process and save human effort, an AI-powered interactive design environment would allow a system to automatically generate new designs by following users' command through a multi-modal interactive session.

To reach the ultimate goal of enabling this creative collaboration between designers and algorithms, we propose a new task to approximate the setting and benchmark models - interactive image editing, where a system can generate new images by engaging with users in an interactive sequential setting. Figure 1 illustrates an interactive image editing system, which supports natural language communication with a user for customizing shape, color, size, texture of a visual

design through conversations. Users can provide feedback on intermediate results, which in turn allows the system to further refine the images. Potential applications of such a system can go beyond visual design and extend to language-guided visual assistance/navigation.

There are some related studies that explored similar tasks. For example, [3, 6, 11, 31, 49] proposed approaches that allow systems to take keyword input (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. Allowing only keywords 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 modify images following textual descriptions, and a neural state tracker to ensure the consistency of sequential context. In each turn, the generator generates a new image by taking into account both the history of previous textual descriptions and previously generated images. To fully preserve the sequential information in the image editing process, the model is trained end-to-end with full sequence sessions. To achieve better fine-grained image quality and coherent region-specific refinement, the model also uses an attention mechanism and a multimodal regularizer based on image-text matching score.

As this is a newly proposed task, we introduce two new datasets, Zap-Seq and DeepFashion-Seq, which were collected via crowdsourcing in a real-world application scenario. In total, there are 8,734 collected sessions in Zap-Seq and 4,820 in DeepFashion-Seq. Each session consists of a sequences of images, with slight variation in design, accompanied by a sequence of sentences describing the difference between each pair of consecutive images. Figure 1 shows an application powered by interactive image editing.

Experiments on these two datasets show that the proposed SeqAttnGAN framework achieves better performance than state-of-the-art techniques. In particular, by incorporating context history, SeqAttnGAN is able to generate high-quality images, beating all baseline models on metrics over 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 taking only keywords or visual attributes as input, which was used in previous approaches. The contributions of our work can be summarized as follows:

- We propose a new task - interactive image editing, which allows an agent to interact with a user for iterative image editing via multi-turn sequential interactions.
- We introduce two new datasets for this task, Zap-Seq and DeepFashion-Seq. Consisting of image sequences paired with free-formed descriptions in diverse vocabularies, the two sets provide new benchmarks for measuring sequential image editing models.
- We propose a new conditional GAN framework, SeqAttnGAN, which can fully utilize context history to synthesize images that conform to users' iterative feedback in a sequential fashion.

2 RELATED WORK

2.1 Image Generation and Editing

Language-based image editing [6, 27] 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 given an editing command. To achieve this, the system is required to have a comprehensive understanding of both natural language information and visual clues. For example, Hu *et al.* [17] focused on language-based image segmentation task, taking phrase as the input. Manuvinakurike *et al.* [27] developed a system using simple language to modify the image, where a classification model is used to understand the user intent.

Since the introduction of GAN [13, 22], there has been growing interest in image generation. In the conditional GAN space, there are studies on generating images from source image [18, 25, 33, 47], sketch [36, 41, 48], scene graph [2, 19], object layout [23, 46], or text (e.g., captions [34], attributes [11], long-paragraph [24]). There is also exploration on how to parameterize the model and training framework [29] beyond the vanilla GAN [32]. Zhang *et al.* [44] stacked several GANs for text-to-image synthesis, with different GANs generating images of different sizes.

AttnGAN [42] introduced 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. FashionGAN [49] aimed at creating new clothing over a human body based on textual descriptions. The text-adaptive GAN [31] 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 applications.

2.2 Sequential Vision Tasks

There are many vision+language tasks that lie in the intersection between computer vision and natural language processing, such as visual question-answering [1], visual-semantic embeddings [39], grounding phrases in image regions [35], and image-grounded conversation [30].

Most approaches have focused on end-to-end neural models based on the encoder-decoder architecture and sequence-to-sequence learning [4, 12, 37]. Specifically, Das *et al.* [8] proposed the visual dialog task, where the agent aims to answer questions about an image in an interactive dialog. Vries *et al.* [9] introduced the GuessWhat?! game, where a series of questions are asked to pinpoint a specific object in an image. However, these dialogue settings are mainly text-based, where visual features only play a complementary role. Manuvinakurike *et al.* [28] investigated building dialog systems that can help users efficiently explore data through visualization. Guo *et al.* [14] introduced an agent that presents candidate images to the user and retrieves new images based on user's feedback. Another piece of related work is Benmalek *et al.* [3] on interactive image generation by encoding dialog history information. Different from these studies, in our work, text information is heavily relied on for guiding the image editing process throughout each image editing session.

Dataset	Zap-Seq	DeepFashion-Seq
# session	8,734	4,820
# turns per session	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.

Compared to recent work on continuous image editing such as ChatPainter [38], our new datasets are designed for multi-turn image generation instead of single-turn. In addition, our data are derived from real fashion images, while CoDraw [20] is based on cartoon images (Abstract Scenes dataset).

3 BENCHMARKS: ZAP-SEQ AND DEEPFASHION-SEQ

The interactive image editing task is defined as follows: in the t-th turn, the system presents a generated image $\hat{x}_t$ to the user, who then provides a textual feedback o_t to describe the change he/she likes to make to realize a target design. The system then takes into account the user's feedback and generates a new image by modifying the previously generated image from the last turn. This process carries on iteratively until the user is satisfied with the result rendered by the system, or the maximum number of editing turns has been reached.

Existing image generation datasets are mostly single-turned, thus not suitable for this sequential editing task. To provide reliable benchmarks for the new task, we introduce two new datasets - Zap-Seq and DeepFashion-Seq, collected through crowdsourcing via Amazon Mechanical Turk (AMT) [5].

The optimal way to construct a sequential dataset for this task is collecting continuous dialog turns via a real interactive system. However, continuous dialog turns are difficult to collect, as the inherent nature of crowdsourcing limits the quality of complex data collection tasks. Real-time interactive interface on crowdsourcing platforms such as AMT is also difficult to control among a large pool of annotators. To circumvent the cumbersome and costly process of collecting pseudo human-machine interactions, we leverage two existing datasets - UT-Zap50K [43] and DeepFashion [26]. UT-Zap50K contains 50,025 shoe images collected from Zappos.com, and DeepFashion contains around 290,000 clothes images from different settings (e.g., store layouts, street snapshots). Each image is accompanied with a list of reference attributes. We first retrieve sequences of images from Zappos and DeepFashion datasets, with each sequence containing 3 to 5 image and every pair of consecutive images slightly different in certain attributes, as candidate sequences from interactive image editing sessions [45]. The attribute manipulation procedure focuses on significant changes (colors, styles) first and finer-grained features (patterns, accessories) later, which is close to human behavior in realistic iterative image editing. As a result, a total of 8,734 image sequences were extracted from UT-Zap50K and 4,820 sequences from DeepFashion as the image sequence pool.

After collecting the image sequences, the second step is to pair them with natural language descriptions that can capture the differences between each image pair, in order to mimic sequential interactive editing sessions. For this, we resort to crowdsourcing via AMT. Specifically, each annotator was asked to provide a free-formed sentence to describe the differences between any two given images. To promote specific and relevant textual descriptions and mimic the real natural language interactive environment, we also provide sentence prefixes for the annotators to select and complete when composing their responses to a retrieved image.

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 "interactive sequences" for the task.

Figure 9 provides some image sequence examples with textual annotations collected from the turkers (more examples are provided in Appendix). To provide robust datasets for benchmarking, we also randomly select a subset of images from the two original datasets as additional sequences to be annotated in AMT, which make up to 10% of the whole datasets.

After manually removing low-quality (e.g., non-descriptive) or duplicate annotations, we obtained a total of 18,497 descriptions for the image sequences from 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 of the description set is 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 Zappos and DeepFashion datasets, these natural language descriptions include fine-grained refinement details on the visual design in each image. More details on 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 Sequential Attention GAN (SeqAttnGAN) model to generate a sequence of images $\hat{x}_1, \ldots, \hat{x}_T$, given an initial input image x_0, and a sequence of natural language descriptions $o_1, \ldots, o_T$. The input image x_0 is encoded into a feature vector $\mathbf{v}_0$ using ResNet-101 [15] pre-trained on ImageNet [10]. Each textual description o_t is encoded via a bi-directional LSTM (BiLSTM), where each word corresponds to two hidden states, one

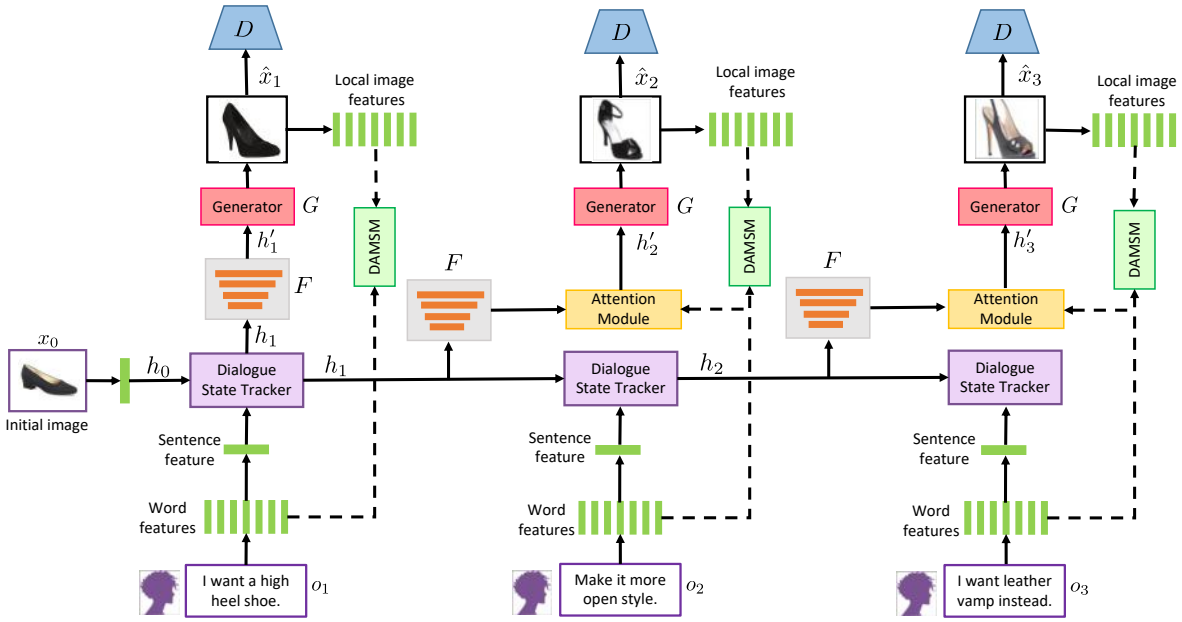

Figure 3: The framework of SeqAttnGAN. The neural state tracker keeps track of the contextual information that has been passed on during the sequential image editing process. The attention module absorbs previous context for refining different sub-regions of the image. The DAMSM regularizer provides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

for each direction. We concatenate its two hidden states to obtain the word feature matrix $\mathbf{e}_t \in \mathbb{R}^{d_e \times L}$, where L is the number of words in a sentence, and the ℓ-th column $\mathbf{e}_t^{(\ell)}$ is the feature vector for the ℓ-th word. Meanwhile, the last hidden states of the BiLSTM are concatenated into a sentence feature vector, denoted as $\bar{\mathbf{e}}_t \in \mathbb{R}^{d_e}$.

As illustrated in Figure 3, in the t-th turn ($t \geq 2$), (i) the Neural State Tracker fuses the sentence feature vector $\bar{\mathbf{e}}_t$ of the current textual description o_t with the hidden state $\mathbf{h}_{t-1}$, to obtain an updated hidden state $\mathbf{h}_t$; (ii) the Attention Module, together with the Up-sampling Module, fuses the word features $\mathbf{e}_t$ of o_t with the feature map that is up-sampled from $\mathbf{h}_{t-1}$, to obtain a context-aware image feature set $\mathbf{h}'_t$; (iii) the Generator generates the current image $\hat{x}_t$ based on $\mathbf{h}'_t$. The following sub-sections introduce each individual component in detail.

4.1 Neural State Tracker

The neural state tracker is modeled as a Recurrent Neural Network (RNN) with the Gated Recurrent Unit (GRU) [7]. The initial hidden state $\mathbf{h}_0 = \text{MLP}(\mathbf{v}_0) \in \mathbb{R}^{d_h}$, where MLP$(\cdot)$ denotes a one-layer MLP layer. In the t-th step, the hidden state $\mathbf{h}_t$ is updated via:

$$\mathbf{h}_t = \text{GRU}(\mathbf{h}_{t-1}, \bar{\mathbf{e}}_t), \tag{1}$$

where $\mathbf{h}_t$ is considered as the context vector that absorbs all the information from the preceding turns.

4.2 Attention Module

The context vector $\mathbf{h}_{t-1} \in \mathbb{R}^{d_h}$ is first upsampled into a feature map $\tilde{\mathbf{h}}_{t-1} \in \mathbb{R}^{d_h \times N}$, where N is the number of sub-regions in an image. This feature map is then combined with the word feature matrix $\mathbf{e}_t \in \mathbb{R}^{d_e \times L}$ of the current textual description o_t via the Attention Module $F_{\text{attn}}(\cdot)$ to obtain $\mathbf{h}'_t \in \mathbb{R}^{d_h \times N}$, which is used for generating image $\hat{x}_t$. Specifically,

$$\mathbf{h}'_t = F_{\text{attn}}(\mathbf{e}_t, F(\mathbf{h}_{t-1})), \quad \hat{x}_t = G(\mathbf{h}'_t, \epsilon_t), \tag{2}$$

where $F(\cdot)$ is the up-sampling module that transforms $\mathbf{h}_{t-1}$ into $\tilde{\mathbf{h}}_{t-1}$, ϵ_t is a noise vector sampled in each step t from a standard normal distribution, and $G(\cdot)$ is the image generator that takes $\mathbf{h}'_t$ and ϵ_t as input.

The attention module F_{attn} is used to perform compositional mapping [40, 42, 49], i.e., enforcing the model to produce regions and associated features that conform to the textual description. Specifically, a word-context vector is computed for each sub-region of the image based on its hidden features $\tilde{\mathbf{h}}_{t-1}$. For the i-th sub-region of the image (i.e., the i-th column of $\tilde{\mathbf{h}}_{t-1}$, denoted as $\tilde{\mathbf{h}}_{t-1}^{(i)}$), its word-context vector $\mathbf{c}_t^{(i)}$ is obtained via:

$$\mathbf{c}_t^{(i)} = \sum_{j=0}^{L-1} \beta_{i,j} \mathbf{e}_t^{(j)}, \text{where } \beta_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{L-1} \exp(s_{i,k})}, \tag{3}$$

where $s_{i,j} = \tilde{\mathbf{h}}_{t-1}^{(i)} \mathbf{e}_t^{(j)}$, and $\beta_{i,j}$ indicates the weight the model attends to the j-th word when generating the i-th sub-region of the image. Finally, the attention module produces a word-context matrix $\mathbf{h}'_t = (\mathbf{c}_t^{(0)}, \mathbf{c}_t^{(1)}, \ldots, \mathbf{c}_t^{(N-1)}) \in \mathbb{R}^{d_h \times N}$, which is passed to the image generator G to generate an image $\hat{x}_t$ in the t-th step.

Compared with AttnGAN [42], our model employs the attention module in a sequence, where all the sequence turns share the same image generator G and discriminator D, while AttnGAN has disjoint generators and discriminators for different scales. Hence, we name our model *Sequential Attention GAN* (SeqAttnGAN). The objective of SeqAttnGAN is defined as the joint conditional-unconditional

losses over the discriminator and the generator. With the supervision of the real image x_t in the t-th turn, the loss of the generator G is defined as:

$$\mathcal{L}_G = -\frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log D(\hat{x}_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log D(\hat{x}_t, \bar{\mathbf{e}}_t)], \quad (4)$$

and the loss of the discriminator D is calculated by:

$$\begin{aligned}\mathcal{L}_D = &-\frac{1}{2}\mathbb{E}_{x_t \sim P_d}[\log D(x_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log(1 - D(\hat{x}_t))] \\ &-\frac{1}{2}\mathbb{E}_{x_t \sim P_d}[\log D(x_t, \bar{\mathbf{e}}_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log(1 - D(\hat{x}_t, \bar{\mathbf{e}}_t)],\end{aligned} \quad (5)$$

where x_t is from the true data distribution P_d and $\hat{x}_t$ is from the model distribution P_G. The above loss is summed over all the sequence turns and paired data samples.

4.3 Deep Multimodal Similarity Regularizer

In addition to the above GAN loss, an image-text matching loss is also introduced into SeqAttnGAN. Specifically, we adopt the Deep Attentional Multimodal Similarity Model (DAMSM) developed in Xu *et al.* [42], which aims to match the similarity between the synthesized images and user input sentences, acting as an effective regularizer to stabilize the training of the image generator and boost model performance.

Given a training sample, which is a sequence of $\{x_0, x_1, o_1, \ldots, x_T, o_T\}$, we first transform it into T image-text pairs. Specifically, for each $t = 1, \ldots, T$, we use x_t as the input image, and then use the concatenation of the image-attribute value of x_{t-1} (provided in the original datasets) and its associated textual description o_t as the paired text $\hat{o}_t$. Note that here we combine image attributes and user textual input as the new text, which is different from [42, 44]. In this way, one training sample is transformed into T image-text pairs $\{x_t, \hat{o}_t\}_{t=1}^{T}$. Following [16, 42], during training, given a mini-batch of M image-text pairs $\{x_i, \hat{o}_i\}_{i=1}^{M}$, the posterior probability of text $\hat{o}_i$ matching image x_i is defined as:

$$P(\hat{o}_i|x_i) = \frac{\exp(\gamma R(x_i, \hat{o}_i))}{\sum_{j=1}^{M} \exp(\gamma R(x_i, \hat{o}_j))}, \quad (6)$$

where γ is a smoothing factor, $R(\cdot,\cdot)$ is the word-level attention-driven image-text matching score (i.e., the attention weights are calculated between the sub-region of an image and each word of its corresponding text. See [42] for details). The loss function for matching the images with their corresponding text is:

$$\mathcal{L}_{\text{DAMSM}}^{i \to t} = -\sum_{i=1}^{M} \log P(\hat{o}_i|x_i). \quad (7)$$

Symmetrically, we can also define the loss function for matching textual descriptions with their corresponding images (by switching $\hat{o}_i$ and x_i). Combining these two, the regularizer is:

$$\mathcal{L}_{\text{DAMSM}} = \mathcal{L}_{\text{DAMSM}}^{i \to t} + \mathcal{L}_{\text{DAMSM}}^{t \to i}. \quad (8)$$

By bringing in the discriminative power of the regularizer, 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. The final objective of the generator G is defined as:

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{\text{DAMSM}}, \quad (9)$$

where λ is the hyperparameter to balance the two loss functions. $\mathcal{L}_G$ and $\mathcal{L}_{\text{DAMSM}}$ are summed over all the sequence turns and data samples.

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
TAGAN	**9.83**	**47.25**	8.26	56.49
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. IS: higher is better; FID: lower is better.

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, human evaluation is also included.

5.1 Datasets and Baselines

All the experiments are performed on the Zap-Seq and DeepFashion-Seq datasets with the same splits: 85% images are used for training, 5% for validation, and the model is evaluated on a held-out test set from the rest 10%. We compare our approach with several baselines: (*i*) **StackGAN** [44] (StackGAN v1 was used due to the relatively low resolution of images in Zap-Seq and DeepFashion-Seq); (*ii*) **AttnGAN** [42]; and (*iii*) **TAGAN** [31]. For the three baselines, the hyper-parameter settings and training details remain the same as in the original paper.

For training, data augmentation is used 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 [21] for training. The mini-batch size M is set to 50. λ in Eqn. (9) is set to 2 on Zap-Seq and 2.5 on DeepFashion-Seq, respectively. d_e and d_h are set to 300 and 128, respectively. The setting of γ follows [42]. DAMSM is used only during training. Baseline model training follows standard conditional-GAN training procedure.

5.2 Quantitative Evaluation

In this section, we provide quantitative evaluation and analysis. For each sequence turn in the test set, we randomly sampled one image from each model, then calculated Inception Score (IS) and Frechet Inception Distance (FID) scores by comparing each selected sample with the ground-truth image. The averaged numbers are presented in Table 2. On Zap-Seq, SeqAttnGAN performs slightly worse than TAGAN, while on DeepFashion-Seq, our model achieves the best performances.

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 show that the images yielded by our model are more consistent with the ground-truth than all the baselines. This indicates that our proposed model can generate images with higher contextual coherency.

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 four approaches: SeqAttnGAN, TAGAN, AttnGAN and StackGAN. To save space, we only display key phrases of each description.

Dataset	StackGAN	AttnGAN	TAGAN	SeqAttnGAN
Zap-Seq	0.437	0.527	0.512	**0.651**
DF-Seq	0.316	0.405	0.428	**0.498**

Table 3: Comparison of SSIM score between our model and the baselines. DF-Seq is short for DeepFashion-Seq.

Model	Zap-Seq		DeepFashion-Seq	
	Vis.	Rel.	Vis.	Rel.
StackGAN	3.34	3.26	3.24	3.19
AttnGAN	2.69	2.54	2.75	2.62
TAGAN	2.14	2.48	2.43	2.52
SeqAttnGAN	**1.97**	**1.83**	**1.70**	**1.69**

Table 4: Results from the user study in terms of both visual quality (Vis.) and context relevance (Rel.). A lower number indicates a higher rank.

Figure 4 and Figure 5 present a few examples comparing all the models with the ground-truth (more examples are provided in Appendix). In each example, it is observable that our model generates images more consistent with the ground-truth images as well as the reference descriptions than the baselines. Specifically, SeqAttnGAN can generate: (*i*) better regional changes (e.g., session (a) in Figure 4, session (c) in Figure 5); and (*ii*) more consistent global changes in color, texture, etc. (e.g., 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. Both AttnGAN and TAGAN are 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 on either visual quality or image sequence consistency (i.e., the generated image has drastic design change from the previous image). This observation is consistent with the quantitative study.

5.3 Human Evaluation

We also conduct a human evaluation comparing our model with baselines 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 model from which each image is generated from is hidden from the annotators. The participants were asked to rank the quality of the generated image sequences based on two aspects independently: (*i*) consistency to the given description and the previous image, and (*ii*) visual quality and naturalness.

Table 4 provides the ranking comparison between SeqAttnGAN and the other methods. For each approach, we computed the average ranking (1 is the best and 4 is the worst). The standard deviation is small, and omitted due to space limit. Results show that our approach achieves the best rank on all dimensions, indicating our proposed method 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, category). 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

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

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 DAMSM	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.

	Single phrase	Composition of phrases	Propositional phrases
IS	10.13	9.82	9.41
FID	46.59	51.02	52.78

Table 6: Ablation study on using different types of phrase on Zap-Seq, measured by IS and FID.

Figure 6: 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 DAMSM.

descriptions (e.g., “thinner",“more open"). Overall, these results show potential of applying the proposed SeqAttnGAN model to real-world applications.

5.4 Ablation Study

We conduct ablation study to validate the effectiveness of the two main components in the proposed SeqAttnGAN model: the attention module and the DAMSM regularizer. We first compare the IS, FID and SSIM scores of SeqAttnGAN with/without attention and DAMSM. Table 5 shows that both attention and DAMSM can improve the model performance with a margin. Figure 6 provides some examples generated by SeqAttnGAN without attention and DAMSM. As shown in these examples, the ablated models generate images that are drastically different from the previous image, losing contextual consistency, and the textual descriptions are not well reflected in the generated images either.

Figure 8 provides an illustration of the sequential attention process, where in each step, the targeted region corresponding to the attribute change is attended. This demonstrates that the attention module can help improve image-text consistency. Similar observations can been seen on DeepFashion-Seq as well.

We also provide additional quantitative analysis on the Zap-Seq dataset, analysing three types of feedback phrases grouped by text structure (single, composition and propositional phrases). Table 6 with IS and FID scores on each type shows that single phrases are the easiest to handle, and our model also achieves good scores on the composition of phrases as well as propositional phrases.

6 DATA COLLECTION AND ANALYSIS

In the section, we explain the details on how we collected Zap-Seq and DeepFashion-Seq datasets and provide insights on the dataset

Single Phrase (43%)	Composition of Phrases (56%)	Propositional Phrases (35%)
is brown	is has a buckle and is darker brown	has a strap that goes around the ankle
has a flatter sole	is closed in and leather	has two more straps across the foot, and slingback strap
has a chevron pattern	has laces and animal print	is a wedge with a strap toe
has a reddish lining	has a tall wedge heel, a peep toe, no thong, and a heel strap	has a strap that goes over the foot instead of in the middle

Table 7: Examples of captions. More than half of the captions contain composite feedback on more than one types of visual feature.

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.

Figure 8: An example on the visualization of attended image regions in each step.

properties. we want the collected descriptions to be concise and relevant for retrieval and avoid casual and non-informative phrases. To this end, as shown in Figure 9, we design the data collection interface to ask annotators to construct the feedback description. The collection procedure can achieve a balance between lexical flexibility and avoiding irrelevant phrases. After manual data cleaning, we are able to get 18,497 descriptions for Zap-Seq and 12,765 for DeepFashion-Seq.

Figure 10 shows the length distribution of the collected captions. Most captions are very concise (3-8 words). In Table 7, we provide some examples in different types. We can see many feedback expressions consist of compositions of multiple phrases, including spatial or structural details. Understanding the semantic meaning of the captions is one of the most challenging part for this task.

Figure 9: AMT annotation interface. Annotators need to complete the rest of the response message given the scenario.

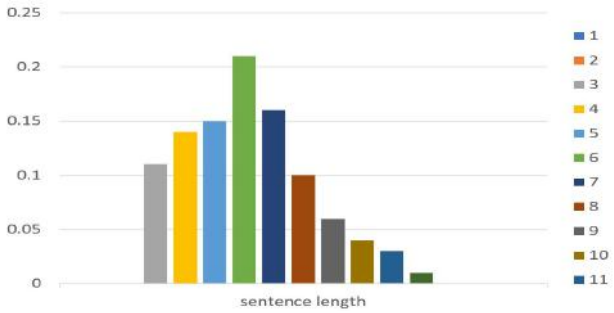

Figure 10: Length distribution of the caption data.

7 CONCLUSION

We present interactive image editing, a novel task that resides in the intersection of computer vision and language. To provide benchmarks, we introduce two new datasets, Zap-Seq and DeepFashion-Seq, which contain image sequences accompanied by textual descriptions. A SeqAttnGAN model is proposed for this task. Experiments on the two datasets demonstrate that SeqAttnGAN outperforms baseline methods across visual quality, image-text relevance and image sequence consistency. For future work, we plan to apply the proposed model to other image types and explore how to generate more consistent image sequences by disentangling learned representations into attributes and other factors. In addition, understanding semantic meanings (e.g., “in front", “on side") of user feedbacks is also important. We also plan to investigate models to support more robust natural language interactions, which requires techniques such as user intent understanding and co-reference resolution.

REFERENCES

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

[2] O. Ashual and L. Wolf. Specifying object attributes and relations in interactive scene generation. In *ICCV*, 2019.

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

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

[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*, 2011.

[6] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu. Language-based image editing with recurrent attentive models. In *CVPR*, 2018.

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

[8] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *CVPR*, 2017.

[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*, 2017.

[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.

[11] M. Dixit, R. Kwitt, M. Niethammer, and N. Vasconcelos. Aga: Attribute guided augmentation. In *CVPR*, 2017.

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

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[14] X. Guo, H. Wu, Y. Cheng, S. Rennie, and R. S. Feris. Dialog-based interactive image retrieval. In *NeurIPS*, 2018.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] X. He, L. Deng, and W. Chou. Discriminative learning in sequential pattern recognition. *IEEE Signal Processing Magazine*, 2008.

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

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

[19] J. Johnson, A. Gupta, and L. Fei-Fei. Image generation from scene graphs. In *CVPR*, 2018.

[20] J. Kim, D. Parikh, D. Batra, B. Zhang, and Y. Tian. Codraw: Visual dialog for collaborative drawing. *CoRR*, abs/1712.05558, 2017.

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. Mmd gan: Towards deeper understanding of moment matching network. In *NeurIPS*, pages 2203–2213, 2017.

[23] Y. Li, Y. Cheng, Z. Gan, L. Yu, L. Wang, and J. Liu. Bachgan: High-resolution image synthesis from salient object layout. *arXiv preprint arXiv:2003.11690*, 2020.

[24] 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. In *CVPR*, 2019.

[25] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017.

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

[27] R. Manuvinakurike, T. Bui, W. Chang, and K. Georgila. Conversational Image Editing: Incremental Intent IdentiïňĄcation in a New Dialogue Task. In *SIGDIAL*, Melbourne, Australia, 2018.

[28] R. Manuvinakurike, D. DeVault, and K. Georgila. Using Reinforcement Learning to Model Incrementality in a Fast-Paced Dialogue Game. In *SIGDIAL*, Saarbruecken Germany, 2017.

[29] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[30] 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.

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

[32] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, 2017.

[33] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[34] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

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

[36] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*, 2017.

[37] 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.

[38] S. Sharma, D. Suhubdy, V. Michalski, S. E. Kahou, and Y. Bengio. Chatpainter: Improving text to image generation using dialogue. *CoRR*, abs/1802.08216, 2018.

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

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

[41] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays. Texturegan: Controlling deep image synthesis with texture patches. *arXiv preprint arXiv:1706.02823*, 2017.

[42] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.

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

[44] 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.

[45] B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *CVPR*, 2017.

[46] B. Zhao, L. Meng, W. Yin, and L. Sigal. Image generation from layout. In *CVPR*, 2019.

[47] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[48] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017.

[49] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017.