

# Seq-SG2SL: Inferring Semantic Layout from Scene Graph through Sequence to Sequence Learning

Boren Li\* Boyu Zhuang Mingyang Li Jian Gu

Alibaba AI Labs

{boren.lbr\*, po-yu.zby, gujian.gj}@alibaba-inc.com

mingyangli009@gmail.com

## Abstract

Generating semantic layout from scene graph is a crucial intermediate task connecting text to image. We present a conceptually simple, flexible and general framework using sequence to sequence (seq-to-seq) learning for this task. The framework, called Seq-SG2SL, derives sequence proxies for the two modality and a Transformer-based seq-to-seq model learns to transduce one into the other. A scene graph is decomposed into a sequence of semantic fragments (SF), one for each relationship. A semantic layout is represented as the consequence from a series of brick-action code segments (BACS), dictating the position and scale of each object bounding box in the layout. Viewing the two building blocks, SF and BACS, as corresponding terms in two different vocabularies, a seq-to-seq model is fittingly used to translate. A new metric, semantic layout evaluation understudy (SLEU), is devised to evaluate the task of semantic layout prediction inspired by BLEU. SLEU defines relationships within a layout as unigrams and looks at the spatial distribution for  $n$ -grams. Unlike the binary precision of BLEU, SLEU allows for some tolerances spatially through thresholding the Jaccard Index and is consequently more adapted to the task. Experimental results on the challenging Visual Genome dataset show improvement over a non-sequential approach based on graph convolution.

## 1. Introduction

Learning the relation from semantic description to its visual incarnation leads to important applications, such as text-to-image synthesis [24] and semantic image retrieval [11]. It remains a challenging and fundamental problem in computer vision [10]. Recent researches have gradually formalized the structured representations of the two modality,

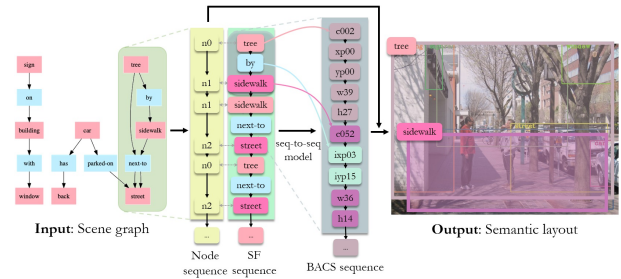


Figure 1: The Seq-SG2SL framework for inferring semantic layout from scene graph.

scene graph [11][13] for semantic description and semantic layout [9][26] for image. Therefore, our goal in this work solves the underlying task, inferring semantic layout from scene graph, for connecting text to image.

Most existing works infer semantic layout from text [7][9][20]. However, leading methods still struggle with complex text inputs depicting multiple objects owing to the unstructured nature of text. Hence, Johnson *et al.* [10] pioneered to infer semantic layout from scene graph as a separate task, isolated from semantic parsing [18]. Scene graph is adopted because it is a powerful structured representation that efficiently conveys scene contents in text [13]. As a notable step forward, they observed that semantic layout is largely constrained by objects within relationships. As such, they developed a graph convolution network to embed a scene graph containing only objects within relationships into respective object feature vectors that further dictate semantic layout through an object layout network. However, they inferred all object feature embeddings simultaneously from a scene graph that comprises exponential variability of object and relationship combinations. It is extremely challenging for a model to express such prohibitive diversity.

We view this task from a novel perspective to avoid combinatorial explosion that largely restricted the model expressiveness in the past. Inferring semantic layout from

\*Corresponding author.

scene graph can be compared to constructing a building from its blueprint. It is unwise to offer corpus of blueprint-to-building correspondences to directly train a learner how to construct a building from its blueprint. Instead, teaching the basic actions to stack the building blocks based on their counterparts in the blueprint is much more feasible. What determines the building blocks? It is the *relationship*.

We propose a conceptually simple, flexible and general framework using sequence to sequence (seq-to-seq) learning to infer semantic layout from scene graph (Figure 1). The framework, called Seq-SG2SL, derives sequence proxies for the two modality and a Transformer-based seq-to-seq model learns to transduce one into the other. A scene graph is decomposed into a sequence of semantic fragments (SF), one for each relationship. A semantic layout is the consequence from a series of brick-action code segments (BACS), dictating the position and scale of each object bounding box in the layout. Viewing the two building blocks, SF and BACS, as corresponding terms in two different vocabularies, a seq-to-seq model is fittingly used to *translate*. Seq-SG2SL is an intuitive framework that learns BACS to drag-and-drop and scale-adjust the two bounding boxes of subject and object in a relationship to the layout supervised by its SF counterpart.

Direct and automated evaluation for semantic layout prediction is another challenging problem unto itself. A new metric, semantic layout evaluation understudy (SLEU), is devised for this purpose inspired by BLEU [15]. SLEU defines relationships within a layout as unigrams and looks at the spatial distribution for  $n$ -grams. Unlike the binary precision of BLEU, SLEU allows for some tolerances spatially through thresholding the Jaccard Index and is consequently more adapted to the task. Mean-SLEU over a large corpus is a proper metric for evaluation.

We experiment on the challenging Visual Genome (VG) dataset [13] that provides human annotated pairs of scene graph and semantic layout for each image. We first show qualitative results from Seq-SG2SL and rationalize SLEU intuitively. Further quantitative comparison shows the advantages of Seq-SG2SL over a non-sequential approach based on graph convolution [10], especially in the aspect of model expressiveness. We show further that this advantage originates from our sequential formulation, not merely from the Transformer model. Various aspects of Seq-SG2SL are studied extensively from additional ablation experiments.

The key contributions are:

- Seq-SG2SL is the first framework to infer semantic layout from scene graph through seq-to-seq learning and outperforms the non-sequential state-of-the-art model by a significant margin.
- SLEU is the first automatic metric to directly evaluate the performance of semantic layout prediction, allowing results reproducibility.

## 2. Related Works

**Scene Graph:** A scene graph is a directed graph representing a scene, where nodes are objects and edges give relationships between objects. Johnson *et al.* [11] first introduced the notion of scene graph as query input for semantic image retrieval. They predicted the most likely semantic layout from a query scene graph as intermediate result for the end retrieval task through a conditional random field (CRF) model. Meanwhile, Schuster *et al.* [18] complemented their work by introducing an automatic approach to create a scene graph from unstructured natural language scene descriptions. Having these brilliant pioneering attempts using scene graph, Krishna *et al.* [13] constructed the VG dataset that aimed to bridge language and vision using dense image annotations. Scene graph and semantic layout were adopted as intermediate representations for the two modality. With the advent of VG, scene graph further shows its value from successive researches, such as in predicting grounded scene graph for images [14][23], in evaluating image captioning [1], and in image generation [10].

**Semantic Description to Semantic Layout:** Semantic layout was first officially given as spatial distribution of clip arts in abstract scene proposed by Zitnick *et al.* [27]. This representation initially aimed to study directly for inferring high-level semantics from images. By contrast, Zitnick *et al.* [28] formulated the opposite problem of predicting an abstract scene from its textual description and proposed a solution using CRF. As clip arts in abstract scene can be easily generalized to object bounding boxes in semantic layout, this concept extends to real images [20].

Predicting a semantic layout from text is usually posed as an intermediate step for complex image generation [7][9]. A complex image refers to the one containing multiple interactive objects. Unlike the family of methods [16][24] that can give stunning results on limited domains, such as fine-grained descriptions of birds or flowers, a semantic layout is usually necessary for complex image generation to dictate multi-object spatial distribution depicted in a text. Johnson *et al.* [10] pioneered to structure text into scene graph for further complex image generation. This work is the closest to ours. We adopt the same idea but only focus on its sub-task of semantic layout prediction from scene graph. The rest of the task, image generation from layout, can be separately solved as [26]. In contrast to the closest work that proposed a non-sequential approach based on graph convolution, Seq-SG2SL views the task from a novel perspective and formulates the problem in a seq-to-seq manner.

During quantitative studies, all existing works did not perform direct evaluations on semantic layout prediction. Instead, they applied *indirect metrics*, such as the inception score or the image captioning score derived from the generated image. Though human evaluations are all incorporated for further evaluation, these results are very expensive to

be *reproduced*, which logjam the way for fruitful research ideas in this field. The circumstance is exceptionally similar to that before BLEU [15] was first introduced to the field of machine translation.

**Sequence to Sequence Learning:** RNN, LSTM [8] and GRU [5] are firmly established for sequence modelling and transduction problems, such as language modelling and machine translation [4][19]. Attention mechanisms further become integral parts of compelling sequence modelling and the transduction models, allowing modelling dependencies without regarding to the distance in the input or output sentences [2]. Vaswani *et al.* [21] popularized the Transformer architecture that can boost the performance of machine translation by a significant margin. Larger-scale architecture exploration specifically for machine translation was conducted using Transformer to further converge towards the optimal settings [3]. Similar conclusion was also drawn by [12]. Seq-to-seq learning is still evolving rapidly. Beneficial experiences can be borrowed for our purpose.

### 3. Seq-SG2SL

Seq-SG2SL is conceptually simple: it predicts a series of actions to form a consequent layout from a corresponding sequence of symbolic triplets derived from relationships in a scene graph. Next, we introduce the design of sequence proxies for the two modality that is key in our work.

#### 3.1. Sequence Proxies

A scene graph encodes objects, attributes and relationships, while its resulting layout is constrained merely by the objects within *relationships*. Therefore, a scene graph is first preprocessed to drop all attributes and independent objects not within any relationship.

A relationship in scene graph is represented by a symbolic triplet: subject, predicate, object. The sequence proxy for a triplet is a tuple, named SF. It is a successive concatenation of the three elements. The preprocessed scene graph can be then decomposed into a sequence of SF, one for each relationship. To fully preserve information in the scene graph, a node sequence, consisting of corresponding object node IDs, is additionally maintained. Note that our Seq-SG2SL framework is flexible to transfer object attributes from a scene graph to a layout through this sequence.

The visual incarnation of an SF in semantic layout contains a pair of object bounding boxes, each for the subject and object. They are referred to as *visual subject* and *visual object*. Its sequence proxy is a BACS. The design requirements are threefold: first, the series of BACS must *uniquely* determine a layout; second, a BACS should correspond to an SF such that the direction of *causality* is clear; third, vocabularies in BACS must be *representable* and *repeatable* such that they can concisely represent any layout.

Type	Functionality
c	set class index of bbox
xp	set xmin of subject bbox
yp	set ymin of subject bbox
ixp	increase xmin of object bbox from subject
ixn	decrease xmin of object bbox from subject
iyp	increase ymin of object bbox from subject
iy n	decrease ymin object bbox from subject
w	set width of bbox
h	set height of bbox
imgar	set aspect ratio index of semantic layout

Table 1: The BACS types and functionalities. The minimum values for x and y are respectively denoted as xmin and ymin. The bounding box is abbreviated as bbox.

A semantic layout, space-quantized into  $H \times W$  square grids, is called a quantized layout where all BACS are defined. It requires 5 types of actions to form an object bounding box in the layout: four to specify the location and scale, and the other to set its class index. The bounding box location for the subject is represented in absolute coordinates, whereas that for the object uses the *relative position* to the subject. This relativity aims to encode the *visual predicate* in a relationship. We show by an experiment that this relative position encoding is key for good performance.

The BACS types and functionalities are formalized in Table 1. A BACS is composed of 10 consecutive words and corresponds to a 3-word SF. Types for the 10 words are sequentially: c, xp, yp, w, h, c, ixp(n), iyp(n), w, h. The first and last 5 words in a BACS form the visual subject and object, respectively. A BACS sequence is the direct concatenation of individual BACS with identical relationship order to its corresponding SF sequence. Optionally, imgar is added to the front of a BACS sequence in case the aspect ratio of semantic layout is of interest.

#### 3.2. Sequence-to-Sequence Model

Given an SF sequence and its corresponding BACS sequence, a seq-to-seq model is fittingly used to translate. We adopt the most recent Transformer model with 6 stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, exactly identical to the one in [21]. This model is used because of its superior performance in machine translation whose formulation is identical to ours. We show by an experiment that the advantage of model expressiveness by Seq-SG2SL originates from our sequential formulation, not merely from Transformer.

#### 3.3. Semantic Layout Restoration

Having predicted BACS sequence from its input SF sequence, the *alignment* is first verified by checking brick action types for each word sequentially. If aligned, the BACS

sequence corresponds to both the input SF and node sequences. The predicted brick actions are then successively executed to form the restored layout.

Note that with the subsidiary node sequence, it is sufficient to derive which bounding boxes in the restored layout should be merged to one as given in scene graph. If the bounding boxes to be merged have identical predicted class index, the merged bounding box is computed simply as their mean value. Otherwise, it picks the one with median bounding box area. In fact, the predicted class indices for these bounding boxes are rarely distinct. More careful merging strategy is left for future investigation.

### 3.4. Implementation Details

**Semantic Layout Encoding:** The maximum side length of a quantized layout is set at 40. Larger values lead to more BACS vocabularies, making seq-to-seq model harder to generalize. Smaller values, however, result in imprecise bounding box localization and scaling. The chosen value is a trade-off. The aspect ratio of semantic layout is also uniformly quantized. The quantization interval and minimum value are 0.05 and 0.5, respectively.

**Data Augmentation:** If a scene-graph corresponding layout is valid, its subgraph counterpart, containing a subset of the relationships, is still acceptable. We augment a scene graph to more pairs of sequences in the two modality by applying this property. The concatenation order of relationships in the pairs of sequences is arbitrary, leading to another freedom for augmentation. To balance training data, each scene graph is augmented to at most 50 correspondences. To limit the maximum length for both sequences, we only preserve at most 9 relationships in a scene graph.

**Training:** We set hyper-parameters exactly as the Transformer work [21], except the warm-up steps is set at 8000. We use the Transformer implementation by OpenNMT [12] for our purpose. We train on a single Tesla P100 GPU for one million iterations (2 days) under batch size of 1024.

**Inference:** We use beam search with beam size of 4 and length penalty  $\alpha = 0.6$  [22]. The inference time is about 132ms on a single Tesla P100 GPU.

## 4. SLEU Metric

Our goal is to design an *automatic* metric to *directly* quantify the success of semantic layout prediction from scene graph. The premise behind automatic evaluation is: the closer a prediction is to human-prepared references, the better it is. Thus, the problem becomes: how to design a metric that can measure the *similarity* between a predicted layout against a set of references.

### 4.1. Notation

Let  $\mathcal{L} = \{\langle r \rangle_k\}_{k \in [1, K]}$  denote a layout with  $K$  relationships, where  $\langle r \rangle_k = (\mathbf{s}_k, \mathbf{o}_k)$  represents a *visual relation-*

*ship*.  $\mathbf{s}_k$  and  $\mathbf{o}_k$  respectively denote visual subject and object, where  $\mathbf{s}_k = [c_k^{[s]}, \mathbf{b}_k^{[s]}]$  and  $\mathbf{o}_k = [c_k^{[o]}, \mathbf{b}_k^{[o]}]$ .  $c_k$  denotes the class index.  $\mathbf{b}_k = [x_k, y_k, w_k, h_k]$  dictates the bounding box.

Our objective is to evaluate how *close* a predicted layout  $\hat{\mathcal{L}}$  is to a collection of reference layouts  $\{\mathcal{L}_j\}_{j \in [1, M]}$ .  $(\hat{\cdot})$  represents predicted values later on.

### 4.2. Metric Design

SLEU is devised inspired by BLEU [15] in machine translation. The cornerstone for BLEU is  $n$ -gram that refers to a contiguous sequence of  $n$  items from a text. Here, the item is word. Evaluating only  $n$ -grams for machine translation is justified under the Markov assumption that the appearance probability for the current word is merely determined by its previous  $(n - 1)$  words, independent from the  $n^{\text{th}}$  one. The  $n$ -gram concept is generalized for SLEU, where the item is *relationship*, instead of *word*. By analogy to BLEU, evaluating  $n$ -grams in a semantic layout assumes the placement for a visual relationship depends only on a maximum of  $(n - 1)$  other relationships.

SLEU evaluates a semantic layout from two perspectives: intra-relationship adequacy as *unigram accuracy*; and inter-relationship fidelity as *n-gram accuracy*. These accuracies are finally combined to a single-number metric.

#### 4.2.1 Unigram Accuracy

---

**Algorithm 1** Unigram accuracy against a single reference

---

**Input:** A predicted layout  $\hat{\mathcal{L}}$  and a reference  $\mathcal{L}$

**Output:** Unigram accuracy  $p_1$

```

1: function  $f_1(\hat{\mathcal{L}}, \mathcal{L})$ :
2:   set  $c = 0$  ▷  $c$ : count of matches
3:   for each  $\langle r \rangle_k$  in  $\mathcal{L}$  do ▷  $K$  iterations
4:     if  $\hat{c}_k^{[s]} = c_k^{[s]}$  and  $\hat{c}_k^{[o]} = c_k^{[o]}$  then
5:       compute  $\tilde{\mathbf{t}}_k = [x_k^{[s]} - \hat{x}_k^{[s]}, y_k^{[s]} - \hat{y}_k^{[s]}]$ 
6:        $\hat{\mathbf{b}}_k^{[s]} \leftarrow \hat{\mathbf{b}}_k^{[s]} \oplus \tilde{\mathbf{t}}_k$ 
7:        $\hat{\mathbf{b}}_k^{[o]} \leftarrow \hat{\mathbf{b}}_k^{[o]} \oplus \tilde{\mathbf{t}}_k$ 
8:       compute  $J_k^{[s]} = IoU(\hat{\mathbf{b}}_k^{[s]}, \mathbf{b}_k^{[s]})$ 
9:       compute  $J_k^{[o]} = IoU(\hat{\mathbf{b}}_k^{[o]}, \mathbf{b}_k^{[o]})$ 
10:      if  $J_k^{[o]} \geq T_{IoU}$  and  $J_k^{[s]} \geq T_{IoU}$  then
11:         $c \leftarrow c + 1$ 
12:      end if
13:    end if
14:  end for
15:  return  $p_1 = c/K$ 
16: end function

```

---

The unigram accuracy  $p_1$  quantifies matching of individual relationships in a predicted layout to a reference, as

shown in Algorithm 1. It simply compares the pairs of visual relationships and counts the number of matches.  $p_1$  is then the count of the matched-pair divided by the total number of relationships.

To compare a pair of visual relationships, one first needs to compute a shift vector  $\tilde{\mathbf{t}}_k$  that aligns the center of  $\hat{\mathbf{b}}_k^{[s]}$  to  $\mathbf{b}_k^{[s]}$ . Then  $\hat{\mathbf{b}}_k^{[s]}$  and  $\hat{\mathbf{b}}_k^{[o]}$  are center-shifted by this vector, where  $\oplus$  denotes this operation. The two Jaccard Indices for the two pairs of bounding boxes are then computed. These Jaccard Indices, each for visual subject and object, give rise to the binary decision of matching through thresholding by  $T_{IoU}$  that allows some tolerances spatially.

#### 4.2.2 $n$ -gram Accuracy

---

**Algorithm 2**  $n$ -gram accuracy against a single reference

---

**Input:** A predicted layout  $\hat{\mathcal{L}}$  and a reference  $\mathcal{L}$

**Output:**  $n$ -gram accuracy  $p_n$

```

1: function  $f_n(\hat{\mathcal{L}}, \mathcal{L})$ :
2:   set  $c = 0$  ▷  $c$ : count of matches
3:   compute  $\mathbb{P}_n(K)$ 
4:   for each  $\mathcal{P}_n$  in  $\mathbb{P}_n$  do ▷  $|\mathbb{P}_n|$  iterations
5:     compute  $\tilde{\mathbf{t}}_q$ 
6:     set  $f = 1$  ▷  $f$ : flag of match
7:     for each  $\langle r \rangle_i$  in  $\mathcal{P}_n$  do ▷  $n$  iterations
8:        $\hat{\mathbf{b}}_i^{[s]} \leftarrow \hat{\mathbf{b}}_i^{[s]} \oplus \tilde{\mathbf{t}}_q$ 
9:       compute  $J_i^{[s]} = IoU(\hat{\mathbf{b}}_i^{[s]}, \mathbf{b}_i^{[s]})$ 
10:      if  $J_i^{[s]} < T_{IoU}$  or  $\hat{c}_i^{[s]} \neq c_i^{[s]}$  then
11:         $f = 0$ 
12:      break
13:    end if
14:  end for
15:  if  $f = 1$  then
16:     $c \leftarrow c + 1$ 
17:  end if
18: end for
19: return  $p_n = c/|\mathbb{P}_n|$ 
20: end function

```

---

The  $n$ -gram accuracy  $p_n$  ( $n \in [2, N]$ ) quantifies similarity of spatial distributions with  $n$  visual relationships between the predicted and reference layout, as elaborated in Algorithm 2.  $\mathbb{P}_n(K) = \{\mathcal{P}_n\}$  denotes a collection that consists of all  $n$ -relationship subsets of  $\mathcal{L}$ .  $|\cdot|$  represents the cardinality of a set. To compute  $p_n$ , it first compares each pair of the  $n$ -relationship subsets and counts the number of matches.  $p_n$  is then the matched proportion of these pairs of  $n$ -relationship subsets.

To measure similarity between a pair of  $n$ -relationship subsets, we only use their  $n$  visual subjects. This is because the relative distribution from a visual object to its subject

has been encoded in unigram accuracy. To compare two spatial distributions, each with  $n$  visual subjects, a shift vector  $\tilde{\mathbf{t}}_q$  is first computed to align the centroids of  $\{\hat{\mathbf{b}}_i^{[s]}\}_{i \in [1, n]}$  and  $\{\mathbf{b}_i^{[s]}\}_{i \in [1, n]}$ . All elements in  $\{\hat{\mathbf{b}}_i^{[s]}\}_{i \in [1, n]}$  are then center-shifted by  $\tilde{\mathbf{t}}_q$ . If all the shifted bounding boxes pass the Jaccard-Index thresholding and class-alignment check, this pair of  $n$ -relationship subsets is considered match.

#### 4.2.3 SLEU Score

SLEU combines unigram and  $n$ -gram accuracies as a single-number indicator. For a predicted layout and a reference, the unigram and  $n$ -gram accuracies are separately computed. Similar to the observation by BLEU,  $n$ -gram accuracy decays roughly exponentially with  $n$ . Hence, SLEU adopts the same averaging scheme: a weighted average of logarithm accuracies.

SLEU can also measure similarity between a prediction and *multiple* references. One first needs to compare a prediction with each reference, obtaining a combined accuracy each. Then, the highest value, corresponding to the closest reference to the prediction, is simply designated as SLEU.

SLEU is formally defined as

$$SLEU = \max_{j \in [1, M]} \left( e^{\sum_{n=1}^N w_n \ln(p_n^j)} \right), \quad (1)$$

where  $N = 3$  and  $w_n = 1/N$  as *uniform weights*.  $N = 3$  is chosen experimentally to make SLEU more distinguishable since larger  $n$  leads to negligible small  $n$ -gram accuracies.

SLEU ranges from 0 to 1. Few prediction can attain a score of 1 unless it is very similar to one of the references. Involving more reasonable reference layouts corresponding to a scene graph leads to a higher value of SLEU. Thus, one must be cautious to *the number of reference* during making comparisons on evaluations. The quality of SLEU can be enhanced by adding the quantity of reference.

#### 4.2.4 Mean-SLEU Metric

One may still wonder the SLEU capability to measure the prediction performance, especially when only one layout is available in the reference corpus. The design of SLEU strictly follows the idea from the well-known BLEU. We just extend the concept from 1D to 2D, which can lead to the binary decision of *visual relationship matching* and allow some tolerances spatially. Therefore, SLEU is largely justified from its similarity to BLEU.

From the work of BLEU [15], using a single-reference test corpus for evaluation is valid if the corpus size is large and the reference translations are not all from the same translator. In our case, the corpus size is maintained at least several thousands and all the samples are crowd-sourced. Hence, the *mean-SLEU* over a large set can be justifiable for evaluation, though its correlation with human judgement is still desirable for future investigation.

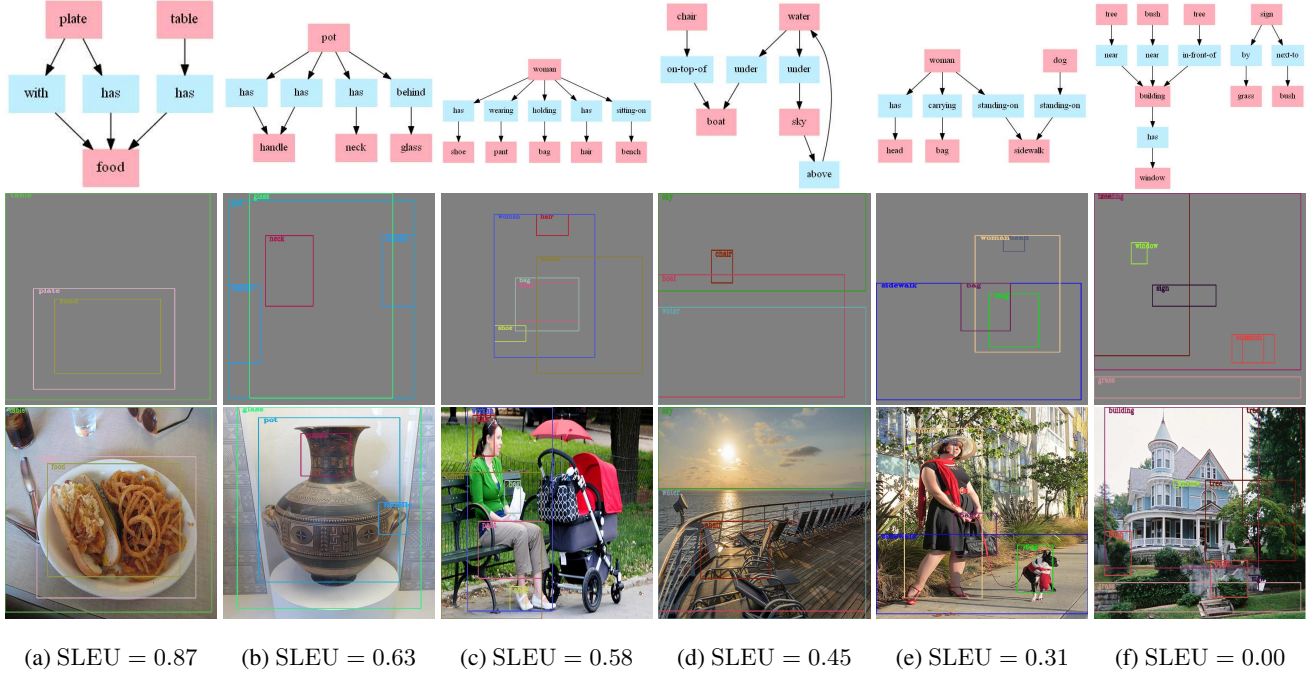


Figure 2: Examples of predicted layouts from Seq-SG2SL on the test set of Visual Genome: the first row is the input scene graph; the second row is the predicted layout; the third row is a reference layout overlaid on its corresponding image.

## 5. Experiments

### 5.1. Dataset

We experiment on the VG dataset [13]. VG comprises 108,077 images, each contains a scene graph and a corresponding layout. To facilitate later comparison with [10], the dataset is organized exactly following its public implementation. We use the same dataset division: 80% training, 10% validation, and 10% test. We use object and relationship classes occurring at least 2000 and 500 times respectively in the training set, leaving 178 object and 45 relationship types. We discard tiny bounding boxes with side length shorter than 32 pixels. We preserve images with between 3 to 30 objects and at least one relationship. This leaves us with 62,565 training, 5,498 validation, and 5,096 test images with an average of ten objects and five relationships per image. Different from [10], we limit the maximum number of relationships in a scene graph to 9 to adapt for our Seq-SG2SL setting. Here, for a scene graph with more relationships, we simply keep the first 9 while discard the rest.

### 5.2. Qualitative Results

This experiment aims to visualize results from Seq-SG2SL and rationalize SLEU intuitively. Figure 2 demonstrates these results from test set and their SLEU scores under  $T_{IoU} = 0.5$ . Qualitatively, Seq-SG2SL is capable to predict a semantic layout from a scene graph containing

multiple interactive objects.

We analyze by examples from Figure 2 to intuitively understand how SLEU can distinguish a good prediction from a bad one. A high SLEU score greater than 0.8, such as (a), definitely indicates a good prediction since the predicted layout is sufficiently close to one of the references. For those with medium SLEU scores between 0.4 to 0.8, as shown in (b)-(d), most predictions are still comparable to a reference. Objects with smaller bounding boxes may appear at apparently irregular positions, such as the *chair* in (d), whereas larger ones are more likely to be reasonable. For cases with low SLEU scores smaller than 0.4, as shown in (e) and (f), the predictions are basically different from the reference. A low SLEU score either suggests a bad prediction or a reasonable prediction that far deviates from any of the given references. Provided more references, SLEU can be more representable, just as the rationale of BLEU.

### 5.3. Quantitative Comparison

This experiment compares Seq-SG2SL with a baseline approach [10]. The pretrained model of the baseline is applied to generate layouts for benchmarking. Our Seq-SG2SL model is trained on exactly the same set as the baseline. Mean-SLEU is employed for evaluation on both the training and test set. Each sample in the two sets consists of a single reference layout corresponding to a scene graph.

Table 2 shows the comparison of mean-SLEU scores under various  $T_{IoU}$  thresholds between the two approaches on



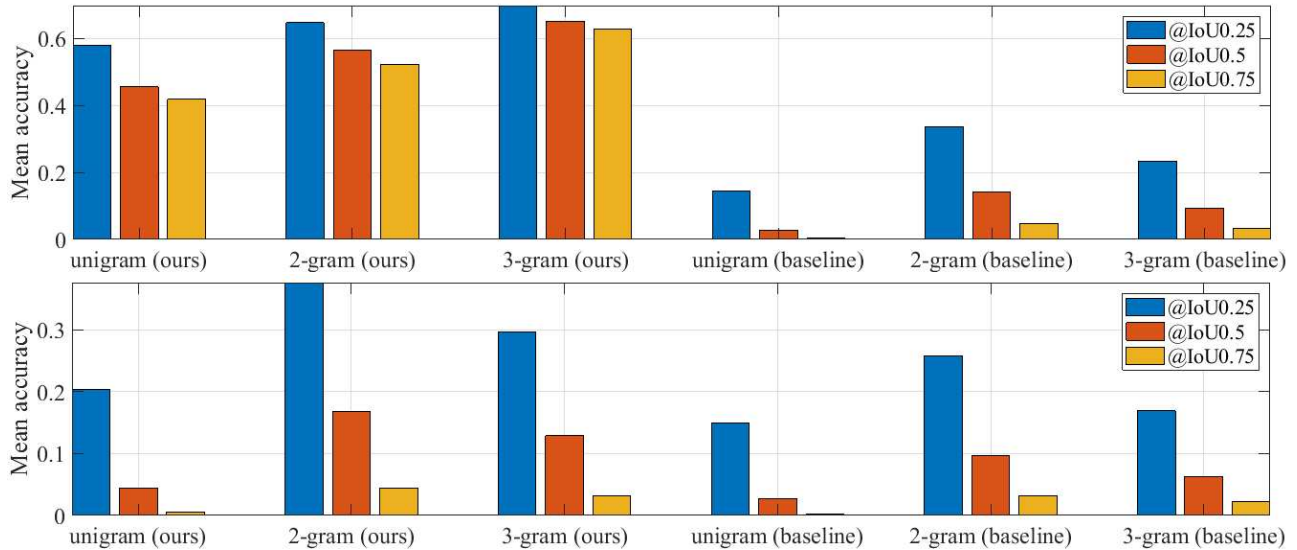


Figure 3: Comparison of mean n-grams accuracies under different  $T_{IoU}$  thresholds between our Seq-SG2SL and the baseline approach on the two sets. The top and bottom figure respectively show results on the training and test set.

	IoU0.0	IoU0.25	IoU0.5	IoU0.75
Baseline-train	1.0000	0.1209	0.0217	0.0036
<b>Ours-train</b>	<b>1.0000</b>	<b>0.5012</b>	<b>0.4174</b>	<b>0.3942</b>
Baseline-test	1.0000	0.1163	0.0205	0.0020
<b>Ours-test</b>	0.9998	<b>0.1665</b>	<b>0.0344</b>	<b>0.0039</b>

Table 2: Comparison of mean-SLEU scores under different  $T_{IoU}$  thresholds between the baseline and our Seq-SG2SL approach on the training and test set.

the training and test set. The larger values, indicating better performances, are highlighted. As demonstrated, Seq-SG2SL outperforms the baseline by a significant margin on the training set. Note that evaluations on the training set are not trivial but offer insight for model expressiveness. By analogy with machine translation, we cannot expect a trained model to produce exactly the same output as a given reference, even on the training set, since the same input inherently results in several different reasonable outputs. But for training, the closer, the better. Therefore, Seq-SG2SL has shown its advantage in model expressiveness over the baseline. This advantage originates from our sequential formulation in avoiding combinatorial explosion. On the test set, Seq-SG2SL still has higher scores than the baseline, suggesting its better generalization capability, except that the score under  $T_{IoU} = 0.0$  is slightly worse. This score only considers the matching of object classes with no evaluation on the spatial distribution of bounding boxes in the layout. This score from Seq-SG2SL is also very close to

1, showing its capability in predicting the correct object classes. In rare cases, the predicted BACS sequence cannot find perfect alignment with the input SF sequence, resulting in such negligible gap.

Figure 3 demonstrates the mean  $n$ -grams accuracies for further analyses, where (a) and (b) respectively show comparisons on the training and test set. As shown, Seq-SG2SL outperforms the baseline in each metric. However, the performance gap during training and testing for Seq-SG2SL is phenomenal, which can be explained from two perspectives. First, only a single reference is given for each scene graph during both training and testing. During training, the model is guided to memorize these particular references and turns out to balance performances on all samples. More supervision implies a higher chance the prediction is closer to its reference. That explains why  $n$ -gram with larger  $n$  has better accuracy on the training set. During testing, however, a given single reference may differ significantly from that during training, resulting in an underestimated score. Thus, adding more references for each test sample is one way to mitigate this training-inference gap. Second, this discrepancy also originates from *exposure bias* for seq-to-seq problem. We refer readers to [17] for more detailed explanations. Recent work [25] may be incorporated in the future to bridge this training-inference gap.

## 5.4. Ablation Experiments

### 5.4.1 Relative vs. Absolute Position Encoding

Seq-SG2SL applies *relative* position encoding. This relativity aims to represent the *visual predicate* in a relation-

	IoU0.0	IoU0.25	IoU0.5	IoU0.75
Abs-train	1.0000	0.3760	0.2698	0.2387
<b>Ours-train</b>	<b>1.0000</b>	<b>0.5012</b>	<b>0.4174</b>	<b>0.3942</b>
Abs-test	0.9996	0.1591	0.0334	0.0038
<b>Ours-test</b>	<b>0.9998</b>	<b>0.1665</b>	<b>0.0344</b>	<b>0.0039</b>

Table 3: Comparison of mean-SLEU scores under different  $T_{IoU}$  thresholds between encoding methods using absolute and relative position on the training and test set.

	IoU0.0	IoU0.25	IoU0.5	IoU0.75
LSTM-train	1.0000	0.4134	0.3822	0.3719
<b>Ours-train</b>	<b>1.0000</b>	<b>0.5012</b>	<b>0.4174</b>	<b>0.3942</b>
LSTM-test	0.9984	0.1012	0.0222	0.0031
<b>Ours-test</b>	<b>0.9998</b>	<b>0.1665</b>	<b>0.0344</b>	<b>0.0039</b>

Table 4: Comparison of mean-SLEU scores under different  $T_{IoU}$  thresholds between our Transformer-based model and an LSTM-based model on the training and test set.

ship. This experiment aims to clarify the significance to use relative position encoding. We compare the results with that from *absolute* position encoding where locations of visual subject and object are represented in absolute coordinates without considering the visual incarnation of predicate. The two models are trained following exactly the same way except the position encoding method. Table 3 shows the results of this comparison. As demonstrated, relative position encoding is much better, suggesting that the training for visual predicate in a relationship is key for good performance.

#### 5.4.2 Network Architecture

This experiment compares results from our Transformer-based model and a 4-layer LSTM-based model with attention mechanism as shown in Table 4. The two models are trained on exactly the same set. As expected, the LSTM-based model performs slightly worse than our Transformer-based model in each metric. However, its performance on the training set is still far better than the non-sequential baseline (See Table 2). Therefore, the advantage of model expressiveness mainly comes from our Seq-SG2SL framework in avoiding combinatorial explosion, not only originates from the superiority of the Transformer architecture.

#### 5.4.3 Greedy vs. Beam Search

This experiment compares results from greedy search and beam search. Table 5 shows the top-4 results from beam search in addition to the greedy decoding baseline. As shown, on the training set, the best mean-SLEU score

	Greedy	Top-1	Top-2	Top-3	Top-4
train	0.3958	<b>0.4174</b>	0.3531	0.3396	0.3357
test	0.0335	0.0344	0.0325	0.0343	<b>0.0367</b>

Table 5: Comparison among results from greedy search and beam search using mean-SLEU score under  $T_{IoU} = 0.5$  from Seq-SG2SL on the training and test set.

is in accordance with the top-1 choice from beam search, which justifies the use of the top-1 prediction. The top-2 to top-4 scores appear almost the same but distinctively lower than the top-1 score. These results may enrich the diversity of semantic layout predictions, though they are more likely to be unreasonable. The score of the top-1 prediction from beam search is higher than that from greedy search as expected, showing the advantage to use beam search. On the test set, however, all scores are generally low and resemble each other. The single-reference test corpus may not be sufficient to conclude with these minor discrepancies.

## 6. Conclusions And Future Works

This paper has presented Seq-SG2SL, a conceptually simple, flexible and general framework for inferring semantic layout from scene graph through seq-to-seq learning. Seq-SG2SL outperforms a non-sequential state-of-the-art approach by a significant margin, especially in the aspect of model expressiveness. This advantage mainly originates from our sequential formulation in avoiding combinatorial explosion. Relative position encoding to represent visual predicate in a relationship is also key for good performance. Beam search during inference can further enhance the quality of prediction compared with greedy search. The top-1 result from beam search is the best layout prediction. SLEU, an automatic metric to directly evaluate semantic layout prediction from scene graph, has been devised in a similar fashion to BLEU. The rationale of SLEU and the justification of using mean-SLEU metric over a large set for evaluation have been discussed.

Direct future works are twofold: first, as for SLEU, its quality can be further enhanced by adding more references. Human evaluations on SLEU may also be of interest. More evaluation metrics on similar tasks, such as text-to-layout synthesis, may be derived by adopting our  $n$ -gram analogy regarding *relationship* as unigram; second, as for Seq-SG2SL, alternative BACS designs may be investigated, such as altering the parameterization of bounding box similar to CenterNet [6]. Additive layout generation may also be investigated due to the sequential nature of our framework.

## References

- [1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image cap-



- tion evaluation. In *ECCV*, pages 382–398, 2016.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, arXiv:1409.0473, 2014.
  - [3] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc V. Le. Massive exploration of neural machine translation architectures. *CoRR*, arXiv:1703.03906, 2017.
  - [4] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, arXiv:1406.1078, 2014.
  - [5] Junyoung Chung, Çağlar Gülçehre and KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, arXiv:1412.3555, 2014.
  - [6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. *CoRR*, arXiv:1904.08189, 2019.
  - [7] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. In *ICLR*, 2019.
  - [8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 8(8):1735–1780, Nov. 1997.
  - [9] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *CVPR*, pages 7986–7994, June 2018.
  - [10] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, pages 1219–1228, June 2018.
  - [11] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678, June 2015.
  - [12] Guillaume Klein, Yoon Kim, Yuntian Deng, Josep Maria Crego, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. *CoRR*, arXiv:1709.03815, 2017.
  - [13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, Feb. 2017.
  - [14] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016.
  - [15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.
  - [16] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, June 2019.
  - [17] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *CoRR*, arXiv:1511.06732, 2015.
  - [18] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *EMNLP*, 2015.
  - [19] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.
  - [20] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *CVPR*, pages 6710–6719, June 2019.
  - [21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.
  - [22] Yonghui Wu, Mike Schuster, and Zhifeng et al. Chen. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, arXiv:1609.08144, 2016.
  - [23] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, July 2017.
  - [24] 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 *ICCV*, Oct. 2017.
  - [25] Wen Zhang, Yang Feng, Fandong Meng, Di You, and Quan Liu. Bridging the gap between training and inference for neural machine translation. In *ACL*, July 2019.
  - [26] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *CVPR*, June 2019.
  - [27] C. L. Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, pages 3009–3016, June 2013.
  - [28] C. L. Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the visual interpretation of sentences. In *ICCV*, pages 1681–1688, Dec. 2013.