A. Illustration of Different Decoders

We present the details of different decoders in Fig. 1. The formalization is provided in the main paper. In the baseline decoder, the context visual feature is calculated by weighting feature vectors from each region of the image and fed into the LSTM in each time step. For simplification, we omit the hidden state initialisation and the attention-based context visual feature calculation in the figure. In the parallel decoder, $N_{\text{topic}}$ baseline decoders are employed as sub-decoders. Apart from the visual feature, it uses the topic label as an additional input, which is responsible for selecting the corresponding decoder for the given topic. Finally, for the conditional decoder, the input topic label is first embedded into a topic embedding. Then, the topic embedding is concatenated to the visual feature and the previous hidden state for each time step. We omit the visual feature concatenation operation in the conditional decoder figure for simplification. After the whole decoding process, the generated masked sentence is fed to a topic classifier to ensure that the sentence belong to the correct topic.

B. Implementation Details

We implement all our models with PyTorch [7]. We optimize the topic decoder with the Adam [6] with a learning rate of $5 \times 10^{-4}$, which decays at a rate of 0.8 every 10 epochs. The batch size is set to 32. We extract $L = 14 \times 14$ with $D = 2,048$ feature maps from the layer before the last pooling layer of a pre-trained ResNet101 [4]. For predicting artistic attributes, we use a four-branch attribute predictor model [3]. The dimensions of the LSTM-based decoder’s hidden states and word embeddings are fixed to 512 for all of the models discussed herein. In the topic conditional decoder, the dimensionality of the topic embedding is set to 20. DrQA and BERT hyperparameters are set as in [1] and [2], respectively. At test time, we employ the beam search for generating text, where a beam size of 5 is empirically selected for all the topic decoder variants.

C. Training Details

For the image encoder, we use a pre-trained ResNet [4] that does not need to be trained. For the decoders, the baseline decoder is trained as the standard captioning model [8], where the whole description is used as ground truth caption for an image. While during training the topic decoder, the ground truth description for an image is split into $N_{\text{topic}}$ parts. Sentences with the same topic label are appended together as a topic-specific description. In the parallel decoder, each sub-decoder is trained independently with its topic-specific description. In training the conditional decoder, the topic-specific description are selected according to the topic label input to the decoder. In the topic classifier part, we employ the continuous approximation technique proposed by Hu et al. [5] to avoid sampling words from a probability distribution, so that the decoder and classifier can be trained in an end-to-end manner. Not all the comments contain the three topics. During training, if a comment does not span the e.g., form topic, the form decoder is not trained with that image.

In the knowledge retrieval part, both attributes prediction model and object detection model are pre-trained. While
the DrQA [1] knowledge retriever adapts a non-machine-learning method. Thus, no optimization is needed in this part. In the knowledge extraction and filling part, BERT is trained with art descriptions. The input is a masked sentence and a list of candidate words, where the masks are generated by replacing the named-entities with their entity type, and candidate words are the named-entities that being replaced. The ground truth is the original sentence before masking. Note that to avoid trivial solutions, the candidate words are extracted from the whole paragraph of description while the input sentence is one short sentence.

D. Knowledge Retrieval Module Evaluation

For evaluating the knowledge retrieval module, we annotate a small number of paintings (150) with their corresponding Wikipedia article. Not all the images possess exact associated Wikipedia article. However, articles related to the painting’s author or theme can also provide useful information. Considering these factors, we first prepare several candidates per painting. The DrQA [1] knowledge retriever adapts a non-machine-learning method. Thus, no optimization is needed in this part. In the knowledge extraction and filling part, BERT is trained with art descriptions. The input is a masked sentence and a list of candidate words, where the masks are generated by replacing the named-entities with their entity type, and candidate words are the named-entities that being replaced. The ground truth is the original sentence before masking. Note that to avoid trivial solutions, the candidate words are extracted from the whole paragraph of description while the input sentence is one short sentence.

The DrQA [1] knowledge retriever adapts a non-machine-learning method. Thus, no optimization is needed in this part. In the knowledge extraction and filling part, BERT is trained with art descriptions. The input is a masked sentence and a list of candidate words, where the masks are generated by replacing the named-entities with their entity type, and candidate words are the named-entities that being replaced. The ground truth is the original sentence before masking. Note that to avoid trivial solutions, the candidate words are extracted from the whole paragraph of description while the input sentence is one short sentence.

We evaluate the accuracy of the knowledge retrieval module by comparing the sorted list of articles from our retriever with the annotated Wikipedia articles, and find the position in which the annotated article is returned. In this way, we measure recall at k (R@k) metric with different values of k (e.g., k = 1, 5, 10). R@k represents the percentage of samples whose annotated article is returned within the top k positions by our retriever. As we have different labels for the annotated articles, we calculate the metrics for the different type of articles.

Table 1 shows the evaluation results using attributes and objects words as query, as in the main paper. We can observe that the useful articles from our retriever mostly come from the author articles. We have also explored to incorporate the generated masked sentences into the query, whose results are shown in Table 2. Comparing the two tables, we find that the incorporation of masked sentences has a negative impact in the knowledge retriever, as these sentences occupy a large proportion in the query but do not contain much specific information.

E. More Qualitative Results

Here we show the generated sentences by all the methods evaluated in the main paper and provide more qualitative results of our proposed method. Figure 2 shows, the qualitative comparison of different methods in Section 4.2. In Figure 3, three more examples of descriptions generated by our method are shown.

Table 1. Knowledge retrieval. Using attributes and objects words as query.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Num. Articles</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Top-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>3.4</td>
</tr>
<tr>
<td>Theme</td>
<td>3</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Author</td>
<td>107</td>
<td>13.7</td>
<td>36.7</td>
<td>46.0</td>
</tr>
<tr>
<td>All articles</td>
<td>150</td>
<td>13.8</td>
<td>36.6</td>
<td>45.5</td>
</tr>
</tbody>
</table>

Table 2. Knowledge retrieval. Using attributes and objects words, as well as generated masked sentences as query.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Num. Articles</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Top-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>3.4</td>
</tr>
<tr>
<td>Theme</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Author</td>
<td>107</td>
<td>3.6</td>
<td>9.5</td>
<td>16.8</td>
</tr>
<tr>
<td>All articles</td>
<td>150</td>
<td>5.0</td>
<td>10.5</td>
<td>17.5</td>
</tr>
</tbody>
</table>

We evaluate the accuracy of the knowledge retrieval module by comparing the sorted list of articles from our retriever with the annotated Wikipedia articles, and find the position in which the annotated article is returned. In this way, we measure recall at k (R@k) metric with different values of k (e.g., k = 1, 5, 10). R@k represents the percentage of samples whose annotated article is returned within the top k positions by our retriever. As we have different labels for the annotated articles, we calculate the metrics for the different type of articles.

Table 1 shows the evaluation results using attributes and objects words as query, as in the main paper. We can observe that the useful articles from our retriever mostly come from the author articles. We have also explored to incorporate the generated masked sentences into the query, whose results are shown in Table 2. Comparing the two tables, we find that the incorporation of masked sentences has a negative impact in the knowledge retriever, as these sentences occupy a large proportion in the query but do not contain much specific information.

E. More Qualitative Results

Here we show the generated sentences by all the methods evaluated in the main paper and provide more qualitative results of our proposed method. Figure 2 shows, the qualitative comparison of different methods in Section 4.2. In Figure 3, three more examples of descriptions generated by our method are shown.

References

This painting is one of a series of four representing the four seasons.

This painting is one of the most famous landscape painters of the Dutch countryside. The estuary with and other figures in the foreground.

This painting is one of a pair of winter landscapes by Jan van de and Jan van .

While in the 1640s most of his paintings were seascapes, van Beyeren began to develop as a skilled still life painter of fish. In the 1650s and 1660s he started to focus on pronkstillevens, i.e. still lifes with fine silverware, Chinese porcelain, glass and selections of fruit. Van Beyeren was likely familiar with the other Dutch painters of pronkstillevens such as Pieter Claesz and Willem Claeszoon Heda who were specialists in monochrome banquet still lives.

This painting depicts a rive landscape with skaters in the foreground. This painting is one of a series of views of the .

The painting depicts a river landscape with skaters and a rowing boat in the foreground. This painting is a typical example of Beyeren’s landscapes that he had to be seen in his own lifetime and he was a good example of his contemporaries. This painting is one of the earliest known works by Beyeren.

This painting depicts a still life still life and signed and dated at lower right.

This painting depicts a still life of flowers in a glass vase with a bee and other insects in. This painting is painted with a variety of colour and flowers in the centre of the composition and the arrangement of the flowers set against the dark background. This painting is one of the most important examples of the artist Arellano's early period.

The painting depicts a young woman in a white dress with a and a woman holding a sword. This painting is a fine example of the artist’s late style and he was influenced by Solomon. This is one of the most important works by Eriksen in Russia.

The painting depicts a Dutch landscape with a shepherd and a horsedrawn cart in the foreground. This painting is a typical example of Martszen’s work in the use of the composition and the use of light as in the foreground. This painting belongs to a series of two representing the ten paintings of the Spanish infantry.

Figure 2. Quantitative comparison with different methods.

Figure 3. More quantitative results produced by our framework.