

# Control Image Captioning Spatially and Temporally

Kun Yan<sup>†\*</sup>, Lei Ji<sup>‡§¶</sup>, Huaishao Luo<sup>||</sup>, Ming Zhou<sup>¶</sup>, Nan Duan<sup>¶</sup>, Shuai Ma<sup>†</sup>

<sup>†</sup>SKLSDE Lab, Beihang University, Beijing, China

<sup>‡</sup>Institute of Computing Technology, CAS, Beijing, China

<sup>§</sup>University of Chinese Academy of Sciences, Beijing, China

<sup>¶</sup>Microsoft Research Asia, Beijing, China

<sup>||</sup>Southwest Jiaotong University, Chengdu, China

<sup>†</sup>{kunyan,mashuai}@buaa.edu.cn <sup>¶</sup>{leiji,mingzhou,nanduan}@microsoft.com

<sup>||</sup>huaishaoluo@gmail.com

## Abstract

Generating image captions with user intention is an emerging need. The recently published Localized Narratives dataset takes mouse traces as another input to the image captioning task, which is an intuitive and efficient way for a user to control what to describe in the image. However, how to effectively employ traces to improve generation quality and controllability is still under exploration. This paper aims to solve this problem by proposing a novel model called LoopCAG, which connects Contrastive constraints and Attention Guidance in a **Loop** manner, engaged explicit spatial and temporal constraints to the generating process. Precisely, each generated sentence is temporally aligned to the corresponding trace sequence through a contrastive learning strategy. Besides, each generated text token is supervised to attend to the correct visual objects under heuristic spatial attention guidance. Comprehensive experimental results demonstrate that our LoopCAG model learns better correspondence among the three modalities(vision, language, and traces) and achieves SOTA performance on trace controlled image captioning task. Moreover, the controllability and explainability of LoopCAG are validated by analyzing spatial and temporal sensitivity during the generation process.

## 1 Introduction

Image captioning is a fundamental task to examine whether an intelligent system can understand the visual world by letting the system describe it with natural language. Generating a reasonable caption requires the model to link linguistic tokens to objects, relationships, scenes of the visual world in the input image. Thus, a great captioning model will help us better understand what characteristics promise a good joint visual-linguistic representation.

\*Contribution during internship at MSRA.



Figure 1: A showcase of Trace Controlled Image Caption. Given an image together with a mouse trace representing user intention, the task is to generate the corresponding captions aligned with each part of the trace. In this case, the trace and the caption marked with the same color correspond to each other.

Most previous attempts aim to describe the image indicating the salient objects and relations without considering user intention. To generate controllable and explainable captions, recent works dedicated to establishing a new *controllable image captioning task* to generate the caption at will. The captioning process can be controlled by POS tagging (Deshpande et al., 2018), sentiment (You et al., 2018), length (Deng et al., 2020), bounding boxes (Cornia et al., 2019), and mouse traces (Pont-Tuset et al., 2020).

In this paper, we mainly investigate trace-controlled image captioning, since it is not only a more natural and interactive paradigm for real web applications, e.g. automatic presentation or help people with visual difficulties but also a new perspective for us to better understand how the long-pursued cross-modality alignment is performed in deep learning models. Figure 1 presents a showcase of the scenario. Given an image, users can easily draw a trace to ask the AI agent to describe the scene in the image along the trace automatically.

In the Localized Narratives dataset (Pont-Tuset et al., 2020), the annotators describe the image

while drawing the traces of their attention movement, which presents a spatial alignment between visual objects and caption tokens as well as a temporal alignment between user intention(by trace) and caption sentences. From Figure 1, we see that the caption tokens, e.g. “person”, “horse”, “trees” can be grounded to the visual objects spatially, and the order of caption sentences can be arranged to align to the order of traces temporally. Although it is easy for humans to recognize which visual object is indicated by the traces, it is a challenge for the agent to recognize, emphasize and arrange visual semantics solely based on several tracepoints’ coordinates. Thereby, we mainly devote our effort to the spatial grounding and temporal controllability of image captioning.

Inspired by the above observation, we design two novel approaches to tackle the above challenges. Specifically, we design sentence-level contrastive constraints to align the generated sentences to the corresponding trace sequences temporally. Besides, we design a type of heuristic spatial attention guidance to supervise each generated text tokens to attend to the correct visual objects. Composing the above together, We propose a novel trace-controlled image captioning model called LoopCAG and demonstrate its superior capability on captioning quality and flexible controllability.

Our contribution can be summarized as:

1) We propose a novel model LoopCAG, which learns the caption tokens’ spatial grounding through attention guidance and temporal localization between trace input and the caption sentences through contrastive constraints in an end-to-end loop manner among the three modalities(vision, language, and traces).

2) The quantitative results show that our LoopCAG model can generate better trace-controlled captions and achieve SOTA performance on automatic criteria. The qualitative results present that our model can generate highly relevant captions given users’ trace inputs.

3) We intensively study the controllability and explainability of trace-controlled image captioning.

## 2 Preliminary

### 2.1 Task Definition

For image captioning, the task is to generate a text description  $\mathbf{y}$  given an image  $\mathbf{I}$ . We first apply a pre-trained visual object detector on the image and get an object level visual feature set

$\mathbf{V} = \{v_1, \dots, v_N\}$ , in which  $v_i \in \mathbb{R}^{2048}$  is the  $i$ -th object visual feature, and  $N$  is the number of visual objects. The text description sequence is  $\mathbf{y} = \{y_1, \dots, y_l\}$ , in which  $y_j$  is the  $j$ -th token and  $l$  is the text sequence length. The output is conditioned on model parameters  $\theta$ , and the optimization process can be formulated as the following maximum likelihood form:

$$\theta^* = \arg \max_{\theta} \log p(\mathbf{y} | \mathbf{V}; \theta). \quad (1)$$

For trace-controlled image captioning, the raw trace input is a sequence of tracepoints coordinates with timestamps. To reduce those tracepoints to an acceptable length due to the limit of GPU memory, we segment the tracepoints sequences uniformly by the same time window  $\tau$ , and then each trace segment is converted to its minimal bounding rectangle. Every bounding rectangle can be represented by a 5D vector which contains normalized coordinates of the top-left and bottom-right corners, and the area ratio with respect to the whole image. We denote the trace input as  $\mathbf{T} = \{t_1, \dots, t_M\}$ , where  $t_i \in \mathbb{R}^5$ . The trace controlled captioning objective can be formulated as follow:

$$\theta^* = \arg \max_{\theta} \log p(\mathbf{y} | \mathbf{V}, \mathbf{T}; \theta) \quad (2)$$

## 3 Method

Our method consists of three components: the caption generation module with a transformer encoder-decoder backbone, the attention guidance for object-level spatial grounding, and the contrastive constraints for sentence-level temporal alignment. The overall model structure is illustrated in Figure 2. The model is trained by jointly optimizing the three objectives listed in the following subsections.

### 3.1 Caption Generation

The caption generation backbone is a transformer-based encoder-decoder proposed by Vaswani et al. (2017), which mainly employs a multi-head attention mechanism and achieves top-tier performance in many sequential related tasks. Here, we highlight several task-oriented modifications.

**Vision-Trace Encoder** The visual embeddings  $\mathbf{V}$  and traces embeddings  $\mathbf{T}$  are encoded separately and then concatenated together as a single input sequence feeding into a transformer encoder.

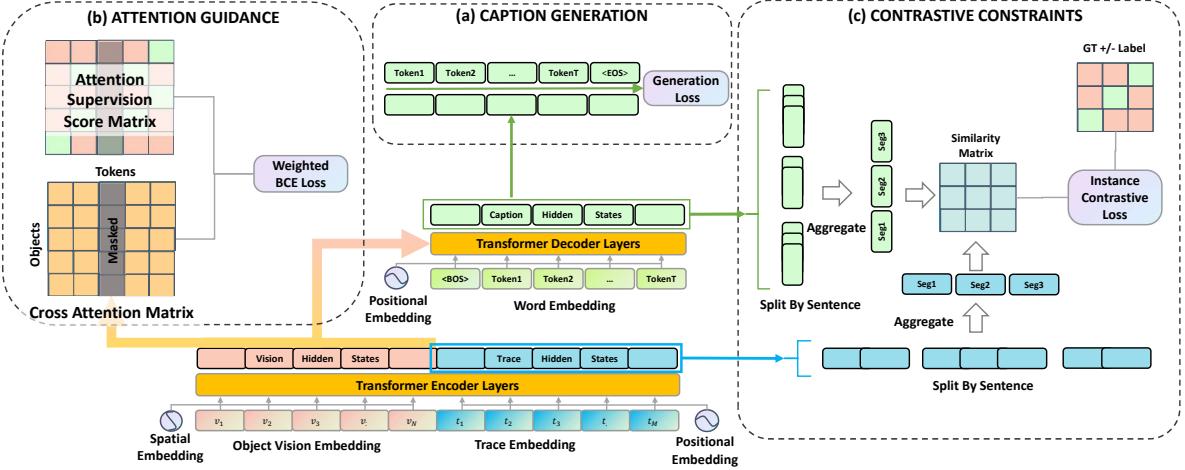


Figure 2: Model Architecture Overview. The model consists of three modules: (a) **Caption Generation**: We directly concatenate the visual object embedding and the trace embedding as encoder input, and then employ a transformer decoder for caption generation. (b) **Attention Guidance**: We use a heuristic supervision attention score matrix to supervise the vision-linguistic cross-attention generated by the transformer backbone, grounding the caption tokens to visual objects spatially. (c) **Contrastive Constraints**: We split the hidden states of caption tokens and traces by sentence respectively and then apply the contrastive loss to make the representations of the sentence and trace segment with same order indices closer, thereby aligning caption sentences to trace segments temporally.

- **Object visual embedding:** We first represent the spatial info of each object proposal by a 5D vector (in the same way as the traces), then project it into a spatial embedding  $p_i \in \mathbb{R}^d$ , where  $d$  is the embedding size across the model. Each object visual feature  $v_i$  is projected into a lower dimension vector  $\hat{v}_i \in \mathbb{R}^d$ . The final visual embedding is  $\tilde{\mathbf{V}} = \{\tilde{v}_1, \dots, \tilde{v}_N\}$ , where  $\tilde{v}_i = \hat{v}_i + p_i$ .
- **Trace Embedding:** Each trace input item  $t_i$  is projected into  $\hat{t}_i \in \mathbb{R}^d$ . We also generate Sinusoidal Positional Embeddings (Vaswani et al., 2017)  $o_i$  to capture the temporal order of the traces. The final trace embedding  $\tilde{\mathbf{T}} = \{\tilde{t}_1, \dots, \tilde{t}_M\}$ , where  $\tilde{t}_i = \hat{t}_i + o_i$ .

**Caption Decoder** Caption decoder combines vision and trace information using cross attention connected to the hidden states of Vision-Trace Encoder’s last layer. Using a casual mask to encode generated token progressively, the transformer decoder ensures that the predictions for position  $i$  can depend only on the known outputs at positions less than  $i$ . During training, the ground truth caption tokens are shifted right, and a special token  $\langle BOS \rangle$  (begin of the sentence) is inserted into the first position. A cross-entropy generation loss  $\mathcal{L}_{gen}$  is then computed with the logits transformed from the last decoder layer’s hidden states and un-shifted ground

truth caption token ids with a special token  $\langle EOS \rangle$  (end of the sentence) appended.

$$\mathcal{L}_{gen} = - \mathbb{E}_{\hat{y}_i \sim \hat{y}} \log p(\hat{y}_i | \hat{y}_{<i}, \tilde{\mathbf{T}}, \tilde{\mathbf{V}}; \theta). \quad (3)$$

It is noted that  $\hat{y}$  is the masked version of the ground-truth caption  $y$ . To make a fair comparison with the baseline (Pont-Tuset et al., 2020), we apply the same setting and do not employ common techniques such as label smoothing (Szegedy et al., 2016) or self-critical training (Rennie et al., 2017).

### 3.2 Attention Guidance for Spatial Ground

**Attention Supervision Construction** To explicitly guide the attention for object-level spatial grounding, we align the semantic caption tokens with the visual object by taking trace as an intermediate bridge. In this way, we construct a supervision matrix to guide the attention between the caption tokens and visual objects by the two following steps.

- 1) Language-trace temporal alignment. In the Localized Narrative dataset, the caption utterances<sup>1</sup>  $u$  and mouse traces are highly temporal-aligned, i.e., every utterance  $u$  has a

<sup>1</sup>We are following the naming tradition of Pont-Tuset et al. (2020), where an utterance means one or several adjacent tokens, not a whole sentence.

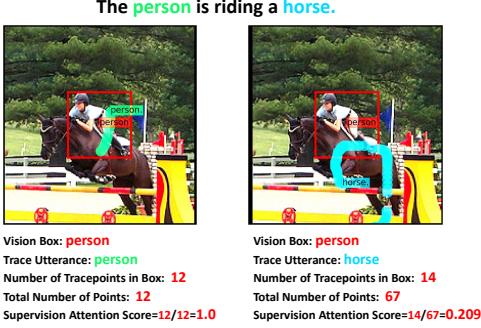


Figure 3: A showcase of spatial attention scoring

corresponding time window, every tracepoint  $p$  has a timestamp. To leverage this information, we first assign each tracepoint  $p$  to a unique utterance  $u$ , where the tracepoint timestamp is in the utterance time window. Thus, every utterance  $u$  is aligned to a series of tracepoints  $P_u = \{p_1, \dots, p_{k_u}\}$ .

- 2) Language-vision spatial alignment. Give the utterance  $u$  and corresponding  $P_u$ , we calculate the alignment score considering the spatial overlap between tracepoints  $P_u$  and each vision object  $v_i$ . Every visual object  $v_i$  has a corresponding spatial bounding box  $b_i = (x_i^1, y_i^1, x_i^2, y_i^2)$ , and the  $x_i^1, y_i^1, x_i^2, y_i^2$  are top-left and bottom-right horizontal and vertical coordinates respectively. We set the alignment score  $s_{(u_j, b_i)}$  between utterance  $u_j$  and bounding box  $b_i$  as,

$$s_{(u_j, b_i)} = \frac{\sum_{p \in P_{u_j}} I_{b_i}(p)}{|P_{u_j}|} \quad (4)$$

where  $I$  is an indicator of whether point  $p$  is in the bounding box  $b_i$ :

$$I_{b_i}(p) = \begin{cases} 1 & \text{if } x_i^1 < x_p < x_i^2 \\ & \text{and } y_i^1 < y_p < y_i^2 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$x_p$  and  $y_p$  are the coordinates of each tracepoint in  $P_u$ . An example of the alignment score calculation is illustrated in Figure 3.

By calculating the alignment score, we establish the spatial grounding supervision between caption tokens and auto-detected visual objects. For every word  $y_i$  in the same utterance  $u$ , the  $s_{(y_i, b_j)} = s_{(u_j, b_j)}$ . Eventually, we get the supervision score matrix  $\mathcal{S} \in [0, 1]^{N \times T}$  and  $\mathcal{S}_{ij} = s_{(y_i, b_j)}$ .

**Attention-guided Grounding** A cross-attention matrix is generated in shape  $(N, T, L, H)$  during the transformer’s decoding steps. Here  $N$  denotes the number of pre-detected visual objects,  $T$  denotes the number of tokens in a caption sentence after padding,  $L$  denotes the number of transformer layers, and  $H$  denotes the number of attention heads in transformer layers. Two linear projections and layer normalization (Ba et al., 2016) are applied sequentially on dimension  $L$  and  $H$ , respectively reducing the dimension to 1. Thus, for a single instance, we eventually calculate an attention matrix  $\mathcal{A} \in \mathbb{R}^{N \times T}$ .

To train the model, the goal can be achieved by minimizing the following attention guidance loss function  $\mathcal{L}_{att}$ :

$$\mathcal{L}_{att} = - \mathbb{E}_{a \sim \mathcal{A}, s \sim \mathcal{S}} [s \cdot [s \log a + (1 - s) \log (1 - a)]], \quad (6)$$

which is a weighted Binary Cross Entropy between  $\mathcal{A}$  and  $\mathcal{S}$ . Noted that we also choose to mask out some stop-words columns of the matrix  $\mathcal{A}$  and  $\mathcal{S}$  to avoid introducing too much annotation noise.

### 3.3 Contrastive Constraints for Temporal Alignment

As illustrated on the left side of Figure 4, we first use a “split by sentence” procedure to build a sentence-level alignment between caption and traces, and then employ contrastive loss to constrain the temporal order of the generation process.

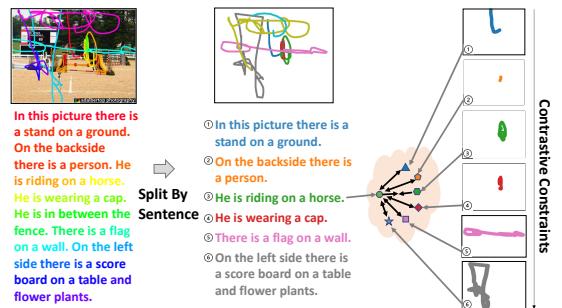


Figure 4: A showcase of split by sentence and contrastive constraints for temporal alignment

**Split by Sentence** An annotated instance consists of an image, a tracepoint list, and a caption paragraph consisting of a list of ordered caption sentences. Here, we define a caption sentence as a

series of utterances segmented out by a period(')'). In section 3.2, we already maintain an alignment between utterances and tracepoints. Following this setting, we can unite a list of ordered utterance  $U = \{u_1, \dots, u_k\}$  in the same caption sentence, and then orderly unite a list of tracepoints corresponding to  $U$ 's elements into a so-called trace segment. The alignment between caption sentences and trace segments can be established by simply uniting the association between utterances and tracepoints with respect to the above sentence split. We call this procedure as **split by sentence**.

**Temporal Contrastive Constraints** According to the split mentioned above, we aggregate the transformer's last layer hidden states of trace segments and caption sentences respectively, and denote them as  $H_{ts} = \{h_{ts}^1, \dots, h_{ts}^n\}$  and  $H_{cs} = \{h_{cs}^1, \dots, h_{cs}^n\}$ . Here  $n$  is the number of caption sentences.

We adopt the NCE loss to learn to discriminate the positive from negative trace-caption pairs. The positive is defined as all the temporal aligned corresponding caption sentence and trace segment pairs i.e. with the same order indices, and other pairs without temporal alignment in the same image as negative samples. This contrastive loss function  $\mathcal{L}_{cts}$  is defined as follows,

$$\mathcal{L}_{cts} = - \mathbb{E}_{h_{ts} \sim H_{ts}} \log \frac{\exp(s(h_{ts}^i, h_{cs}^i))}{Z}, \quad (7)$$

$$Z = \sum_{j=1}^n \exp(s(h_{ts}^i, h_{cs}^j)) \quad (8)$$

where  $s(\cdot, \cdot)$  means two linear layers and an L2 normalization applied on the elements respectively, and a dot production between them. By minimizing the  $\mathcal{L}_{cts}$ , we force the model to learn a representation being aware of sentence-level temporal ordering, which leads to more precise captioning.

### 3.4 Loss

Finally, the model is trained with three losses  $\mathcal{L}_{gen}$ ,  $\mathcal{L}_{att}$ , and  $\mathcal{L}_{cts}$ , where  $\mathcal{L}_{gen}$  is the caption generation loss,  $\mathcal{L}_{att}$  is the spatial attention guidance loss, and  $\mathcal{L}_{cts}$  is the temporal contrastive loss. We jointly optimize our model by minimizing all losses added together:

$$\mathcal{L}_{all} = \mathcal{L}_{gen} + \mathcal{L}_{att} + \mathcal{L}_{cts}. \quad (9)$$

## 4 Experiments

### 4.1 Dataset

We use the annotated COCO subset of Localized Narratives to evaluate our method. We call this dataset split as **LN-COCO** for short. Each image has one or several pairs of the captioning paragraph and corresponding mouse traces. Every single pair is a so-called localized narrative. The training and validation splits are identical to Pont-Tuset et al. (2020)'s setting. There are 134,272 localized narratives in the training set and 8,573 in the validation set. We train on the whole training set and evaluate our model performance against the identical validation set.

### 4.2 Implementation Details

For the visual feature, we adopt Faster-RCNN(Ren et al., 2015) to extract 100 bounding box proposals. For trace feature, we use  $\tau = 0.4s$  to extract trace segment for feature extraction. The embedding size  $d$ , number of transformer layers, hidden size of the transformer feed-forward layer are 768, 2, and 768, respectively. The number of attention heads is 8, and the dropout rate is 0.1. We adopt the AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate of 7e-4(which is the best performance setting of baseline, and adopted widely for other trials), and set two momentum parameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.99$ . We set the batch size to 256. All models are trained on 4 Tesla V100 GPUs with 32GB memory for 10 to 12 hours.

### 4.3 Evaluation Metrics

This generation task adopts the traditional image captioning evaluation metric using the open-source tool<sup>2</sup> with a minor modification<sup>3</sup> to suit with LN-COCO, including BLEU(Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin and Och, 2004), ROUGE-1-F1(Pont-Tuset et al., 2020), and CIDEr-D (Vedantam et al., 2015).

### 4.4 Results

**Baseline and +Trace methods** The Baseline and +Trace methods are our re-implementations following (Pont-Tuset et al., 2020)'s method description. The Baseline method only takes image feature as input while the +Trace model take image feature

<sup>2</sup><https://github.com/tylin/coco-caption>

<sup>3</sup>We add an additional id to every trace-image-caption triplet and adjust some code of the standard evaluation tool to meet the "1 trace-vs-1 caption" evaluation need.

and trace both as input. They employ the architecture in Changpinyo et al. (2019) with a few minor differences. First, they set the number of Transformers’ layers for both the encoder and the decoder to 2 instead of 6. Second, their projection layers also consist of layer normalization(Ba et al., 2016). Third, they set the maximum number of iterations to 150k. Finally, they allow the maximum number of target captions to be as long as 225 to account for the narration’s longer nature.

**LoopCAG methods** Our model comprises of four components: 1) the transformer encoder-decoder framework; 2) the trace input; 3) Attention Guidance(+AG for short) grounding loss; 4) Contrastive constraints(+C for short).

**Main Results** The Table 1 shows the overall performance comparison on the LN-COCO dataset. To reduce the deviation caused by different implementation details, we first present our implementations’ performance (with \*), which have a higher score than Pont-Tuset et al. (2020) reported. Thus, we have a more strict baseline to evaluate the improvement purely coming from our innovative method. Compared to Baseline\* method, the performance on all metrics improves significantly when controlling captioning using the mouse trace (+Trace\*), it indicates that using the mouse trace enables the system to describe better those user intended parts of the image.

Most importantly, the results indicate that our LoopCAG method achieves state of the art on all automatic criteria, outperforming the previous state-of-art model by **2.4** and **7.5** on BLEU-4 and CIDEr-D, respectively. This demonstrates our proposed Attention Guidance method helps the model generate better spatially grounded and more precise captions. When considering the **2.0** rising on ROUGE-L score, we can conclude that Contrastive constraints can help the model better align the order of generated sentence to the user intent because ROUGE-L mainly employs an order matter longest common sequence F-measure.

**Ablations** We perform three ablations to verify the most improvements in-deed come from the Attention Guidance and Contrastive constraints. Starting from standard captioning (Baseline\*), we add the Attention Guidance to help the model better spatially ground visual objects and caption tokens (Table 2, “+ Ag”). This affects performance, suggesting that the model does benefit from knowing

where to find the highly semantic related appearance feature in the image. Next, we add the trace feature (Table 2, “+ Trace”). This introduces user intention to the model. We also take this line to show the performance lift caused by Contrastive constraints fairly. Then we add the contrastive module (Table 2, “+C”) and see a good improvement on almost all criteria. Hence, we verify the significance of the positive influence of temporal contrastive constraints. Moreover, in the last line is our full LoopCAG model. We can see the two proposed methods are not exclusive to each other.

## 4.5 Quantitative Analysis

### Controllability Analysis on Temporal Order

We also design an experiment to further demonstrate LoopCAG’s superior controllability on the caption sentences’ temporal order. Specifically, we split each localized narrative input by sentence as described in Sec3.3, and reverse the sequential order of the splits, i.e., the last sentence of a caption paragraph will become the first one, the same processing is applied to trace segments, too. We conduct an evaluation on the sentence&segment reverted dataset, and the performance comparison is shown in Table 3. With the Contrastive constraints mechanism’s help, the LoopCAG model is much more robust to trace input reversing, even competitive with the model trained on reverted data. In contrast, the base models all face a dramatic drop on almost all metrics when the input trace order is reversed. This also implies there are some biased habits of human annotators. For example, they always describe the salient objects first and end with a sentence about the background of the image.

**Controllability Analysis on Temporal Frequency** Then, we analyze the controllability of the temporal frequency  $\tau$  to present whether the coarse-grained or fine-grained tracepoints (sampling rate, in other words) affects the generation performance. As the Table 4 shows, we change the temporal frequency  $\tau$  from 0.4 to 1.2. A performance drop is impressive with the  $\tau$  getting larger. The purpose of this experiment for various  $\tau$  is to simulate the trace drawing speed of users in a real application scenario, and a larger  $\tau$  is equivalent to a faster drawing speed. As Deng et al. (2020) has demonstrated, the length is one of the critical facts that impact quantitative performance. This result implies we can further decide to generate either a coarse-grained or fine-grained caption by

Method	ROUGE-L	ROUGE-1-F1	BLEU-1	BLEU-4	CIDEr-D	METEOR
Baseline(Pont-Tuset et al., 2020)	31.7	47.9	32.2	8.1	29.3	-
+Trace(Pont-Tuset et al., 2020)	48.3	60.7	52.2	24.6	106.5	-
Baseline*	34.1	54.0	36.0	10.3	29.5	16.4
+Trace*	49.0	68.1	55.4	25.0	107.9	25.2
LoopCAG(our)	<b>50.3</b>	<b>69.8</b>	<b>57.2</b>	<b>27.0</b>	<b>114.0</b>	<b>26.0</b>

Table 1: Comparison with baseline methods results: Baseline means an encoder-decoder model without taking trace as input. +Trace means concatenating encoded trace feature to the encoder input, i.e., trace controlled caption performance. LoopCAG is our complete model. The results with \* are the baseline performance re-implemented by ourselves

Method	ROUGE-L	ROUGE-1-F1	BLEU-1	BLEU-4	CIDEr-D	METEOR
Baseline*	34.1	54.0	36.0	10.3	29.5	16.4
+AG	34.7(+0.6)	55.5(+1.5)	37.4(+1.4)	10.5(+0.2)	30.1(+0.6)	16.6(+0.2)
+Trace*	49.0	68.1	55.4	25.0	107.9	25.2
+Trace+C	50.1(+1.1)	69.3(+1.2)	56.7(+1.3)	26.4(+1.4)	113.6(+5.7)	25.9(+0.7)
LoopCAG	<b>50.3(+1.3)</b>	<b>69.8(+1.7)</b>	<b>57.2(+1.8)</b>	<b>27.0(+2.0)</b>	<b>114.0(+6.1)</b>	<b>26.0(+0.8)</b>

Table 2: Ablation study results: Baseline means an encoder-decoder model without taking trace as input. +AG means using attention guidance. +Trace means concatenating trace feature to the encoder input, i.e., trace controlled caption performance. +C means applying the contrastive constraints method. LoopCAG is our complete model. The results with \* are the baseline performance re-implemented by ourselves

Method	Reverse	Reverse	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D
	Trained	Evaluated					
Baseline*		✓	36.0	10.1	16.3	28.7	29.1
+Trace*		✓	50.8	15.5	19.9	33.2	43.4
+Trace*	✓		50.2	16.4	20.1	36.4	45.2
+Trace*	✓	✓	53.4	19.6	21.6	38.2	55.1
LoopCAG	✓		53.7	18.6	21.7	34.6	52.2

Table 3: Analysis on temporal order results: Model performance on caption sentence and trace segment reversed evaluation dataset. The results with \* are the baseline performance re-implemented by ourselves.

$\tau$	BLEU-4	METEOR	ROUGE-L	CIDEr-D
0.4	26.9	25.5	47.2	91.7
0.6	26.7	25.5	46.9	91.1
0.8	26.1	25.3	46.2	88.3
1.0	24.8	24.7	44.9	82.4
1.2	24.1	24.4	44.3	79.1

Table 4: Analysis on temporal frequency results.

controlling the time-frequency  $\tau$ .

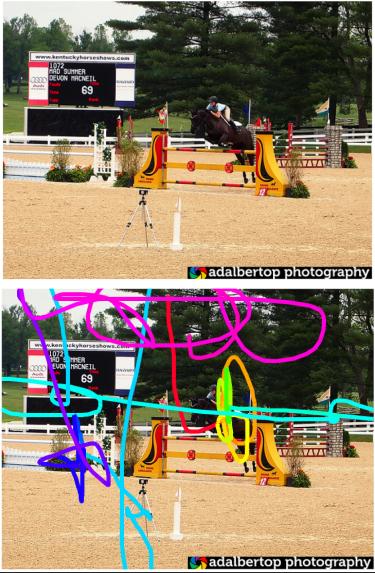
**Controllability Analysis on Spatial Semantic Grounding** One of our important purposes of using attention guidance is introducing more interpretability to the model while improving the caption performance. When generating each token, the model is forced to show which visual elements are the most effective reason for the current generation. And this effectiveness is supervised by our pseudo

attention label. In this way, we can hopefully obtain better visual-linguistic joint representation. In appendix A, we showcase the attention values comparison of models w/wo attention guidance. We find that the AG model has a more diverse distribution across all different types of tokens. A "neater" activation is observed in Appendix A (a) compared with (c), e.g., activations of "who", "is" and "on" are clearly suppressed. We observe that these suppressions happen on most function word, so we add this illustration for further discussion and exploration by our research community.

#### 4.6 Qualitative Case Study

We present a showcase of a captioning result of different methods in Figure 7. We can easily find that the Baseline captioning describes the image in random order while the +Trace Captioning and

LoopCAG Captioning almost have the same order as Ground Truth Captioning. It is also awe-inspiring that the Baseline captioning and +Trace Captioning both consist of some preposterous description highlighted in red color. In contrast, the LoopCAG captioning is all reasonable. This is evidence of superior fact grounding advantage brought by our Attention Guidance Method.



Ground Truth Captioning

In this picture there is a stand on a ground. On the backside there is a person. He is riding on a horse. He is wearing a cap. He is in between the fence. There is a flags on a wall. On the left side there is a score board on a table and flower plants.

We can see in the background sky. trees.

Baseline Captioning

In this image I can see a horse which is in white color, at left there is a person sitting on the horse, at the back ground there are some people standing, in the background there are few buildings, trees and sky.

+Trace Captioning

This picture might be taken outside of the city. in this image, in the middle there is a man sitting on horse and holding the collar rope of a horse. on the right side, we can also see another horse and a person is riding it. In the background, there are group of people, flags, trees, plants, metal fence, hoardings, trees. on top there is s a sky, at the bottom there are some grass and a land.

LoopCAG Captioning

There is a person sitting on a horse. he is holding a horse thread and he is wearing a cap. there are flags, board on the left side. we can see in the background sky, trees.

**Figure 5: Controlled Captioning Qualitative Examples**  
1: Ground Truth Captioning by annotator versus Baseline captioning where the input is only the image, captioning controlled by mouse traces where the mouse traces are also an input to the model (+Trace and LoopCAG Captioning). Gradient indicates time.

## 5 Related Work

**Controllable Image Captioning** is an emerging research direction. Previous works aim to control the captioning by Part-Of-Speech tagging(Deshpande et al., 2018), sentiment, (You et al., 2018), length (Deng et al., 2020), bounding box (Cornia et al., 2019) etc. Those works either tried to describe a semantic guided captioning. Other works relied on predefined categories, e.g., bounding box or sentiment classes. Similar works (Yu et al., 2018; Cornia et al., 2019) control the caption by a sequence of ordered topics and bounding boxes. However, those methods limit the captioning on the pre-defined or recognized objects in the bounding box and hard to scale out. Besides, the trace is a more natural way to input than the bounding box. The most similar work (Pont-Tuset et al., 2020) proposed a trace-controlled image captioning task and designed a simple benchmark by directly concatenating the mouse trace coordinates and size into a self-attention module. Although mouse trace is flexible and interactive, it is easy for humans to understand the trace’s semantic representation but hard for AI agents. Unlike previous works, we propose a novel trace-controlled model for capturing the semantic representation of trace from both fine-grained and coarse-grained spatial and temporal characteristics.

**Contrastive Learning** Recently, contrastive learning has been widely studied in unsupervised representation learning for vision, (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Caron et al., 2020; Chen and He, 2020), language (Mikolov et al., 2013; Saunshi et al., 2019; Chi et al., 2020; Fang and Xie, 2020; Giorgi et al., 2020; Kong et al., 2020; Gunel et al., 2021), or multi-modal (Sun et al., 2019; Luo et al., 2020). The goal is to learn semantic representation between two views by allowing the positive sample to be similar (in semantic space) and negatives to be dissimilar semantically simultaneously. CLIP (Radford et al.) and MIL-NCE (Miech et al., 2020) has demonstrated the effectiveness for learning the semantic mapping between vision and language. Previous attempts mainly exploit the InfoNCE (Oord et al., 2018) objective to maximize a lower bound of the mutual information. This paper extends the multi-modal contrastive learning between the trace in the image and captioning sentence. In the same image, they correspond to each other semantically. This motivates us to design a contrastive loss for better

alignment between the trace and language.

## 6 Conclusion

In this paper, we focus on the controlled image captioning task and find mouse traces provide an intuitive and efficient way for a user to control the description. We propose a novel caption generation model with contrastive constraints and attention guidance called LoopCAG to control the captioning process spatially and temporally. The experimental results demonstrate the our model’s effectiveness, and our work will inspire more future research on vision-linguistic understanding and generation.

## 7 Acknowledgement

We thank Botian Shi, Rongcheng Tu for helpful discussions. This work is supported in part by National Key R&D Program of China 2018AAA0102301 and NSFC 61925203.

## References

- Jimmy Ba, J. Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*.
- Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. *ArXiv*, abs/1909.02097.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119, pages 1597–1607.
- Xinlei Chen and Kaiming He. 2020. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*.
- Zewen Chi, L. Dong, Furu Wei, N. Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He yan Huang, and M. Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.
- M. Cornia, L. Baraldi, and R. Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8299–8308.
- Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. 2020. Length-controllable image captioning. In *Computer Vision – ECCV 2020*, pages 712–729, Cham. Springer International Publishing.
- Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David A Forsyth. 2018. Diverse and controllable image captioning with part-of-speech guidance.
- Hongchao Fang and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.
- John M Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaoan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*.
- Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *ICLR*.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738.
- Lingpeng Kong, Cyprien de Masson d’Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. In *ICLR*.
- Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*, page 605.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.
- Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.
- Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.
- Jordi Pont-Tuset, Jasper Uijlings, Beer Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. [Connecting vision and language with localized narratives](#). In *ECCV*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’15, page 91–99, Cambridge, MA, USA. MIT Press.
- Steven J. Rennie, E. Marcheret, Youssef Mroueh, J. Ross, and V. Goel. 2017. Self-critical sequence training for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.
- Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5628–5637.
- Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.
- C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. [Rethinking the inception architecture for computer vision](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.
- Quanzeng You, Hailin Jin, and Jiebo Luo. 2018. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121*.
- Niange Yu, Xiaolin Hu, Binhe Song, Jian Yang, and Jianwei Zhang. 2018. Topic-oriented image captioning based on order-embedding. *IEEE Transactions on Image Processing*, 28(6):2743–2754.

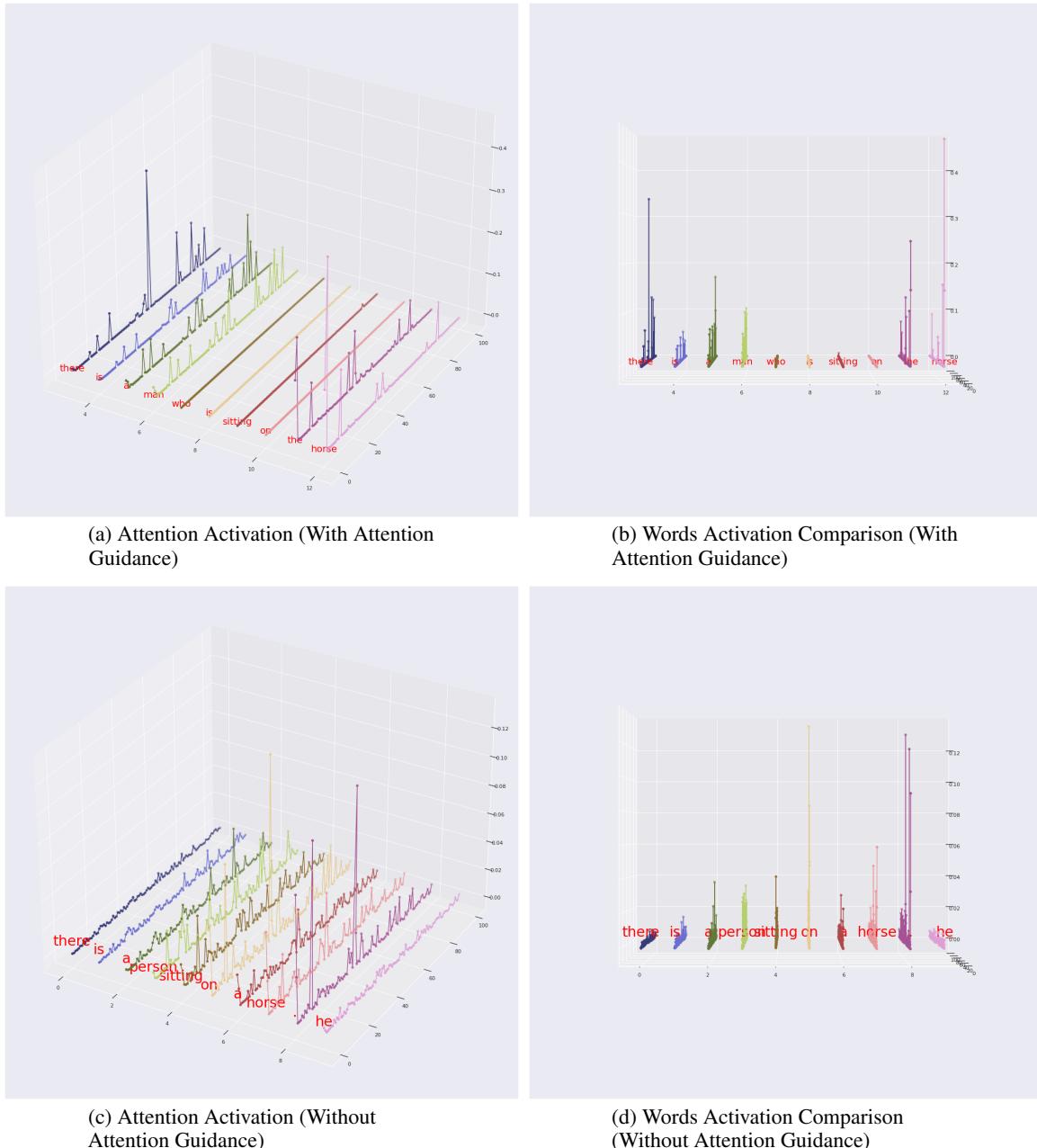


Figure 6: Appendix A: Controllability Analysis on Spatial Semantic Grounding



Ground Truth Captioning

In this image i can see a person wearing white shirt, blue tie, blue blazer, skirt and black shoes is standing and holding a black colored bag in his hand. In the background i can see the white colored wall.

Baseline Captioning

in this picture we can see a man standing and holding a **mobile in his hand**, in the background we can find a wall.

+Trace Captioning

in the middle of the image a man is standing and smiling and he is holding a **tennis racket**. behind him there is a **cloth on the red color wall**. **bottom left side of the room there are two shoes**.

LoopCAG Captioning

in this image i can see a person wearing blue coat, black pant and black shoe is standing and holding a black colored bag in his hand. in the background i can observe the white colored wall.

Figure 7: **Appendix B: Controlled Captioning Qualitative Examples 2:** Ground Truth Captioning by annotator versus Baseline captioning where the input is only the image (top left), captioning controlled by mouse traces where the mouse traces are also an input to the model (+Trace and LoopCAG Captioning). Gradient  indicates time.