

# **SocialGPT: Prompting LLMs for Social Relation Reasoning via Greedy Segment Optimization**

Wanhua Li<sup>\*,1</sup> Zibin Meng<sup>\*,1,2</sup> Jiawei Zhou<sup>3</sup> Donglai Wei<sup>4</sup> Chuang Gan<sup>5,6</sup> Hanspeter Pfister<sup>1</sup>

<sup>1</sup>Harvard University <sup>2</sup>Tsinghua University <sup>3</sup>Stony Brook University <sup>4</sup>Boston College <sup>5</sup>MIT-IBM Watson AI Lab <sup>6</sup>UMass Amherst

## Outline



### > Introduction

- > Framework
- **Greedy Segment Prompt Optimization (GSPO)**
- > Visualization
- > Experiments
- > Conclusion

## Introduction

### 3

Social relation reasoning aims to identify relation categories such as friends, spouses, and colleagues from images.

NEURAL INFORMATION PROCESSING SYSTEM

- Current methods adopt the paradigm of training a dedicated network end-to-end using labeled image data, they are limited in terms of generalizability and interpretability.
- To address these issues, we present a simple yet well-crafted framework named SocialGPT, which combines the perception capability of Vision Foundation Models (VFMs) and the reasoning capability of Large Language Models (LLMs) within a modular framework, providing a strong baseline for social relation recognition.



## Framework



### > Main Process

> Social relation recognition takes an image I and two bounding boxes  $b_1$  and  $b_2$  of two interested individuals as inputs, and requires a model that outputs the social relationship y.



Figure 2: The framework of SocialGPT, which follows the "perception with VFMs, reasoning with LLMs" paradigm.

## Framework

## NEURAL INFORMATION PROCESSING SYSTEMS

### Perception with Vision Foundation Models

- ➢ Use SAM to segment the image to obtain all different object masks, and then send individual objects by masking out others to BLIP-2 to obtain descriptions of each object.
- Ask specific questions related to social identities by using the BLIP-2 dialog functionality to extract more specific information depending on object types. (the age and gender of individuals, as well as the social scene and activity)

### Social Story Generation

➢ Fuse the raw information into a coherent *social story* in textual format, denoted as S, which can be best reasoned with LLMs.



**Objects with Symbols** 

a. Image Resolution: 1024X768

b. Image Caption: a man and a woman standing next to a boy nearby a golf cart

c. Image Scene: A family enjoying a day on the golf coursed. People and Objects Semantic:

{<symbol>:[P1], <coordinate>:[103, 96, 59, 181], <caption>:[a boy in a black shirt and shorts holding a golf club], <age>:[This is a child], <gender>:[male]};

{<symbol>:[P2], <coordinate>:[143, 24, 79, 257], <caption>:[a man in shorts and a hat], <age>:[He is a man in his mid-50s], <gender>:[male]};

{<symbol>:[P3], <coordinate>:[89, 50, 79, 213], <caption>:[a woman
with sungLasses], <age>:[She is in her mid-40s], <gender>:[female]};
{<symbol>:[01], <coordinate>:[93, 130, 184, 114], <caption>:[a red
and black lawn mower]};

{<symbol>:[02], <coordinate>:[0, 93, 383, 193], <caption>:[a green
golf course]);
(courbel>:[02] coordinate>:[0 0 202 210] coordinate>:[Uuch

{<symbol>:[03], <coordinate>:[0, 0, 383, 210], <caption>:[Lush
bush]};

Dense Captions with Symbols

In this picturesque scene at the golf course, a delightful family is clearly enjoying a day of outdoor fun. At the center of the image, a young <P1> boy, dressed in a black shirt and shorts, stands confidently, gripping a golf club with enthusiasm. To his right, a middle-aged <P2> man, wearing shorts and a hat, mirrors the boy's excitement. On the left side of the frame, a <P3> woman stands next to the aforementioned man. Their presence complements each other, emphasizing the familial bond shared by this trio. Adding to the liveliness of the setting, various objects become integral to the overall ambiance. A striking red and black lawn mower <O1> positioned nearby evokes the meticulous maintenance of the greens, enhancing the serene atmosphere. A green golf course <O2> and lush bush <O3> add a natural touch to the family gathering. Overall, this captivating image captures the essence of family togetherness, blending elements of recreational joy with the beauty of the golf course surroundings.

Social Story

Figure 3: An example of social story generation.

Input Image

## Framework

## NEURAL INFORMATION PROCESSING SYSTEMS

### Reasoning with Large Language Models

- After obtaining the mapping from image to social story: *I* → *S*, feed both *S* and bounding box queries (*b<sub>i</sub>*, *b<sub>j</sub>*), converted to textual queries *q* with referencing symbols *P<sub>i</sub>*, *P<sub>j</sub>*, into LLMs to obtain interpretable answers *a*.
- Since LLM performance is highly sensitive to prompt variations, design social relation reasoning prompt with four segments, which is called **SocialPrompt**.
  - System (denoted as *o*) This is the system prompt provided by many LLMs to steer their behavior.
     We utilize it to explicitly define several core rules for our task of social reasoning.
  - Expectation (denoted as r) This is the instruction that we give to the model to set expectations of the anticipated outcomes. This helps avoid vague or unexpected outputs. To do so, we construct a role assignment and task description prompt, where we explicitly assign the role of a social relation expert to the LLM and provide a detailed elaboration of the task's input and output.
  - Context (denoted as c) This provides sufficient contextual information to help the LLMs understand the background of the problem. As a classification task, we provide specific definitions for each social relationship category.
  - Guidance (denoted as e) This offers an exemplar to show the LLMs how to respond to a query based on a social story. We manually construct an in-context example prompt, to better guide LLMs in performing social relationship reasoning in the desired format.

## **Greedy Segment Prompt Optimization (GSPO)**

- Motivation of Designing GSPO
  - Different ways of prompt rephrasing and demonstration example variations can significantly impact the LLM reasoning performance.
  - Manually searching for the optimal prompt is time-consuming and labor-intensive, thus automatic prompt tuning is desired.
- Tuning Object
  - > We aim to find the optimal prompt  $\{o^*, r^*, c^*, e^*\}$  that maximize the probability of LLMs generating the correct answer *a* for any given sample z = (S, q).
  - We assume that the ground truth answer a for sample z takes the following form:  $a = [a^0, a^1, a^2, ...]$ , where  $a^0$  denotes the first sentence of a,  $a^1$  is the second sentence, and so forth. We specify  $a^0$  to have the following fixed format:  $a^0 = "The final answer is str(y)"$ , where str(y) represents the string representation of class label y. Then we can define the objective:

$$\mathcal{L}(\boldsymbol{o}, \boldsymbol{r}, \boldsymbol{c}, \boldsymbol{e}; \boldsymbol{z}, \boldsymbol{y}) = -\mathbb{E}_{(\boldsymbol{z}, \boldsymbol{a}^0)}[\log p(\boldsymbol{a}^0 | \boldsymbol{o}, \boldsymbol{r}, \boldsymbol{c}, \boldsymbol{e}; \boldsymbol{z})]$$

## **Greedy Segment Prompt Optimization (GSPO)**

### Long Prompt Optimization

- > Propose a candidate set  $W_m$  consisting of alternative prompts for each segment, and the algorithm searches over the combination of different candidates.
- ▶ The gradient is computed as:  $\nabla_{h_{w_m}} \mathcal{L}(w_{1:M}) \in \mathbb{R}^{|W_m|}$ , where  $h_{w_m}$  represents the one-hot representation of selecting  $w_m$  from the set  $W_m$ .

Algorithm 1 Greedy Segment Prompt Optimization

Input: Initial segments  $w_{1:M}$ , training dataset  $\mathcal{T}$ , iteration number NBuild the candidate set  $\mathcal{W}_m$  for each segment  $w_m$ repeat N timesRandomly sample a batch of data  $\mathcal{D}$  from  $\mathcal{T}$ for  $m = 1, \ldots, M$  do $\mathcal{U}_m := \operatorname{Top-} k(-\sum_{z \in \mathcal{D}} \nabla_{h_{w_m}} \mathcal{L}(w_{1:M}; z))$  $\vdash$  Compute top-k promising segment substitutionsfor  $b = 0, 1, \ldots, K * M - 1$  do $\tilde{w}_{1:M}^{(b)} := w_{1:M}$  $\tilde{w}_i^{(b)} := \mathcal{U}_i(\lfloor b/M \rfloor)$ , where  $i = (b \mod M) + 1$  $w_{1:M} := \tilde{w}_{1:M}^{(b^*)}$ , where  $b^* = \operatorname{argmin}_b \sum_{z \in \mathcal{D}} \mathcal{L}(\tilde{w}_{1:M}^{(b)}, z)$  $\vdash$  Compute best replacementOutput: Optimized segments  $w_{1:M}$ 

## Visualization



### **Reasoning Process and Interpretability**



In this heartwarming scene, a father lovingly holds his baby while a young boy joyfully eats his food nearby. The father, depicted by  $\langle P3 \rangle$ , wearing an orange shirt, cradles the baby, represented by  $\langle P1 \rangle$ . The baby, captured in a tender moment, lays peacefully, with arms outstretched towards his father, symbolizing a baby reaching out for connection and exploration. Meanwhile,  $\langle P2 \rangle$ , a happy young boy in a red and blue plaid shirt, eats his food. Together, these individuals form a harmonious family unit, with the father providing nourishment and love, the baby representing innocence and vulnerability, and the young boy embodying curiosity and playfulness. Surrounding them, various objects add depth to the scene. A plate filled with food, <01>, signifies the importance of nutrition and nourishment ...

In this captivating image with a resolution of 1024X777, we find ourselves immersed in a lively scene where a group of people are joyfully engaged in playing music. The atmosphere is electric as they come together, their instruments in hand, creating a harmonius symphony. At the forefront, we are drawn to the presence of a woman, denoted as a point, perched on a chair, deeply engrossed in her instrument. She is a teenager, exuding youthful energy, while nearby, <P2> captivates us with her vibrant floral dress. Her youthful demeanor suggests she is a young woman, adding to the diverse age range of participants. Standing confidently with hands on her hips, <P3> commands attention in her elegant black dress, representing a woman in her mid-20s. On a stool, <P4> catches our eye, dressed smartly in a tie and shirt ...

Social Story

and <P3>: [The final answer is <fatherchild>]. Based on the description of the image, the most likely social relationship between \*1 and P3 is that of a father and child. The description emphasizes the father's loving interaction with the baby, as he cradles him. The presence of a young boy playing nearby suggests that he is likely the older sibling of the baby, further supporting the father-child relationship.

Based on the description of the image, the most likely social relationship between and P2 is that they are friends. The lively atmosphere, their engagement in playing music together, and the diverse age range of participants suggest a close and enjoyable friendship rather than a familial or professional relationship.

#### Interpretable Answer

Figure 4: Visualization results of interpretability. We show the SocialGPT perception and reasoning process. We see that our model predicts correct social relationships with plausible explanations.

### Generalization on Different Image Styles



#### [GR<sup>2</sup>N] <P1>-<P2>: Colleagues

[Ours] <P1>-<P2>: [The final answer is <classmates>]. Based on the description of the image, the most likely social relationship between <P1> and <P2> is that they are classmates. The setting of a classroom and the presence of other students indicate an educational context. The description does not provide any indication of a familial or romantic ...



#### [GR<sup>2</sup>N] < > - < > : grandma-grandchild

[Ours] (P2>-(P3>: [The final answer is <father-child>]. Based on the description of the image, the most Likely social relationship between <P2> and <P3> is that of a father and child. The presence of a family playing board games in the living room suggests a familial context. The description emphasizes the young age and playful demeanor of <P2>, indicating a child ...

Figure 5: Results when applying SocialGPT to sketch and cartoon images. The images are generated by GPT-4V. Our method generalizes well on these novel image styles.

## **Experiments**

## NEURAL INFORMATION PROCESSING SYSTEMS

### Datasets

- > PIPA dataset
  - The PIPA dataset categorizes 16 types of social relationships, including family bonds (like parent-child, grandparent-grandchild), personal connections (friends, loves/spouses), educational and professional interactions (teacher-student, leader-subordinate), and group associations (band, sport team, colleagues).
- > PISC dataset
  - The PISC dataset categorizes social relationships into 6 types: commercial, couple, family, friends, professional, and no-relation.
- > Metric
  - > For both datasets, we measure **classification accuracy** as our evaluation metric.

## **Experiments**



### **Zero-Shot Performance on PIPA Dataset**

Methods	ZS	Acc (%)	
All attributes + SVM [1]	X	57.2	
Pair CNN [13]	×	58.0	
Dual-Glance [13]	×	59.6	
SRG-GN [54]	×	53.6	
GRM 6	×	62.3	
MGR [2]	×	64.4	
$GR^2N$ [3]	×	64.3	
TRGAT [14]	×	65.3	
SocialGPT (w/ GPT-3.5)	/	64.1	
SocialGPT (w/ Vicuna-13B)	-	66.7	

Table 1: The comparison results on the PIPA dataset. ZS stands for Zero-Shot.

### Ablation Study on PIPA dataset with Vicuna-7B

Methods	Acc (%)
SocialGPT	61.58
<ul> <li>Dense Captions</li> <li>Task-oriented Captions</li> <li>Symbol → Object Coordinate</li> <li>Symbol → Object Caption</li> <li>Social Story</li> </ul>	52.63 59.89 57.68 59.83 45.31
<ul> <li>SocialPrompt Segment {System}</li> <li>SocialPrompt Segment {Expectation}</li> <li>SocialPrompt Segment {Context}</li> <li>SocialPrompt Segment {Guidance}</li> </ul>	60.23 59.19 61.18 43.56

Table 2: Ablations on components of SocialGPT with Vicuna-7B. The results are obtained on the PIPA dataset with a zero-shot setting.

## **Experiments**

## NEURAL INFORMATION PROCESSING SYSTEMS

### 12

### Zero-Shot Performance on PISC Dataset

Methods	ZS	Acc (%)
Pair CNN [13]	×	46.30
GRM 6	×	64.18
$GR^2N$ [3]	×	64.70
SocialGPT (w/ GPT-3.5)	/	53.43
SocialGPT (w/ Vicuna-13B)	1	65.12

Table 3: The comparison results on the PISC dataset. Previous methods are replicatedwith open-source code to report the accuracy metric. ZS means Zero-Shot.

### > Prompt tuning results with GSPO

### Comparison with existing VLMs on PIPA Dataset

Methods	Acc (%)
BLIP-2 [41]	35.84
LLaVA [ <mark>68</mark> ]	45.12
GPT-4V [55]	59.67
SocialGPT	66.70

Table 4: Comparison with existing Vision-Language Models on the PIPAdataset, with SocialGPT using Vicuna-13B model.

Model	PIPA		PISC			
	SocialGPT	+ GSPO	$\Delta$	SocialGPT	+ GSPO	$\Delta$
Vicuna-7B Vicuna-13B Llama2-7B Llama2-13B	61.58 <b>66.70</b> 31.91 37.86	62.99 <b>69.23</b> 34.07 41.27	+1.41 +2.53 +2.16 +3.41	45.13 65.12 36.71 42.74	49.79 <b>66.19</b> 38.04 48.39	+4.66 +1.07 +1.33 +5.65

Table 5: Prompt tuning results (accuracy in %) with GSPO.

## Conclusion

### > We present SocialGPT, a modular framework with foundation models for social relation reasoning.

- Furthermore, we propose the GSPO for automatic prompt tuning, which further improves the performance.
  - Without additional model training, SocialGPT achieves competitive zero-shot results on two databases while offering interpretable answers, as LLMs can generate language-based explanations for the decisions.
  - Experimental results show that GSPO significantly improves performance, and our method also generalizes to different image styles.
- Our approach opens new avenues for exploring the synergy between vision and language models in highlevel cognitive tasks and offers a promising direction for future advancements in the field of social relation recognition.





## **Thanks For your Attention**