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

Wanhua Li[*,1] Zibin Meng[*,1,2] Jiawei Zhou[3] Donglai Wei[4] Chuang Gan[5,6] Hanspeter Pfister[1]

[1]Harvard University [2]Tsinghua University [3]Stony Brook University
[4]Boston College [5]MIT-IBM Watson AI Lab [6]UMass Amherst

Outline

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

Introduction

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

Figure 1: (a) End-to-end Learning-Based Framework (b) Modular Framework with Foundation Models

Framework

NEURAL INFORMATION PROCESSING SYSTEMS

- **Main Process**
  - Social relation recognition takes an image $\boldsymbol{I}$ and two bounding boxes $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ of two interested individuals as inputs, and requires a model that outputs the social relationship $\boldsymbol{y}$.

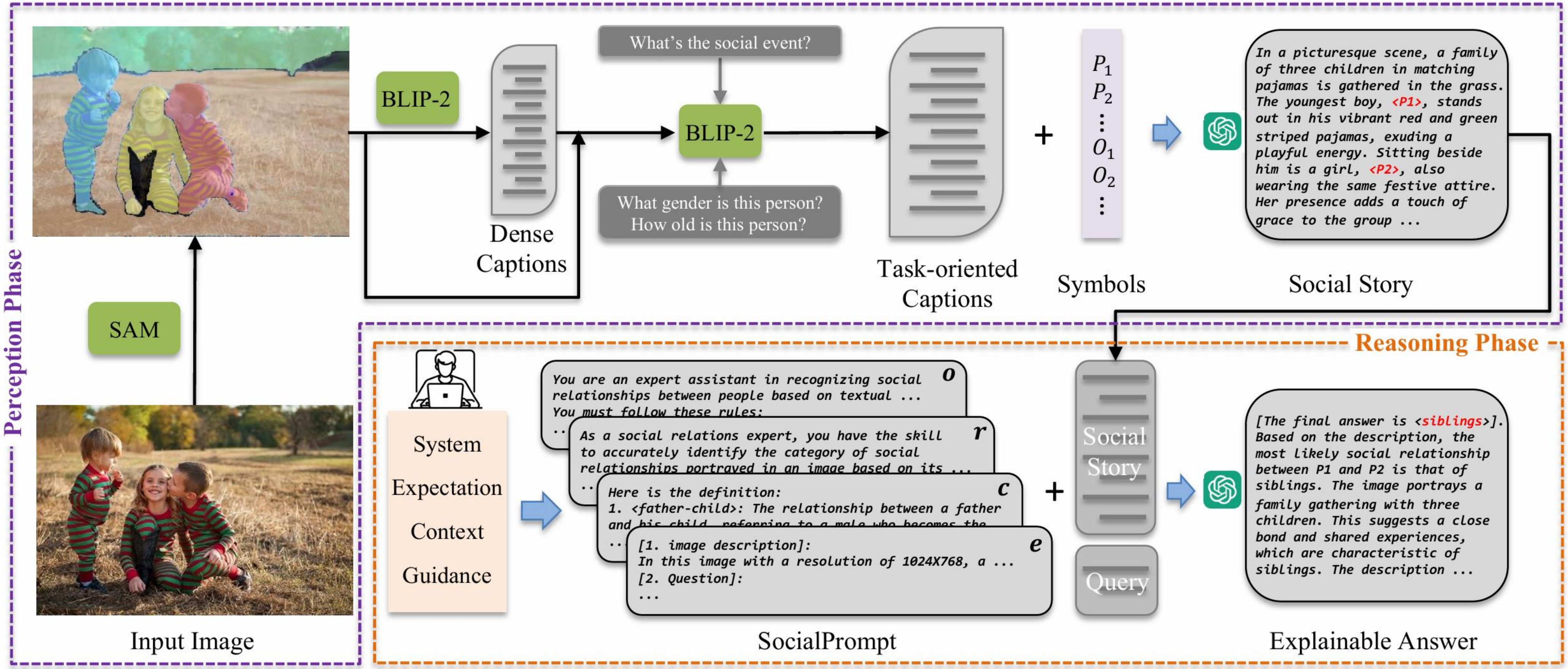

Figure 2: The framework of SocialGPT, which follows the "perception with VFMs, reasoning with LLMs" paradigm.

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.

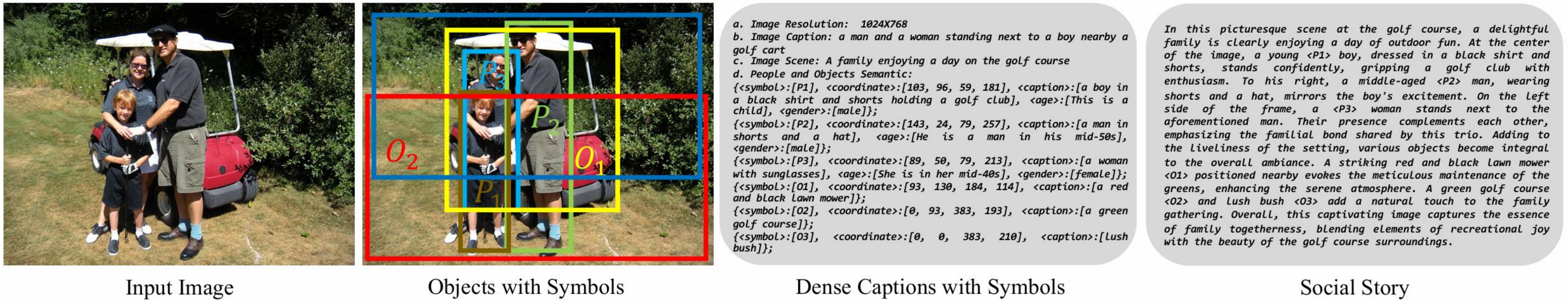

Figure 3: An example of social story generation.

Framework

- **Reasoning with Large Language Models**
  - After obtaining the mapping from image to social story: $\boldsymbol{I} \rightarrow \boldsymbol{S}$, feed both $\boldsymbol{S}$ and bounding box queries $(\boldsymbol{b}_i, \boldsymbol{b}_j)$, converted to textual queries $\boldsymbol{q}$ with referencing symbols P_i, P_j, into LLMs to obtain interpretable answers $\boldsymbol{a}$.
  - 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 $\boldsymbol{a}$ for any given sample $\boldsymbol{z} = (\boldsymbol{S}, \boldsymbol{q})$.
  - We assume that the ground truth answer $\boldsymbol{a}$ for sample $\boldsymbol{z}$ takes the following form: $\boldsymbol{a} = [\boldsymbol{a}^0, \boldsymbol{a}^1, \boldsymbol{a}^2, \ldots]$, where $\boldsymbol{a}^0$ denotes the first sentence of $\boldsymbol{a}$, $\boldsymbol{a}^1$ is the second sentence, and so forth. We specify $\boldsymbol{a}^0$ to have the following fixed format: $\boldsymbol{a}^0 =$ "$The\ final\ answer\ is\ str(\boldsymbol{y})$", where $str(\boldsymbol{y})$ represents the string representation of class label $\boldsymbol{y}$. Then we can define the objective:

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

Greedy Segment Prompt Optimization (GSPO)

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

Algorithm 1 Greedy Segment Prompt Optimization

Input: Initial segments $\boldsymbol{w}_{1:M}$, training dataset $\mathcal{T}$, iteration number N

Build the candidate set $\mathcal{W}_m$ for each segment $\boldsymbol{w}_m$

repeat N times

Randomly sample a batch of data $\mathcal{D}$ from $\mathcal{T}$

for $m = 1, \ldots, M$ **do**

$\mathcal{U}_m :=$ Top-$k(-\sum_{\boldsymbol{z}\in\mathcal{D}} \nabla_{h_{w_m}} \mathcal{L}(\boldsymbol{w}_{1:M}; \boldsymbol{z}))$

▷ *Compute top-k promising segment substitutions*

for $b = 0, 1, \ldots, K * M - 1$ **do**

$\tilde{\boldsymbol{w}}^{(b)}_{1:M} := \boldsymbol{w}_{1:M}$ ▷ *Initialization*

$\tilde{w}^{(b)}_i := \mathcal{U}_i(\lfloor b/M \rfloor)$, where $i = (b \bmod M) + 1$

▷ *Select one replacement segment*

$\boldsymbol{w}_{1:M} := \tilde{\boldsymbol{w}}^{(b^\star)}_{1:M}$, where $b^\star = \mathrm{argmin}_b \sum_{\boldsymbol{z}\in\mathcal{D}} \mathcal{L}(\tilde{\boldsymbol{w}}^{(b)}_{1:M}, \boldsymbol{z})$ ▷ *Compute best replacement*

Output: Optimized segments $\boldsymbol{w}_{1:M}$

Visualization

Reasoning Process and Interpretability

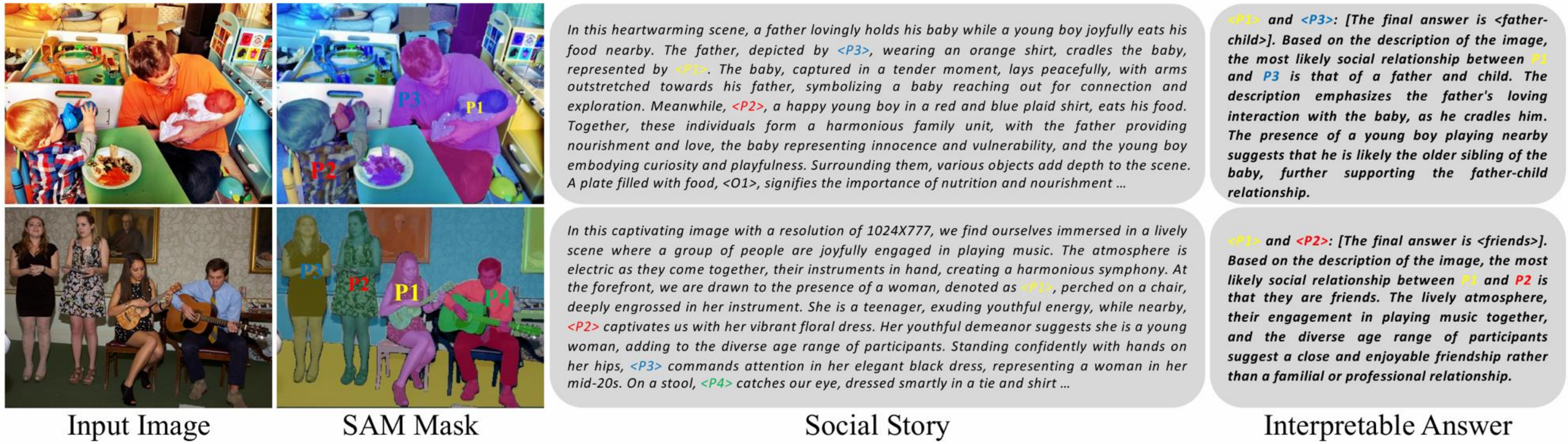

Figure 4: Visualization results of interpretability. We show the SocialGPT perception and reasoning process. We see that our model predicts correct social relationships with plausible explanations.

Generalization on Different Image Styles

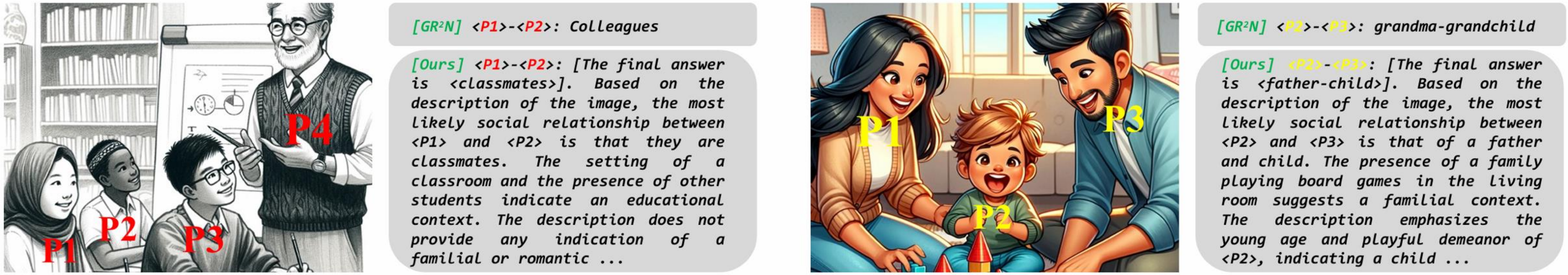

Figure 5: Results when applying SocialGPT to sketch and cartoon images. The images are generated by GPT-4V. Our method generalizes well on these novel image styles.

Experiments

- **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

NEURAL INFORMATION PROCESSING SYSTEMS

Zero-Shot Performance on PIPA Dataset

Methods	ZS	Acc (%)
All attributes + SVM [1]	✗	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**
- Dense Captions	52.63
- Task-oriented Captions	59.89
- Symbol $\rightarrow$ Object Coordinate	57.68
- Symbol $\rightarrow$ Object Caption	59.83
- Social Story	45.31
- SocialPrompt Segment {System}	60.23
- SocialPrompt Segment {Expectation}	59.19
- SocialPrompt Segment {Context}	61.18
- SocialPrompt Segment {Guidance}	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

Zero-Shot Performance on PISC Dataset

Methods	ZS	Acc (%)
Pair CNN [13]	✗	46.30
GRM [6]	✗	64.18
GR^{2}N [3]	✗	64.70
SocialGPT (w/ GPT-3.5)	✓	53.43
SocialGPT (w/ Vicuna-13B)	✓	**65.12**

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

Comparison with existing VLMs on PIPA Dataset

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

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

Prompt tuning results with GSPO

Model	PIPA			PISC		
	SocialGPT	+ GSPO	Δ	SocialGPT	+ GSPO	Δ
Vicuna-7B	61.58	62.99	+1.41	45.13	49.79	+4.66
Vicuna-13B	**66.70**	**69.23**	+2.53	**65.12**	**66.19**	+1.07
Llama2-7B	31.91	34.07	+2.16	36.71	38.04	+1.33
Llama2-13B	37.86	41.27	+3.41	42.74	48.39	+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 high-level cognitive tasks and offers a promising direction for future advancements in the field of social relation recognition.

Thanks For your Attention