Semantic Heat Guided Relational Privacy Inference Based on Panoptic Scene Graph

1st Qi Hao
School of Cyber Science and Engineering
Southeast University
Nanjing, China
230229538@seu.edu.cn

2nd Jie Huang*
School of Cyber Science and Engineering
Southeast University
Purple Mountain Laboratories
Nanjing, China
jhuang@seu.edu.cn

3rd Changhao Ding
School of Cyber Science and Engineering
Southeast University
Nanjing, China
230239281@seu.edu.cn

4th Zeping Zhang
School of Cyber Science and Engineering
Southeast University
Nanjing, China
zhangzp9970@seu.edu.cn

***Abstract*—Privacy is a subjective concept that depends on human perception and contextual interpretation, based on interaction between subject and object. With the increasing occurrence of privacy leaking incidents, awareness about implicit risks has been evolving. The leakage and misuse of relational information between critical objects emerge as core issues in such security events, defined in this study as "relational privacy". In images, relational privacy primarily manifests through semantic relations between object pairs. To identify pairs with higher privacy potential, we propose the concept of "semantic heat". To ensure interpretability and avoid rigid logical judgments, Probabilistic Soft Logic (PSL) is employed to construct semantic heat levels. Scene graph, providing structured semantic representations of image contents, is highly suitable for investigating relational privacy. We utilize panoptic scene graphs to mitigate noise introduced by traditional bounding boxes and leverage contextual information between object pairs. Additionally, a mask cross-attention mechanism guided by textual instruction is proposed to extract interactive features between objects effectively. Finally, a two-stage relation decoder based on a Large Multi-modal Model (LMM) is designed to perform open-set relation prediction and strength judgment. Experimental results demonstrated that the proposed method achieved performance close to the state-of-the-art and showed certain advantages in recall rate, enabling a more comprehensive detection of relational privacy.**

***Index Terms*—Relational Privacy, Semantic Heat, Panoptic Scene Graph Generation, Open-set**

I. Introduction

Privacy is a relatively subjective concept that varies depending on individual perceptions and contextual scenarios. Strictly speaking, privacy is not entirely independent but rather relational, existing between subjects and objects. Privacy objects typically refer to specific entities of concern, such as addresses or identification documents, while privacy subjects denote the entities—individuals or organizations—to whom these objects belong. This inherent subjectivity in privacy stems from variations in the significance attributed to privacy objects by different subjects. Generally, individuals possess vague notions of privacy, yet their judgments regarding privacy breaches are typically decisive, categorized into binary outcomes of private or non-private. However, the escalating number of privacy leakage incidents has shifted public perception from a binary viewpoint to a continuous spectrum between zero and one. A critical observation from numerous privacy breaches indicates that privacy is predominantly inferred rather than merely identified. Prominent objects in images, such as faces, addresses, or documents, are easily recognized and thus preemptively protected. However, implicit semantic information, particularly relational data between critical objects, is frequently overlooked despite being a significant source of privacy risks. We define this implicit semantic information as *relational privacy*. Relational privacy specifically pertains to relations existing between object pairs, inherently founded on object-relation-object triplets, requiring inference based on interactions between paired objects. Notably, relational privacy itself is not necessarily a direct cause of privacy breaches but represents potential risks.

The aforementioned triplet structure naturally aligns with the concept of scene graphs, which express semantic content within images through structured representations. Traditional scene graph generation (SGG) methods detect objects and their relations using bounding boxes, structuring semantic understanding of images. As scene graphs have found broader application [1, 2] in various downstream tasks, limitations of bounding box-based approaches have become evident [3]. Bounding boxes frequently include irrelevant pixels, and isolated object detection neglect contextual interactions between objects and backgrounds. To address these limitations, panoptic scene graph generation (PSG) was proposed, providing pixel-level segmentation of images, significantly reducing errors from irrelevant visual information and enhancing compre-

hensive semantic understanding. Nevertheless, the relatively low recall rates observed in scene graph detection tasks highlight differential capabilities in comprehending diverse relations. Additionally, most current PSG methodologies operate within closed-set environments. Despite high-quality panoptic datasets optimized from COCO annotations, these methods inevitably remain constrained in relational understanding [4]. Meanwhile, relational privacy inference emphasizes meaningful interactions between crucial object pairs. Clearly, not every object pair warrants analysis; for instance, the relation between "man and woman" demands greater attention compared to "laptop and chair." Therefore, determining which object pairs hold greater semantic analytical value becomes essential for relational privacy inference. Recent advancements in Large Multi-modal Models (LMM) have demonstrated robust semantic reasoning capabilities [5–11], providing promising solutions for open-set relation prediction and thus merit exploration in relational privacy inference.

To address these challenges, this paper proposed the concept of *semantic heat*, inspired by the application of Probabilistic Soft Logic (PSL) in social networks. Semantic heat provides a structured metric to assess and filter object pairs with higher relational relevance as subjects for relational privacy inference. Additionally, we propose a novel relation feature extractor guided by textual instructions, designed to capture visual interaction features between paired objects. Lastly, a two-stage multi-modal relation decoder based on LMM is developed, employing generative and evaluative instructions to predict candidate relations and relational tightness for open-set inference.

In summary, the primary contributions of this paper are as follows:

- To address the selection of critical object pairs and the filtering of irrelevant pairs within images, we proposed Semantic Heat as a measure of relational importance among object pairs. Semantic heat levels, influenced by object categories, mask information, and relation strength, were employed to assess the significance of object pairs. Furthermore, we utilized Probabilistic Soft Logic (PSL) to construct semantic heat rankings, thereby ensuring interpretability and scalability in our inference framework. The final semantic heat level allowed for effective identification of important subject-object pairs.
- For more accurate extraction of visual interaction features between object pairs, we proposed an instruction-guided interaction feature extraction method. Specifically designed textual instructions explicitly defined the query tasks, guiding the feature extractor. Additionally, mask cross-attention mechanism was employed to focus the model's attention on interaction regions between object masks, enabling the extraction of representative interaction features.
- To achieve open-set prediction of object pairs' relation and evaluate their semantic tightness, we proposed a two-stage multi-modal relation inference method. It utilized a relation decoder based on a LMM, employing generation and judgment instructions in two stages to predict relations and assess their semantic tightness in order to reduce the occurrence of occasional hallucinations, respectively. Finally, panoptic scene graphs were constructed, achieving relational privacy inference among significant object pairs.

II. Related Work

The present study is inspired by prior research on image privacy and privacy-leaking image detection. Given that relations among objects in an image are increasingly becoming a primary source of privacy leakage, this work focuses on the relations between critical objects. Scene graph, which provides a structured representation of object attributes and inter-object relations, align well with the aim and is thus adopted as the foundational task. However, conventional scene graphs based on bounding boxes fail to accurately capture the visual features of objects and inevitably introduce cluttered noise. Therefore, panoptic scene graph is utilized to offer more precise data support for the mining of relational privacy. Considering the complexity of relations in real-world scenes and the varying sensitivities of semantically similar relations, this work explores a language-model-based open-set relation prediction method built upon the conventional closed-set paradigm.

A. Image Privacy Detection

Previous methods for image privacy detection can be categorized into image-level detection and region/object-level detection based on their granularity. Image-level detection primarily classifies an entire image as either "private" or "public" according to holistic visual features. For instance, Reference [12] utilized deep visual features extracted from various pretrained neural networks (e.g., AlexNet, GoogLeNet, VGG-16, and ResNet), combined with user-generated and automatically generated labels, to perform binary classification of images. Reference [13] extracted visual features from images and combined these with historical user privacy preferences and behavioral data as inputs into a multimodal variational autoencoder (MVAE). Using a Product-of-Experts (PoE) inference network, this approach integrates information from multiple modalities to learn a joint representation between images and users. Earlier studies mainly relied on traditional visual features and metadata, inherently limiting themselves to shallow semantic information within images.

In contrast, region/object-level detection refines semantic understanding at the sub-image level, demonstrating a trend towards privacy inference rather than simple detection. Reference [14] proposed a dynamic region-aware graph convolutional network (GCN), generating region-aware feature maps by clustering spatially correlated feature channels. Self-attention mechanisms and GCNs were utilized to dynamically model correlations between regions, identifying critical areas such as objects, scenes, and textures, to assess potential privacy risks. Reference [15] automatically identified sensitive regions within images based on personalized user privacy preferences.

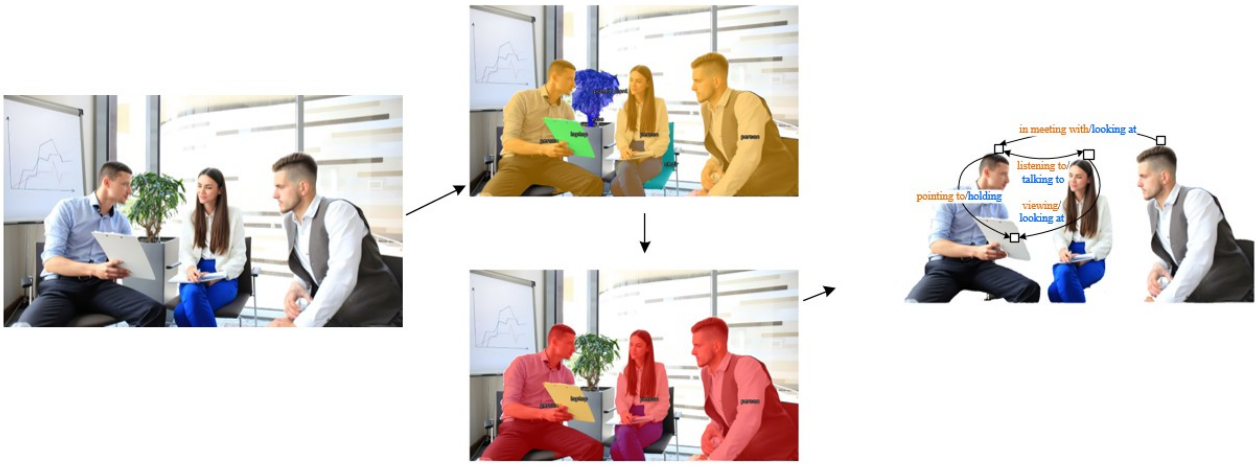

Fig. 1: The left image is the original input. The top-center image illustrates the basic semantic heat constructed from panoptic segmentation. The bottom-center image shows the inferred relational semantic heat-map. The right image presents the final relation prediction results, where orange denotes open-set predictions and blue denotes closed-set ones.

This approach involved constructing a two-layer heterogeneous semantic graph, integrating multi-level semantic features to recognize sensitive content. Reference [16] proposed a street-view image privacy detection and protection method based on Mask R-CNN, leveraging instance segmentation capabilities to automatically detect sensitive content (e.g., faces and license plates) and produce precise segmentation masks, thereby facilitating masking or blurring operations to protect privacy. Reference [17] developed a Graph-based neural networks for Image Privacy (GIP) model, employing graph neural networks (GNN) for explainable inference of image privacy risks. The method primarily focuses on objects and their relations within images, constructing object graphs for privacy risk prediction. Despite progress towards privacy inference, these studies still demonstrate limitations in their understanding and delineation of privacy. There remains insufficient attention given to distinguishing critical relational information from irrelevant data within images, and these methods heavily depend on the quality and accuracy of original datasets and their annotations.

B. Panoptic Scene Graph Generation

Panoptic Scene Graph Generation (PSG) task was first proposed by Yang *et al.* [18], aiming to replace traditional bounding boxes with panoptic segmentation to achieve a more comprehensive and structured representation of scene semantics. This approach integrates both "thing" and "stuff" categories and establishes a high-quality PSG dataset containing approximately 49,000 images. Additionally, four two-stage baseline models were proposed, significantly advancing research in this domain.

To address predicate annotation biases prevalent within the PSG dataset, a semantics-prototype learning strategy was proposed by Li *et al.* [19]. By constructing semantic prototypes, this approach mitigates biases stemming from annotators' linguistic preferences and predicate semantic overlaps, enhancing the discriminative ability of models in relation prediction tasks.

Wang *et al.* [20] proposed Pair-Net, featuring a Pair Proposal Network (PPN) module explicitly designed to learn and select sparse subject-object pairs. Utilizing a lightweight matrix learner, this model directly predicts pairwise relations, resulting in substantial performance improvements in relation prediction within PSG.

The HiLo framework, proposed by Zhou *et al.* [21], addresses the issue of long-tail distributions of relation categories in the PSG task. This method employs separate network branches dedicated to high-frequency and low-frequency relations, respectively. Predictions from these branches are combined during inference, effectively alleviating prediction bias caused by imbalanced data distributions.

Moreover, the VLPrompt model proposed by Zhou *et al.* [22] leverages linguistic cues from large language models (LLMs) to facilitate relation prediction, particularly improving performance in scenarios involving rare relations. This model integrates visual information with linguistic knowledge through an attention-based prompt network, further enhancing the effectiveness of PSG.

However, the above-mentioned studies primarily operate within closed-set scenarios, lacking the capability to handle diverse and complex relations prevalent in real-world situations.

C. Open-Set Scene Graph Generation

A unified framework termed OvSGTR was proposed by Chen *et al.* [23], aiming to achieve fully open-vocabulary scene graph generation. By employing visual-concept alignment and knowledge distillation techniques, this method enables models to recognize previously unseen object and re-

lation categories, thus extending the boundaries of traditional scene graphs.

A two-stage approach proposed by He *et al.* [24] first involves pretraining on large-scale region-caption datasets, followed by prompt-based finetuning without updating model parameters. This technique facilitates inference on novel object categories absent from the training dataset, effectively addressing open-vocabulary scenarios.

The COACHER framework proposed by Kan *et al.* [25] enhances zero-shot relation prediction by integrating external commonsense knowledge graphs. Specifically, a novel graph mining pipeline was developed to model the neighborhoods and relational paths of entities within commonsense knowledge graphs, significantly improving the accuracy of relation prediction in unseen scenarios.

The CaCao framework, described by Yu *et al.* [26], leverages a visually-prompted language model to generate diverse and fine-grained predicates, thereby enhancing the representation of tail predicates in low-resource settings. Additionally, the proposed Epic method facilitates zero-shot generalization of predicates in an open-world context.

Yu *et al.* [27] proposed a zero-shot scene graph generation method that integrates knowledge graph completion techniques. Utilizing structural information from knowledge graphs, the method enables effective inference of unseen objects and relations, improving performance in zero-shot scenarios.

Visual-semantic space (VSS) pretraining was utilized by Zhang *et al.* [28] for language-supervised, open-vocabulary scene graph generation. By parsing linguistic descriptions of images to construct semantic graphs, and mapping these into a visual-semantic embedding space, the approach allows scene graph generation with novel objects and relations even in the absence of precise annotations.

Zhou et al. [29] propose OpenPSG, a framework that addresses open-set panoptic scene graph generation by integrating LMMs. The method combines an open-set panoptic segmentation model with an autoregressive relation decoder, enabling the prediction of both known and novel object relations.

Inspired by the above research, this paper adopts a two-stage method based on LMM to realize open-set relation prediction.

III. THREAT MODEL

This section defines the threat model underlying the proposed relational privacy inference method. The model assumes an adversary capable of analyzing publicly available images to infer sensitive relationships between critical object pairs. The threat model is organized into the following subsections.

A. Adversary Assumptions

The adversary is assumed to have access to images that are publicly shared on social media platforms. It is further assumed that the adversary possesses advanced computational capabilities, including the use of panoptic segmentation models, semantic reasoning frameworks, and LLMs. The adversary does not require any additional metadata or contextual user information beyond the image content itself.

B. Adversarial Goals

The primary objective of the adversary is to uncover sensitive relational information between objects depicted in the image. These relations may include private social, spatial, or other interactions that are not explicitly disclosed. The attack targets implicit relations that can be inferred by analyzing visual interaction features of object pairs within the scene.

C. Attack Procedure

The adversarial process consists of four main stages:

- **Segmentation:** The adversary first performs panoptic segmentation to accurately delineate object instances and background regions, providing a unified understanding of the scene structure.
- **Object Pair Filtering:** A specific module is used to evaluate the significance of each object pair based on category type, occurrence frequency, and mask area, enabling the prioritization of potentially sensitive object pairs.
- **Visual Interaction Feature Extraction:** For each selected pair, features are extracted, focusing on their joint interaction region. This captures context-specific cues essential for inferring relations.
- **Relation Decoding:** A LLM decodes the visual features into structured relation descriptions, revealing implicit or private associations between object pairs.

D. Privacy Risk Implications

The threat lies in the ability to infer non-obvious, sensitive relations that are not explicitly present in the image. By leveraging powerful multi-modal models, the adversary can bridge low-level visual cues with high-level semantic reasoning, resulting in privacy breaches that extend beyond traditional object detection or classification threats.

E. Limitations

The proposed threat model has several limitations:

- **Dependence on Available High-Quality data:** The approach assumes access to publicly shared images with sufficient resolution and object density. In cases of occlusion, low quality, or minimal object presence, the performance of segmentation and interaction analysis may degrade significantly.
- **Subjectivity of Semantic Heat Modeling:** The semantic heat computation is based on predefined rules incorporating category frequency and mask area. These rules may not generalize across different datasets.
- **Reliance on Pre-trained Language Models:** The relation decoding process relies heavily on pre-trained LLMs, which may not accurately infer context-specific or rare relations due to limitations in training data coverage.
- **Lack of Auxiliary Contextual Information:** The model operates solely on visual data and does not incorporate

textual captions, geo-tags, or user profiles, which are often available and leveraged in real-world attacks. Thus, the threat scope is limited to visual inference only.

- **Incomplete Relation Verification Mechanism:** The approach generates relational descriptions without a dedicated verification step. As such, the results may include hallucinated or semantically incorrect relations, especially in ambiguous contexts.

IV. METHODS

A. Task Definition

The inference of relational privacy can be reformulated as the task of generating a panoptic scene graph that explicitly includes semantically important object pairs. The task is defined as follows.

Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width of the image respectively, the objective is to form semantic heat and extract an open-set scene graph that captures semantically salient object relations within the image. Semantic heat serves as an intuitive representation of the semantic importance of objects within an image.

The desired output [29] is an open-set scene graph $\mathcal{G} = (\mathcal{O}, \mathcal{R})$, where:

- $\mathcal{O} = \{o_i\}_{i=1}^{N}$ represents the set of filtered objects in the image. Each object o_i is characterized by:
  - A segmentation mask $\mathbf{M}_i$.
  - A class label c_i.
  - A class frequency f_i.
  - A mask area a_i.
- $\mathcal{R} = \{(o_i, r_{ij}, o_j)\}$ denotes the set of relational triplets, where r_{ij} is the relation label between object pair (o_i, o_j), predicted by a LMM.

For each object pair (o_i, o_j), a related semantic heat h_{ij} is inferred, comprising:

- Base Semantic heat h_i: Derived from the object's class label c_i, class frequency f_i, and mask area a_i.
- Semantic relevance h_{ij}^{r}: Inferred based on the pair's combination type, mask center distance d_{ij}, and relation strength h_{rel}, via PSL rules.

B. Overall Framework

The proposed method comprises four components: panoptic segmentation, semantic heat generation, interaction feature extraction, and relation prediction. As illustrated in Fig 2.

1) Segmentation: A pretrained panoptic segmentation model is utilized to process the input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$. This model outputs:

- Global visual features: $\mathbf{V} \in \mathbb{R}^{H'\times W'\times C}$, where H' and W' denote the spatial dimensions, and C is the number of channels.
- Object class labels: $\{c_i\}_{i=1}^{N}$, where c_i represents the class label of object o_i.
- Object masks: $\{\mathbf{M}_i\}_{i=1}^{N}$, where $\mathbf{M}_i \in \{0,1\}^{\frac{H'}{s}\times\frac{W'}{s}}$ denotes the binary mask for object o_i.

The object class labels $\{c_i\}$ and masks $\{\mathbf{M}_i\}$ are subsequently used to compute semantic heat. The global visual features $\mathbf{V}$ are transformed into visual tokens $\mathbf{T} \in \mathbb{R}^{L\times D}$ through a single convolutional layer with kernel size $s \times s$ and stride s, where $L = \frac{H'}{s} \times \frac{W'}{s}$ is the number of tokens, and D is the token dimension. For each selected object mask $\mathbf{M}_i \in \{0,1\}^{H\times W}$, first apply nearest-neighbor interpolation to resize it to match the spatial dimensions of the global visual feature map $\mathbf{V} \in \mathbb{R}^{H'\times W'\times C}$. The resized mask is denoted as $\tilde{\mathbf{M}}_i \in \{0,1\}^{H'\times W'}$. Subsequently, $\tilde{\mathbf{M}}_i$ is reshaped into a one-dimensional binary vector $\mathbf{m}_i \in \{0,1\}^{L}$, where $L = \frac{H'}{s} \times \frac{W'}{s}$. Collectively, the set of mask vectors for all selected objects forms the mask sequence $\mathbf{M}_{\text{seq}} = \{\mathbf{m}_i\}_{i=1}^{N'}$, where N' denotes the number of selected object masks. These visual tokens $\mathbf{T}$ and mask sequences $\mathbf{M}_{\text{seq}}$ are then input into the Relation Feature Extractor for further processing.

2) Semantic Heat Formation: Semantic Heat (h_{ij}): The overall semantic heat for an object pair (o_i, o_j) is determined utilizing PSL rules through a two-stage process.

PSL is a statistical learning framework designed to reason about the intrinsic connections within data by using weighted logical rules [30]. Unlike traditional binary logic, PSL allows uncertainty and ambiguity by associating continuous confidence levels with logical statements, thus facilitating soft, probabilistic inference.

In PSL, logical rules are formulated as weighted implications, having the general form:

$$w : A \wedge B \rightarrow C \tag{1}$$

where A, B, and C represent logical predicates or atoms, and w denotes the weight of the rule, reflecting its confidence, importance, or relative reliability.

Formally, PSL operates on continuous truth values within the interval $[0, 1]$, unlike traditional logic that uses discrete binary values. The truth of a PSL rule is measured by its distance to satisfaction, quantified by a hinge-loss function. Specifically, for each PSL rule $r : A \rightarrow B$ with weight w_r, the hinge-loss function is defined as:

$$\text{loss}_r = w_r \cdot \max 0, ; \text{truth}(A) - \text{truth}(B) \tag{2}$$

The overall PSL optimization objective is to minimize the total loss across all rules and predicates, which can be formulated as:

$$\min \sum_{r\in\mathcal{RL}} w_r \cdot \max 0, ; \text{truth}(A_r) - \text{truth}(B_r) \tag{3}$$

where $\mathcal{RL}$ denotes the set of all logical rules in the model.

The inference in PSL is performed through convex optimization techniques, resulting in a globally optimal solution efficiently. Due to its probabilistic and flexible structure, PSL is particularly suitable for relational learning tasks where uncertainty and partial truth are inherent to the data.

In this work, PSL is leveraged to systematically adjust semantic heat levels based on characteristics of object pairs.

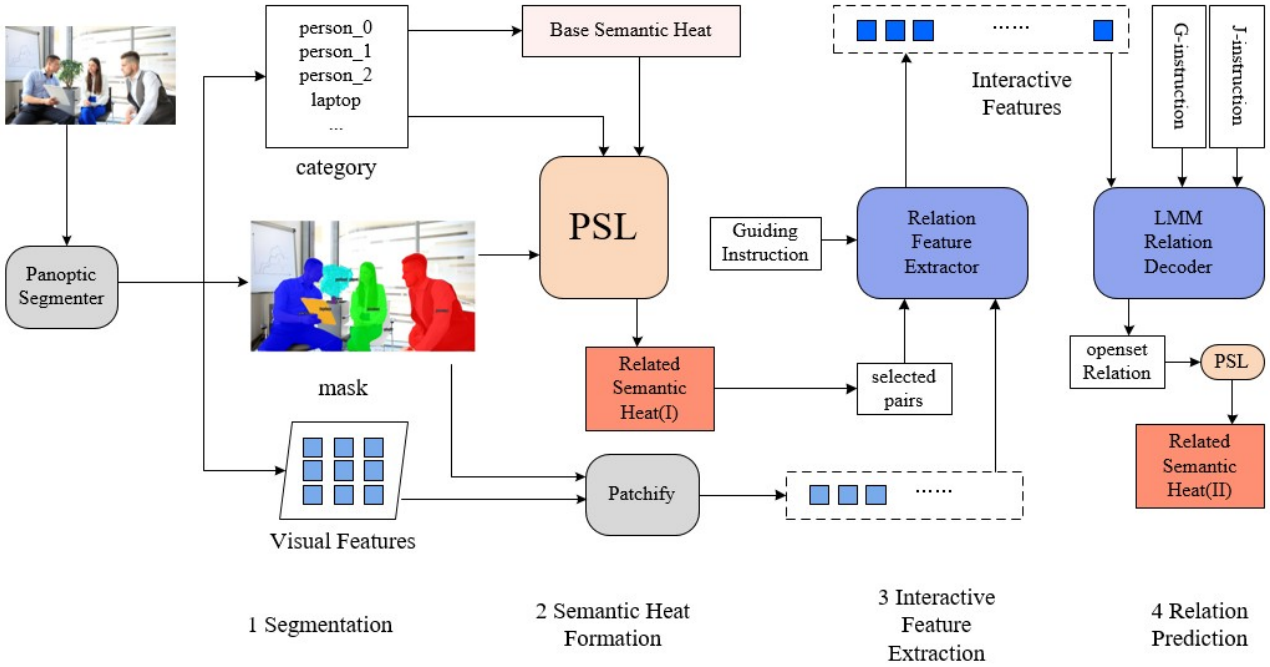

Fig. 2: The overall framework of the method, which comprises four components:panoptic segmentation, semantic heat generation, interaction feature extraction, and relation prediction

a) Base Semantic heat: The base semantic heat for an object o_i is determined by its class label c_i, class frequency f_i, and mask area a_i. The rules are as follows:

- Class Label Category: Classes are manually categorized into five subsets: High (H), Medium High (MH), Medium (M), Medium Low (ML), and Low (L) heat, each assigned an initial semantic heat level $h_{\text{init}}(c_i)$.
  This level division is derived from a series of preliminary experimental tests. The results indicate that a five-level classification is a relatively robust choice, balancing discriminative ability, utilization, and computational efficiency. Fewer levels would result in a crowded semantic heat distribution among object pairs, reducing distinguishability; whereas more levels could lead to fragmented distributions and unnecessary computational overhead.
- Normalization: If the highest category among all classes is below H, all categories are adjusted upward so that the highest becomes H.
- Class Frequency Adjustment: If the class frequency f_i exceeds a predefined threshold f_θ, the heat level is decreased by one level:

$$h_i' = \begin{cases} h_{\text{init}}(c_i) - 1, & \text{if } f_i \geq f_\theta \\ h_{\text{init}}(c_i), & \text{otherwise} \end{cases} \tag{4}$$

The default initial weight is 1, and it will be adjusted during training. PSL rule:

$$\mathsf{InitialHeat}(o,h) \wedge \mathsf{ClassFrequency}(o,f) \wedge \mathsf{GEQ}(f,f_\theta) \rightarrow \mathsf{AdjustedHeat}(o,h-1) \tag{5}$$

This rule enforces that frequently occurring object categories are assigned lower semantic heat levels unless they already have the minimum possible heat score.

- Mask Area Adjustment: If the mask area a_i exceeds a predefined threshold a_θ, the heat level is increased by one level:

$$h_i = \begin{cases} h_i' + 1, & \text{if } a_i \geq a_\theta \\ h_i', & \text{otherwise} \end{cases} \tag{6}$$

- Clipping: Ensure that the final heat level h_i remains within the predefined bounds:

$$h_i = \min(\max(h_i, \mathrm{L}), \mathrm{H}) \tag{7}$$

PSL rule:

$$\mathsf{MaskArea}(o,a) \wedge \mathsf{GEQ}(a,a_\theta) \rightarrow \mathsf{AdjustedHeat}(o,h+1) \tag{8}$$

This rule reflects the intuition that objects occupying larger areas in the image are more visually salient and thus may carry greater semantic significance in relation to privacy inference. Consequently, such objects are assigned a higher heat level if their area surpasses the threshold.

b) Semantic relevance:

1) Initial Semantic Heat Level: Assign the initial semantic heat level based on the lower of the base semantic heat levels of the two objects:

$$h_{ij}^{(0)} = \min(h_i, h_j) \tag{9}$$

PSL rule:

$$\mathsf{BaseHeat}(o_1, h_1) \wedge \mathsf{BaseHeat}(o_2, h_2) \wedge \mathsf{LEQ}(h_1, h_2) \rightarrow \mathsf{PairHeat}(o_1, o_2, h_1) \tag{10}$$

This rule ensures that the pairwise heat level is initialized based on the less significant of the two objects, under the assumption that a relation is only as sensitive as its least sensitive component.

2) Adjustments Based on PSL Rules:
   - Combination Type Adjustment: Modify $h_{ij}^{(0)}$ based on the combination type of the object pair [31]:

$$h_{ij}^{(1)} = \begin{cases} h_{ij}^{(0)}, & \text{if thing-thing} \\ h_{ij}^{(0)} - 1, & \text{if thing-stuff} \\ h_{ij}^{(0)} - 2, & \text{if stuff-stuff} \end{cases} \tag{11}$$

   Ensure that $h_{ij}^{(1)}$ remains within the predefined heat level bounds.
   - Center Distance Adjustment: If the normalized center distance between the object masks exceeds a threshold d_θ, decrease the heat level by one:

$$h_{ij}^{(2)} = \begin{cases} h_{ij}^{(1)} - 1, & \text{if } \frac{d_{ij}}{l_i} \geq d_\theta \\ h_{ij}^{(1)}, & \text{otherwise} \end{cases} \tag{12}$$

   Again, ensure that $h_{ij}^{(2)}$ remains within the predefined heat level bounds.
   - Relation Strength Adjustment: After obtaining the relation strength s_{ij} from the LMM, adjust the heat level accordingly:

$$h_{ij} = \begin{cases} h_{ij}^{(2)} + 1, & \text{if } s_{ij} = \text{tight} \\ h_{ij}^{(2)}, & \text{otherwise} \end{cases} \tag{13}$$

   Ensure that h_{ij} remains within the predefined heat level bounds. The corresponding PSL rules follow a similar formulation as in the previous setup.

3) Interaction Feature Extraction: The **Relation Feature Extractor (RFE)** is designed to extract interaction features [32] for each selected object pair (o_i, o_j) by leveraging the serialized visual tokens $\mathbf{T}$, the corresponding mask sequences $\mathbf{M}_{\text{seq}}$, and guiding instructions. This process enables the model to focus on regions pertinent to the object pair's interaction.

a) Input Preparation::

- **Visual Tokens:** The global visual features $\mathbf{V} \in \mathbb{R}^{H' \times W' \times C}$ are transformed into a sequence of visual tokens $\mathbf{T} \in \mathbb{R}^{L \times D}$ through a convolutional layer with kernel size $s \times s$ and stride s, where $L = \frac{H'}{s} \times \frac{W'}{s}$ and D is the token dimension.
- **Mask Sequences:** Each object mask $\mathbf{M}_i$ is resized via nearest-neighbor interpolation to match the spatial dimensions of $\mathbf{V}$, resulting in $\tilde{\mathbf{M}}_i \in \{0,1\}^{H' \times W'}$. These are then reshaped into one-dimensional binary vectors $\mathbf{m}_i \in \{0,1\}^L$, forming the mask sequence $\mathbf{M}_{\text{seq}} = \{\mathbf{m}_i\}_{i=1}^{N'}$.
- **Pair Masks:** For each selected object pair (o_i, o_j), a combined mask $\mathbf{m}_{ij} = \mathbf{m}_i \vee \mathbf{m}_j$ is computed, highlighting the union of their spatial regions.
- **Guiding Instructions:** For each object pair, a textual instruction is crafted to guide the pair feature extraction. For example: *"Extract interactive features of* {*subject*}-{*object*} *based on visual features and the masks."* The instruction is tokenized and embedded into $\mathbf{I}_{ij} \in \mathbb{R}^{L_I \times D}$, where L_I is the instruction length and D is the embedding dimension.

b) Feature Extraction Process:

1) **Query Initialization:** For each object pair (o_i, o_j), a learnable pair feature extraction query $\mathbf{Q}_{ij}^{\text{feat}} \in \mathbb{R}^{E \times D}$ is initialized, where E is the number of query tokens.
2) **Self-Attention Encoding:** The pair query and instruction embeddings are concatenated along the sequence dimension and fed into a self-attention layer:

$$\mathbf{F}_{\text{SA}}^{\text{feat}} = \text{Trunc}\left(\text{SA}\left(\text{Concat}\left(\mathbf{Q}_{ij}^{\text{feat}}, \mathbf{I}_{ij}\right)\right), E\right) \tag{14}$$

   where $\text{SA}(\cdot)$ denotes the self-attention operation, and $\text{Trunc}(\cdot, E)$ extracts the first E features corresponding to the updated query tokens.
3) **Mask-Guided Cross-Attention:** The updated query features attend to the visual token sequence $\mathbf{T} \in \mathbb{R}^{L \times D}$, guided by the combined object pair mask $\mathbf{m}_{ij}$:

$$\mathbf{F}_{\text{CA}}^{\text{feat}} = \text{MaskCA}\left(\mathbf{F}_{\text{SA}}^{\text{feat}}, \mathbf{T}, \mathbf{m}_{ij}\right) \tag{15}$$

   where $\text{MaskCA}(\cdot)$ denotes the mask-guided cross-attention mechanism.
4) **Feed-Forward Refinement:** The attended features are passed through a feed-forward network to obtain the final interaction feature vector:

$$\mathbf{f}_{ij} = \text{FFN}\left(\mathbf{F}_{\text{CA}}^{\text{feat}}\right) \tag{16}$$

   where $\text{FFN}(\cdot)$ denotes a two-layer feed-forward network with non-linear activation.

This interaction feature $\mathbf{f}_{ij}$ encapsulates the contextual and spatial information necessary for subsequent relation prediction tasks.

4) Relation Prediction: In this stage, a Relation Decoder based on a LMM is employed to predict open-set relations between object pairs. Utilizing the interaction features $\mathbf{f}_{ij}$ obtained from the Relation Feature Extractor (RFE), two types of instructions are designed to guide the Relation Decoder:

Generation Instruction (G-instruction) and Judgment Instruction (J-instruction). These instructions enable the decoder to perform relation prediction in an auto-regressive manner.

The candidate relations used for the stage are generated through the following procedure. Firstly, a set of preliminary relation candidates is produced by the Relation Decoder using the Generation Instruction, as described previously. Specifically, given the interaction features $\mathbf{f}_{ij}$ for each object pair (o_i, o_j), the Relation Decoder autoregressively generates possible relations conditioned on these features and the corresponding instruction embeddings. The output relations are organized as a candidate set, potentially containing multiple relations per pair, separated by a special delimiter token.

Subsequently, the Judgment Instruction is applied to each candidate relation individually to verify its semantic tightness. By iteratively evaluating each candidate relation through the judgment stage, implausible or weakly-supported relations are filtered out. The remaining candidate relations constitute the final set of predicted relations $\mathcal{R}$ for each object pair.

This two-step candidate generation and filtering process ensures comprehensive exploration of possible relations while maintaining high semantic precision.

a) Generation Instruction: The Generation Instruction prompts the Relation Decoder to generate potential relations between object pairs. For each object pair (o_i, o_j), a textual prompt is constructed, such as:

"Infer the relation between {object i} and {object j} based on their interaction features."

This instruction is tokenized into embeddings $\mathbf{I}_{ij}^{\mathrm{G}} \in \mathbb{R}^{L_{\text{gen}} \times D}$, where L_{gen} denotes the length of the tokenized instruction. The interaction feature $\mathbf{f}_{ij}$ and the instruction embeddings $\mathbf{I}_{ij}^{\mathrm{G}}$ are concatenated and input into the Relation Decoder:

$$\hat{r}_{ij}^{\mathrm{G}} = \mathrm{RD}([\mathbf{f}_{ij}; \mathbf{I}_{ij}^{\mathrm{G}}]) \tag{17}$$

The decoder outputs a sequence of relation labels, potentially containing multiple relations separated by a special token (e.g., [SEP]).

b) Judgment Instruction: The Judgment Instruction guides the Relation Decoder to determine whether a specific relation r between an object pair is semantically tight. A prompt is constructed as:

"Judge whether the relation {relation} between {object i} and {object j} is tight."

This instruction is tokenized into embeddings $\mathbf{I}_{ij}^{\mathrm{J}}(r) \in \mathbb{R}^{L_{\text{judge}} \times D}$. Combined with $\mathbf{f}_{ij}$, it is input into the decoder to yield a binary prediction:

$$\hat{y}_{ij}^{\mathrm{J}}(r) = \mathrm{RD}([\mathbf{f}_{ij}; \mathbf{I}_{ij}^{\mathrm{J}}(r)]) \tag{18}$$

Here, $\hat{y}_{ij}^{\mathrm{J}}(r) \in [0, 1]$ denotes the predicted confidence score for whether relation r is tight between o_i and o_j.

By iterating over a set of candidate relations, the set of predicted relations $\mathcal{R}$ for the image is constructed.

This approach leverages the Relation Decoder's capability, built upon the LMM, to understand complex interactions and perform open-set relation prediction effectively.

Compared to a unified decoding approach that attempts to predict relation types directly in a single step, this two-stage formulation explicitly separates candidate generation and relation verification. This separation brings several advantages:

1) more interpretable intermediate outputs;
2) better integration with mask-guided cross attention and instruction tuning;
3) more flexible and extensible verification strategies, especially in open-set settings.

This design allows the system to reject implausible relations even when object features are visually similar, which would be more difficult for a single-stage decoder to handle robustly.

5) Loss Function: The proposed model adopts a two-stage reasoning mechanism guided by Generation Instructions (G-inst) and Judgment Instructions (J-inst). Accordingly, the loss function is formulated as a combination of two supervised objectives.

a) Relation Generation Loss (G-inst): To optimize the model's ability to generate relation sequences under the guidance of generation instructions, we apply a cross-entropy loss between the predicted relation sequence $\hat{r}_{ij}^{\mathrm{G}}$ and the ground-truth label r_{ij} for each object pair (o_i, o_j):

$$\mathcal{L}_{\text{gen}} = \sum_{(i,j)} \mathrm{CE}(\hat{r}_{ij}^{\mathrm{G}}, r_{ij}) \tag{19}$$

Here, $\mathrm{CE}(\cdot)$ denotes the token-level cross-entropy loss over the decoder outputs conditioned on G-inst.

b) Relation Judgment Loss (J-inst): To determine whether a generated relation r reflects a semantically tight connection between objects o_i and o_j, a binary classification task is performed. The decoder, under J-inst guidance, predicts a binary confidence score $\hat{y}_{ij}^{\mathrm{J}}(r)$, supervised with binary cross-entropy:

$$\mathcal{L}_{\text{judg}} = \sum_{(i,j),r} \mathrm{BCE}(\hat{y}_{ij}^{\mathrm{J}}(r), y_{ij}^{r}) \tag{20}$$

where $y_{ij}^{r} \in \{0, 1\}$ denotes whether relation r is considered tight between objects o_i and o_j.

c) Overall Loss: The final objective is a weighted combination of generation and judgment losses:

$$\mathcal{L} = \lambda_{\text{gen}} \cdot \mathcal{L}_{\text{gen}} + \lambda_{\text{judg}} \cdot \mathcal{L}_{\text{judg}} \tag{21}$$

where λ_{gen} and λ_{judg} are hyperparameters balancing the two loss components.

V. Experiments

A. Dataset

In this study, experiments were conducted using the Panoptic Scene Graph (PSG) dataset. This dataset was selected due to its higher annotation accuracy and reliability compared to the Visual Genome (VG) dataset. Although VG provided

extensive annotations, its relatively lower precision in annotations made it unsuitable for the requirements of this research. To better address the challenges of relational privacy, we further refined the original PSG dataset. Statistical analysis revealed that approximately 55% of images contained the category "person". We reduced the total number of images from 49,000 to 41,000, selectively retaining samples to ensure that over 65% of the images in both the training and validation sets contain at least one "person" object. This step filtered the data by querying the segmentation information in psg.json file.

B. Tasks and Metrics

a) Tasks: In the general setting of scene graph generation, three sub-tasks are commonly defined: predicate prediction, scene graph classification, and scene graph detection. However, since panoptic segmentation models are unable to utilize ground-truth object masks for classification, this study focuses on the following two tasks:

- **Predicate Prediction:** Given the ground-truth object masks and class labels, the model is required to predict the correct relations between object pairs.
- **Scene Graph Detection:** The model simultaneously detects object segments and predicts their relations from the input image.

Following the open-set evaluation setting proposed in OpenPSG, the full set of relation categories is split into two subsets: **base relations** and **novel relations**, with a 7:3 ratio. The model is trained using only base relations and evaluated on both base and novel relations, in order to assess its generalization ability in open-set relation prediction.

b) Metrics: To comprehensively evaluate model performance, the following metrics are adopted:

- **Recall@K (R@K):** Measures the proportion of ground-truth relations that are correctly retrieved within the top-K predictions.
- **mean Recall@K (mR@K):** Calculates the average Recall@K across all relation categories, giving equal weight to each class regardless of its frequency. This metric is especially useful for evaluating the model's performance on long-tailed or rare relations.

Precision@K and F1@K metrics were not included in our evaluation primarily due to the open-set nature of the panoptic scene graph generation task, where the ground truth annotations do not exhaustively cover all possible valid relations. This inherently results in ambiguity regarding false positives, making Precision@K and F1@K potentially misleading. Additionally, due to the high annotation uncertainty and inconsistency in labeling subtle or implicit relations, these metrics might not reliably reflect actual model performance. Instead, Recall@K and mean Recall@K are preferred as they provide a clearer indication of the model's ability to capture relevant relations without being adversely affected by incomplete annotations.

TABLE I: Comparison on Predicate Classification with excellent methods considering closed-set and open-set

	Predicate Classification closed-set		
Method	R/mR@20	R/mR@50	R/mR@100
Motifs	42.5/19.3	48.1/21.3	50.3/22.6
VCTree	42.9/20.1	48.8/21.9	50.6/22.9
OpenPSG	53.5/36.9	68.2/50.6	76.5/61.0
SemH	51.4/36.4	65.6/51.8	75.9/62.7
	Predicate Classification open-set		
OpenPSG	44.3/28.7	53.9/37.6	60.2/44.8
SemH	43.6/28.3	54.2/38.3	62.4/46.0

C. Implementation Configuration

This study adopted OpenSeeD [33] as the panoptic segmenter. The choice of OpenSeeD as the panoptic segmenter was motivated by its state-of-the-art performance and robust generalization capabilities in open-vocabulary scenarios. OpenSeeD leveraged a unified segmentation approach capable of accurately handling both semantic and instance segmentation tasks, providing high-quality panoptic segmentation outputs essential for reliable relation prediction. Furthermore, OpenSeeD's design integrated seamlessly with language-guided mechanisms, facilitating effective integration with the downstream Relation Decoder based on Language Models (LMM). Its proven effectiveness in recent benchmarks and studies made it an optimal choice to ensure accurate and generalizable segmentation results, directly contributing to the robustness and efficacy of the subsequent semantic heat analysis and relation inference processes.

In semantic heat modeling, the category frequency threshold f_θ was set to 0.6, the mask area threshold a_θ was set to 40000, and the normalized spatial distance threshold d_θ was set to 600. These thresholds are determined by Preliminary statistical analysis to balance category frequency, object saliency, and spatial proximity, ensuring the discriminability and robustness of heat level assignment. The kernel size s of the single convolutional layer was set to 8. For pair feature extraction, the number of tokens was set to 32. We selected OpenLLaMA-3B as the core component of the Relation Decoder [34]. The hyperparameters λ_{gen} and λ_{judge} were set to 1 and 2, respectively. The model was optimized using the AdamW [35] optimizer with a learning rate of 1×10^{-5} and a weight decay of 0.05. Training was conducted for a total of 8 epochs. The experiments used `gpu_v100` queue with a total of six NVIDIA V100 GPUs, each equipped with 32 GB of memory. The CUDA compiler version used in experiments was 11.7.

D. Comparison with State-of-the-Art Methods

Tables I and II compare the proposed **SemH** method with several state-of-the-art baselines(Motifs [36], VCTree [37], PSGTR [18], PairNet [20]) on the PSG dataset, evaluated under both closed-set and open-set settings. Two core tasks are considered: *Predicate Classification (PredCls)* and *Scene Graph Detection (SGDet)*.

TABLE II: Comparison on Scene Graph Detection with excellent methods considering closed-set and open-set

	Scene Graph Detection closed-set		
Method	R/mR@20	R/mR@50	R/mR@100
Motifs	19.7/8.8	20.5/9.4	21.0/9.6
VCTree	20.1/9.4	21.1/10.0	21.6/10.1
PSGTR	27.9/15.8	33.3/19.6	35.1/21.4
PairNet	29.5/23.9	35.5/27.7	39.6/30.0
OpenPSG	37.2/32.0	45.8/40.0	51.5/48.0
SemH	32.5/27.6	37.2/33.4	42.1/37.9
	Scene Graph Detection open-set		
OpenPSG	25.5/19.2	31.0/23.8	35.9/25.5
SemH	20.1/15.5	24.6/17.1	27.3/18.0

Predicate Classification. As shown in Table I, SemH achieves competitive performance in the closed-set setting. Compared to Motifs and VCTree, which struggle with long-tailed relation distributions, SemH significantly improves mean Recall (mR), and achieves comparable or better results than OpenPSG. Notably, SemH outperforms OpenPSG on mR@50 (51.8 vs. 50.6) and mR@100 (62.7 vs. 61.0), demonstrating its effectiveness in identifying less frequent but semantically important relations through semantic heat-based object pair filtering.

In the open-set setting, where unseen relation categories appear during evaluation, SemH maintains strong generalization. Although its overall Recall is marginally lower than OpenPSG at R@20 (43.6 vs. 44.3), SemH consistently achieves higher mean Recall across all thresholds (e.g., 46.0 vs. 44.8 at mR@100), indicating that the semantic heat prior helps the model generalize better to rare or unseen relations.

Scene Graph Detection. As reported in Table II, the performance trend shifts in the SGDet task. While SemH still surpasses earlier models such as Motifs, VCTree, and PSGTR in both R and mR, it underperforms compared to OpenPSG and PairNet. This performance gap is primarily due to the fact that SemH explicitly filters object pairs based on semantic heat before relation prediction. Although this filtering can be relaxed by adjusting the heat threshold to include more candidate combinations, SemH prioritizes precision in discovering truly meaningful relations over exhaustive recall. As a result, in SGDet—where ground-truth objects and their combinations must be recovered without prior knowledge—SemH sacrifices some coverage, leading to a noticeable drop in mR (e.g., 37.9 vs. 48.0 at mR@100). This trade-off reflects a deliberate design decision: emphasizing semantic significance and reducing redundant or low-quality pairings, which is more aligned with the goals of high-level scene understanding rather than raw detection completeness.

In summary, SemH demonstrates superior performance in relation-centric evaluation (PredCls), especially in handling long-tail and open-set challenges. While its SGDet performance declines due to aggressive object pair filtering, the method remains competitive and interpretable, offering a principled balance between precision and recall guided by semantic priors.

In an additional experiment, we further evaluate the effectiveness of our proposed SemH method in detecting sensitive relations from images containing human subjects. Based on common sense and practical observations, we categorized the relations in the dataset into two types: sensitive relations (e.g., "kissing," "touching," and "talking to") and non-sensitive relations (e.g., "beside," "running on," and "eating"). We randomly selected 100 images containing "person" objects from social networks and performed relation detection using both OpenPSG and SemH methods. The detected relations were classified into sensitive or non-sensitive categories using a similar two-stage classification approach previously employed.

The results, summarized in Figure 4, indicated that OpenPSG detected a total of 845 relations, among which 367 were sensitive relations. In comparison, SemH detected fewer total relations (793) but identified more sensitive relations (419). This suggests that SemH is more focused and effective in capturing sensitive semantic interactions within scenes involving human subjects. The higher proportion of sensitive relations identified by SemH (approximately 52.8%) compared to OpenPSG (approximately 43.4%) further confirms the improved sensitivity and relevance of SemH in detecting contextually important semantic interactions in privacy-critical scenarios.

E. Ablation Study

To investigate the impact of the panoptic segmenter on the overall PSG performance, we conducted an ablation study by replacing OpenSeeD with a standard Mask2Former [38] under the same model configuration. As shown in Table III, we observed that OpenSeeD outperforms Mask2Former in both Predicate Classification and Scene Graph Detection metrics, indicating more accurate instance and stuff segmentation.

More importantly, this improvement in segmentation translated into stronger performance on relation prediction. OpenSeeD yielded consistent gains in both Recall and Mean Recall across different top-K settings. These results suggested that the segmentation quality provided by OpenSeeD was better aligned with downstream relational reasoning, and the scene graph generator benefited from more precise object boundaries and semantics.

As reported in Table IV, the complete **SemH** model is compared against four ablated variants: (1) without semantic heat filtering (*No SemHeat*), (2) without the use of guiding instructions (*No guiding instr*), (3) with the judgment loss disabled by setting $\lambda_{\text{judge}} = 0$, and (4) replacing the mask-guided cross attention module with a simpler mask pooling strategy (*No Mask-A*).

Removing the semantic heat mechanism leads to a significant decrease in both recall and mean recall. For instance, mR@50 drops from 51.8 to 39.4, and mR@100 from 62.7 to 49.5. This indicates that semantic heat-based filtering effectively eliminates uninformative object pairs, thereby enhancing

Fig. 3: Visualization results. The left image shows the original input, the middle one highlights the important object pairs identified through semantic heat inference, and the right one presents the results of open-set relation prediction.

TABLE III: Ablation study of segmenter

	Predicate Classification				Scene Graph Detection		
Segmenter	PQ	R/mR@20	R/mR@50	R/mR@100	R/mR@20	R/mR@50	R/mR@100
SemH	55.4	51.4/36.4	65.6/51.8	75.9/62.7	32.5/27.6	37.2/33.4	42.1/37.9
Mask2Former	51.7	50.5/35.8	64.8/50.1	74.3/62.4	30.2/24.6	35.5/32.7	39.1/35.0

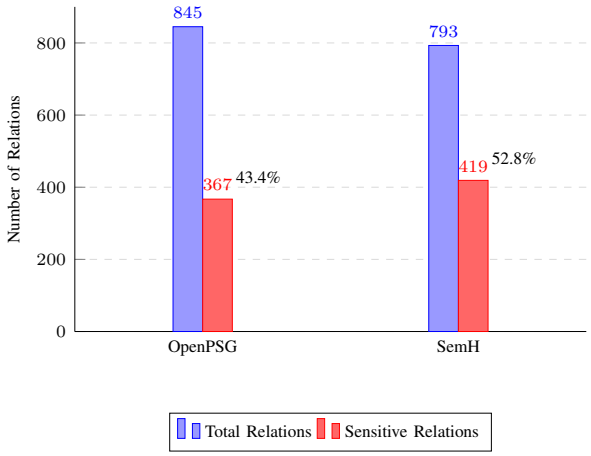

Fig. 4: Grouped bar chart comparing total and sensitive relations detected by OpenPSG and SemH. Sensitive ratios are shown beside the red bars.

TABLE IV: Ablation study of modules

	Predicate Classification		
Method	R/mR@20	R/mR@50	R/mR@100
SemH	51.4/36.4	65.6/51.8	75.9/62.7
No SemHeat	38.5/28.9	50.1/39.4	61.3/49.5
No Mask-A	45.3/31.2	60.0/46.9	71.4/58.8
No guiding instr	44.3/31.7	58.6/44.3	67.4/55.0
λ_{judge}=0	48.3/35.4	62.1/49.5	73.0/61.2

the model's ability to capture meaningful and long-tailed relational patterns.

Excluding guiding instructions results in a consistent decline in performance across all thresholds. Without structured language prompts, the model's ability to align semantic and visual cues is weakened, leading to a drop in mR@50 from 51.8 to 44.3. This highlights the importance of instruction-guided reasoning in improving relation understanding.

Disabling the judgment loss branch by setting $\lambda_{\text{judge}} = 0$ causes a moderate reduction in mean recall (e.g., mR@50 decreases from 51.8 to 49.5), suggesting that judgment-based supervision plays a supportive role in refining relation predic-

tions, especially in open-set configurations.

Replacing the mask-guided cross attention mechanism with a baseline mask pooling strategy (*No Mask-A*) also leads to a noticeable performance drop. In this variant, instead of applying cross-attention between the textual query and the region-specific masked features, we extract object-level features by computing the average of visual features within each object mask region. These pooled features are then used for relation decoding. The resulting performance, with mR@50 dropping from 51.8 to 46.9, demonstrates that explicitly modeling cross-modal interactions at the masked region level yields more discriminative features than simple spatial pooling.

These results collectively demonstrate that semantic heat filtering, instruction-driven reasoning, mask-guided cross attention, and dual-branch loss design each contribute meaningfully to the performance of the proposed model. The complete configuration achieves the most favorable balance between precision and generalization.

F. Visualization

As shown in Fig 3, the proposed method effectively selects more semantically meaningful object pairs from the image and performs open-set relation prediction.

VI. Conclusion

In this paper, we proposed a novel framework for relational privacy inference by leveraging panoptic scene graph generation guided by semantic heat. The concept of relational privacy was defined to explicitly characterize privacy risks arising from semantic interactions between object pairs in images. To quantify and select the more significant object pairs, we proposed the method of semantic heat, employing Probabilistic Soft Logic (PSL) to ensure interpretability and flexibility in semantic heat ranking.

To facilitate accurate inference of relational privacy, we developed a Relation Feature Extractor guided by textual instructions, employing mask cross-attention to extract interaction-focused visual features between objects. Additionally, a two-stage LMM-based relation decoder was implemented, which leverages generation and judgment instructions (G-inst and J-inst) to perform open-set relation prediction and semantic tightness evaluation.

Experimental results validated that the proposed method effectively addresses challenges in open-set relation prediction, demonstrating performance competitive with state-of-the-art approaches, particularly excelling at low-frequency relation prediction and precise identification of critical object interactions. Additionally, a comparison based on randomly selected social network images containing human subjects demonstrated that the proposed SemH method, despite detecting fewer total relations compared to OpenPSG, identified a higher number and proportion of sensitive relations. This highlights SemH's enhanced capability in focusing on semantically important and privacy-sensitive interactions.

Future research will focus on refinement of semantic heat formation, exploring more sophisticated logic rules within PSL, and relational privacy among multiple images.

Acknowledgment

We thank Southeast University and Purple Mountain Laboratory. We thank the Big Data Computing Center of Southeast University for sup porting computing resources. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9962–9971.

[2] M. Hildebrandt, H. Li, R. Koner, V. Tresp, and S. Günnemann, "Scene graph reasoning for visual question answering," *arXiv preprint arXiv:2007.01072*, 2020.

[3] X. Dong, T. Gan, X. Song, J. Wu, Y. Cheng, and L. Nie, "Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 427–19 436.

[4] T. Gupta, A. Kamath, A. Kembhavi, and D. Hoiem, "Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 399–16 409.

[5] J. Huang, J. Zhang, K. Jiang, H. Qiu, and S. Lu, "Visual instruction tuning towards general-purpose multimodal model: A survey," *arXiv preprint arXiv:2312.16602*, 2023.

[6] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 084–14 093.

[7] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[8] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Open and efficient foundation language models," *Preprint at arXiv. https://doi. org/10.48550/arXiv*, vol. 2302, no. 3, 2023.

[11] L. Yao, J. Han, X. Liang, D. Xu, W. Zhang, Z. Li, and H. Xu, "Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 497–23 506.

[12] A. Tonge and C. Caragea, "Image privacy prediction using deep neural networks," *ACM Transactions on the Web (TWEB)*, vol. 14, no. 2, pp. 1–32, 2020.

[13] N. Vishwamitra, Y. Li, H. Hu, K. Caine, L. Cheng, Z. Zhao, and G.-J. Ahn, "Towards automated content-based photo privacy control in user-centered social networks," in *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*, 2022, pp. 65–76.

[14] G. Yang, J. Cao, Q. Sheng, P. Qi, X. Li, and J. Li, "Drag: Dynamic region-aware gcn for privacy-leaking image detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 11, 2022, pp. 12 217–12 225.

[15] R. Jiao, L. Zhang, and A. Li, "Ieye: Personalized image privacy detection," in *2020 6th International Conference on Big Data Computing and Communications (BIGCOM)*. IEEE, 2020, pp. 91–95.

[16] J. Yu, M. Wu, C. Li, and S. Zhu, "A street view image privacy detection and protection method based on mask-rcnn," in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 9. IEEE, 2020, pp. 2184–2188.

[17] G. Yang, J. Cao, Z. Chen, J. Guo, and J. Li, "Graph-based neural networks for explainable image privacy inference," *Pattern Recognition*, vol. 105, p. 107360, 2020.

[18] J. Yang, Y. Z. Ang, Z. Guo, K. Zhou, W. Zhang, and Z. Liu, "Panoptic scene graph generation," in *European Conference on Computer Vision*. Springer, 2022, pp. 178–196.

[19] L. Li, W. Ji, Y. Wu, M. Li, Y. Qin, L. Wei, and R. Zimmermann, "Panoptic scene graph generation with semantics-prototype learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 4, 2024, pp. 3145–3153.

[20] J. Wang, Z. Wen, X. Li, Z. Guo, J. Yang, and Z. Liu, "Pair then relation: Pair-net for panoptic scene graph generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[21] Z. Zhou, M. Shi, and H. Caesar, "Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 637–21 648.

[22] ——, "Vlprompt: Vision-language prompting for panoptic scene graph generation," *arXiv preprint arXiv:2311.16492*, 2023.

[23] Z. Chen, J. Wu, Z. Lei, Z. Zhang, and C. W. Chen, "Expanding scene graph boundaries: fully open-vocabulary scene graph generation via visual-concept alignment and retention," in *European Conference on Computer Vision*. Springer, 2024, pp. 108–124.

[24] T. He, L. Gao, J. Song, and Y.-F. Li, "Towards open-vocabulary scene graph generation with prompt-based finetuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 56–73.

[25] X. Kan, H. Cui, and C. Yang, "Zero-shot scene graph relation prediction through commonsense knowledge integration," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*. Springer, 2021, pp. 466–482.

[26] Q. Yu, J. Li, Y. Wu, S. Tang, W. Ji, and Y. Zhuang, "Visually-prompted language model for fine-grained scene graph generation in an open world," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 560–21 571.

[27] X. Yu, R. Chen, J. Li, J. Sun, S. Yuan, H. Ji, X. Lu, and C. Wu, "Zero-shot scene graph generation with knowledge graph completion," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[28] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2915–2924.

[29] Z. Zhou, Z. Zhu, H. Caesar, and M. Shi, "Openpsg: Open-set panoptic scene graph generation via large multimodal models," in *European Conference on Computer Vision*. Springer, 2024, pp. 199–215.

[30] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor, "Hinge-loss markov random fields and probabilistic soft logic," *Journal of Machine Learning Research*, vol. 18, no. 109, pp. 1–67, 2017.

[31] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1209–1218.

[32] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[33] H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Yang, and L. Zhang, "A simple framework for open-vocabulary segmentation and detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1020–1031.

[34] X. Geng and H. Liu, "Openllama: An open reproduction of llama," May 2023. [Online]. Available: https://github.com/openlm-research/open_llama

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[36] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5831–5840.

[37] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6619–6628.

[38] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

SUPPLEMENTARY MATERIAL

A. Division of relations in PSG dataset

To evaluate the open-set relation prediction performance, we partition the original 56 relations in the PSG dataset into base relations and novel relations with a ratio of 7:3. In this setting, the novel relations are selected to be semantically more complex compared to the base relations.

```
base_rel
Spatial relations:
    on, in, beside, attached to, over, in front of
action:
    holding, wearing, sitting on, standing on, walking on, running on,
    lying on, leaning on, carrying, looking at, guiding, feeding, biting,
    catching, picking, playing with, chasing, climbing, touching,
    pushing, pulling, opening, driving, riding, parked on, driving opening
other:
    hanging from, on the back of, falling off, going down, painted on

novel_rel
action:
    cooking, talking to, throwing, slicing, jumping over, jumping from, kissing,
    eating, drinking, cleaning, playing, about to hit, kicking, swinging
other:
    entering, exiting, enclosing

all_rel=base_rel+novel_rel
```

Fig. 5

B. Guiding instruction

We utilize six different guiding instructions to enhance the model's robustness. During training, one instruction is randomly selected for each iteration, while the last instruction is consistently used during evaluation.

```
"Identify the interaction between {subject} and {object} using visual cues and mask-based regions."

"Focus on the region-specific interaction of {subject} and {object} guided by their segmentation masks."

"Extract the relation-relevant features between {subject} and {object} by attending to their visual overlap and regions."

"Analyze how {subject} and {object} interact by utilizing their mask regions and image features."

"Capture the context-aware interaction pattern of {subject}-{object} using segmentation and visual representation."

"Extract interactive features of \{subject\}-\{object\} based on visual features and the masks"
```

Fig. 6

C. Generation instruction

We define six distinct generation instructions to guide the relation prediction process. During training, one instruction is randomly selected for each instance. During evaluation, the last instruction is consistently used for inference.

```
"What does the image suggest about the interaction between {object i} and {object j}?"

"Describe the scene-level connection involving {object i} and {object j}."

"Determine how {object i} and {object j} are related in this visual context."

"Based on their appearance and positioning, what is the relationship between {object i} and {object j}?"

"In the context of the image, explain the link between {object i} and {object j}."

"Infer the relation between \{object i\} and \{object j\} based on their interaction features"
```

Fig. 7

D. Judgment instruction

The same as Generation instruction

```
"Is the relation {relation} between {object i} and {object j} valid and semantically strong?"

"Should the relation {relation} be considered tight between {object i} and {object j}?"

"Does {relation} accurately describe a meaningful relation between {object i} and {object j}?"

"Is it reasonable to say that {object i} and {object j} are tightly connected by {relation}?"

"Can the relation {relation} be confidently applied to {object i} and {object j}?"

"Judge whether the relation \{relation\} between \{object i\} and \{object j\} is tight."
```

Fig. 8