Attention over Scene Graphs: Indoor Scene Representations Toward CSAI Classification

Artur Barros[1]
artur.barros@students.ic.unicamp.br
Carlos Caetano[1]
caetanoc@unicamp.br
João Macedo[2, 3]
joaomacedo@gmail.com
Jefersson A. dos Santos[4]
j.santos@sheffield.ac.uk
Sandra Avila[1]
sandra@ic.unicamp.br

[1] Instituto de Computação (IC)
Universidade Estadual de Campinas (UNICAMP)
Campinas, São Paulo, Brazil

[2] Departamento de Ciência da Computação (DCC)
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Minas Gerais, Brazil

[3] Polícia Federal (PF)
Belo Horizonte, Minas Gerais, Brazil

[4] School of Computer Science
University of Sheffield
Sheffield, England, United Kingdom

Abstract

Indoor scene classification is a critical task in computer vision, with wide-ranging applications that go from robotics to sensitive content analysis, such as child sexual abuse imagery (CSAI) classification. The problem is particularly challenging due to the intricate relationships between objects and complex spatial layouts. In this work, we propose the **A**ttention over **S**cene **Gr**aphs for Sensitive Content **A**nalysis (ASGRA), a novel framework that operates on structured graph representations instead of raw pixels. By first converting images into Scene Graphs and then employing a Graph Attention Network for inference, ASGRA directly models the interactions between a scene's components. This approach offers two key benefits: (i) inherent explainability via object and relationship identification, and (ii) privacy preservation, enabling model training without direct access to sensitive images. On Places8, we achieve 81.27% balanced accuracy, surpassing image-based methods. Real-world CSAI evaluation with law enforcement yields 74.27% balanced accuracy. Our results establish structured scene representations as a robust paradigm for indoor scene classification and CSAI classification. Code is publicly available at https://github.com/tutuzeraa/ASGRA.

1 Introduction

Scene classification is a fundamental problem in computer vision, especially for indoor environments [18]. The goal is to categorize an image according to predefined scene types (e.g., bedroom, living room), with indoor scene classification focusing on interior spaces.

Beyond general-purpose applications, indoor scene classification has also proven useful in highly sensitive domains, such as Child Sexual Abuse Imagery (CSAI) classification [7, 26]. Prior studies involving interviews with law enforcement agents have shown that environmental and object-based contextual cues within a scene can serve as critical indicators of inappropriate content [12].

Despite the availability of powerful approaches for indoor scene classification [28], significant limitations persist [18]. The task is particularly demanding due to the high complexity and variability of indoor scenes, characterized by a sheer diversity of objects, textures, and colors. Additionally, inherent ambiguity and inter-class similarity, such as the subtle distinction between a bedroom and a child's room, make this task particularly challenging.

Scene Graphs (SGs) [10] offer a promising approach, representing scenes as structured graphs where objects are nodes and their relationships are edges. Each scene is modeled as a set of triplets in the form (`subject`, `predicate`, `object`), such as (`bed`, `next to`, `window`), which explicitly encode semantic and spatial interactions between entities. Unlike traditional image representations, SGs explicitly encode semantic and spatial interactions between objects, offering a compact and interpretable abstraction of the scene. This structured format can be particularly valuable in indoor environments where object relationships distinguish visually similar scenes.

We introduce the **A**ttention over **S**cene **Gr**aphs for Sensitive Content **A**nalysis (ASGRA), a novel framework that leverages the inherent structure of the scene, modeled via SGs, to improve indoor scene classification. We use the extracted SGs as input to a Graph Attention Network [27] (GAT), which effectively weighs the importance of each triplet within the scene graph, thereby enhancing the model's robustness in discerning similar classes.

One of the primary motivations for adopting this pipeline is its suitability for handling highly sensitive content, such as CSAI. Since direct training on CSAI datasets is ethically prohibitive and legally constrained, our approach offers a practical solution: using only scene graph representations and corresponding high-level labels (e.g., "CSAI", "Not CSAI") from law enforcement agents. This enables effective model training while maintaining strict adherence to ethical and legal standards, without exposing researchers to harmful content.

2 Related Work

Pioneering work by Quattoni and Torralba [20] showed that indoor scenes possess unique characteristics that make their classification inherently more difficult than outdoor environments. Consequently, specialized datasets such as MIT Indoor Scenes [20] and Places8 [26] emerged, promoting focused research on indoor contexts.

More recent approaches leverage deep convolutional neural networks to capture local and global features [22, 23, 32, 33], while methods employing Graph Neural Networks (GNNs) explicitly model object relationships and spatial layout within scenes [1, 4, 9, 19]. In particular, these graph-based methods have demonstrated improved performance by encoding relational semantics and structural information explicitly, thereby effectively addressing ambiguities inherent to indoor scenes.

Most recently, Valois et al. [26] and Coelho et al. [6] have developed complementary methods for scene classification, evaluated on Places8. Valois et al. [26] present a comprehensive study of self-supervised approaches, showing that a ResNet-50 fine-tuned on the Barlow Twins protocol can leverage unlabeled data to achieve strong performance. In contrast, Coelho et al. [6] propose a few-shot learning framework based on Vision Trans-

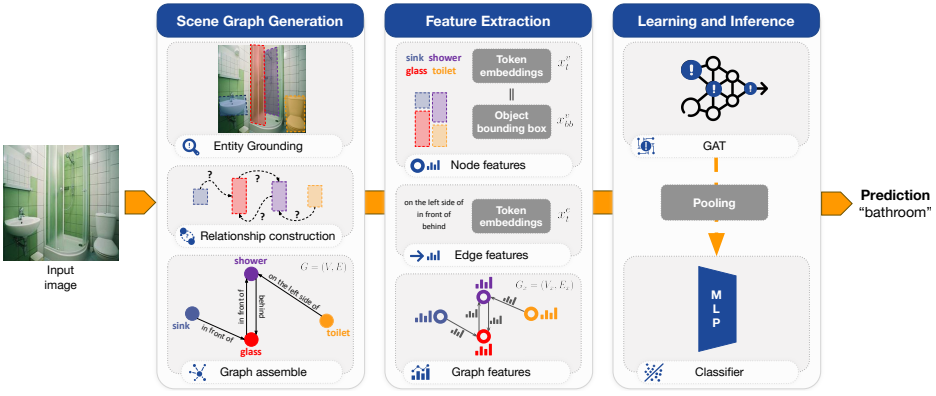


Figure 1: The ASGRA framework processes input images through a pre-trained SGG model to generate structured graph representations. Detected objects and bounding boxes become node features while relations form edge features. A GAT performs learning and inference, with attention pooling, and a multilayer perceptron (MLP) predicts the indoor scene category.

formers (ViTs), demonstrating competitive results with only five annotated examples per class. However, both approaches rely on image-based representations rather than structured graph representations, limiting their capacity to model complex spatial and semantic object relationships critical for disambiguating similar indoor scene categories.

3 Our Approach

Our problem is formulated as follows: given an arbitrary indoor image I, our objective is to classify it into one of the predefined indoor scene categories $y \in C$. Instead of operating directly on pixel data from the image I, we extract an SG representation, $G = (V, E)$, where V is a set of nodes representing the objects and E is a set of directed edges representing the relationships between them. After, we treat it as a graph classification problem: given the scene graph G, the goal is to predict its scene label $y \in C$.

Our proposed ASGRA framework consists of three main steps: (i) scene graph generation, (ii) feature extraction, and (iii) learning and inference. Fig. 1 illustrates our pipeline.

For the first step, we leverage Pix2Grp [12], an off-the-shelf SGG model based on a Vision-Language Model architecture. Pix2Grp has an entity grounding module with the aim of predicting the bounding boxes and labels of the objects in the scene. With the objects in hand, a relationship construction module generates spatial and category labels to create the relation triplets. Finally, we can perform a graph assembly that creates the scene graph G.

In the feature extraction step, node features are constructed by concatenating two types of information for each node $v \in V$: x_t^v, derived from the token embeddings of the object labels; and x_{bb}^v, the normalized bounding box coordinates of the detected object. The final node representation is a concatenation of the aforementioned features: $x_{t\|bb}^v = x_t^v \parallel x_{bb}^v$. For each edge $e \in E$, we extract features x_t^e, derived from the token embeddings of the predicate (relation). Finally, our graph features can be represented as $G_x = (V_x, E_x)$, where $V_x = \{x_{t\|bb}^0, x_{t\|bb}^1, \ldots, x_{t\|bb}^v\}$ is the set of node features and $E_x = \{x_t^0, x_t^1, \ldots, x_t^e\}$ is the set of edge features.

In the learning and inference step, we employ the GATv2 [2]. In this step, the graph

is processed by computing attention coefficients for each edge, dynamically weighting the influence of neighboring nodes during message passing. This attention mechanism allows the model to focus on the most relevant triplets that drive the prediction. Finally, a graph pooling layer aggregates the node representations into a single graph-level vector, which is passed to an MLP for the final scene classification.

4 Experimental Results

4.1 Experimental Setup

Datasets. We evaluate our benchmark on the Places8 dataset [26]. Places8 is a curated subset of the Places365 dataset [33], consisting of 407,640 images (256×256 pixels) selected from 23 of the original 365 scene classes. The authors remapped these classes into 8 indoor scene categories, chosen for their relevance to frequently encountered environments in CSAI. We follow the train/val/test experimental protocol proposed by the authors[1].

For sensitive media evaluation, we collaborate with law enforcement to use the Region-based Annotated Child Pornography Dataset (RCPD) [17], a private dataset maintained by the Brazilian Federal Police. Originally designed for forensic analysis, the dataset lacks standard train/test splits and prohibits direct training access by non-law-enforcement personnel.

Through formal police collaboration, authorized agents perform SGG internally on RCPD images, providing only resulting scene graphs to our research team. This ensures no sensitive image access during the study. We train models using 5-fold cross-validation on these graph representations, enabling effective evaluation without direct CSAI data access.

Vision-Language Model Baseline. To establish a solid point of comparison, we adopted a Vision-Language Model (VLM) configured for a Visual Question Answering (VQA) task, tailored to scene classification. VQA is a task in which a model receives an image alongside a textual question and must generate an appropriate answer, combining visual perception with natural language understanding. Our baseline uses the Large Language and Vision Assistant (LLaVA) [15], which combines a vision encoder (e.g., CLIP [21]) with a language decoder (e.g., Vicuna [6]). LLaVA's multimodal instruction tuning enables highly accurate, context-aware responses, making it an effective choice for our scene recognition task.

One of the primary reasons for adopting a VLM in this role is its versatility. Trained on large-scale multimodal datasets, these models can adapt to a wide variety of visual contexts and maintain performance even in challenging conditions [30, 31]. Additionally, using language in the form of VQA provides a flexible and robust evaluation protocol [25].

The overall pipeline operates in a simple yet effective manner. Images are processed individually, each paired with the corresponding textual prompt. The model's outputs are parsed to extract the predicted category, which is then compared with the ground truth labels for evaluation. This approach offers a robust and easily reproducible benchmark, enabling a fair assessment of more specialized methods introduced in later sections.

Scene Graph Generation and Feature Extraction. We employ Pix2Grp [14] to generate scene graphs due to its robust performance. The model outputs triplets comprising predicted subjects, objects, and their relationships. We used the model weights pre-trained on VG150 [13], one of the most widely adopted datasets for evaluating SGG [14, 24, 29].

[1] Dataset splits are available at https://doi.org/10.5281/zenodo.13910525.

Consequently, the predicted triplet labels are constrained to the 150 object classes and 50 relationship classes defined in VG150, which do not contain object classes (e.g., intimate body parts) or relationship classes (e.g., hugging, kissing, touching) directly related to CSAI.

For feature representation, each node is encoded by concatenating its label index with the detected bounding box coordinates, while edge features correspond to the relation token ID. Importantly, we avoid explicit image features to ensure privacy preservation, as incorporating such features could enable reconstruction of sensitive images from trained models [3], which is undesirable for sensitive media applications.

Implementation Details. All experiments are conducted on 5 NVIDIA RTX5000 GPUs. Hyperparameter optimization is performed using the Optuna framework to efficiently explore optimal configurations.

For the baseline, we use LLaVA version 1.6 with the Vicuna decoder and provided hyperparameters. For our experiments, the model receives along the input image the following prompt: *"Classify the received image into one of the following 8 categories: (0) bathroom; (1) bedroom; (2) child's room; (3) classroom; (4) dressing room; (5) living room; (6) studio; or (7) swimming pool. Answer with only the number of the corresponding category provided."* This prompt format ensures that predictions are both interpretable, constrained to the desired label set and easy to parse.

The GAT model is trained using early stopping with a patience of 10 epochs based on validation loss, with a maximum of 120 epochs. Cross-entropy loss is optimized via the Adam optimizer [11], with an initial learning rate of 1×10^{-4} and weight decay of 3×10^{-5}. A batch size of 8 is used. To mitigate overfitting, a dropout rate of 0.2 is applied across all GATv2 layers. The final architecture comprises two GATv2 layers, 364 hidden dimensions, and 4 attention heads.

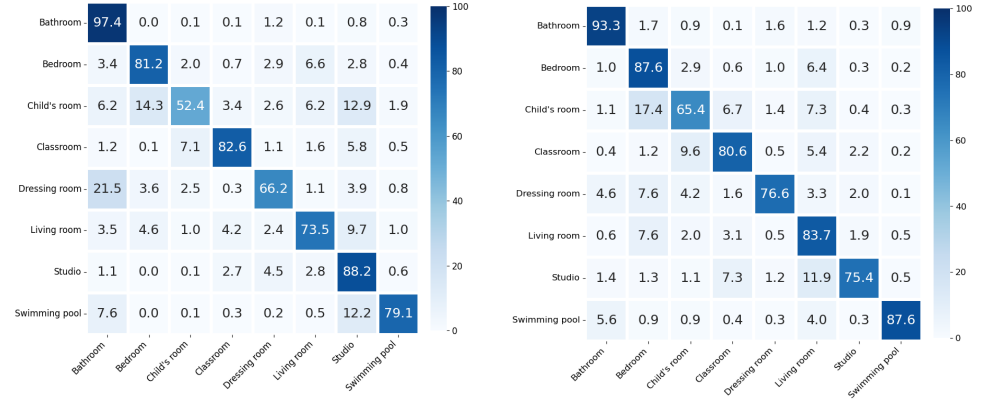
For the CSAI classification task, each fold of the 5-fold cross-validation protocol is trained for 20 epochs, with a learning rate of 3.8×10^{-4}, batch size of 8, and weight decay of 1.4×10^{-5}. The final CSAI model consists of two GATv2 layers, 128 hidden dimensions, and 8 attention heads.

4.2 Results and Analysis

Quantitative Analysis. On Places8, the VQA-baseline achieved 77.69% balanced accuracy on the test split, surpassing previous approaches: self-supervised learning (71.60% [26]) and few-shot learning (73.50% [6]). Despite being computationally heavy (7B parameters), this baseline establishes a strong comparison point. Our proposed ASGRA achieved superior performance at 81.27% balanced accuracy using only 242 million parameters, demonstrating efficiency and scalability. Table 1 presents detailed comparisons.

Method	Core Component(s)	Input Modality	#Params	Acc. (%)
VQA-baseline [14]	LLaVA v1.6 + Vicuna	Image & Text	7B	77.69
Few-shot [6]	ViT-Small	Image	21.7M	73.50
Self-supervised [26]	ResNet-50	Image	23.5M	71.60
ASGRA (ours)	Pix2Grp + GATv2	Scene Graph	242M	**81.27**

Table 1: Comparative results on the Places8 test set. Our ASGRA framework is benchmarked against state-of-the-art methods. ASGRA operates on scene graphs, a fundamentally different input modality from image-based approaches.



(a) VQA-baseline (b) ASGRA

Figure 2: Confusion matrices on the Places8 test split.

Detailed performance breakdowns are shown in Fig. 2. Considering the VQA-baseline, Fig. 2 (a), the most frequent misclassification occurs between *child's room* and *bedroom*, as well as between *child's room* and *studio*. Another notable confusion is between *dressing room* and *bathroom*, where the baseline incorrectly predicts *bathroom* in 21.5% of *dressing room* cases. Additionally, the model tends to confuse *living room* and *studio*. However, in this case, the baseline is more accurate for *studio* and less accurate for *living room*, which is the opposite pattern of our proposed framework. While the baseline shows slightly less confusion between *bedroom* and *child's room* than our proposed ASGRA framework, its true positive rate for *child's room* is considerably lower (52.4% vs. 65.4%). Overall, the baseline exhibits lower accuracy for key classes compared to ASGRA.

For the ASGRA framework, Fig. 2 (b), we highlight the achieved high true positive rates for categories with distinctive objects and spatial arrangements, such as *bathroom*, *bedroom*, *classroom*, *living room*, and *swimming pool*. This suggests that the proposed framework is effective in identifying discriminative triplets that characterize these environments. However, ASGRA's most prominent confusion is between *child's room* and *bedroom*, which is intuitive given their shared core objects (e.g., beds, pillows, windows). A smaller but still noticeable confusion occurs between *living room* and *studio*. Furthermore, we highlight that ASGRA achieved higher accuracy than the baseline in *bedroom*, *child's room*, *dressing room*, *living room*, and *swimming pool*, showing its advantage in both overall classification and in challenging class pairs. These specific confusion patterns are further explored in the qualitative analysis section.

The performance achieved by our framework in Places8 motivated its application to the sensitive-media scenario, where such discriminative cues could help capture context-specific patterns. To that end, we conducted RCPD experiments using 5-fold cross-validation. Beyond balanced accuracy, we employed recall given its critical importance for sensitive content, as recall directly measures the model's ability to detect CSAI, ensuring such instances are not overlooked.

We began by training the best-performing architecture from Places8 from scratch on binary classification, achieving 72.42% balanced accuracy and 70.61% recall. Subsequently, we investigated the potential benefits of transfer learning by fine-tuning the best pre-trained

	Training from scratch	Fine-tuning (head)	Fine-tuning (network)	Optimization
Acc. (%)	72.42	71.28	73.25	**74.27**
Recall (%)	70.61	71.44	73.27	**76.55**

Table 2: 5-fold cross-validation results on the RCPD dataset with our ASGRA framework.

Class	Top-10 Objects	Top-5 Relations
bathroom	sink, toilet, handle, door, window, towel, cabinet, tile, shelf, counter	has, on, near, in front of, in
bedroom	pillow, window, table, chair, bed, door, lamp, room, curtain, [unk_obj]	has, near, on, in front of, with
child's room	flower, pillow, [unk_obj], window, table, chair, shelf, bed, box, bear	has, on, with, near, in front of
classroom	man, boy, woman, girl, person, table, hair, shirt, chair, [unk_obj]	on, has, in front of, with, wearing
dressing room	shelf, door, woman, man, handle, person, window, shirt, bag, [unk_obj]	on, has, in front of, with, under
living room	chair, window, table, door, [unk_obj], room, lamp, shelf, pillow, man	on, has, with, near, in front of
studio	man, woman, person, shirt, hair, head, window, hand, chair, [unk_obj]	on, has, in front of, wearing, holding
swimming pool	window, chair, door, person, table, man, pole, tree, [unk_obj], building	on, has, near, with, in front of

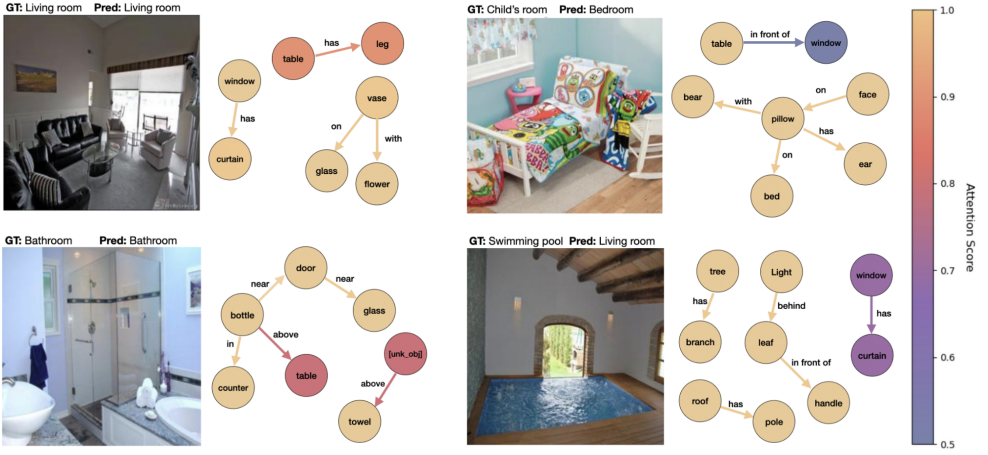
Table 3: Top-10 most influential objects (nodes) and Top-5 most influential relations (edges) per class, ranked by importance. They were aggregated from the GATv2 model's attention weights across all correct predictions in the validation set. The [unk_obj] token represents objects not recognized in the vocabulary.

Places8 model. Fine-tuning only the classification head yielded 71.28% balanced accuracy and 71.44% recall, while fine-tuning the entire network improved to 73.25% balanced accuracy and 73.27% recall. These results suggest a significant domain shift between Places8 indoor scenes and the RCPD context, making naive transfer learning suboptimal. Finally, new hyperparameter optimization with Optuna achieved our best results: **74.27%** balanced accuracy and **76.55%** recall, underscoring the importance of architectural adaptation for specialized real-world data. Table 2 summarizes these results.

Qualitative Analysis. Beyond the quantitative metrics, our framework provides inherent explainability through Scene Graphs with Graph Attention Networks. By analyzing the attention coefficients computed by the GATv2 layers, we can move beyond simply assessing accuracy and begin to understand the model's decision-making process. This analysis allows us to pinpoint which objects and relationships contributed most to a given prediction.

First, we performed an aggregated analysis to understand the general patterns learned for each class. By accumulating the attention scores across all correctly classified images, we identified the most important features for each scene category (Table 3). Results confirm the model learns intuitive, human-understandable patterns. For example, *bathroom* is strongly characterized by objects like `sink`, `toilet`, and `towel`. Interestingly, for *classroom*, the model learned that the presence of people (`man`, `boy`, and `woman`) was a more reliable indicator than specific furniture, which can overlap with scenes like *studio* or *living room*.

The attention mechanism effectively diagnoses failure modes. Fig. 3 shows four predictions (two correct, two errors) with SG subsets and attention scores. In the correct cases



(a) Correct predictions (b) Wrong predictions

Figure 3: Qualitative results of ASGRA on Places8. Column (a) shows correctly classified scenes, while column (b) shows misclassifications. Each image includes its scene graph with GATv2 attention scores for nodes and edges. These visualizations showcase the model's ability to identify key semantic components for correct predictions and provide transparent analysis of failure cases, including confusion between similar scenes.

(Fig. 3 (a)), the model assigns higher attention to class-defining objects and their incident relations (lighter color bar tones), e.g., `window`/`table` for *living room* and `sink`/`towel`/`counter` for *bathroom*. Attention also propagates to neighbors of important nodes, reflecting the model's focus on triplets relating to high-attention objects.

In Fig. 3 (b), we observe a typical source of error: upstream SGG hallucinations. The detector incorrectly inserts a `curtain` next to a `window`, yielding the high-attention triplet $\langle$`window`,`has`,`curtain`$\rangle$, which steers the classifier toward *living room* contexts. Thresholding SGG triplets by their confidence did not improve accuracy, as erroneous triplets often receive high scores and survive filtering. This suggests robustness depends more on improving SGG quality and vocabulary than on simple score pruning.

In the first case of Fig. 3 (b), the model confuses *child's room* with *bedroom*, a frequent ambiguity in Places8. Although the graph includes cues such as `bear`, `face`, and `pillow`, the closed vocabulary cannot express concepts like `toy`, `cartoon`, or other child-specific attributes, limiting the semantic signal available to GATv2. Moving to an open-vocabulary SGG or enriching the label space with attributes (e.g., `toy`, `crib`, `cartoon pattern`) should help disambiguate these classes by capturing the semantics that distinguish a *child's room* from a *bedroom*.

In collaboration with law enforcement, we performed the same analysis for CSAI classification. Table 4 showcases the most important objects and relationships for discerning between categories, enabling nuanced scene understanding even without directly viewing images. For detecting CSAI, `hand` was the most important object with `holding` and `near` as key relationships. This aligns with what expert law enforcement agents may identify during manual screening, indicating the model's capability to assign attention where most relevant.

For each evaluated image, a large set of object relations was generated. Although the

Class	Top-10 Objects	Top-5 Relations
CSAI	hand, woman, head, girl, boy, [unk_obj], man, person, arm, leg	has, on, near, holding, in front of
Not CSAI	woman, girl, [unk_obj], hand, head, hair, man, shirt, boy, leg	has, on, behind, near, wearing

Table 4: Top-10 most influential objects and Top-5 most influential relations per category, ranked by importance, for the RCPD dataset.

Category	Acc. (%)	Description
CSAI	76.97	Child sexual abuse imagery
Not CSAI – *child*	74.29	Images containing children
Not CSAI – *adult*	77.34	Images containing adults
Not CSAI – *suspicious*	72.55	Children/adolescents in underwear, swimwear, or shirtless
Not CSAI – *pornography*	72.41	Pornographic content
Not CSAI – *normal*	84.62	No nudity
CSAI and Not CSAI – *global*	76.69	Overall dataset

Table 5: Accuracy of ASGRA in CSAI classification on RCPD, reported by image category. While the task remains the same, results are broken down by image content.

Step	Model	#Images	Energy (kWH)	CO_2-eq (kg)
SGG	Pix2Grp [14]	407,640	1.0136	0.3960
GNN	GATv2 [3]	364,806	1.3520	0.1330

Table 6: Average energy consumption and equivalent CO_2 emissions (CO_2-eq) for each step of our method.

relation confidence was not the analysis focus, we observed that among the most relevant relations (considering both score and the top-10 most influential classes for CSAI classification), some were consistent with the actual image content. However, we also identified implausible or semantically inconsistent relations. This reflects domain shift and the presence of close-ups and task-specific CSAI content containing visual elements that are presumably absent from the base vocabulary used for object and relation tokenization, highlighting natural limitations in visual element coverage.

Table 5 shows model accuracy for CSAI classification by image category. Accuracy varies across subsets, with the *normal* category reaching 84.62%, well above the global average, showing effective classification when no sexual or suggestive content is present. In contrast, performance drops to around 72% for the more challenging *suspicious* and *pornography* categories, which involve higher visual variability and ambiguous boundaries. For *CSAI*, results are comparable to the *global* accuracy (76.69%), indicating that despite object-relation vocabulary limitations and the presence of many irrelevant relations, the model successfully exploits the most informative ones to produce adequate final decisions.

Environmental Impact Analysis. We also assess the environmental impact of our experiments by quantifying the energy consumption and carbon footprint. To that end, we employed the *CodeCarbon* tool [8], which tracks the power usage of our hardware configuration and estimates the carbon emissions generated during the computational processes involved in SGG and GNN training. Table 6 details energy consumption metrics.

5 Conclusions

In this work, we introduced ASGRA, a novel framework for indoor scene classification that leverages the semantic and the structure of Scene Graphs. Using a Graph Attention Network, our method achieves state-of-the-art 81.27% balanced accuracy on Places8. In collaboration with law enforcement, we evaluated ASGRA on real-world CSAI datasets, obtaining 74.27% balanced accuracy for CSAI classification, showcasing practical utility in digital forensics. The framework's primary strengths are its inherent explainability through attention weight analysis for error diagnosis, and privacy-preserving architecture suitable for sensitive applications like CSAI analysis.

While promising, performance depends on upstream Scene Graph Generation quality and is constrained by closed-set vocabulary. Future work will focus on integrating advanced, open-vocabulary SGG models and enriching the graph's node and edge features to overcome these limitations.

Acknowledgments

This work is partially funded by FAPESP 2023/12086-9, FAEPEX/UNICAMP 2597/23, and the Serrapilheira Institute R-2011-37776. Artur Barros (2024/09372-2), Carlos Caetano (2024/01210-3), and Sandra Avila (2023/12865-8, 2020/09838-0, 2013/08293-7) are also funded by FAPESP. Sandra Avila is also funded by H.IAAC 01245.003479/2024-10 and CNPq 316489/2023-9.

References

[1] Nassim Belmecheri, Arnaud Gotlieb, Nadjib Lazaar, and Helge Spieker. Explainable scene understanding with qualitative representations and graph neural networks. In *IEEE Intelligent Vehicles Symposium*, 2025.

[2] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations (ICLR)*, 2022.

[3] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX security symposium (USENIX Security 21)*, 2021.

[4] Gongwei Chen, Xinhang Song, Haitao Zeng, and Shuqiang Jiang. Scene recognition with prototype-agnostic scene layout. *IEEE Transactions on Image Processing (TIP)*, 29:5877–5888, 2020.

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[6] Thamiris Coelho, Leo S. F. Ribeiro, João Macedo, Jefersson A dos Santos, and Sandra Avila. Transformers-based few-shot learning for scene classification in child sexual

abuse imagery. In *IEEE Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 8–14, 2024.

[7] Thamiris Coelho, Leo S. F. Ribeiro, João Macedo, Jefersson A. dos Santos, and Sandra Avila. Minimizing risk through minimizing model-data interaction: A protocol for relying on proxy tasks when designing child sexual abuse imagery detection models. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1543–1553, 2025.

[8] Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, et al. mlco2/codecarbon: v2.4.1, 2024. URL https://doi.org/10.5281/zenodo.11171501.

[9] Kanglong Fan, Wei Liu, Xiaowen Chen, Bharath Ramesh, and Cheng Xiang. An interpretable scene understanding framework via graph learning. *SSRN 4238333*, 2022.

[10] Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. Devil's on the edges: Selective quad attention for scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18664–18674, 2023.

[11] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

[12] Juliane A. Kloess, Jessica Woodhams, and Catherine E. Hamilton-Giachritsis. The challenges of identifying and classifying child sexual exploitation material: Moving towards a more ecologically valid pilot study with digital forensics analysts. *Child Abuse & Neglect*, 118:105166, 2021.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[14] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28076–28086, 2024.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 34892–34916, 2023.

[16] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[17] Joao Macedo, Filipe Costa, and Jefersson A. dos Santos. A benchmark methodology for child pornography detection. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 455–462, 2018.

[18] Tanvi A Patel, Vipul K Dabhi, and Harshadkumar B Prajapati. Survey on scene classification techniques. In *IEEE International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 452–458, 2020.

[19] Jiayan Qiu, Yiding Yang, Xinchao Wang, and Dacheng Tao. Scene Essence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8318–8329, 2021.

[20] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 413–420, 2009.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[22] Mehdi S. M. Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, and Klaus Greff. RUST: Latent neural scene representations from unposed imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17297–17306, 2023.

[23] Chuanxin Song and Xin Ma. Srrm: Semantic region relation model for indoor scene recognition. In *International Joint Conference on Neural Networks (IJCNN)*, 2023.

[24] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9002–9011, 2019.

[25] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: a fully open, vision-centric exploration of multimodal llms. In *Advances on Neural Information Processing Systems (NeurIPS)*, 2024.

[26] Pedro H. V. Valois, João Macedo, Leo S. F. Ribeiro, Jefersson A dos Santos, and Sandra Avila. Leveraging self-supervised learning for scene classification in child sexual abuse imagery. *Forensic Science International: Digital Investigation*, 53:301918, 2025.

[27] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[28] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14408–14419, 2023.

[29] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[31] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15134–15186, 2025.

[32] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 487–495, 2014.

[33] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1452–1464, 2017.