Introspection in Learned Semantic Scene Graph Localisation

Manshika Charvi Bissessur, Efimia Panagiotaki, Daniele De Martini
Mobile Robotics Group (MRG), University of Oxford, UK

***Abstract*— This work investigates how semantics influence localisation performance and robustness in a learned self-supervised, contrastive semantic localisation framework. After training a localisation network on both original and perturbed maps, we conduct a thorough post-hoc introspection analysis to probe whether the model filters environmental noise and prioritises distinctive landmarks over routine clutter. We validate various interpretability methods and present a comparative reliability analysis. Integrated gradients and Attention Weights consistently emerge as the most reliable probes of learned behaviour. A semantic class ablation further reveals an implicit weighting in which frequent objects are often down-weighted. Overall, the results indicate that the model learns noise-robust, semantically salient relations about place definition, thereby enabling explainable registration under challenging visual and structural variations.**

I. Introduction

Conventional localisation methods that rely on extracting geometric features from input sensor data, are often susceptible to perceptual variations. Factors such as changes in lighting and weather conditions, or the presence of dynamic objects can significantly alter the appearance of traffic scenes, making these methods less reliable. In contrast, semantic information offers greater robustness, as it is inherently invariant to variations in environmental conditions. For instance, a desk remains identifiable as a desk regardless of whether it is observed in daylight or darkness, making it a stable and reliable landmark for localisation. Notably, this is consistent with how humans navigate an environment, mainly relying on high-level semantic cues – such as rooms, furniture, and structural layouts – rather than low-level geometric attributes or "pixel-level" appearance features.

Reflecting this, recent works in place recognition and global localisation increasingly leverage higher-level visual features that correspond more closely to the semantic structure of the environment [1], [2], [3], [4], [5]. Beyond leveraging semantics, humans naturally filter out environmental "noise" to localise themselves, instinctively distinguishing scene elements that concretely define a place. While objects offer limited localisation value, persistent structures provide stable and reliable reference points.

Prior research has demonstrated that semantic elements can be effectively exploited for localisation while also investigating their role in localisation performance and decision making [3], [6]. Building on this foundation, we extend the scope of the analysis by systematically examining model introspection. Specifically, we propose a semantics-driven framework in which a model performs place registration from high-level semantic layers, followed by thorough post-hoc introspection that quantifies the influence of each object class. This class-level importance analysis enables us to verify that the model prioritises stable, meaningful features over transient clutter, mirroring the filtering strategies employed by humans. Because these explanations are generated after inference, they provide an independent and interpretable audit trail, which is crucial for diagnosing failure modes early and building trust in the robot's decisions, especially in safety-critical and dynamic environments.

In this work, we address a critical need in current robotics research, advancing towards localisation systems that combine robustness and efficiency with transparency and interpretability. Our key contributions are as follows:

1) A perturbation-based class-importance analysis, assessing both place registration performance degradation and attribution shifts;
2) A rigorous introspection analysis, along with fidelity tests, demonstrating that Integrated Gradients and Attention Weights provide the most faithful object-importance attribution signals.

II. Related Work

A. Semantic Localisation

A common strategy for semantic localisation is to augment individual features [7] or bag-of-words (BoW) representations [8] with semantic information as a post-processing step. Going further, [9] learns a single descriptor that fuses semantics and geometry. Hybrid approaches such as SemSegMap [10] leverage both cues and generally achieve better localisation than geometry alone.

There are also settings where only semantic information is available for the query [11]. Text2Loc [12] addresses this by aligning rich textual embeddings extracted with a frozen T5 model with 3D submap embeddings derived from point clouds using PointNet++, trained via a cross-modal contrastive objective.

B. Importance of Different Semantic Classes

Neural networks excel in a wide range of tasks, yet their distributed feature representations remain semantically opaque. Since the features leveraged by these models often lack a direct correspondence to human-interpretable concepts, explaining their decision-making processes can be challenging. In the domain of geolocalisation, [13] introduces

The project was supported by the EPSRC Programme Grant "From Sensing to Collaboration" (EP/V000748/1). Manshika Charvi Bissessur was supported by the State of Mauritius Science Scholarship, and Efimia Panagiotaki by the Oxford-DeepMind Scholarship. Corresponding author: Manshika Charvi Bissessur, manshika.bissessur@reuben.ox.ac.uk

a concept-influence metric that estimates the contribution of semantic visual concepts (e.g., "building", "mountain") to a model's prediction – rather than specific pixels or regions. Using Integrated Gradients with SmoothGrad to reduce noise, they find that localised concepts (e.g., tower, bridge) are highly informative. In contrast, generic concepts (e.g., grass, car) contribute little due to their ubiquity.

We extend this line of research to introspective localisation on scene graphs. The closest prior work is SEM-GAT [3], an attention-based GNN method that integrates semantic and geometric cues to identify reliable correspondences for point cloud registration also introducing an explainability component by analysing the model's attention weights [6]. Our work focuses on indoor environments and investigates analogous semantic–structural relationships on scene graphs. We conduct a more comprehensive introspection analysis that goes beyond attention weights, proposing and evaluating a range of post-hoc explainability methods to identify reliable objects that can serve as strong landmarks for interpretable and accurate localisation.

C. Model Explainability for Graphs

Graph neural networks (GNNs) integrate rich node and edge features with graph topology, yielding strong performance but often lacking transparency in their decision-making processes. We focus on post-hoc explanations, leaving the trained model unchanged, and follow the taxonomy of [14], which categorises methods into: (i) gradient/feature-based, estimating importance from input gradients or feature magnitudes; (ii) perturbation-based, measuring prediction changes when features, nodes, or edges are modified or removed; (iii) decomposition-based, tracing contributions backward through the network; and (iv) surrogate-based, fitting simple, interpretable models (e.g.,decision trees) to approximate the behaviour of GNNs. Within this space, *Saliency* [15], *Integrated Gradients* [16], and *Shapley Value Sampling* [17] are employed. Attention weights are also considered as model-intrinsic signals in architectures such as GATs [18]. While early work cautioned that attention alone may be insufficient or misleading as an explanation [19], subsequent studies argue it can be informative when validated [20]. In this vein, [6], [21] recommend verifying that changes in attention correlate with performance before treating attention as an explanation.

III. Overview

In this section, we formulate the problem addressed and describe the scene graphs and dataset used. A 3D scene graph models an environment as a hierarchy of spatial concepts (nodes) and their relations (edges), providing a compact abstraction that captures both semantic cues and spatial organisation.

Our work leverages the hierarchical scene graphs proposed in Hydra [22] which employs five layers of hierarchy: (L1) Metric-Semantic Mesh, a dense 3D mesh with per vertex geometry and semantic labels; (L2) Objects & Agents, entities (including the robot) with spatial and semantic attributes, each linked to the mesh elements composing them; (L3) Places, free space nodes connected by traversability edges into a topological map and linked to nearby L2 nodes; (L4) Rooms, nodes with location and bounding boxes connected to their constituent places; (L5) Buildings, nodes connected to all constituent rooms. Our experiments use only the L2-L3 layers to stress the model's semantic reliance.

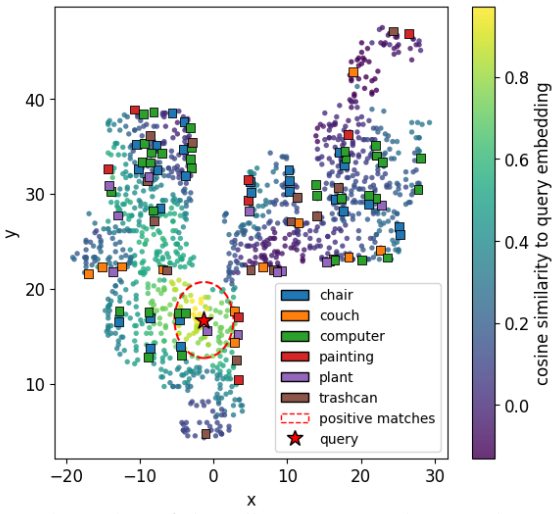


Fig. 1: Floor-plan of the *office* scene. Each L2 and L3 nodes are plotted in its metric coordinate. L3 nodes are coloured by cosine similarity to a query.

The uHumans2 dataset [23] provides photorealistic, synthetic indoor environments with complete metric–semantic annotations, making it a standard benchmark for spatial-perception pipelines such as Hydra, which produces a scene graph comprising about 1,300 place nodes and 110 object nodes on the `office` scene, depicted in Figure 1.

TABLE I: Breakdown of object instances in the uHumans2 office scene.

Semantic Label	Semantic Class	Semantic Instances
5	Chair (CH)	28
8	Couch (CO)	11
10	Computer (CP)	35
11	Plant (PL)	13
12	Painting (PA)	8
18	Trash-can (TC)	15

Given the class distribution in Table I and its potential effect on localisation, we expect:

1) **Broad but moderate signals from frequent classes:** computers and chairs should offer wide coverage while contributing only moderate confidence to mitigate perceptual aliasing.
2) **High salience of distinctive landmarks:** rarer categories such as paintings and couches should act as strong, location-specific anchors for disambiguation.
3) **Down-weighting of mobile objects over time:** the relative importance of chairs and trash cans should decrease across traversals as positional variability is detected.

TABLE II: Validation and test performance on uHumans2 (`office`).

Model	Loss	Validation					Testing				
		PR-AUC	F1	Recall@N (%)			PR-AUC	F1	Recall@N (%)		
				1	5	10			1	5	10
Ours	Contrastive	0.667	0.649	8.92	39.82	63.62	0.721	0.680	17.22	63.12	85.12
	Triplet	0.524	0.554	8.86	37.21	59.53	0.562	0.571	17.59	59.82	78.54
	InfoNCE	0.699	0.685	8.61	40.53	65.64	0.718	0.685	18.21	65.72	86.14
BoW	–	0.247	0.338	5.32	23.79	38.37	0.243	0.360	8.60	32.68	51.15

These hypotheses are assessed via class-removal perturbations and post hoc attribution analyses.

IV. Semantic Localisation

A. Problem Statement

Let a large-scale 3D hierarchical scene graph $M = (\mathcal{P} \cup \mathcal{O}, \mathcal{E}^t \cup \mathcal{E}^v)$ be a subset of Hydra's [22], containing places $\mathcal{P} = \{p_i\}$ and objects $\mathcal{O} = \{o_i\}$, connected by traversability edges $\mathcal{E}^t = \{e^t_{i,k}\}$ between places i and k, and visibility edges $\mathcal{E}^v = \{e^v_{i,j}\}$ between a place i and an object j. Each object and place are denoted by their Cartesian coordinates (x, y) and each object also by its semantic class $c \in C$, where C is the set of semantic classes.

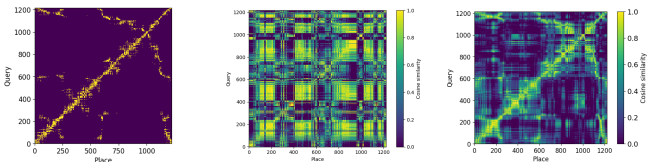
Let T denote a query place. The goal is to learn an encoder F that maps both T and the reference places in M into a metric space, such that the nearest neighbour of T in M corresponds to its true location. Additionally, we want to assign to each class $c \in C$ an *attribution* score that quantifies its contribution to localisation, both at the instance level and in expectation over the dataset.

B. Semantic Localisation

We employ a GNN backbone relying on message passing and attention, trained contrastively on two perturbed variants of the `office` dataset to form place–query pairs.

1) GNN Backbone: We *intentionally* represent each object node with only the semantic class associated with it and treat each place node as an aggregation point, discarding geometric information, to force our model to focus on the semantic cues. Semantic labels are linearly projected into 64-dimensional hidden states and passed through an ELU activation. Two subsequent message-passing layers aggregate neighbour information with a `sum` operator, each followed by tanh nonlinearity and batch normalisation. Long-range context is captured with a GATv2 convolution using three attention heads, and a final linear layer maps hidden states into 32-dimensional embeddings for localisation.

2) Dataset Splits and Perturbations: Experiments use the `office` subset of uHumans2 with two repeats of the same trajectory. Place nodes are partitioned into 70-20-10 split for train, validation, and test respectively, with the test set unseen during training. For data augmentation, object nodes are displaced according to class-specific mobility (e.g., small shifts for highly mobile classes such as trash cans and chairs), thereby generating additional query–target pairs and reducing overfitting to the original layout. Since Hydra often produces closely spaced places, a matching radius of 4 m is used: all map nodes within this radius of the ground-truth location are treated as positives, allowing multiple positive matches per query.



(a) Ground truth. (b) BoW similarity. (c) Model similarity.

Fig. 2: Place–query matrices. Diagonal structure indicates correct matches; our model sharpens the diagonal and suppresses false positives.

3) Metrics: Evaluation is conducted from two complementary views. *Classification*: we report F1 and PR–AUC, as positives are sparse and the class distribution is highly imbalanced. *Retrieval*: we report Recall@ N , which reflects practical usage where a shortlist is inspected and quantifies the probability that a correct match appears among the top-N candidates. Table II summarises our results on the uHumans2 dataset when applying different losses – contrastive [24] with margin of 1, triplet loss [25] with hard mining, and InfoNCE ($T = 0.7$) [26]. Among these, InfoNCE yields the most balanced performance across PR-AUC, F1, and Recall@N. Moderate absolute scores are expected under a semantics-only regime on a small dataset, as discarding geometry removes precise 3D anchors, label aliasing occurs when visually distinct objects share the same class, and limited training diversity restricts generalisation. *The semantics-only constraint is intentional*, allowing an isolated assessment of semantic cues. For context, table II compares the model with a bag-of-words (BoW) baseline that represents each location by per-class object counts; the learned encoder improves PR-AUC by a large margin and consistently increases Recall@N. Although absolute recall remains modest, the relative gains indicate a more discriminative semantic embedding for retrieval.

Qualitative results support these findings. The similarity matrices in Figs. 2a–2c reveal a clear diagonal of correct matches; the learned encoder sharpens this diagonal and suppresses off-diagonal responses, improving separation of non-matching pairs. Unlike the BoW baseline, which primarily responds to frequent classes and thus conflates semantically similar yet spatially distinct locations, the learned localiser adaptively reweights object configurations around each query, amplifying distinctive object–object relations (e.g., a specific couch–painting pairing) while down-weighting ubiquitous items (e.g., computers). Table III shows that removing the GAT block induces a marked performance drop that is not recovered by adding MPNN [27].

V. Semantic Class Attribution

The role of object semantics in localisation is quantified through complementary analyses, including class ablation, post-hoc attribution, and fidelity analysis, with no retraining process.

TABLE III: Ablation study: impact of model design choices on PR-AUC (↑) and Recall@1 (↑). Best score in **bold**.

Variant	# MPNN	Hidden Dim	Heads	PR-AUC	Recall@1
Number of layers	1	64	3	0.34 ± 0.03	0.14 ± 0.05
	3	64	3	0.54 ± 0.11	0.15 ± 0.01
Hidden dimension	2	32	3	0.63 ± 0.08	0.16 ± 0.01
	2	128	3	**0.73 ± 0.03**	**0.17 ± 0.01**
Attention heads	2	64	1	0.58 ± 0.05	0.14 ± 0.01
	2	64	2	0.65 ± 0.07	0.14 ± 0.02
No GAT block	2	64	–	0.09 ± 0.01	0.02 ± 0.003
	3	64	–	0.61 ± 0.02	0.16 ± 0.01
Our model	**2**	**64**	**3**	0.72 ± 0.02	**0.17 ± 0.02**

Class ablation For each semantic class c, all instances are removed, and we measure the change in performance as the difference in PR-AUC with and without the objects of class c, normalised by class frequency.

Post-hoc attribution Node-level importance is estimated using Saliency, Integrated Gradients, Shapley Value Sampling, and Attention Weights. Let p^{pre} and p_c^{post} denote the distributions of node-importance scores before and after perturbing class c, respectively. The shift is quantified by the Jensen–Shannon divergence (JSD) between p^{pre} and p_c^{post}, normalised by class frequency. A larger JSD indicates a greater criticality of class c (i.e., its removal substantially shifts the attribution scores) and reveals how its absence reweights the remaining nodes.

Fidelity Fidelity of the explanation [28], i.e.whether highlighted nodes influence the output, is assessed via *necessity* and *sufficiency* tests, also called *fidelity*$^+$ and *fidelity*$^-$, adapted to the contrastive objective, using changes in embedding similarity rather than predicted probabilities.

Let $\mathbf{z} : \mathcal{G} \to \mathbb{R}^d$ denote the learned encoder that maps a place graph G to a d-dimensional embedding. For a place graph P and its perturbed query Q, let the baseline similarity s_{full} be the scalar product of $\mathbf{z}$ applied to P and Q. We can then define a budget ρ to be the fraction of nodes of Q we can keep, ranked by an explainer. We can then recalculate the similarity between $\mathbf{z}$ applied to P and Q containing only the top nodes selected by ρ or every node excet them. We the two metrics with $s_{\text{keep}}(\rho)$ and $s_{\text{drop}}(\rho)$ respectively.

$s_{\text{keep}}(\rho)$ and $s_{\text{drop}}(\rho)$ can be used to calculate the fidelity$^+$ (necessity) and fidelity$^-$ (sufficiency) of $\mathbf{z}$ as:

$$\begin{aligned} \Delta^+(\rho) &= \left| s_{\text{drop}}(\rho) - s_{\text{full}} \right| \\ \Delta^-(\rho) &= \left| s_{\text{full}} - s_{\text{keep}}(\rho) \right| \end{aligned} \tag{1}$$

Large $\Delta^+(\rho)$ indicates that the removed top-ranked nodes are necessary, while small $\Delta^-(\rho)$ indicates that the retained top-ranked nodes are sufficient.

Finally, at a fixed budget $\rho_\star$ we aggregate necessity and sufficiency into a single *characterisation score* using a weighted harmonic mean of Δ^+ and $1 - \Delta^-$:

$$\text{charact}(w_+, w_-; \rho_\star) = \frac{w_+ + w_-}{\frac{w_+}{\Delta^+(\rho_\star)} + \frac{w_-}{1 - \Delta^-(\rho_\star)}} \tag{2}$$

with $w_+, w_- \in [0, 1]$ and $w_+ + w_- = 1$. Higher charact indicates that the explainer simultaneously exhibits strong necessity and sufficiency at budget $\rho_\star$.

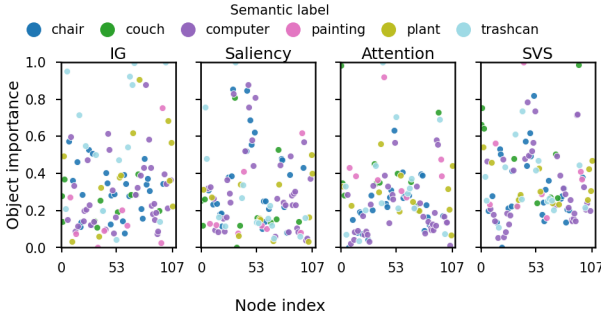


Fig. 3: Importance of each object node averaged over all place embeddings in the office scene, as determined by four explainability methods: Saliency, Integrated Gradients, Attention Weights, and Shapley Value Sampling. All methods highlight a small subset of highly influential objects, but notable discrepancies in mid-range scores reveal method-specific differences in assessing node relevance.

VI. Results

A. Class Ablation

As shown in table IV, leave-one-class-out ablation indicates that `trash can` and `couch` cause the most significant frequency-normalised decrease in PR-AUC, whereas the most frequent classes, `computer` and `chair`, exert only a minor effect. This pattern is consistent with contrastive sentence encoders that down-weight ubiquitous tokens, such as "the" [29].

B. Post-hoc Attribution

Table IV summarises class-level importance from all explainers, compared using the previously defined JSD to assess whether frequent classes receive consistently lower attributions. The methods broadly agree on the classes most (`couch`, `plant`, `trash can`) and least (`chair`, `computer`) useful for localisation. Most methods rank `painting` relatively low, with Shapley Value Sampling being the primary exception. Figure 4 verifies that Attention Weights correlate with perturbation-induced changes in localisation performance, indicating their effectiveness as feature-importance scores, described in section II-C. Attention shows a clear correlation [21], [6] and thus is adopted as an explanatory signal.

Agreement is strongest at the extremes and is stable across runs for Integrated Gradients and Attention Weights. Saliency and Shapley exhibit greater run-to-run variability under environmental perturbations, yet their highest- and lowest-ranked classes typically align with the consensus. To identify the most reliable explainer, each method is further evaluated using fidelity metrics.

C. Fidelity

Overall, the methods agree at the extremes but diverge in mid-rankings. To identify the most reliable explainer, each method is further evaluated using fidelity metrics.

TABLE IV: Semantic-class rankings per method and run. Entries are listed in descending order of importance.

Method	Run	1st	2nd	3rd	4th	5th	6th
Single-Class Ablation	1	TC	CO	PA	PL	CH	CP
	2	TC	CO	PL	PA	CH	CO
	3	TC	CO	PL	CH	PA	CP
Saliency	1	PL	TC	CH	CP	CO	PA
	2	PL	CH	CP	TC	CO	PA
	3	PL	TC	CO	CH	CP	PA
Integrated Gradients	1	TC	PL	CH	CP	CO	PA
	2	TC	PL	CH	CO	CP	PA
	3	TC	PL	CH	CO	CP	PA
Shapley Value Sampling	1	CO	PA	TC	PL	CP	CH
	2	CO	PA	TC	CP	CH	PL
	3	CP	CO	CH	TC	PA	PL
Attention	1	CO	PL	TC	PA	CH	CP
	2	CO	PL	TC	PA	CH	CP
	3	CO	PL	TC	PA	CH	CP

Figures 5a and 5b plot $\Delta^{-}(\rho)$ and $\Delta^{+}(\rho)$ averaged over all place–query pairs in run 1, and fig. 5c presents the characterisation scores across different node selection levels (ρ) for each explainability method. Integrated Gradients consistently achieves the highest characterisation scores, indicating that its top-ranked nodes are both indispensable – i.e. their removal leads to a significant drop in embedding similarity – and sufficient, in that only retaining them preserves most of the original similarity. Attention Weights achieve the second-best scores at lower selection levels, indicating that they are particularly effective in identifying the most important nodes early on. However, their performance diminishes at higher selection levels.

These results demonstrate that Integrated Gradients and Attention Weights are the most effective explainers for our purposes, especially as their rankings align most closely with the class ablation. Attention Weights, in particular, offer the added benefits of clear interpretability and minimal computational overhead – since they are part of the localisation network itself – making them the preferred method when inference speed or resource constraints are paramount.

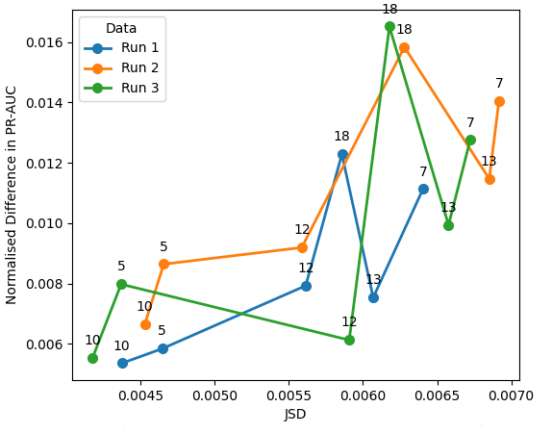


Fig. 4: Normalised change in PR-AUC resulting from removal of each semantic class from the scene graph, measured against the change in node importance distribution obtained from attention weights. The positive correlation confirms that attention can serve as an effective attribution method.

D. Relationship with class frequency

Semantic class rankings from Integrated Gradients, attention and class-ablation ranking (table IV) can be compared with raw object frequencies (table I). Integrated Gradients assigns the highest importance to `trash can` and `plant`, and the lowest to `computer` and `painting`. Attention, by contrast, ranks `couch` and `plant` highest, and `chair` and `computer` lowest. The class ablation likewise places `trash can`, `couch`, and `plant` at the top, with `chair` and `computer` at the bottom.

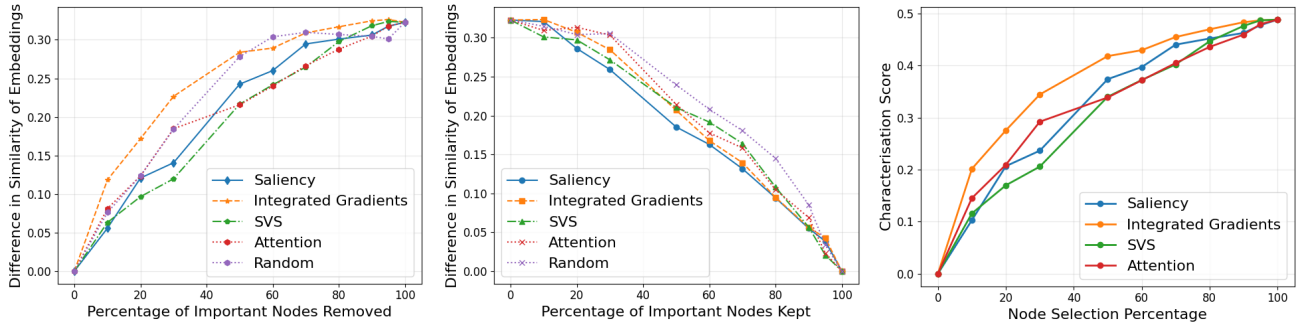
Table I shows that `computer` and `chair` are the most frequent objects, whereas `painting`, `couch`, and `plant` are less common. The results thus indicate an inverse relationship between class frequency and assigned importance, analogous to TF–IDF weighting, whereby high-frequency classes are downweighted and low-frequency classes are upweighted. Notable exceptions persist: `trash can` remains highly important despite moderate frequency, and Integrated Gradients assigns relatively low weights to `painting` instances. Although random perturbations were introduced to `trash can` and `chair` during training to model potentially-moveable landmarks, `trash can` still receives high importance scores.

Overall, object frequency is a *strong but not exclusive* importance indicator, demonstrating that even in a simple localisation model, such as the one we presented, spatial layout and relational context also contribute to the observed importance profiles.

VII. Conclusion, Limitations, and Future Work

We presented a lightweight, interpretable pipeline for coarse place registration on semantic scene graphs. By embedding an unperturbed "map" graph and its perturbed "query" into a shared latent space, the model aligns matching place–query pairs. Notably, the study places explainability at the centre of semantics-based localisation. A comprehensive analysis, combining fidelity-based evaluations and semantic-class perturbations, identifies Integrated Gradients and Attention Weights as the most faithful explanatory signals for this task. Class-importance patterns reveal a TF–IDF–like bias, where common objects are often down-weighted. The high weight of `trash can` indicates that contextual relations, not just frequency, influence performance.

The results come with certain limitations: the semantics-only setting omits geometry, which enables a focused attribution analysis but caps absolute retrieval performance, and the evaluation is confined to a single synthetic domain (uHumans2 office). Future work will integrate lightweight geometric cues, explore more complex model architectures, and evaluate our approach across diverse real-world datasets and semantic taxonomies to stress-test the stability and utility of explanations in deployed localisation systems.



(a) (b) (c)

Fig. 5: Fidelity+ (a, necessity of removed nodes), fidelity- (b, sufficiency of kept nodes) and combined characterisation score (c) curves on Run 1. Generally, Integrated Gradients performs best overall, as indicated by its higher characterisation score, where Attention also exhibits a high initial slope, suggesting strong performance in identifying the most critical nodes.

References

[1] Y. Liu, Y. Petillot, D. Lane, and S. Wang, "Global localization with object-level semantics and topology," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 4909–4915.

[2] M. Yoshida, K. Tanaka, R. Yamamoto, and D. Iwata, "Active semantic localization with graph neural embedding," in *Pattern Recognition*, H. Lu, M. Blumenstein, S.-B. Cho, C.-L. Liu, Y. Yagi, and T. Kamiya, Eds. Cham: Springer Nature Switzerland, 2023, pp. 216–230.

[3] E. Panagiotaki, D. De Martini, G. Pramatarov, M. Gadd, and L. Kunze, "Sem-gat: Explainable semantic pose estimation using learned graph attention," in *2023 21st International Conference on Advanced Robotics (ICAR)*, 2023, pp. 367–374.

[4] G. Pramatarov, D. De Martini, M. Gadd, and P. Newman, "Boxgraph: Semantic place recognition and pose estimation from 3d lidar," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7004–7011.

[5] G. Pramatarov, M. Gadd, P. Newman, and D. De Martini, "That's my point: Compact object-centric lidar pose estimation for large-scale outdoor localisation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 276–12 282.

[6] E. Panagiotaki, D. De Martini, and L. Kunze, "Semantic interpretation and validation of graph attention-based explanations for gnn models," in *2023 21st International Conference on Advanced Robotics (ICAR)*, 2023, pp. 375–380.

[7] N. Kobyshev, H. Riemenschneider, and L. V. Gool, "Matching features correctly through semantic understanding," in *2014 2nd International Conference on 3D Vision*, vol. 1, 2014, pp. 472–479.

[8] R. Arandjelović and A. Zisserman, "Visual vocabulary with a semantic twist," 11 2014, pp. 178–195.

[9] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6896–6906.

[10] A. Cramariuc, F. Tschopp, N. Alatur, S. Benz, T. Falck, M. Brühlmeier, B. Hahn, J. Nieto, and R. Siegwart, "Semsegmap – 3d segment-based semantic localization," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 1183–1190.

[11] E. Panagiotaki, D. D. Martini, L. Kunze, P. Newman, and P. Veličković, "Nar-*icp: Neural execution of classical icp-based pointcloud registration algorithms," 2024. [Online]. Available: https://arxiv.org/abs/2410.11031

[12] Y. Xia, L. Shi, Z. Ding, J. F. Henriques, and D. Cremers, "Text2loc: 3d point cloud localization from natural language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 958–14 967.

[13] J. Theiner, E. Müller-Budack, and R. Ewerth, "Interpretable semantic photo geolocation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022.

[14] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5782–5799, 2023.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Workshop at International Conference on Learning Representations*, 2014.

[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," ser. ICML'17. JMLR.org, 2017.

[17] E. Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *J. Mach. Learn. Res.*, vol. 11, pp. 1–18, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:14451872

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[19] B. Wen, K. P. Subbalakshmi, and F. Yang, "Revisiting attention weights as explanations from an information theoretic perspective," 2022. [Online]. Available: https://arxiv.org/abs/2211.07714

[20] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Nov. 2019.

[21] Y. Fan, Y. Yao, and C. Joe-Wong, "Gcn-se: Attention as explainability for node classification in dynamic graphs," in *2021 IEEE International Conference on Data Mining (ICDM)*, 2021, pp. 1060–1065.

[22] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," *Robotics: Science and Systems (RSS)*, 2022.

[23] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1689–1696.

[24] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:8281592

[25] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. 9, pp. 207–244, 2009. [Online]. Available: http://jmlr.org/papers/v10/weinberger09a.html

[26] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *ArXiv*, vol. abs/1807.03748, 2018.

[27] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 1263–1272.

[28] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 764–10 773.

[29] H. Kurita, G. Kobayashi, S. Yokoi, and K. Inui, "Contrastive learning-based sentence encoders implicitly weight informative words," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 932–10 947. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.729/