

Entropy-Aware Keyframe Selection for Semantic Topological Mapping

Sanaullah Khaskheli*, Mohammad Asif Khan, Muhammad Ismail Mangrio, Sajid Khan, Zakria

¹Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan

*Corresponding author: sanaullah@iba-suk.edu.pk

Abstract

One of the key challenges in autonomous mobile robot navigation is the construction of compact, semantically coherent topological maps. Current geometric simultaneous localization and mapping (SLAM) methods, including ORB-SLAM2, produce semantically empty and metrically accurate SLAM maps that are vulnerable to light changes, dynamic occlusion and perceptual aliasing. Conversely, vision-language foundation models trained on web-scale data encode rich semantic priors, yet their utility for online topological map construction remains insufficiently characterized. This paper presents a framework for entropy-aware keyframe selection in semantic topological mapping. We utilize Shannon entropy, computed over the latent embeddings of three state-of-the-art models (DINOv2, CLIP, and SigLIP), as an information-theoretic utility function to govern map node insertion. Extensive evaluations across four sequences of the KITTI Visual Odometry Benchmark (encompassing 9,504 frames) demonstrate that SigLIP achieves a high compression ratio (CR), yielding a 96.91% frame reduction on high-speed trajectories at the cost of larger spatial gaps between nodes. Conversely, DINOv2 yields the most consistent entropy profile across all environments ($\mu = 3.31 - 5.72$), while SigLIP drops to 2.19 on featureless highway sequences, indicating significantly reduced spatial discriminability under semantic redundancy. The results show that DINOv2 achieves the highest gap reliability on three of four sequences, SigLIP achieves the highest compression, and CLIP achieves the highest informativeness, giving insights for the design of perception systems in autonomous mobile robotics.

Keywords—Topological Mapping, Vision Foundation Models, DINOv2, CLIP, SigLIP, Keyframe Selection, Shannon Entropy, KITTI Benchmark

1 INTRODUCTION

THE capacity of an autonomous mobile robot to construct and maintain a coherent internal representation of its operating environment is foundational to safe and efficient navigation. The evolution of these representations has followed two classical paradigms, with hybrid approaches emerging more recently: metric SLAM, which reconstructs geometry relative to a global coordinate frame [1], [2], and topological mapping, which models the environment as a discrete graph where nodes represent navigable places and edges denote traversability [3]. Although metric SLAM systems exhibit high precision in structured settings, they remain susceptible

to perceptual degradation caused by environmental dynamics in urban and peri-urban environments—challenges thoroughly documented in the KITTI benchmark [4].

Large-scale vision foundation and vision-language models—specifically DINOv2 [5], CLIP [6], and SigLIP [7]—offer powerful feature representations for semantic topological mapping. The models were pre-trained on hundreds of millions of images, using self-supervised or contrastive objectives, and encode high-quality transferable representations that generalize across a wide range of visual tasks without task-specific fine-tuning. In this work, all models are used in frozen inference mode, with no task-specific fine-tuning. Leveraging these models for topological place representation offers visual maps invariant to changes in illumination, viewpoint, and dynamic scenes, a property un-

achievable by traditional geometric SLAM pipelines utilizing handcrafted feature descriptors.

Nevertheless, a primary and under-studied complication of implementing these models within on-line topological mapping architectures is the information bottleneck generated by continuous, high-rate sensor streams. A mobile robot operating at standard video frame rates (25–30 Hz) generates roughly 90,000–108,000 frames per hour, introducing massive information redundancy. Fixed-rate temporal sub-sampling and velocity-based decimation schemes do not account for the semantic content of a scene, and can therefore discard critical transition frames while retaining redundant data [8].

This information bottleneck is solved here by proposing an Entropy-Aware Utility Function, which measures the informational novelty of each incoming frame with respect to the current set of keyframes directly in the latent space of the vision foundation model. The frame is added to the topological map only if the joint utility computed by the frame is above a threshold set to control a target level of information. We test this framework systematically on DINOv2, CLIP, and SigLIP, on four KITTI sequences consisting of 9,504 frames, uncovering characteristic behaviors in each model according to its objective in training and architecture.

The principal contributions of this work are: (1) a formal information-theoretic framework for keyframe selection in semantic topological mapping, grounded in Shannon entropy [9] over foundation model latent spaces; (2) a systematic comparative evaluation of DINOv2, CLIP, and SigLIP on the KITTI benchmark across compression ratio, spatial reliability, and average informativeness, powered by empirical processing logs; and (3) empirical demonstration that training objective governs the compression–reliability trade-off: self-supervised patch distillation (DINOv2) minimizes max topological gap (37.77 m on Seq 00) at moderate compression (57%), while sigmoid language alignment (SigLIP) maximizes compression (96.91%) at the cost of navigational safety (375.67 m gap on Seq 01).

2 RELATED WORK

Our work is contextualized across four research strands: geometric SLAM limitations, learned place recognition, the emergence of vision foundation models (VFMs), and information-theoretic map management.

2.1 Geometric SLAM and its Limitations

SLAM has evolved from first generations of filter formulation to today’s graph-based optimization. Existing work, such as PTAM [10], demonstrated that processing an informative subset of frames allows for computational tractability and is keyframe-based SLAM. The ORB-SLAM family of algorithms [1], [2], [11] are the current state-of-the-art in feature-based metric SLAM. Although these systems have low relative pose error in controlled conditions, they are systematically degraded by changes in illumination, and perceptual aliasing, which is a common event in the KITTI dataset [4]. Direct methods, such as LSD-SLAM [12] and DSO [13], have been found to be more robust in datasets with few or no textured scenes, but are sensitive to changes in exposure. Other approaches like RTAB-Map [14] or Cartographer [15] use loop-closure in combination with graph optimization, however they use hand-made representations which are not semantically generalizable [16].

2.2 Visual Place Recognition and Learned Descriptors

Early topological mapping used bag-of-words vocabularies to represent place signatures, exemplified by the FAB-MAP system [3]. To mitigate perceptual aliasing, SeqSLAM [17] took advantage of the temporal coherence of image sequences. Lowry et al. [18] identified that handcrafted techniques are not effective at handling seasonal and viewpoint changes, and need to be semantically discriminative. The advent of deep learning revolutionized the field with the introduction of NetVLAD [19] (end-to-end learned global descriptors). This was further improved by Patch-NetVLAD [20] and CosPlace [21]. However, these methods require a lot of labeled training data and domain-specific supervision to transfer to new domains [22].

2.3 Vision Foundation Models for Spatial Reasoning

Large-scale Vision Foundation Models (VFMs) have transformed robotic perception. Contrastive vision-language pre-training via CLIP [6] enables zero-shot transfer to navigation tasks like LM-Nav [23]. SigLIP [7] optimized this via a sigmoid pairwise loss, while AnyLoc [24] demonstrated the potential of VFMs for universal place recognition. Additionally, the Visual Navigation Transformer (ViNT) [25] showed generalization across diverse robot

platforms. While Firoozi et al. [26] review VFM applications in robotics, the relationship between their learned embedding structure and information-theoretic keyframe selection criteria remains understudied.

2.4 Information-Theoretic Keyframe Management

Information theory provides a principled basis for active perception [27]. Traditional active SLAM formulations pose map refinement as the maximization of mutual information between sensor observations and the underlying map state [28]. Recent work by Asgharivaskasi and Atanasov [29] presented a framework for calculating Shannon mutual information in semantic OcTree maps. Our work diverges by using Shannon entropy as a semantic filter for the information bottleneck of high-rate sensor streams rather than for path planning. As noted by Das and Waslander [8], information-theoretic criteria for keyframe admission remain unexplored in the context of foundation model latent spaces.

Research gap. The preceding review reveals that, while VFMs have been individually applied to visual place recognition tasks [24], and while information-theoretic criteria have been used for active SLAM path planning [27], [29], no prior work has applied Shannon entropy computed directly over the Perceptual Proximity Distribution of a foundation model keyframe buffer as an online admission criterion for topological map nodes [8]. Neither a comparative characterisation of DINOv2, CLIP, and SigLIP with respect to information-theoretic map compression nor an analysis of how VFM training objectives govern the compression–reliability trade-off has been reported.

3 THEORETICAL FRAMEWORK

3.1 Information-Theoretic Modeling of Latent Spaces

To quantify the informational value of a sensor stream, we map each incoming frame I_t to a high-dimensional embedding space via a frozen VFM denoted by ϕ . Let $\phi(I_t) = \mathbf{z}_t \in \mathbb{R}^d$ represent the unit-normalized latent embedding.

We define a *Perceptual Proximity Distribution* p_t over the existing map keyframes via a softmax-normalized cosine similarity index:

$$p_t(i) = \frac{\exp\left(\frac{\cos(\mathbf{z}_t, \mathbf{v}_i)}{\tau_{\text{soft}}}\right)}{\sum_{\mathbf{v}_j \in \mathcal{K}} \exp\left(\frac{\cos(\mathbf{z}_t, \mathbf{v}_j)}{\tau_{\text{soft}}}\right)} \quad (1)$$

where τ_{soft} represents the softmax scaling temperature controlling the focus of the neighborhood search space (see Eq. (1)). A high entropy value indicates that the current frame’s similarity is distributed broadly across existing nodes, suggesting that it does not closely match any single map node and therefore carries higher novelty.

3.2 The Joint Utility Function

We propose a Joint Utility Function $U(\mathbf{z}_t)$ that balances global embedding novelty with semantic density (see Eq. (2)). This formulation ensures that the topological map remains sparse along redundant linear trajectories but increases node density at complex environmental transitions:

$$U(\mathbf{z}_t) = \alpha \cdot \underbrace{\left(1 - \max_{\mathbf{v}_k \in \mathcal{K}} \cos(\mathbf{z}_t, \mathbf{v}_k)\right)}_{\text{Embedding Novelty}} + \beta \cdot \underbrace{\left(\frac{-\sum_{i=1}^k p_t(i) \log_2 p_t(i)}{\log_2(k)}\right)}_{\text{Semantic Complexity}} \quad (2)$$

where $\alpha = 0.75$ and $\beta = 0.25$ act as operational weights. The term $\log_2(k)$ normalizes the Shannon entropy relative to the current map size k , normalizing the entropy term to $[0, 1]$ (for $k = 1$ the term is set to zero). A frame is inserted as a keyframe node if and only if $U(\mathbf{z}_t) > \tau$. Nodes are not subsequently removed, as the focus of this work is on online insertion.

4 SYSTEM ARCHITECTURE

Our feature extraction module selects the appropriate forward pass for each model architecture (CLS token for DINOv2; pooled visual embedding for CLIP and SigLIP). Experiments were conducted on Cloud TPU v3-8. The released code includes a GPU-compatible inference path for broader reproducibility.

Fig. 1 illustrates the workflow of the proposed method for selecting nodes from the sequence of images captured by the onboard camera. For each incoming image, a ViT descriptor is generated and compared with the descriptors of the existing nodes in the topological map. A new node is added to the map when the predefined node-creation criteria are satisfied; otherwise, the system continues processing the subsequent images.

Algorithm 1 Entropy-Aware Keyframe Selection for Topological Mapping

Input: Stream $\mathcal{I} = \{I_1, \dots, I_N\}$, $\alpha, \beta, \tau, \tau_{\text{soft}}$
Output: Topological Map Buffer $\mathcal{K}$

```

1: Initialize  $\mathcal{K} \leftarrow \{\phi(I_1)\}$ 
2: for  $t = 2$  to  $N$  do
3:    $\mathbf{z}_t \leftarrow \text{ExtractEmbedding}(I_t)$ ;  $k \leftarrow |\mathcal{K}|$ 
4:   for each  $\mathbf{v}_i \in \mathcal{K}$  do
5:      $p_t(i) \leftarrow \frac{\exp(\cos(\mathbf{z}_t, \mathbf{v}_i)/\tau_{\text{soft}})}{\sum_j \exp(\cos(\mathbf{z}_t, \mathbf{v}_j)/\tau_{\text{soft}})}$ 
6:   end for
7:    $H_t \leftarrow -\sum_{i=1}^k p_t(i) \log_2 p_t(i)$ 
8:    $\text{Nov} \leftarrow 1 - \max_{\mathbf{v}_j \in \mathcal{K}} \cos(\mathbf{z}_t, \mathbf{v}_j)$ 
9:    $U_t \leftarrow \alpha \cdot \text{Nov} + \beta \cdot H_t / \log_2(k)$ 
10:  if  $U_t > \tau$  then
11:     $\mathcal{K} \leftarrow \mathcal{K} \cup \{\mathbf{z}_t\}$ ; Add node & edge to  $G(V, E)$ 
12:  end if
13: end for
14:
15: return  $\mathcal{K}$ 

```

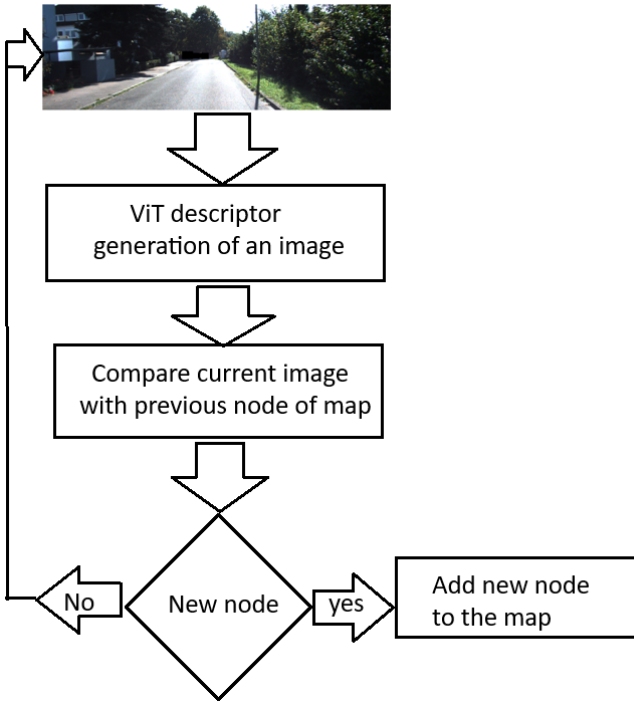


Fig. 1: Workflow of the proposed framework for sequential image processing, keyframe selection, and topological map construction.

4.1 Implementation and Tensor Extraction

The extraction pipeline operates on raw RGB image arrays resized to structural baselines: 224×224 pixels for CLIP and SigLIP geometries, and 518×518 pixels for DINOv2 layers. All models are used

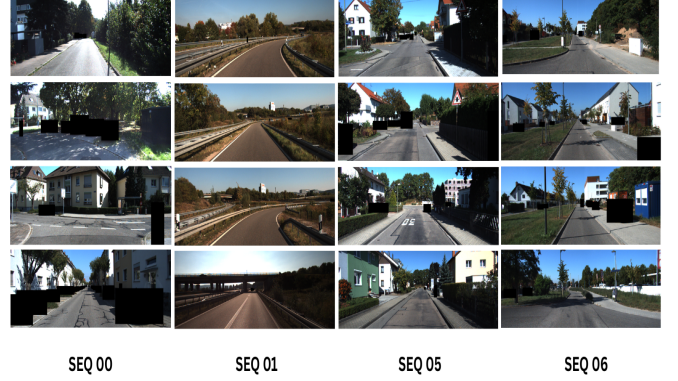


Fig. 2: Diverse operating environments from the KITTI dataset: Seq 00 (Suburban loop), Seq 01 (High-speed highway), Seq 05 (Residential area), Seq 06 (Urban intersections).

in frozen inference mode (weights fixed; no fine-tuning), ensuring that downstream evaluations reflect the raw, unmodified representational geometry of the pre-trained embedding spaces (see Algorithm 1).

5 EXPERIMENTAL SETUP

5.1 KITTI Visual Odometry Benchmark

We evaluate four KITTI sequences spanning diverse environments (Fig. 2), which is rigorous in testing VFM-based spatial perception [4], [30]:

- **Seq 00:** Suburban loop with high perceptual aliasing (repetitive houses).
- **Seq 01:** High-speed highway with extreme semantic redundancy.
- **Seq 05:** Residential area with complex foliage and intersections.
- **Seq 06:** Urban intersection sequence with rapid viewpoint transitions.

5.2 Hyperparameters

The hyperparameters are kept fixed for all the sequences to allow a direct cross-sequence and cross-model comparison. The Joint Utility weights (Section 3) are set to $\alpha = 0.75$ and $\beta = 0.25$, obtained through grid search over $\alpha \in \{0.50, 0.60, 0.70, 0.75, 0.80, 0.90\}$ with $\beta = 1 - \alpha$. The selection criterion is $\tau = 0.20$. The softmax temperature $\tau_{\text{soft}} = 0.07$ is used, commonly adopted in contrastive evaluation protocols [6].

Throughout this paper, the Compression Ratio (CR) is defined as:

$$\text{CR} = \left(1 - \frac{N_{\text{kf}}}{N_{\text{total}}}\right) \times 100\% \quad (3)$$

where N_{kf} denotes the number of accepted keyframes and N_{total} the total number of frames. Avg Inf is the mean unnormalized entropy H_t (bits), Max Gap is the largest distance between the ground-truth positions of consecutive keyframes, and Edges counts the links between consecutive keyframes.

6 RESULTS

6.1 Quantitative Performance and Compression Analysis

Table 1 summarises benchmark results across all four sequences and three models. The variation in compression ratio across models (DINOv2: 45–84%, SigLIP: 87–97%) despite identical input sequences confirms that the selection criterion is driven by the representational geometry of the embedding space, not by scene geometry alone.

6.2 Compression Ratio and Mapping Efficiency

The performance of the entropy-sensitive selection mechanism is measured using the compression ratio (CR) values in Table 1. A mean CR of 63.10% is attained in DINOv2, 73.31% in CLIP, and 90.45% in SigLIP.

The spatial characteristics of each operational environment directly dictate the compression profile across all models. On Sequence 01 (High-speed highway), raw sensory streams present severe informational redundancy. Our entropy-aware selection mechanics successfully drop uninformative data blocks, with SigLIP yielding peak compression efficiency at CR = 96.91%, closely shadowed by CLIP at 87.83% as shown in Fig. 3.

However, SigLIP’s high compression ratio carries a navigational cost: on Sequence 06 (urban intersections) its 108-node map yields a Max Gap of 129.73 m, about 3.7 and 3.0 times that of DINOv2 (35.14 m) and CLIP (43.46 m), respectively, making it unsuitable for dense urban navigation. The moderate compression of DINOv2 (mean 63.10%) is balanced by superior gap control, as evidenced by the lowest Max Gap on perceptually complex sequences Seq 00 (37.77 m) and Seq 06 (35.14 m) as seen in Fig. 4.

6.3 Landscape Entropy and Informativeness Profiling

The Average Informativeness (Avg Inf) reveals a major architectural anomaly when considering the results for different training targets. CLIP, aligned

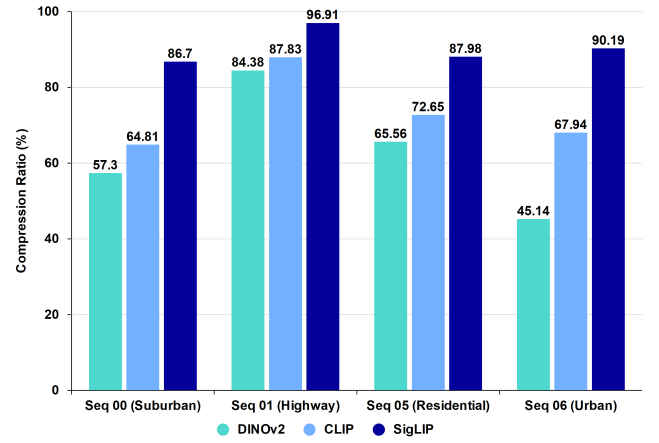


Fig. 3: Keyframe Compression Efficiency Across Models and Sequences.

with language, creates high average entropy signatures on most sequences. Overall, CLIP achieves the highest information metric ($\mu = 6.1044$ on Sequence 00), and DINOv2 is close behind ($\mu = 5.7153$). For all sequences, SigLIP always produces lower and flatter entropy profiles (lowest $\mu = 2.1920$ on Sequence 01, highest $\mu = 4.9720$ on Sequence 00). Indeed, CLIP’s cross-modal semantic space generates cosine similarity profiles that are widely spread, leading to near-uniform proximity distributions (Eq. (1)) and thus to higher values of the calculated landscape entropy measures. Fig. 5 shows the per-frame entropy $H(z|K)$ and joint utility $U(z)$ of all three models on Sequence 01; as the keyframe buffer increases, CLIP and DINOv2 have monotonically increasing entropy, while SigLIP has monotonically decreasing entropy.

7 COMPARATIVE ANALYSIS

A critical distinction between our framework and geometric baselines like ORB-SLAM2 [11] lies in the handling of semantic stasis. On a trajectory where a robot remains stationary—for instance, at a traffic light—distance-based sampling continues to process or pause based solely on spatial metrics. This comparison is conceptual; an empirical evaluation of stationary-scenario behavior is reserved for future work. While our utility function similarly avoids redundant node insertion (where $U \rightarrow 0$ due to lack of shift), it does so by explicitly recognizing information redundancy in the latent space. This allows for a semantically-grounded map that ignores temporal or geometric noise while remaining sensitive to sudden semantic changes, such as the

TABLE 1: Master Benchmark: Entropy-Aware Keyframe Selection on KITTI Paths

Model	Seq	Frames	Nodes	CR%	Avg Inf	Max Gap (m)	Edges
DINOv2	00	4541	1939	57.30	5.7153	37.77	1938
DINOv2	01	1101	172	84.38	3.3101	249.73	171
DINOv2	05	2761	951	65.56	5.0579	112.49	950
DINOv2	06	1101	604	45.14	4.9157	35.14	603
CLIP	00	4541	1598	64.81	6.1044	61.40	1597
CLIP	01	1101	134	87.83	3.4779	289.97	133
CLIP	05	2761	755	72.65	5.0608	82.60	754
CLIP	06	1101	353	67.94	4.6299	43.46	352
SigLIP	00	4541	604	86.70	4.9720	90.04	603
SigLIP	01	1101	34	96.91	2.1920	375.67	33
SigLIP	05	2761	332	87.98	4.1878	90.59	331
SigLIP	06	1101	108	90.19	3.5555	129.73	107

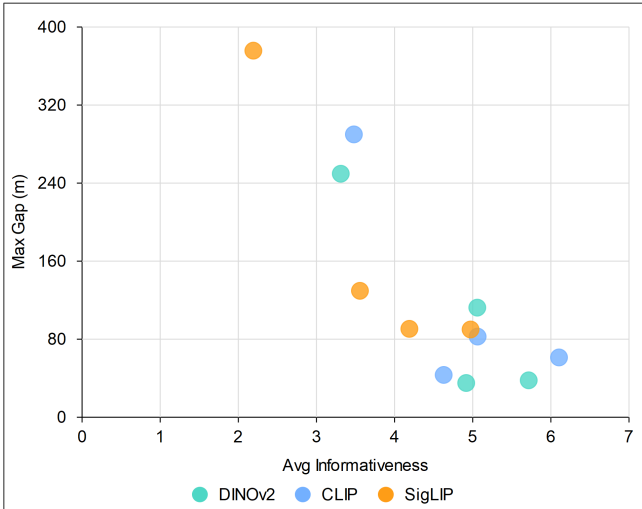


Fig. 4: Max Gap versus Average Informativeness. Note: Some Points Overlap.

appearance of a new landmark. Table 2 summarises the key architectural properties of our framework against representative baselines; among these methods, ours is designed to combine online operation with semantic representations, while its out-of-distribution robustness is inherited from the frozen foundation models.

8 DISCUSSION

8.1 Structural Feature Discrimination vs. Cross-Modal Abstraction

A series of empirical benchmark profiles emerge, highlighting the types of architectural compromises that can be made between self-supervised visual descriptors and language-aligned contrastive spaces. The DINOv2 model is optimized by self-supervised

TABLE 2: Comparative analysis with state-of-the-art mapping paradigms. ORB-SLAM2 generates dense metric points rather than downsampled topological nodes, so a CR is not directly applicable.

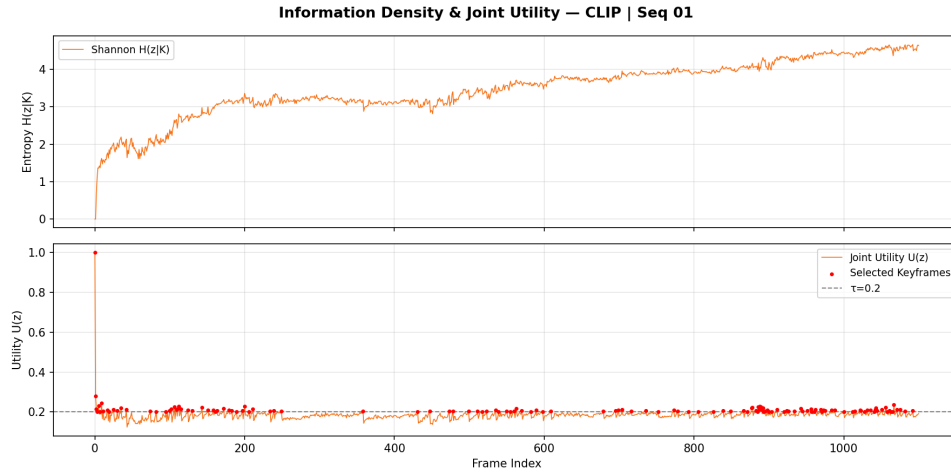
Method	Paradigm	OOD	Online	Sem.	CR
ORB-SLAM2 [11]	Geometric	Low	Yes	No	—
FAB-MAP [3]	Probabilistic	Med.	Yes	Par.	~60% ^b
NetVLAD [19]	Supervised	Low	No	Par.	~75% ^b
AnyLoc [24]	VFM (DINO)	High	No	Yes	~85% ^b
Ours	VFM+H	High^c	Yes	Yes	63–90%^a

^aMean CR range across backbones (Table 1).

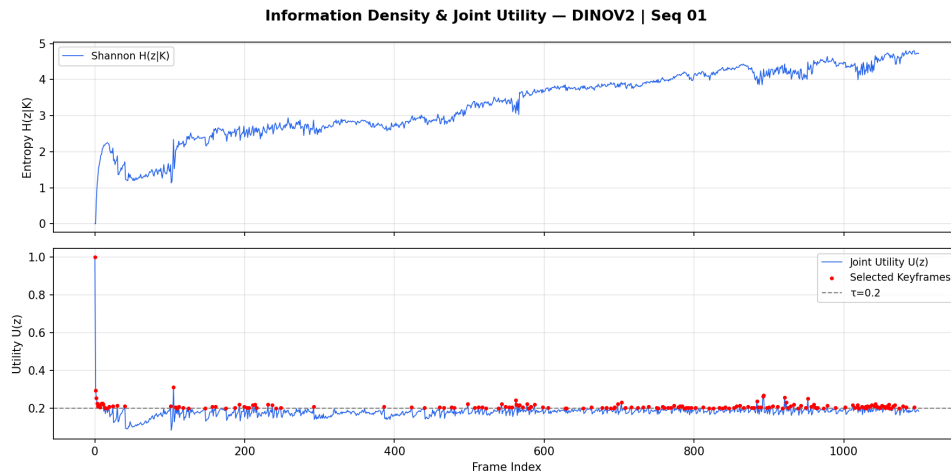
^bIndicative estimates; not reported in the original publications.

^cExpected from the frozen VFM backbones; not separately evaluated.

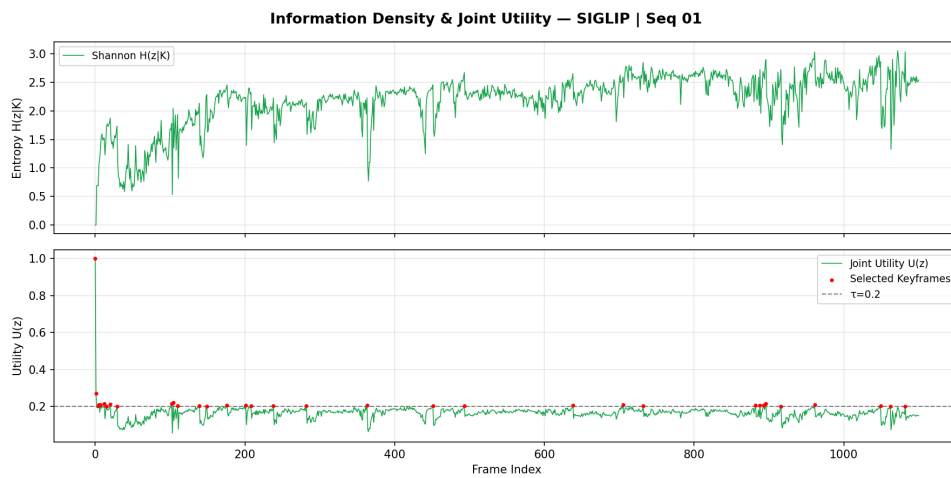
patch-level and image-level objectives and it is very sensitive to fine-grained structural features. On highway stretches, where the structural horizon is uniform (Sequence 01), DINOv2 still retains more observations than CLIP and SigLIP because small changes in the underlying geometric features remain detectable (CR = 84.38%). This is the main reason for DINOv2’s smaller gaps compared to SigLIP and CLIP. In contrast, visual scenes are projected into strongly abstracted semantic spaces in CLIP and SigLIP. Over the tested sequences, Average Informativeness (Avg Inf) values of 3.48–6.10 are obtained with CLIP and SigLIP values of 2.19–4.97 (Table 1). Their latent projections do not have extreme colinearity, thus the tracking engine gains high spatial entropy fluctuations in different environmental structures. This spatial sensitivity also directly correlates to gap safety: DINOv2 has the lowest Max Gap on both sequences (35.14 m on Seq 06, 37.77 m on Seq 00), demonstrating that it is sensitive to locations that are structurally



(a) CLIP — Seq 01



(b) DINOv2 — Seq 01



(c) SigLIP — Seq 01

Fig. 5: Comparison of information density and joint utility profiles for CLIP, DINOv2, and SigLIP on KITTI Sequence 01.

significant, such as intersections, turns, or scene boundaries. The observed differences in behavior are partly due to architectural differences, including patch size (DINOv2: 14×14 vs. CLIP: 32×32), the composition of training datasets, and the training objective. These factors can only be separated in future work using controlled variants of models.

These data-driven insights give clear-cut model selection rules for autonomous robotic perception pipelines. In cases where very high compression is required throughout a repetitive, high-speed route with a relatively large tracking gap, SigLIP is a great standalone filter. But in dense urban settings where visual place recognition needs to be persistent with high utility, the structural sensitivity of DINOv2, followed by CLIP, provides denser tracking updates at distinct scene transitions.

The framework uses several sequences (00, 01, 05, and 06) from the publicly available KITTI dataset. The compression ratio of these sequences differs (see Table 1) because of variations in scene appearance. When color and texture remain consistent, a higher compression ratio is expected compared to more dynamic sequences. KITTI includes diverse conditions, such as lighting changes (direct sunlight and shadowed areas) and dynamic scenes with moving vehicles and changing surroundings. The ViT models we employ handled these conditions across the tested sequences, and their strengths in visual place recognition may benefit robot localization in future evaluations.

While this study utilizes the outdoor KITTI benchmark, the phenomenon of Semantic Over-Abstraction (distinct places mapped to near-identical embeddings) is expected to be more pronounced in highly symmetric indoor environments (e.g., warehouses). Future work will evaluate the RGB modality of the NYU Depth V2 and ScanNet datasets to assess performance in indoor environments, together with a direct empirical comparison against fixed-rate and velocity-based sampling.

9 CONCLUSION

In this paper, an information-theoretic model of entropy-aware keyframe selection in semantic topological mapping was proposed. We applied Shannon entropy over VFM latent spaces to develop a principled mechanism for inserting a map node online. Our results establish a clear compression–reliability spectrum: SigLIP maximises compression (96.91%) at the cost of navigational safety (375.67 m

on Seq 01), while DINOv2’s patch-level discriminability yields the lowest topological gap (35.14 m on Seq 06) at moderate compression (mean 63.10%), making it the preferred backbone for safety-critical deployment.

Ethics approval: Not applicable. This study utilized a publicly available benchmark dataset (KITTI) and did not involve human participants or animal subjects.

Data availability: The data that support the findings of this study are openly available in the KITTI Visual Odometry Benchmark at <http://www.cvlibs.net/datasets/kitti/>.

REFERENCES

- [1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM: A versatile and accurate monocular SLAM system,” *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.
- [2] C. Campos, R. Elvira, J. J. Gómez Rodríguez, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM,” *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.
- [3] M. Cummins and P. Newman, “FAB-MAP: Probabilistic localization and mapping in the space of appearance,” *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, Jun. 2008.
- [4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
- [5] M. Oquab *et al.*, “DINOv2: Learning robust visual features without supervision,” *Trans. Mach. Learn. Res.*, pp. 1–31, 2024.
- [6] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, 2021, pp. 8748–8763.
- [7] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 11975–11986.
- [8] A. Das and S. L. Waslander, “Entropy based keyframe selection for multi-camera visual SLAM,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2015, pp. 3676–3681.
- [9] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
- [10] G. Klein and D. Murray, “Parallel tracking and mapping for small AR workspaces,” in *Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality (ISMAR)*, 2007, pp. 225–234.
- [11] R. Mur-Artal and J. D. Tardós, “ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras,” *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.
- [12] J. Engel, T. Schöps, and D. Cremers, “LSD-SLAM: Large-scale direct monocular SLAM,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 834–849.
- [13] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

- [14] M. Labbé and F. Michaud, “RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation,” *J. Field Robot.*, vol. 36, no. 2, pp. 416–446, 2019.
- [15] W. Hess, D. Kohler, H. Rapp, and D. Andor, “Real-time loop closure in 2D LIDAR SLAM,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2016, pp. 1271–1278.
- [16] C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, “Deep learning for visual localization and mapping: A survey,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 17000–17020, Dec. 2024.
- [17] M. J. Milford and G. F. Wyeth, “SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2012, pp. 1643–1649.
- [18] S. Lowry *et al.*, “Visual place recognition: A survey,” *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.
- [19] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN architecture for weakly supervised place recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 5297–5307.
- [20] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, “Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14141–14152.
- [21] G. Berton, C. Masone, and B. Caputo, “Rethinking visual geo-localization for large-scale applications,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 4878–4888.
- [22] C. Masone and B. Caputo, “A survey on deep visual place recognition,” *IEEE Access*, vol. 9, pp. 19516–19547, 2021.
- [23] D. Shah, B. Osiński, B. Ichter, and S. Levine, “LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action,” in *Proc. 6th Conf. Robot Learn. (CoRL)*, PMLR, vol. 205, 2023, pp. 492–504.
- [24] N. Keetha *et al.*, “AnyLoc: Towards universal visual place recognition,” *IEEE Robot. Autom. Lett.*, vol. 9, no. 2, pp. 1286–1293, Feb. 2024.
- [25] D. Shah *et al.*, “ViNT: A foundation model for visual navigation,” in *Proc. 7th Conf. Robot Learn. (CoRL)*, PMLR, vol. 229, 2023, pp. 711–733.
- [26] R. Firoozi *et al.*, “Foundation models in robotics: Applications, challenges, and the future,” *Int. J. Robot. Res.*, vol. 44, no. 5, pp. 701–739, 2025.
- [27] J. A. Placed *et al.*, “A survey on active simultaneous localization and mapping: State of the art and new frontiers,” *IEEE Trans. Robot.*, vol. 39, no. 3, pp. 1686–1705, Jun. 2023.
- [28] B. J. Julian, S. Karaman, and D. Rus, “On mutual information-based control of range sensing robots for mapping applications,” *Int. J. Robot. Res.*, vol. 33, no. 10, pp. 1375–1392, Sep. 2014.
- [29] A. Asgharivaskasi and N. Atanasov, “Semantic OcTree mapping and Shannon mutual information computation for robot exploration,” *IEEE Trans. Robot.*, vol. 39, no. 3, pp. 1910–1928, Jun. 2023.
- [30] M. El Banani *et al.*, “Probing the 3D awareness of visual foundation models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 21795–21806.